ulm university universität
uulm

Contextual Time-Continuous Emotion Recognition based on Multimodal Data

A doctoral thesis jointly supervised with ITMO University (Russian Federation) and submitted in fulfilment of the requirements for the academic degree of

Dr.rer.nat.

by

Dmitrii Fedotov

born in Krasnoyarsk, Russian Federation

2020

Acting Dean:	Prof. Dr.-Ing. Maurits Ortmanns
Supervisors:	Prof. Dr. Dr.-Ing. Wolfgang Minker
	Prof. Dr.Sc. Alexey A. Karpov
Examiners:	Prof. Dr. Enrico Rukzio
	PD Dr. Friedhelm Schwenker
Examination date:	December 17, 2020

Ульмский Университет

Ульм, Германия

Национальный исследовательский университет ИТМО

Санкт-Петербург, Россия

ulm university universität

ITMO UNIVERSITY

Федотов Дмитрий Валерьевич

Контекстно-зависимое распознавание эмоций на основе многомодальных данных

Диссертация на соискание ученой степени

кандидата технических наук

На правах рукописи

Ульм 2020

Ульмский Университет

Ульм, Германия

Национальный исследовательский университет ИТМО

Санкт-Петербург, Россия

Федотов Дмитрий Валерьевич

Контекстно-зависимое распознавание эмоций на основе многомодальных данных

Специальность 05.13.17

«Теоретические основы информатики»

Диссертация на соискание ученой степени

кандидата технических наук

Научный руководитель:
доктор технических наук
Ульмский Университет
Минкер Вольфганг

доктор технических наук
Национальный исследовательский университет ИТМО
Карпов Алексей Анатольевич

На правах рукописи

Ульм 2020

Диссертация подготовлена в: Ульмский Университет и Национальный исследовательский университет ИТМО.

Научный руководитель:	Доктор технических наук профессор Ульмского Университета Минкер Вольфганг
	Доктор технических наук профессор Университета ИТМО Карпов Алексей Анатольевич
Официальные оппоненты:	Доктор технических наук профессор Института Науки и Технологий Нары Накамура Сатоши
	Доктор технических наук главный научный сотрудник ИПУ РАН Мещеряков Роман Валерьевич

Ulm University

Ulm, Germany

ITMO University

Saint Petersburg, Russia

ulm university universität

Fedotov Dmitrii Valer'evich

Contextual Time-Continuous Emotion Recognition based on Multimodal Data

Specialty 05.13.17

«Theoretical Foundations of Informatics»

Academic dissertation candidate of engineering

Supervisor:
Dr. Dr.-Ing, Professor
Ulm University
Minker Wolfgang

Doctor of Technical Science, Professor
ITMO Univerity
Karpov Alexey Anatol'evich

As a manuscript

Ulm 2020

The research was carried out at: Ulm University and ITMO University

Scientific adviser :

Dr. Dr.-Ing, Professor
Ulm University
Minker Wolfgang

Doctor of Technical Science, Professor
ITMO Univerity
Karpov Alexey Anatol'evich

Official opponents :

Dr. Professor
Nara Institute of Science and Technology
Nakamura Satoshi

Doctor of Technical Science, Chief Researcher
ICS RAS
Meshcheryakov Roman Valer'evich

Contents

Реферат

Общая характеристика работы

Актуальность темы. Интеллектуальные информационные технологии и, в частности, системы человеко-машинного взаимодействия получили значительное развитие за последние десятилетия. Существенное повышение качества систем автоматического распознавания речи позволило создать коммерчески успешные продукты, получившие широкое распространение. Примерами таких систем являются голосовые помощники, встроенные в программное обеспечение смартфонов и отдельных аппаратных продуктов, чаще всего «умных колонок», например, Google Assistant, Apple Siri, Яндекс Алиса (Яндекс.Станция), Amazon Alexa, Microsoft Cortana и др. В то же время, такие системы распознают непосредственные управляющие команды и запросы пользователя и имеют достаточно ограниченный список возможных сценариев поведения. В процессе естественной коммуникации между людьми, помимо вербальной и семантической составляющей, присутствует эмоциональный контекст. При его окраске, отличной от нейтральной, смысл фраз и истинные намерения и желания пользователя могут варьироваться. Этот факт обуславливает высокий интерес к сфере распознавания эмоций. Помимо ожидаемого повышения спроса на технологии распознавания эмоций, актуальность научных разработок в данной области подтверждается многочисленными соревнованиями (ISCA Interspeech ComParE, ACM MM Audio-Visual Emotion Challenge, ACM ICMI EmotiW и другие), специальными сессиями конференций и семинарами (IEEE PerCom Emotion Aware), а также специализированными конференциями (ACII) и журналами мирового уровня (IEEE Transactions on Affective Computing), которые посвящены этой тематике.

Помимо упомянутых выше голосовых помощников и других систем человеко-машинного взаимодействия, сферами применения технологий распознавания эмоций являются медицинское обслуживание (мониторинг состояния пациентов в медицинских учреждениях), рекомендательные системы (повышение точности рекомендаций за счет использования дополнительных источников информации), онлайн-обучение (для мониторинга вовлеченности слушателей и повышения качества обратной связи с преподавателем), «умные» пространства и окружения (расширение возможностей «умных домов» и других пространств за счет использования информации о настроении и эмоциях пользователя).

Описанные выше приложения требуют непрерывного распознавания эмоций на протяжении определенного отрезка времени. Однако большинство разработанных ранее систем распознавания эмоций работают на уровне отдельных высказываний или фраз. Значительное повышение мощности вычислительных машин, позволившее применять более сложные алгоритмы машинного обучения, а также создавать соответствующие базы данных, открыло возможность произвести постепенное

смещение фокуса научных исследований в сторону непрерывного распознавания эмоций.

Несмотря на более гибкую постановку задачи непрерывного распознавания эмоций, многие аспекты до сих пор остались без должного внимания исследователей. Одним из самых перспективных, является анализ контекста поведения пользователя. В большинстве случаев окружение пользователя, например, наличие собеседника и его эмоциональное состояние, а также место, в котором он находится, не учитывается, что приводит к потерям ценной информации. Также на настоящий момент нет однозначного ответа на вопрос об объеме данных от самого пользователя, который необходимо использовать при моделировании для достижения наилучшей точности системы распознавания эмоций.

В диссертации предлагаются методы решения вышеобозначенных проблем на примере распознавания эмоций в целом, а также для конкретного сценария применения.

Степень разработанности темы исследования. Значительный вклад в развитие технологий распознавания эмоций внесли такие исследователи, как Björn Schuller, Rosalind Picard, Maja Pantic, Shrikanth Narayanan, Elisabeth Andre, Anton Batliner, Gerhard Rigoll, Florian Eyben, Carlos Busso, Hatice Gunes, Fabien Ringeval, Michel Valstar, Heysem Kaya и другие. В частности, значительное повышение популярности непрерывного распознавания эмоций было достигнуто благодаря соревнованиям, организованными научными коллективами под руководством Björn Schuller, Fabien Ringeval и Maja Pantic. Однако, несмотря на высокую степень интереса к данной научной области и большое количество проведенных исследований, в настоящее время использование контекстной информации в системах распознавания эмоций является слабо проработанным аспектом, что тормозит развитие области в целом и разработку интеллектуальных приложений, в частности.

Целью данного исследования является повышение эффективности автоматического непрерывного распознавания эмоций человека с использованием контекстной многомодальной информации.

Для достижения данной цели в рамках диссертации были поставлены и решены следующие **задачи**:

1. Анализ и исследование современных подходов к распознаванию эмоций на основе различных представлений данных в условиях функционирования, максимально приближенных к реальным (с использованием спонтанных и непрерывных эмоций).
2. Исследование существующих методов и алгоритмов непрерывного распознавания эмоций, а также этапов дополнительной предобработки многомодальных данных.
3. Разработка и исследование методов гибкого моделирования объема контекста активного пользователя в моделях распознавания эмоций.
4. Разработка и исследование методов интеграции контекстных данных собеседника и его эмоциональных состояний в модель распознавания эмоций пользователя.
5. Разработка и исследование многомодальной системы распознавания эмоционального состояния пользователя в условиях повышенного влияния физического окружения на его настроение (туристический тур).
6. Проведение экспериментальных исследований в условиях классической и кросс-корпусной задачи, с одно- и многомодальными данными, с различными методами объединения модальностей.

Объектом исследования являются эмоциональные состояния активного пользователя (говорящего).

Предметом исследования являются контекстно-зависимые системы автоматического непрерывного распознавания эмоций человека.

Методы исследования. В диссертации применялись методы распознавания образов, машинного обучения, глубокого обучения, корреляционного и статистического анализа данных, объединения моделей и цифровой обработки сигналов.

Научная новизна диссертации отражена в следующих пунктах:

1. Предложен метод гибкого моделирования контекста активного пользователя (говорящего) на основе рекуррентных нейросетевых моделей, характеризующийся способностью обеспечить оптимальную загрузку модели непрерывного распознавания эмоций данными и добиться увеличения производительности (точности) работы системы.
2. Разработаны методы интеграции контекста собеседника, позволяющие производить его объединение с контекстом активного пользователя (говорящего) на различных этапах распознавания, отличающиеся от широко используемых современных методов применимостью в условиях непрерывности данных.
3. Разработана не имеющая аналогов многомодальная автоматическая система комплексного извлечения признаков и распознавания эмоционального состояния пользователя в условиях повышенного влияния физического окружения на его настроение.

Данные результаты соответствуют п. 5 паспорта специальности: «Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений».

Практическая значимость работы заключается в возможности использования методов, алгоритмов и моделей, разработанных в ходе диссертационного исследования, в автоматических системах непрерывного распознавания эмоций для повышения их точности.

Положения, выносимые на защиту:

1. Метод гибкого моделирования объема контекста активного пользователя (говорящего) на основе рекуррентных нейросетевых моделей.
2. Методы интеграции контекста собеседника, позволяющие производить его объединение с контекстом активного пользователя (говорящего) на различных этапах распознавания.
3. Многомодальная автоматическая система комплексного извлечения признаков и распознавания эмоционального состояния пользователя в условиях повышенного влияния физического окружения на его настроение.

Достоверность научных положений, выводов и практических рекомендаций, полученных в рамках данной диссертационной работы, подтверждается корректным обоснованием постановок задач, точной формулировкой критериев, компьютерным моделированием, результатами экспериментальных исследований, нашедших отражение в 14 публикациях в научных журналах и изданиях, индексируемых Scopus и Web of Science, а также представлением основных положений на ведущих международных конференциях.

Апробация результатов исследования. Результаты исследования представлялись для обсуждения на следующих международных научных конференциях: 19th, 20th, 21st

International Conference on Speech and Computer (SPECOM 2017, 2018, 2019); 11th International Conference on Language Resources and Evaluation (LREC 2018); IEEE International Conference on Smart Computing (SMARTCOMP 2018); Annual Conference of the International Speech Communication Association (Interspeech 2018); Workshop on Modeling Cognitive Processes from Multimodal Data, при ACM ICMI 2018; ACM International Joint Conference on Pervasive and Ubiquitous Computing; 9th International Audio/Visual Emotion Challenge and Workshop, при ACM MM 2019; IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom 2019); IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019).

Публикации. По теме диссертации было опубликовано 14 научных работ, в том числе, 14 статей опубликованы в изданиях из базы данных Scopus, 10 статей опубликованы в изданиях из базы данных Web of Science.

Личный вклад автора в работах, выполненных в соавторстве, заключается в:

- [1]: Федотов Д.В. – разработка систем контекстного распознавания эмоций, проведение экспериментов, анализ результатов (80%). Иванько Д.В. – помощь в обработке данных (10%). Сидоров М.Ю., Минкер В. – формализация задачи контекстного непрерывного распознавания эмоций (10%).
- [2]: Федотов Д.В. – сбор данных, извлечение признаков, разработка систем распознавания эмоций, проведение экспериментов, анализ результатов (50%). Матсуда Ю. – сбор данных, извлечение признаков, проведение экспериментов, анализ результатов (30%). Такахаши Ю. – сбор данных, извлечение признаков (10%). Аракава Ю., Ясумото К., Минкер В. – формализация задачи распознавания эмоций в условиях туристического тура (10%).
- [3]: Федотов Д.В. – разработка систем распознавания эмоций, проведение кросс-корпусных экспериментов, анализ результатов (70%). Кайа Х.– анализ результатов, формализация задачи кросс-корпусного распознавания эмоций (20%). Карпов А.А. – формализация задачи кросс-корпусного распознавания эмоций (10%).
- [4]: Федотов Д.В. – разработка концепции применения систем распознавания эмоций в умных окружениях (60%). Матсуда Ю. – разработка концепции применения систем распознавания эмоций в умных окружениях (30%). Минкер В. – формализация задачи распознавания эмоций в умных окружениях (10%).
- [5]: Федотов Д.В. – сбор данных, извлечение признаков, разработка систем распознавания эмоций, проведение экспериментов, анализ результатов (50%). Матсуда Ю. – сбор данных, извлечение признаков, проведение экспериментов, анализ результатов (30%). Такахаши Ю. – сбор данных, извлечение признаков (10%). Аракава Ю., Ясумото К., Минкер В. – формализация задачи распознавания эмоций в условиях туристического тура (10%).
- [6]: Федотов Д.В. – разработка систем распознавания эмоций, проведение экспериментов, анализ результатов (50%). Ким Б. – разработка систем извлечения признаков на основе сверточных нейронных сетей, проведение экспериментов, анализ результатов (40%). Карпов А.А., Минкер В. – формализация задачи распознавания эмоций на основе сверточных нейронных сетей (10%).
- [7]: Федотов Д.В. – извлечение признаков, разработка систем распознавания вовлеченности пользователей, проведение экспериментов, анализ результатов (50%). Перепелкина О. – сбор данных, извлечение признаков, анализ результатов (30%). Казимирова Е., Константинова М. – сбор данных, извлечение признаков (10%)., Минкер В. – формализация задачи распознавания вовлеченности пользователей (10%).

- [8]: Федотов Д.В. – разработка систем распознавания эмоций, проведение экспериментов, анализ результатов (80%). Сидоров М.Ю., Минкер В. – формализация задачи непрерывного распознавания эмоций (20%).
- [9]: Федотов Д.В. – разработка систем кросс-корпусного распознавания эмоций, проведение экспериментов, анализ результатов (20%). Кайа Х. – анализ данных, проведение экспериментов, анализ результатов, формализация задачи кросс-культурного распознавания эмоций, формализация задачи распознавания депрессии (30%). Дресвянский Д.В. – разработка систем кросс-культурного распознавания эмоций, проведение экспериментов (20%). Дойран М. – разработка систем распознавания депрессии, проведение экспериментов (10%). Мамонтов Д.Ю., Маркитантов М.В. – проведение экспериментов (10%). Салах А., Кавцар Е., Карпов А.А., Салах А. – формализация задачи кросс-культурного кросс-корпусного распознавания эмоций, а также задачи распознавания депрессии (10%).
- [10]: Федотов Д.В. – разработка систем кросс-корпусного и кросс-задачного распознавания эмоций, проведение экспериментов, анализ результатов (25%). Кайа Х. – анализ данных, проведение экспериментов, анализ результатов, формализация задачи кросс-корпусного и кросс-задачного распознавания эмоций (35%). Есилканат А.,– извлечение признаков, проведение экспериментов для систем кросс-корпусного и кросс-задачного распознавания эмоций (15%), Верхоляк О. – проведение экспериментов для систем кросс-корпусного и кросс-задачного распознавания эмоций (15%), Джан Я., Карпов А.А. – формализация задачи кросс-культурного кросс-корпусного распознавания эмоций (10%).
- [11]: Федотов Д.В. – сбор данных, извлечение признаков, разработка систем распознавания эмоций, проведение экспериментов, анализ результатов (35%). Матсуда Ю. – сбор данных, извлечение признаков, проведение экспериментов, анализ результатов (45%). Такахаши Ю. – сбор данных, извлечение признаков (10%). Аракава Ю., Ясумото К., Минкер В. – формализация задачи распознавания эмоций в условиях туристического тура (10%).
- [12]: Федотов Д.В. – сбор данных, извлечение признаков, разработка систем распознавания эмоций, проведение экспериментов, анализ результатов (35%). Матсуда Ю. – сбор данных, извлечение признаков, проведение экспериментов, анализ результатов (45%). Такахаши Ю. – сбор данных, извлечение признаков (10%). Аракава Ю., Ясумото К., Минкер В. – формализация задачи распознавания эмоций в условиях туристического тура (10%).
- [13]: Федотов Д.В. – сбор данных, извлечение признаков, разработка систем распознавания эмоций, проведение экспериментов, анализ результатов (35%). Матсуда Ю. – сбор данных, извлечение признаков, проведение экспериментов, анализ результатов (45%). Такахаши Ю. – сбор данных, извлечение признаков (10%). Аракава Ю., Ясумото К., Минкер В. – формализация задачи распознавания эмоций в условиях туристического тура (10%).
- [14]: Федотов Д.В. – разработка систем непрерывного распознавания эмоций, (25%). Верхоляк О. – анализ данных, извлечение признаков, разработка систем двухуровневого непрерывного распознавания эмоций и проведение экспериментов (50%), Кайа Х. – анализ данных, анализ результатов, формализация задачи двухуровнего непрерывного распознавания эмоций (15%), Джан Я., Карпов А.А. – формализация задачи двухуровнего непрерывного распознавания эмоций (10%).

Внедрение результатов работы. Результаты диссертационной работы были внедрены в учебный процесс Университета ИТМО — курс «Распознавание речи», а также использовались при проведении прикладных научных исследований:

1. НИР «Методы, модели и технологии искусственного интеллекта в биоинформатике, социальных медиа, киберфизических, биометрических и речевых системах» (проект 5-100) № 718574.
2. НИР «Разработка виртуального диалогового помощника для поддержки проведения дистанционного экзамена на основе аргументационного подхода и глубокого машинного обучения» № 619423.
3. Грант DAAD по программе «Годовые гранты для аспирантов и молодых ученых» в 2017 г..
4. Совместный грант Министерства образования и науки РФ и Германской службы академических обменов (DAAD) “Михаил Ломоносов 2018 Линия А”, госзадание 2.12795.2018/12.2.
5. Проект немецкого научно-исследовательского общества (DFG): Technology Transfer Project “Do it yourself, but not alone: Companion Technology for DIY support” of the Transregional Collaborative Research Centre SFB/TRR 62 “Companion Technology for Cognitive Technical Systems”.

Личный вклад автора состоит в выполнении представленных в диссертационной работе теоретических и экспериментальных исследований по разработке систем контекстно-зависимого распознавания эмоций. Автором проведен анализ современных подходов к решению задачи непрерывного распознавания эмоций, методов предобработки данных и извлечения признаков. На основании проведенного анализа были предложены и исследованы алгоритмы адаптивного моделирования контекста активного пользователя, а также его собеседника. Разработка компонентов и экспериментальные исследования многомодальной системы распознавания эмоционального состояния пользователей проводились при участии исследователей из СПИИРАН (Санкт-Петербург). Сбор базы данных для определения эмоционального состояния пользователя в условиях повышенного влияния физического окружения на эмоциональное состояние проводился при участии исследователей из Nara Institute of Science and Technology (Икома, Нара, Япония).

Объем и структура диссертации. Диссертационная работа состоит из введения, пяти глав, заключения, трех приложений и списка литературы. Основной материал изложен на 163 страницах, включает 16 таблиц, 86 рисунков и схем. В список использованных источников входят 223 наименований.

Содержание работы

Во **введении** формулируется актуальность исследования, рассматриваются основы компьютерной паралингвистики, контекстного моделирования, «умных» окружений (пространств) и диалоговых систем. Далее формулируются цели и задачи исследования, рассматриваются сферы применения контекстно-зависимого непрерывного распознавания эмоций, а также перечисляются положения, выносимые на защиту.

В **первой** главе представлен обзор современного состояния области автоматического распознавания психоэмоциональных состояний человека. Представлены основные подходы к построению моделей, основанные на

категориальном и непрерывном представлении данных. Далее, контекстное распознавание эмоций рассмотрено с трех основных позиций: контекста активного пользователя (говорящего), диалогового контекста (говорящий и его собеседник), а также контекста окружения. Затем представлен обзор крупнейших ежегодных соревнований по распознаванию эмоций: Interspeech ComParE, AVEC, EmotiW, с указанием изменений и устойчивых трендов в постановках задач. Данные соревнования рассматриваются как отражение развития состояния области распознавания эмоций за последнее десятилетие. Победители соревнований использовали разнообразные современные алгоритмы, и в диссертации проанализированы их подходы к решению поставленных задач. Далее в этой главе описаны основные использованные модели для распознавания эмоций: нейронные сети (полносвязные прямого распространения, сверточные, рекуррентные и с блоками длинной краткосрочной памяти), линейная регрессия с регуляризацией Тихонова, метод опорных векторов для классификации и регрессии, метод градиентного бустинга на деревьях.

Во **второй** главе представлены данные и методы, которые были использованы в диссертации, а также базовые методы предобработки данных. Описаны пять корпусов эмоционально-окрашенной речи и поведения пользователей: RECOLA (французский язык), SEMAINE (английский), SEWA (немецкий и венгерский), IEMOCAP (английский) и UUDB (японский), а также приведен краткий обзор документации по каждому из них. Далее рассмотрены следующие шаги предобработки данных: очистка сигнала от шумов и речи посторонних людей (всех, кроме говорящего), извлечение признаков, согласование аннотаций различных экспертов и коррекция их задержек. В качестве признаков в данной работе использованы экспертные наборы признаков, такие как eGeMAPS для аудиосигналов и коды лицевых движений (Facial Action Units, далее FAU) для видеосигналов, а также представления признаков, полученные с помощью моделей глубокого обучения: предобученная одномерная сверточная нейросетевая модель (Vggish) для аудиосигнала и остаточная сверточная нейросетевая модель (ResNet-50), предобученная на базе данных VGGFace2 и дообученная на базе данных AffectNet, состоящей из 450 000 фото, размеченных с помощью эмоциональных показателей. Далее представлены количественные показатели (метрики), используемые в данной работе для оценивания предложенных моделей по критерию качества распознавания.

В **третьей** главе предложен метод гибкого моделирования контекста активного пользователя на трех этапах: извлечения признаков, предобработки данных и моделирования.

На этапе извлечения признаков имеется возможность варьировать ширину окна сигнала, для которого высчитываются функционалы, тем самым, изменяя длину контекстного окна для модели. На этапе моделирования это можно производить с помощью изменения количества шагов данных, принимаемых моделью в качестве одного примера выборки. Далее в диссертации предлагается простой и эффективный метод гибкого моделирования контекста, основанный на прореживании данных, т.е. отбрасывании промежуточных значений с определенной частотой, но, при сохранении всех данных обучающей выборки за счет адаптивного сдвига между примерами выборки. Данный метод позволяет производить регулировку контекста на этапе предобработки данных, а в сочетании с описанными ранее на всех трех этапах, обеспечивая необходимую гибкость. С помощью данного подхода одно и то же значение объема контекста может быть достигнуто с помощью различных комбинаций параметров, что позволяет исключить влияние модели или набора

признаков на производительность системы, оставив контекст, как единственный фактор.

Рисунок 1 – Применение метода моделирования контекста активного пользователя на различных этапах процесса распознавания эмоций

Проведены эксперименты с применением различных способов моделирования контекста активного пользователя. Точность системы измерялась с помощью взвешенного усредненного корреляционного коэффициента согласованности (concordance correlation coefficient, CCC):

$$\mathrm{CCC}_w = \sum_{r=1}^{N} \left(w_r \times CCC(true_r, pred_r) \right) \quad (1),$$

$$\mathrm{CCC}(y, \hat{y}) = \frac{2 \times cov(y, \hat{y})}{\sigma_y^2 + \sigma_{\hat{y}}^2 + \left(\mu_y - \mu_{\hat{y}}\right)^2} \quad (2),$$

$$w_r = \frac{l_r}{\sum_{i=1}^{N} l_i} \quad (3),$$

где N – общее число записей в выборке, $true_r$ – временной ряд истинных меток для записи r, $pred_r$ – временной ряд предсказаний системы для записи r, w_r – вес значения записи r, определяемый через отношение длины записи r – l_r к общей длине записей в выборке, CCC – корреляционный коэффициент согласованности, $cov(y, \hat{y})$ – ковариация двух временных рядов, σ_y и μ_y– оценки среднеквадратичного отклонения и математического ожидания временного ряда y соответственно.

Сначала были исследованы вероятностные модели на основе методов, не учитывающих непрерывную структуру данных и, соответственно, не способные моделировать контекст самостоятельно: метод опорных векторов для регрессии (SVR), линейная регрессия с регуляризацией Тихонова (Ridge Regression), полносвязные нейронные сети прямого распространения (Feed-forward NN) и метод градиентного бустинга на деревьях (XGBoost). Настройка используемой контекстной информации происходит исключительно на этапе извлечения признаков. В качестве набора признаков были использованы eGeMAPS для аудио и FAU для видеосигналов. Были проверены значения длительности предыдущего (аудио и видео) контекста от 1 до 30 секунд и обнаружены зависимости производительности систем распознавания эмоций от объема данных, используемых для формирования каждого примера. Закономерности были обнаружены при применении всех четырех моделей, как для

видео-, так и для аудиомодальности. Для разных корпусов эмоционально окрашенного поведения пользователя значение объема контекста, обеспечивающее наивысшую производительность системы, отличается, но лежит в интервале от 5 до 20 секунд, причем, видеомодальности требуется меньший объем контекста, чем аудио. Далее рассмотрены рекуррентные нейронные сети с блоками длинной краткосрочной памяти (RNN-LSTM), в которых для регулировки контекстной информации используются количество шагов, на основе которых из изначального двухмерного массива данных формируется трехмерный массив обучающей и тестовой выборок. В качестве набора признаков были так же использованы eGeMAPS и FAU; значения контекста – от 0,1 секунды до длины, соответствующей средней продолжительности одной записи в каждом из корпусов (от 150 до 300 секунд). При применении данных моделей также наблюдается зависимость, описанная ранее, но с оптимальными значениями, смещенными в сторону меньшего контекста, что может быть обусловлено способностью этого типа нейросетевых моделей накапливать информацию о предыдущих значениях. Далее, для исключения вероятности связи обнаруженных зависимостей и набора признаков, использованы представления признаков, полученные с помощью моделей глубокого обучения: предобученная одномерная сверточная нейросетевая модель Vggish для аудиосигнала и остаточная сверточная нейросетевая модель (ResNet-50), предобученная на базе данных VGGFace2 и дообученная на базе данных AffectNet, описанные ранее. Несмотря на совершенно иную форму представления данных и отсутствие экспертных знаний, закономерности повторяют те, что были получены с использованием моделей RNN-LSTM и признаков eGeMAPS или FAU.

Далее представлены эксперименты с методом гибкого моделирования контекста, основанным на прореживании данных. Он позволяет исключить влияние непосредственно особенностей модели и количества шагов, используемых для формирования примеров, оставив временное покрытие, т.е. контекст каждого из них, как единственный фактор. Эксперименты показали, что зависимости, полученные ранее, сохраняются и указывают на одно и то же оптимальное значение объема контекста. При применении прореживания данных, изменение показателей оценки качества системы, при варьировании временного покрытия примеров, происходит более плавно по сравнению с описанным ранее подходом, где изменение происходило непосредственно за счет изменения количества данных, используемых для создания одного примера. Также данный метод позволяет использовать меньшее количество шагов для формирования примеров с идентичным временным покрытием, что увеличивает скорость обучения модели.

Далее в главе рассмотрены дополнительный способ увеличения гибкости контекстного моделирования – изменение частоты данных и ширины окна для извлечения признаков. Эксперименты, проведенные с частотой данных (векторов признаков) в 25, 12.5, 6, 3 и 1.5 Гц, показали схожие результаты, как по закономерностям, так и по значениям показателей качества системы распознавания эмоций.

Кроме этого, в главе рассмотрена применимость данного подхода к кросс-корпусному распознаванию эмоций, когда используются несколько различных корпусов. Для снижения влияния условий записи данных была применена адаптация обучающего и тестового корпусов с помощью методов главных компонент (PCA) и канонического корреляционного анализа (CCA), схема которой представлена на рисунке 2.

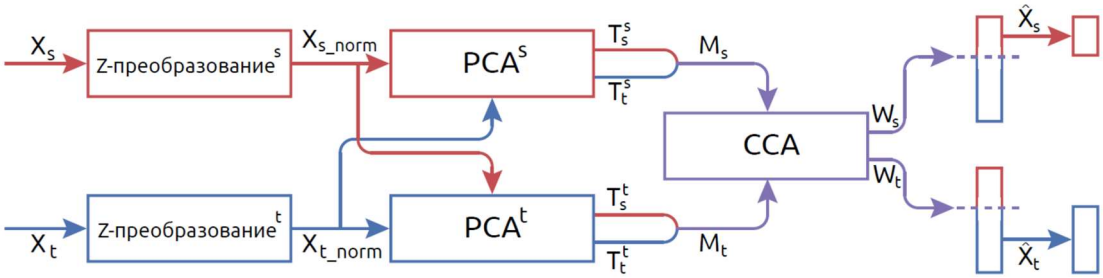


Рисунок 2 – Схема метода кросс-корпусной адаптации с использованием методов главных компонент и канонического корреляционного анализа

Результаты экспериментов показали зависимость оптимального объема контекста при кросс-корпусном обучении как от обучающего, так и от тестового корпусов, в большинстве случаев, находящегося между оптимальными значениями объема контекста для этих корпусов.

В заключение данной главы приведен краткий анализа закономерностей, а также возможных причин различий между корпусами. В частности, рассмотрена зависимость оптимального объема контекста от средней продолжительности высказываний и пауз активного пользователя для аудиомодальности, а также от количества случаев сбоя системы распознавания лиц для видеомодальности. В дополнение предложен способ регулирования контекста с помощью изменения частоты данных вместо их прореживания. Показано, что такой метод работает более стабильно, особенно при больших объемах контекста.

В **четвертой** главе представлены методы объединения данных и эмоциональных состояний активного пользователя и его собеседника для повышения точности распознавания (диалоговый контекст). Рассмотрены два метода к объединению данных: на уровне признаков (раннее объединение) и на уровне гипотез распознавания (позднее объединение), и два метода регулировки зависимости контекстных окон активного пользователя и его собеседника: зависимый (одинаковая ширина окна) и независимый (ширина окон может быть различной).

Зависимое моделирование диалогового контекста может быть проведено с помощью обоих видов объединения данных, что схематично представлено на рисунке 3.

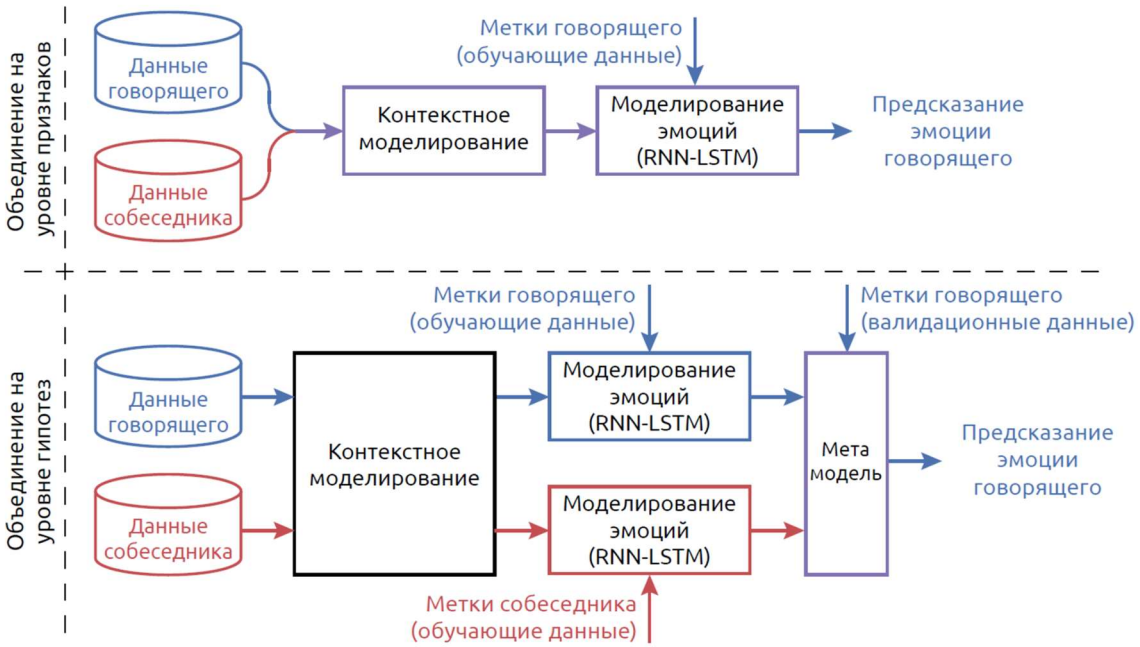


Рисунок 3 – Схемы методов зависимого моделирования диалогового контекста (сверху: объединение данных активного пользователя и его собеседника на уровне признаков; снизу: объединение данных активного пользователя и его собеседника на уровне гипотез)

Используя фиксированную архитектуру рекуррентной нейросетевой модели для установления точности базовой системы распознавания (без использования данных собеседника) и системы, использующей диалоговый контекст, есть возможность сравнить влияние моделирования данных собеседника на точность распознавания эмоций активного пользователя.

Независимое моделирование диалогового контекста с помощью объединения данных на уровне признаков в целом не представляется возможным, т.к. векторы данных в этом случае будут иметь несоответствие в одной из размерностей, отвечающей за количество шагов, используемых для формирования примера. Однако, с помощью применения прореживания данных, представленного в третьей главе, это становится возможным, и коэффициент прореживания выступает в качестве регулятора объема контекстной информации, используемой для активного пользователя и его собеседника. Независимое моделирование диалогового контекста в этом случае распадается на две модели: с фиксированной шириной контекстного окна для активного пользователя и с изменением окна для обоих участников диалога.

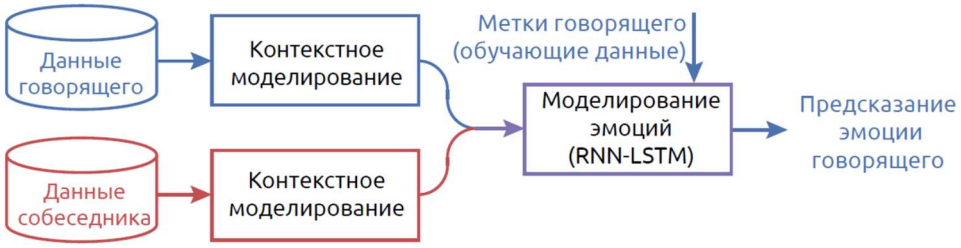


Рисунок 4 – Схема метода независимого моделирования диалогового контекста при объединении данных на уровне признаков

В первом случае для фиксации объема контекста активного пользователя было использовано оптимальное значение этого параметра, полученное в предыдущей

главе. Контекстное окно собеседника варьировалось в пределах от 1 до 60 секунд. Во втором случае ширина окна изменялась как для активного пользователя, так и для его собеседника.

Эксперименты показали увеличение точности системы в 33 из 56 случаев, где один случай представляет собой комбинацию базы данных, модальности и шкалы аннотаций. Для оценки статистической значимости был использован парный выборочный t-тест и отличия признавались значимыми при $p < 0.01$. Представленный подход показал значимое повышение точности в 12 случаях. При рассмотрении результатов по модальностям большинство из улучшений (22 случая, из которых 9 значимых) были получены для аудио; по шкалам аннотация – 19 случаев (5 значимых) для валентности и 14 случаев (7 значимых) для возбужденного состояния. Наибольшее число случаев улучшения было получено для базы IEMOCAP – в более чем 80% случаев, для SEWA и SEMAINE – в приблизительно 50% случаев, для UUDB – в 37.5% случаев. Рассматривая результаты по методам, 10 улучшений (3 значимых) было получено для полностью независимого моделирования с объединением на уровне признаков, по 8 улучшений (по 3 значимых) – для независимого моделирования с объединением на уровне признаков и фиксированным контекстным окном для активного пользователя, а также для зависимого моделирования с объединением на уровне признаков, 7 улучшений (2 значимых) – для зависимого моделирования с объединением на уровне предсказаний. Не было получено ни одного статистически значимого случая ухудшения точности распознавания эмоций.

Таким образом, независимое моделирование контекста в диалоговом сценарии оказалось наиболее результативной моделью, а также было показано, что при интеграции данных собеседника в модель любым из предложенных способов можно добиться улучшения для некоторых корпусов и модальностей. В целом, модели, основанные на аудиоданных, способны извлечь из данных собеседника по диалогу больше, чем модели на видеоданных.

В **пятой** главе понятие контекста пользователя расширяется до его «умного» окружения. Поскольку данный аспект является крайне обширным и с трудом может быть описан моделью достаточно точно, в данной диссертации выбран один конкретный пример влияния окружения на эмоции человека – посредством сценария туристического тура. Для этого в рамках сотрудничества с Nara Institute of Science and Technology (Икома, Нара, Япония) была создана экспериментальная установка, разработана новая методика сбора данных, а также адаптирована схема их аннотирования и разработана многомодальная система распознавания эмоционального состояния пользователя.

Для сбора данных использовались несколько носимых устройств – трекер движения глаз, «умный» браслет для отслеживания пульса и электрической проводимости кожи, а также миниатюрный сенсорный датчик для отслеживания поворотов головы и движений тела. Также был использован смартфон для записи коротких видеороликов и аннотирования данных. Экспериментальная установка представлена на рисунке 5.

Для сбора данных были привлечены 47 человек, которым необходимо было пройти по одному из трех туристических маршрутов, отмечая степень их удовлетворенности увиденными достопримечательностями и испытываемые эмоции. Маршруты составляли от 1.5 до 3.5 км и требовали в среднем от 50 до 110 минут для их прохождения. Два их них находились в Японии, один – в Германии. Большинство участников являлись студентами, приехавшими в страну эксперимента по обмену.

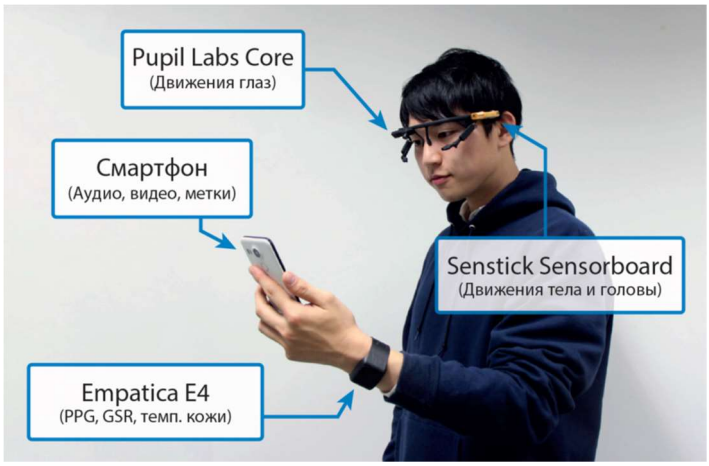


Рисунок 5 – Иллюстрация экспериментальной установки для сбора многомодальных данных о поведении пользователя

Для использования данных, полученных с перечисленных выше устройств, были разработаны алгоритмы обработки сигналов и извлечения интерпретируемых признаков, например, «поворот головы направо/налево», «частота шагов» и т.д. Также из коротких видеороликов были извлечены аудио- и видеопризнаки, описанные в предыдущих двух главах.

Далее, были обучены уни-, би-, три- и многомодальные системы на основе извлеченных признаков. Эксперименты показали, что для каждой из трех постановок задач были получены результаты, значительно превышающие случайный выбор. Унимодальные системы, обученные на признаках движения головы, а также аудиопризнаков показали наилучшие результаты для распознавания эмоций; системы, обученные на признаках движения головы и глаз – для распознавания уровня удовлетворенности; системы, обученные на аудиовизуальных признаках – для определения качества туристического опыта. Объединение модальностей на уровне признаков показало улучшение точности только для системы определения уровня удовлетворенности. Однако взвешенное объединение на уровне предсказаний показало значительно более высокие результаты по сравнению с любой из унимодальных систем, особенно для задачи распознавания эмоций. Веса, подобранные для линейной системы объединения, соответствуют точности унимодальных систем, придавая большие значения модальностям, показавшим наиболее высокие результаты. Схема данной системы представлена на рисунке 6.

В **заключении** сделаны обобщающие выводы диссертационного исследования, приведены основные результаты и возможные направления дальнейших исследований в области контекстно-зависимого непрерывного распознавания эмоций.

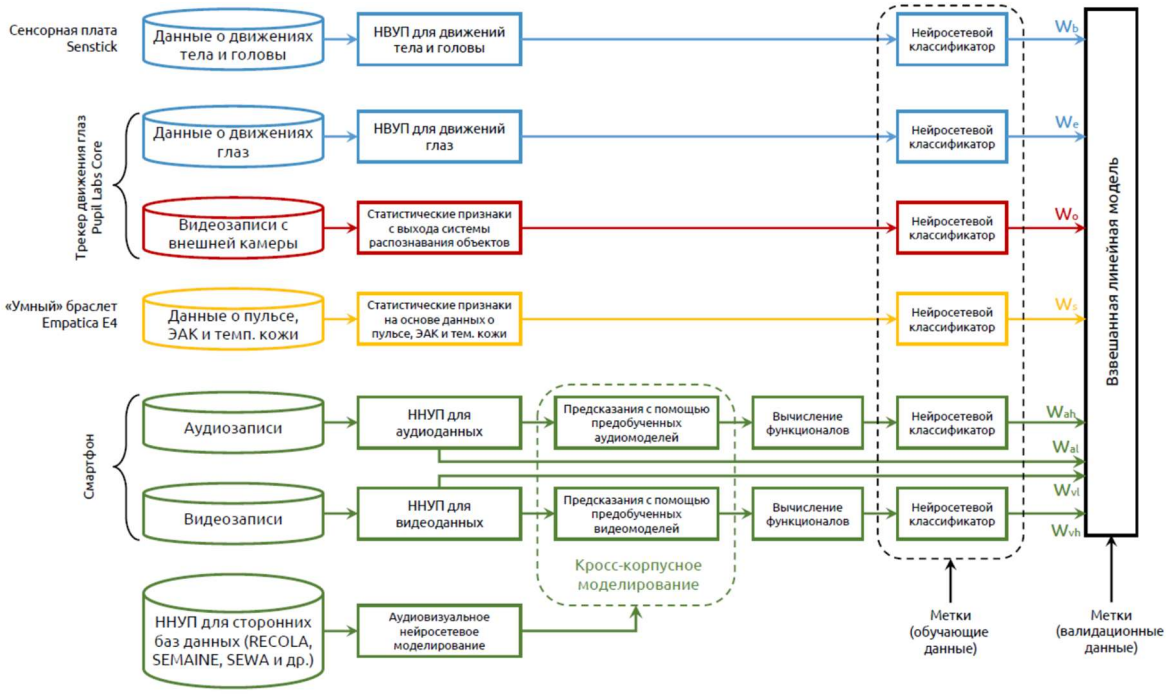


Рисунок 6 – Схема многомодальной автоматической системы комплексного извлечения признаков и распознавания эмоционального состояния пользователя в условиях повышенного влияния физического окружения на его настроение. НВУП – набор высокоуровневых признаков, ННУП – набор низкоуровневых признаков, ЭАК – электрическая активность кожи, w_x – вес для гипотез (выхода) соответствующей унимодальной системы

Заключение

Основным результатом диссертационной работы является разработка эффективных методов интеграции контекстной информации в системы автоматического непрерывного распознавания эмоций.

В рамках данной диссертационной работы были получены следующие основные теоретические и практические результаты:

1. Разработаны методы гибкого моделирования контекста активного пользователя (говорящего) на основе рекуррентных моделей для обеспечения оптимальной загрузки модели данными, позволяющей повысить производительность системы. Предложенные методы позволяют определить зависимость точности модели от использованного контекста и предлагают несколько способов произвести точную настройку его объема. Для разных корпусов эмоционально окрашенного поведения пользователя полученное значение объема контекста, обеспечивающее наивысшую производительность системы, варьируется, но лежит в интервале до 20 секунд, причем, видеомодальности требуется меньший объем контекста, чем аудио. Результаты кросс-корпусных экспериментов показали зависимость оптимального объема контекста как от обучающего, так и от тестового корпусов, в большинстве случаев, находящегося между оптимальными значениями объема контекста для этих корпусов.
2. Разработаны методы интеграции контекста говорящего, позволяющие производить объединение на двух уровнях, как после получения полного набора данных, так и в реальном времени. Также рассмотрены два метода к

объединению данных: на уровне признаков (раннее объединение) и на уровне гипотез распознавания (позднее объединение), и два метода регулировки зависимости контекстных окон активного пользователя и его собеседника: зависимый (одинаковая ширина окна) и независимый (ширина окон может быть различной). Результаты экспериментов показали, что независимое моделирование контекста в диалоговом сценарии оказалось наиболее результативной моделью, а также было показано, что при интеграции данных собеседника в модель любым из предложенных способов можно добиться улучшения для некоторых корпусов и модальностей. В целом, модели, основанные на аудиоданных, способны извлечь из данных собеседника по диалогу больше выгоды для точности, чем модели видеоданных.

3. Разработана многомодальная автоматическая система комплексного извлечения признаков и распознавания эмоционального состояния пользователя в условиях повышенного влияния физического окружения на эмоции пользователя. Данная установка протестирована в трех туристических локациях с использованием 47 участников. Результаты показали работоспособность такой системы и возможность ее применения для решения задач определения влияния физического окружения на эмоциональное состояние пользователя. Наилучшие результаты были получены на аудиовизуальных наборах признаков, а также на наборах, извлеченных из движений глаз и головы.

Список публикаций

В научных журналах и изданиях, входящих в международные реферативные базы данных Scopus и Web of Science:

1. Fedotov D., Ivanko D., Sidorov M., Minker W. Contextual Dependencies in Time-Continuous Multidimensional Affect Recognition // Proceedings of 11th International Conference on Language Resources and Evaluation, LREC 2018, pp. 1220-1224 (Scopus)
2. Fedotov D., Matsuda Y., Takahashi Y., Arakawa Y., Yasumoto K., Minker W. Towards Estimating Emotions and Satisfaction Level of Tourist based on Eye Gaze and Head Movement // Proceedings of 2018 IEEE International Conference on Smart Computing, SMARTCOMP 2018 - 2018, pp. 399-404 (Scopus, Web of Science)
3. Fedotov D., Kaya H., Karpov A. Context Modeling for Cross-Corpus Dimensional Acoustic Emotion Recognition: Challenges and Mixup // Lecture Notes in Computer Science, SPECOM 2018 - 2018, Vol. 11096, pp. 155-165 (Scopus, Web of Science)
4. Fedotov D., Matsuda Y., Minker W. From Smart to Personal Environment: Integrating Emotion Recognition into Smart Houses // IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2019 - 2019, pp. 943-948 (Scopus)
5. Fedotov D., Matsuda Y., Takahashi Y., Arakawa Y., Yasumoto K., Minker W. Towards Real-Time Contextual Touristic Emotion and Satisfaction Estimation with Wearable Devices // IEEE International Conference on Pervasive

Computing and Communications Workshops, PerCom Workshops 2019 - 2019, pp. 358-360 (Scopus)
6. Fedotov D., Kim B., Karpov A., Minker W. Time-Continuous Emotion Recognition Using Spectrogram Based CNN-RNN Modelling // Lecture Notes in Computer Science, SPECOM 2019 - 2019, Vol. 11658, pp. 93-102 (Scopus, Web of Science)
7. Fedotov D., Perepelkina O., Kazimirova E., Konstantinova M., Minker W. Multimodal approach to engagement and disengagement detection with highly imbalanced in-the-wild data // Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data, MCPMD 2018 - 2018, pp. 1-9 (Scopus)
8. Fedotov D., Sidorov M., Minker W. Context-Awared Models in Time-Continuous Multidimensional Affect Recognition // Lecture Notes in Computer Science, SPECOM 2017 - 2017, Vol. 10459, pp. 59-66 (Scopus, Web of Science)
9. Kaya H., Fedotov D., Dresvyanskiy D., Doyran M., Mamontov D., Markitantov M.V., Salah A., Kavcar E., Karpov A., Salah A. Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics // - Proceedings of the 9th International Audio/Visual Emotion Challenge and Workshop AVEC 2019, co-located with ACM Multimedia 2019 - 2019, pp. 27-35 (Scopus, Web of Science)
10. Kaya H., Fedotov D., Yesilkanat A., Verkholyak O., Zhang Y., Karpov A. LSTM based Cross-corpus and Cross-task Acoustic Emotion Recognition // Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH - 2018, pp. 521-525 (Scopus, Web of Science)
11. Matsuda Y., Fedotov D., Takahashi Y., Arakawa Y., Yasumoto K., Minker W. EmoTour: Estimating Emotion and Satisfaction of Users Based on Behavioral Cues and Audiovisual Data // Sensors, 2018, Vol. 18, No. 11, 3978 (Scopus, Web of Science)
12. Matsuda Y., Fedotov D., Takahashi Y., Arakawa Y., Yasumoto K., Minker W. Emotour: Multimodal emotion recognition using physiological and audio-visual features // - Adjunct Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the ACM International Symposium on Wearable Computers UbiComp/ISWC 2018 - 2018 , pp. 946-951 (Scopus, Web of Science)
13. Matsuda Y., Fedotov D., Takahashi Y., Arakawa Y., Yasumoto K., Minker W. Estimating User Satisfaction Impact in Cities using Physical Reaction Sensing and Multimodal Dialogue System // Lecture Notes in Electrical Engineering - 2019, Vol. 579, pp. 177-183 (Scopus, Web of Science)
14. Verkholyak O., Fedotov D., Kaya H., Zhang Y., Karpov A. Hierarchical Two-level Modelling of Emotional States in Spoken Dialog Systems // Processing of IEEE International Conference on Acoustics, Speech and Signal ICASSP 2019 - 2019, pp. 6700-6704 (Scopus, Web of Science)

Synopsis

General description of the work

Relevance. The field of information technology and, in particular, human-machine interaction systems have developed greatly over the past decades. A significant improvement in the quality of automatic speech recognition systems has made it possible to create commercially successful products that have become widespread. Examples of such systems are voice assistants built into the software of smartphones and individual hardware products, e.g. “smart speakers”, such as Google Assistant, Apple Siri, Yandex Alice, Amazon Alexa. At the same time, such systems recognize direct user commands and have a fairly limited list of possible behavior scenarios. In the process of natural communication between people, in addition to the semantic component, there is an emotional context. When colored other than “neutral”, the meaning of the phrases and the true intentions and desires of the user may deviate significantly from their purely logical meaning. This fact introduces a high interest in the field of emotion recognition. According to market analysis, by 2024 the volume is expected to increase from USD 21.6 billion (in 2019) to USD 56.0 billion, i.e. more than 2 times with a compound average annual growth rate of 21%. Also, the relevance of scientific developments in this area is confirmed by numerous competitions in emotion recognition (ISCA Interspeech ComParE, ACM MM Audio-Visual Emotion Challenge, ACM ICMI EmotiW, and others), special sessions of conferences and workshops (IEEE PerCom Emotion Aware), as well as individual conferences (ACII) and top-rated international journals (IEEE Transactions on Affective Computing) that are dedicated to this topic.

In addition to the above-mentioned voice assistants and other systems of human-machine interaction, the fields of application of emotion recognition technologies are medical care (monitoring of patient condition in medical institutions), recommendation systems (increasing the accuracy of recommendations by the introduction of additional information), online training (for monitoring student engagement and improving the quality of feedback to the teacher), “smart environments” (expanding the capabilities of “smart homes” and other environments through the use of information about the user's mood and emotions).

The applications described above require continuous emotion recognition over a certain period of time. However, most of the previously developed emotion recognition systems operate at the level of individual statements or phrases. A significant increase in computational power, which allowed the use of more complex machine learning algorithms, as well as the collection of corresponding databases, opened up the possibility of a gradual shift in the focus of scientific research towards continuous recognition of emotions.

Despite the more flexible formulation of the problem of continuous emotion recognition, many aspects have not yet been adequately addressed. One of the most promising is the

context of user behavior. In most cases, the user's environment, for example, the presence of the interlocutor and his emotional state, as well as the surroundings, are not taken into account, which leads to the loss of valuable information. Also, at the moment there is no unambiguous answer to the question of the amount of data from the user himself that must be used in modeling to achieve the best accuracy of the recognition system.

The dissertation proposes methods for solving the above-mentioned problems for emotion recognition in general, as well as for a specific application scenario.

Research topic elaboration level. Researchers such as Björn Schuller, Rosalind Picard, Maja Pantic, Shrikanth Narayanan, Elisabeth Andre, Anton Batliner, Gerhard Rigoll, Florian Eyben, Carlos Busso, Hatice Gunes, Fabien Ringeval, Michel Valstar, Heysem Kaya and others have made a significant contribution to the development of emotion recognition technologies. In particular, a significant increase in the popularity of continuous emotion recognition has been achieved through competitions organized by scientific groups of Björn Schuller, Fabien Ringeval, and Maja Pantic. However, despite the high degree of interest in this scientific field and a large number of studies conducted, the use of contextual information in emotion recognition systems is currently a poorly developed aspect, which hinders the development of the field in general and the development of high-tech applications in particular.

The **aim** of this work is to increase the performance of automatic time-continuous emotion recognition system by utilizing multimodal contextual information.

To achieve this goal, within the framework of the dissertation, the following **tasks** were set and solved:

1. Analysis of modern approaches to the recognition of emotions based on various representations of data in the close to real (spontaneous and continuous) conditions.
2. Analysis of methods and algorithms for continuous recognition of emotions, as well as stages of additional preprocessing of multimodal data.
3. Development of methods for flexible modeling of the amount of the context of an active user in emotion recognition models.
4. Development of methods for integrating the interlocutor's context and his emotional states into the model for recognizing user emotions.
5. Development of a multimodal system for recognizing the user's emotional state in the specific use case of the increased influence of physical environment on his mood (tourist tour).
6. Carrying out experimental studies in the conditions of the classical and cross-corpus problem, with single and multimodal data, with various methods of combining modalities.

The **object** of the study is emotional states of the user.

The **subject** of the study is contextual systems of automatic time-continuous emotion recognition.

Research methods. The dissertation applied methods of pattern recognition, machine learning, deep learning, correlation, and statistical data analysis, model fusion, and digital signal processing.

The **scientific novelty** of the dissertation is contained in the following:

1. Methods for flexible modeling of the context of an active user (speaker) based on recurrent neural network models have been developed, which allows for optimal loading of the time-continuous emotion recognition model with data and an increase in the performance of the system. In most modern systems, the scope of the context is not taken into account, and either the entire record is used, or it is split into parts of a certain length without justifying its choice.

2. Methods for integrating the interlocutor's context have been developed, allowing it to be combined with the context of an active user (speaker). In modern systems, the information of the interlocutor is recorded at the utterance-level. These methods allow for the data fusion for continuous recognition systems, both after receiving a complete set of data as well as in real-time. Moreover, it allows for variation of the context of the interlocutor and speaker independently of each other.
3. A multimodal feature extraction and emotion recognition system has been developed that serves to recognize the user's emotional state in the use case of an increased influence of the physical environment on his mood. Due to the use of several sources of information, this system allows for the recognition of the user's emotional state in real conditions, both by audiovisual data and by physical signs of behavior.

These points correspond to paragraph 5 of the passport of the specialty: "Development and research of models and algorithms for data analysis, detection of patterns in data and their extractions, development, and research of methods and algorithms for analyzing text, speech, and images."

The **practical relevance** of the work lies in the possibility of using the techniques developed during the dissertation research in automatic systems for continuous recognition of emotions to increase their accuracy.

Principal positions:

1. Methods of flexible modeling of the amount of the context of the active user (speaker) based on recurrent neural network models. It allows to ensure optimal loading of the model with data to increase the performance of the recognition system.
2. Methods for integrating the interlocutor's context, allowing it to be combined with the context of the active user (speaker) at two levels, both after receiving a complete set of data, and in real-time. Moreover, it allows for variation of the context of the interlocutor and speaker independently of each other.
3. Multimodal analysis system, which serves to recognize the user's emotional state in the use case of an increased influence of the physical environment on his mood.

The **credibility of the principal provisions, conclusions, and practical recommendations** obtained within the framework of this dissertation work is confirmed by the correct problem statements, the exact formulation of criteria, computer modeling, the results of experimental research, reflected in 14 publications in scientific journals and publications indexed by Scopus and Web of Science as well as presenting the main points at leading international conferences.

Approbation of research results. The research results were presented for discussion at the following international scientific conferences: 19th, 20th, 21st International Conference on Speech and Computer (SPECOM 2017, 2018, 2019); 11th International Conference on Language Resources and Evaluation (LREC 2018); IEEE International Conference on Smart Computing (SMARTCOMP 2018); Annual Conference of the International Speech Communication Association (Interspeech 2018); Workshop on Modeling Cognitive Processes from Multimodal Data, at ACM ICMI 2018; ACM International Joint Conference on Pervasive and Ubiquitous Computing; 9th International Audio / Visual Emotion Challenge and Workshop, at ACM MM 2019; IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom 2019); IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2019).

Publications. On the topic of the dissertation, 14 scientific papers were published, including 14 articles indexed by the Scopus database, and 10 articles indexed by the Web of Science database.

The **personal contribution** of the author in co-authored publications:

- [1]: Fedotov D.V. – development of systems for contextual recognition of emotions, conducting experiments, analyzing the results (80%). Ivanko D.V. – assistance in data processing (10%). Sidorov M.Yu., Minker W. – formalization of the problem of contextual continuous recognition of emotions (10%).
- [2]: Fedotov D.V. – data collection, feature extraction, development of emotion recognition systems, experiments, analysis of results (50%). Matsuda Y. – data collection, feature extraction, experiments, results analysis (30%). Takahashi Y. – data collection, feature extraction (10%). Arakawa Y., Yasumoto K., Minker W. – formalization of the problem of emotion recognition in a tourist tour (10%).
- [3]: Fedotov D.V. – development of emotion recognition systems, cross-corpus experiments, analysis of results (70%). Kaya H. – analysis of results, formalization of the problem of cross-corpus recognition of emotions (20%). Karpov A.A. – formalization of the problem of cross-corpus recognition of emotions (10%).
- [4]: Fedotov D.V. – development of the concept of using emotion recognition systems in smart environments (60%). Matsuda Y. – development of the concept of applying emotion recognition systems in smart environments (30%). Minker W. – formalization of the problem of emotion recognition in smart environments (10%).
- [5]: Fedotov D.V. – data collection, feature extraction, development of emotion recognition systems, experiments, analysis of results (50%). Matsuda Y. – data collection, feature extraction, experiments, results analysis (30%). Takahashi Y. – data collection, feature extraction (10%). Arakawa Y., Yasumoto K., Minker W. – formalization of the problem of emotion recognition in a tourist tour (10%).
- [6]: Fedotov D.V. – development of emotion recognition systems, experiments, analysis of results (50%). Kim B. – development of features extraction systems based on convolutional neural networks, testing of experiments, analysis of results (40%). Karpov A.A., Minker W. – formalization of the emotion recognition problem based on convolutional neural networks (10%).
- [7]: Fedotov D.V. – feature extraction, development of user engagement recognition systems, experiments, analysis of results (50%). Perepelkina O. – data collection, feature extraction, analysis of results (30%). Kazimirova E., Konstantinova M. – data collection, feature extraction (10%)., Minker W. – formalization of the user engagement recognition problem (10%).
- [8]: Fedotov D.V. – development of emotion recognition systems, experiments, analysis of results (80%). Sidorov M.Yu., Minker W. – formalization of the problem of continuous recognition of emotions (20%).
- [9]: Fedotov D.V. – development of systems for cross-corpus recognition of emotions, conducting experiments, analyzing the results (20%). Kaya H. – data analysis, experiments, analysis of results, formalization of the task of cross-cultural recognition of emotions, formalization of the task of recognizing depression (30%). Dresvyanskiy D.V. – development of systems for cross-cultural recognition of emotions, conducting experiments (20%). Doyran M. – development of depression recognition systems, conducting experiments (10%). Mamontov D.Yu., Markitantov M.V. – conducting experiments (10%). Salah A., Kavcar E., Karpov A.A., Salah A. – formalization of the problem of cross-cultural cross-corpus recognition of emotions, as well as the problem of recognizing depression (10%).
- [10]: Fedotov D.V. – development of systems for cross-corpus and cross-task recognition of emotions, conducting experiments, analyzing the results (25%). Kaya H. – data analysis, experiments, analysis of results, formalization of the problem of cross-corpus and cross-task recognition of emotions (35%). Yesilkanat A., – feature

extraction, experiments for cross-corpus and cross-task emotion recognition systems (15%), Verkholyak O. – conducting experiments for cross-corpus and cross-task emotion recognition systems (15%), Zhang Y., Karpov A.A. – formalization of the problem of cross-cultural cross-corpus recognition of emotions (10%).

- [11]: Fedotov D.V. – data collection, feature extraction, development of emotion recognition systems, experiments, analysis of results (35%). Matsuda Y. – data collection, feature extraction, experiments, results analysis (45%). Takahashi Y. – data collection, feature extraction (10%). Arakawa Y., Yasumoto K., Minker W. – formalization of the problem of emotion recognition in a tourist tour (10%).
- [12]: Fedotov D.V. – data collection, feature extraction, development of emotion recognition systems, experiments, analysis of results (35%). Matsuda Y. – data collection, feature extraction, experiments, results analysis (45%). Takahashi Y. – data collection, feature extraction (10%). Arakawa Y., Yasumoto K., Minker W. – formalization of the problem of emotion recognition in a tourist tour (10%).
- [13]: Fedotov D.V. – data collection, feature extraction, development of emotion recognition systems, experiments, analysis of results (35%). Matsuda Y. – data collection, feature extraction, experiments, results analysis (45%). Takahashi Y. – data collection, feature extraction (10%). Arakawa Y., Yasumoto K., Minker W. – formalization of the problem of emotion recognition in a tourist tour (10%).
- [14]: Fedotov D.V. – development of systems for continuous recognition of emotions, (25%). O. Verkholyak – data analysis, feature extraction, development of systems for two-level continuous emotion recognition and conducting experiments (50%), Kaya H. – data analysis, analysis of results, formalization of the problem of two-level continuous emotion recognition (15%), Zhang Y.., Karpov A.A. – formalization of the problem of two-level continuous recognition of emotions (10%).

Implementation of work results. The results of the dissertation work were introduced into the educational process of the ITMO University – the course "Speech Recognition", and were also used in applied scientific research:

1. Research work “Methods, models, and technologies of artificial intelligence in bioinformatics, social media, cyber-physical, biometric and speech systems” (project 5-100) No. 718574;
2. Research work "Development of a virtual dialogue assistant to support the conduct of a distance exam based on an argumentation approach and deep machine learning" No. 619423;
3. DAAD grant under the program "Annual grants for graduate students and young scientists" in 2017;
4. Joint grant of the Ministry of Education and Science of the Russian Federation and the German Academic Exchange Service (DAAD) “Mikhail Lomonosov 2018 Line A”, state assignment 2.12795.2018 / 12.2;
5. Project of the German Research Society (DFG): Technology Transfer Project “Do it yourself, but not alone: Companion Technology for DIY support” of the Transregional Collaborative Research Center SFB / TRR 62 “Companion Technology for Cognitive Technical Systems”.

The **personal contribution** of the author is the implementation of theoretical and experimental studies presented in the dissertation work on the development of context-dependent emotion recognition systems. The author analyzes modern approaches to solving the problem of continuous recognition of emotions, methods of data processing, and feature extraction. Based on the analysis, algorithms for adaptive modeling of the context of an

active user, as well as his interlocutor, were proposed and investigated. The collection of a database to determine the user's emotional state in the use case of an increased influence of the physical environment on the emotional state was carried out with the participation of researchers from the Nara Institute of Science and Technology (Ikoma, Nara, Japan).

Thesis structure. The dissertation work consists of an introduction, five chapters, a conclusion, three appendices and a bibliography. The material is presented on 163 pages, includes 16 tables, 86 figures and diagrams. The list of sources used includes 223 items.

The content of the work

The **introduction** formulates the relevance of the research, examines the basics of computer paralinguistics, contextual modeling, smart environments and dialogue systems. Further, the goals and objectives of the study are formulated, the scope of application of context-dependent continuous recognition of emotions is considered, and the main contributions are listed.

The **first chapter** provides an overview of the current state of the field of automatic recognition of human emotional states. The main approaches to building models based on categorical and continuous data presentation are presented. Further, the contextual recognition of emotions is considered from three main positions: the context of the active user (the speaker), the dialogue context (the speaker and his interlocutor), and the context of the environment. Then an overview of the largest annual emotion recognition challenges is presented: Interspeech ComParE, AVEC, EmotiW, indicating changes and constant trends in problem setting. These competitions are considered as a reflection of the development of the state of the field of emotion recognition over the past decade. The winners of the competition used a variety of modern algorithms, and their dissertation analyzes their approaches to solving the assigned tasks. Further in this chapter, the main models used for emotion recognition are described: neural networks (fully connected feedforward, convolutional, recurrent and long short-term memory), linear regression with Tikhonov's regularization, support vector machines for classification and regression, gradient boosting on trees.

The **second chapter** presents the data and methods that were used in the thesis, as well as basic data preprocessing methods. Five corpora of emotionally colored speech and user behavior are described: RECOLA (French), SEMAINE (English), SEWA (German and Hungarian), IEMOCAP (English) and UUDB (Japanese), as well as a brief overview of the literature on each of them. Further, the following steps of data preprocessing are considered: cleaning the signal from noise and speech of other people (everyone except the speaker), extracting features, annotation alignment, and reaction lag correction. As features we used expert datasets, such as eGeMAPS for audio signals and Facial Action Units (FAU) for video signals, as well as feature representations obtained using deep learning models: a pretrained one-dimensional convolutional neural network model Vggish for audio signal and residual convolutional neural network model (ResNet-50), pretrained on the VGGFace2 database and retrained on the AffectNet database, which consists of 450 000 photos marked with emotional indicators. Further, quantitative indicators (metrics) used in this work to assess the proposed models by the criterion of recognition quality are presented.

The **third chapter** presents an approach to flexible modeling of the active user context in three stages: feature extraction, data preprocessing, and modeling.

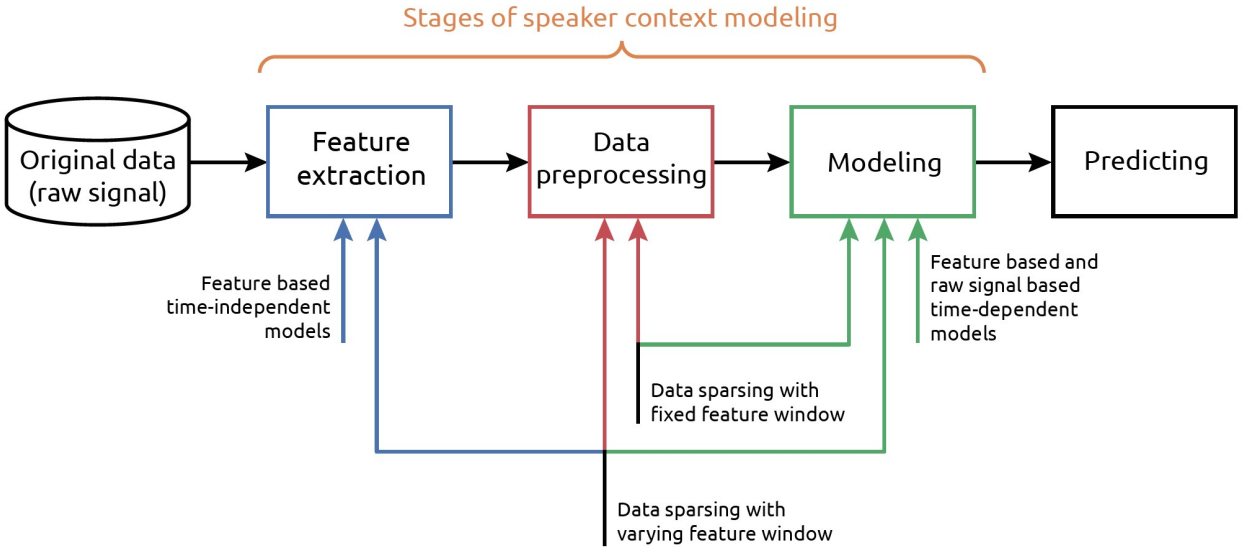


Figure 1 – Speaker context modeling in general pipeline for emotion recognition system

At the stage of feature extraction, it is possible to vary the width of the window for which the functionals are calculated, thereby changing the length of the context window for the model. At the modeling stage, this can be done by changing the number of time steps fed into the model as one sample. Further in the thesis, a simple and effective approach to flexible modeling of the context based on data sparsing is proposed, i.e. skipping intermediate values with a certain frequency, but while preserving all data of the training sample with an adaptive shift between samples. This approach allows adjusting the context at the stage of data preprocessing. In combination with approaches described earlier, one may vary context at all three stages, providing the necessary flexibility. With this methodology, the same value of the amount of context can be achieved using different combinations of parameters, which makes it possible to exclude the influence of the model or set of features on the performance of the system, leaving the context as the only factor.

Experiments have been carried out using various methods of modeling the context of an active user. System accuracy was measured using a weighted average concordance correlation coefficient (CCC):

$$\mathrm{CCC}_w = \sum_{r=1}^{N} \left(w_r \times CCC(true_r, pred_r) \right) \quad (1),$$

$$\mathrm{CCC}(y, \hat{y}) = \frac{2 \times cov(y, \hat{y})}{\sigma_y^2 + \sigma_{\hat{y}}^2 + \left(\mu_y - \mu_{\hat{y}}\right)^2} \quad (2),$$

$$w_r = \frac{l_r}{\sum_{i=1}^{N} l_i} \quad (3),$$

where N – is the total number of recordings in the subset, $true_r$ – is the time series of true labels for the recording r, $pred_r$ – is the time series of the system's predictions for the recording r, w_r – is the weight of the record value r, determined through the ratio of the record's r length l_r to the total length of recordings in the sample, CCC – is the concordance correlation coefficient, $cov(y, \hat{y})$ – is the covariance of two time series, σ_y and μ_y – are the

estimates of the standard deviation and the mathematical expectation of the time series *y*, respectively.

At first, probabilistic models were used based on methods that do not take into account the continuous data structure and, therefore, are not able to model the context on their own: support vector regression (SVR), linear regression with Tikhonov's regularization (Ridge Regression), fully connected neural networks (feed-forward) and gradient boosting on trees (XGBoost). The contextual information is configured exclusively at the stage of feature extraction. eGeMAPS for audio and FAU for the video were used as feature sets. We checked the duration values of the audio and video context from 1 to 30 seconds and found the dependence of the performance of emotion recognition systems on the amount of data used to generate each sample. Patterns were found across all four models for both video and audio modality. For different corpora, the value of the amount of context that provides the highest system performance differs, but lies in the range from 5 to 20 seconds, and video modality requires less context than audio. Next, we consider recurrent neural networks with blocks of long short-term memory (RNN-LSTM), in which the number of steps is used to adjust the context information, which is used to form a three-dimensional array of training and test samples from the initial two-dimensional data array. eGeMAPS and FAU were also used as a feature set; context values - from 0.1 seconds to the full length of recording (corresponding to the average duration of one record in each of the corpora – from 150 to 300 seconds). When applying these models, the dependence described earlier is also observed, but with the optimal values shifted towards a smaller context values, which may be due to the ability of this type of neural network models to accumulate information about previous values. Further, to exclude the chance of a connection between the detected dependencies and a set of features, feature representations obtained using deep learning models were used: a pretrained one-dimensional convolutional neural network model Vggish for an audio signal and a residual convolutional neural network model (ResNet-50), pretrained on the VGGFace2 database and fine-tuned on the AffectNet database described earlier. Despite the completely different form of data representation and lack of expert knowledge in features, the patterns repeat those obtained using the RNN-LSTM models and eGeMAPS or FAU feature sets.

Further, we present experiments with a flexible context modeling technique based on data sparsing. It allows to exclude the influence of the model peculiarities and the number of steps used to generate examples, leaving context coverage as the only factor. Experiments have shown that the dependencies obtained earlier are preserved and indicate the same optimal context value. When using data sparsing, the change in the system performance with varyations the context coverage of examples, occurs more smoothly compared to the previously described approach, where the change occurred directly due to the change in the amount of data used to create one sample. Also, this method allows using fewer steps to generate examples with identical context coverage, which increases the speed of model training.

Further in the chapter, an additional way to increase the flexibility of context modeling is considered – changing the data frequency and window width for feature extraction. Experiments carried out with the data frequency of 25, 12.5, 6, 3 and 1.5 Hz showed similar results, both in terms of patterns and performance criterion of the emotion recognition system.

In addition, in this chapter we discuss the applicability of this approach to cross-corpus emotion recognition when several different corpora are used.

To reduce the influence of data recording conditions, a domain adaptation of the training and test corpus was applied using the methods of principal component analysis and canonical correlation analysis (PCA-CCA), the diagram of which is shown in Figure 2.

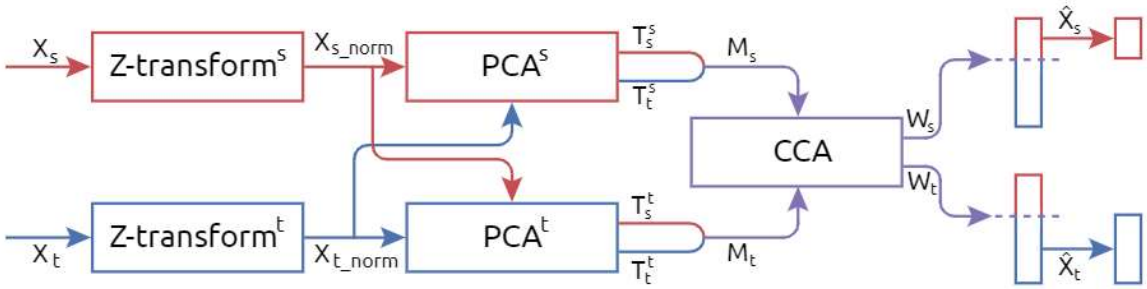


Figure 2 – PCA-CCA approach to cross-corpus domain adaptation

The results of the experiments showed a strong dependence of the optimal amount of context in cross-corpus learning. This amount lies between the optimal context for training and test corpora in most cases.

This chapter concludes with a brief analysis of the patterns and possible reasons for the differences between the corpora. In particular, the dependence of the optimal amount of context on the average duration of utterances and pauses of an active user for audio modality, as well as on the number of failures of the face recognition system for video modality, is considered. In addition, a method is proposed for adjusting the context by changing the data frequency instead of data sparsing. It is shown that this method works more stably, especially with a larger amount of context.

The **fourth chapter** presents strategies for combining data and emotional states of an active user and his interlocutor to improve the recognition accuracy (dialog context). Two methods for combining data are considered: feature-level (early fusion) and at the decision-level (late fusion), and two methods for the dependence of the context windows of the active user and his interlocutor: dependent (the same window width) and independent (the window width can be different).

Dependent modeling of the dialog context can be carried out using both types of data fusion, which is schematically shown in Figure 3.

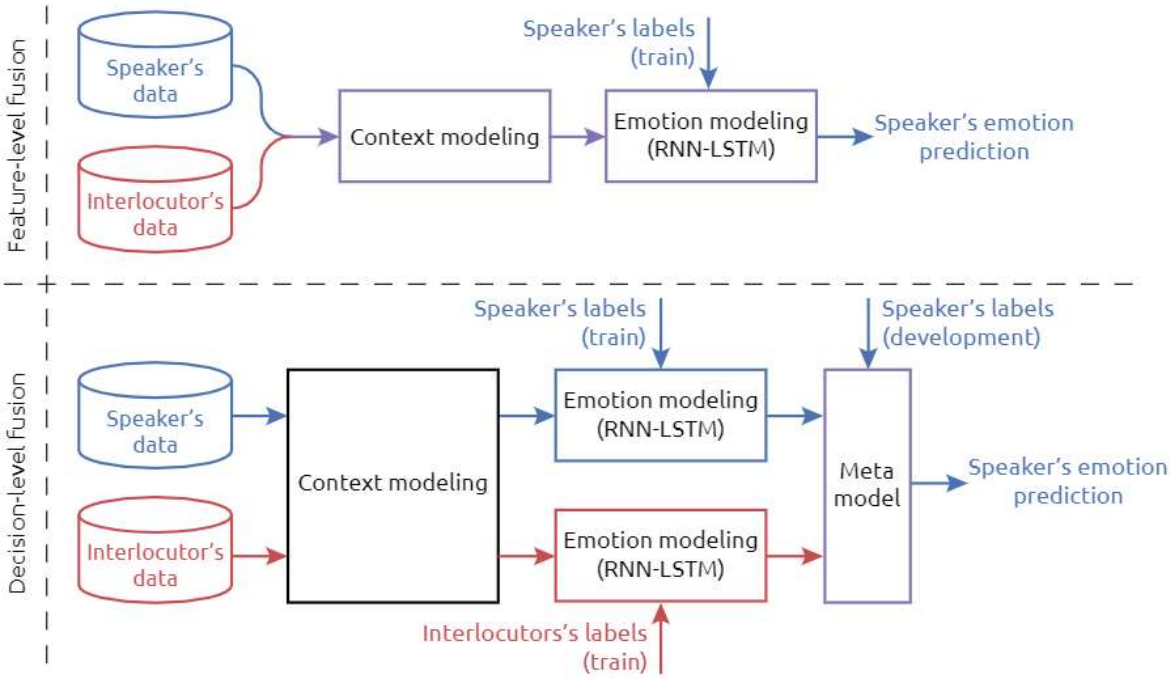


Figure 3 – Pipeline of dependent dyadic context modeling

Using the fixed architecture of a recurrent neural network model to a baseline performance (without using the interlocutor's data), it is possible to compare the effect of integrating the interlocutor's context on the performance of the emotion recognition system for the active user.

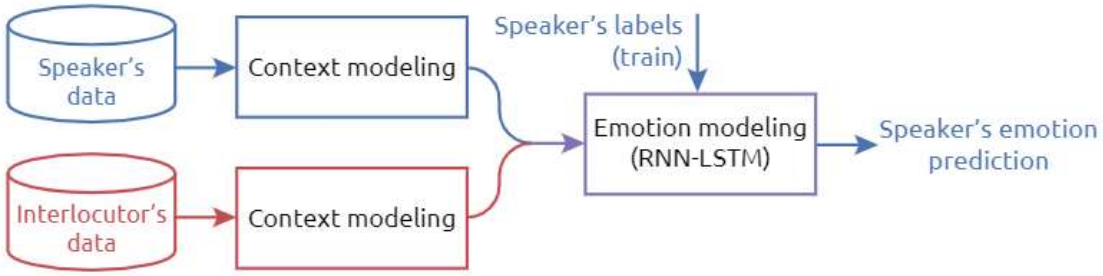


Figure 4 – Pipeline of independent dyadic context modeling

Independent modeling of the dialogue context by combining data at the feature level is basically not possible, since it will raise an issue of mismatch in the data vectors dimension responsible for the number of steps used to generate the example. However, with the use of data sparsing presented in Chapter 3, this becomes possible and the sparsing coefficient acts as a regulator of the amount of contextual information used for the active user and his interlocutor. In this case, independent modeling of the dialogue context splits into two options: with a fixed width of the context window for the active user and with a change in the window for both participants in the dialogue.

In the first case, the optimal value obtained in the previous chapter was used to fix the volume of the active user context. The interlocutor's contextual window varied from 1 to 60 seconds. In the second case, the window width was changed for both the active user and his interlocutor.

Wrapping the results up, we achieved performance gain with applied approaches in 33 out of 56 cases, where one case is one of four presented approaches applied to certain modality-dimension pair of particular database. For significance check, we use paired sample t-test and consider differences compared to speaker-only baseline significant if $p < 0.01$. Our approach resulted in 12 statistically significant cases. If we consider them modality-wise, most of them (22 cases, 9 significant) were obtained with audio features; dimension-wise – 19 cases (5 significant) for valence and 14 (7 significant) for arousal. The highest percentage of improvement achieved on IEMOCAP database – in more than 80% of the application cases of dyadic context modeling provided an improvement over the baseline, on continuously annotated SEWA and SEMAINE it was in approximately 50% of the cases and the worst results are for UUDB with 37.5% of the cases. Considering approaches used, the highest number of improvements (10, 3 significant) was obtained with fully independent context modeling with FLF, 8 (3 significant) – with independent context modeling with FLF and fixed context for of speaker, 8 (4 significant) – with dependent context modeling with DLF and, and finally 7 (2 significant) with dependent context modeling with FLF. There are no significant cases for performance decrease, therefore, utilization of proposed approach either increases the quality of emotion recognition system or does not affect it in a negative way.

Thus, independent modeling of the context in the dialogue scenario turned out to be the most effective model, and it was also shown that by integrating the interlocutor's data into the model in any of the proposed ways, improvements can be achieved for some corpora and

modalities. In general, models based on audio data are able to extract more from the data of a conversation partner than models based on video data.

In the **fifth chapter**, the concept of user context is extended to his environment. Since this aspect is extremely extensive and can hardly be described by the model accurately enough, in this dissertation one specific use case of the influence of the environment on human emotions is selected – a sightseeing tour. For this, in the scope of cooperation with the Nara Institute of Science and Technology (Ikoma, Nara, Japan), an experimental setup, a data collection method was created, and also adapted for annotation, and a multimodal system for feature extraction and emotion recognition was developed.

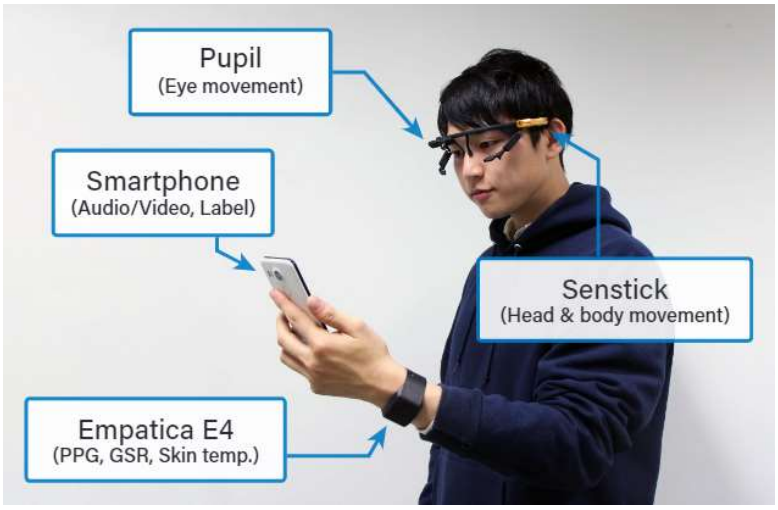


Figure 5 – Device setup for EmoTourDB

Several wearable devices were used to collect the data - an eye tracker, a smart wristband to track heart rate and electrical conduction of the skin, and a miniature sensor to track head turns and body movements. A smartphone was also used to record short videos and annotate data. The experimental setup is shown in Figure 5.

We have collected the data from 47 participants, who had to go along one of three tourist routes, noting the degree of their satisfaction, emotions and specially designed labels – touristic experience quality. The routes ranged from 1.5 to 3.5 km and it took 50 to 110 minutes on average to complete them. Two of them were in Japan, one in Germany. Most of the participants were exchange students in the country of the experiment.

To use the data obtained from the devices listed above, algorithms have been developed for processing raw signals and extracting meaningful features, for example, "head turn to the right / left", "pace", etc. Also, the audio and video features described in the previous two chapters were extracted from short video clips.

Further, we trained systems on each available modality, logically combining them into bi- and trimodal models using feature-level fusion, as well as multimodal system with all available modalities in feature-level and decision-level fusion setup. Results showed performance significantly over chance-level, on each task. Unimodal systems trained on head tilt and audio features, showed highest performance for emotion recognition; trained on head tilt and eye movements – for satisfaction estimation; and trained on audio-visual features – for touristic experience quality estimation. Feature-level fusion of all modalities showed performance gain over the best unimodal system only for satisfaction estimation. However, weighted decision-level fusion showed much higher results, especially for emotion recognition. Weights of built linear meta system cohere with unimodal results, favouring the top performing feature sets and models. This system is presented in Figure 6.

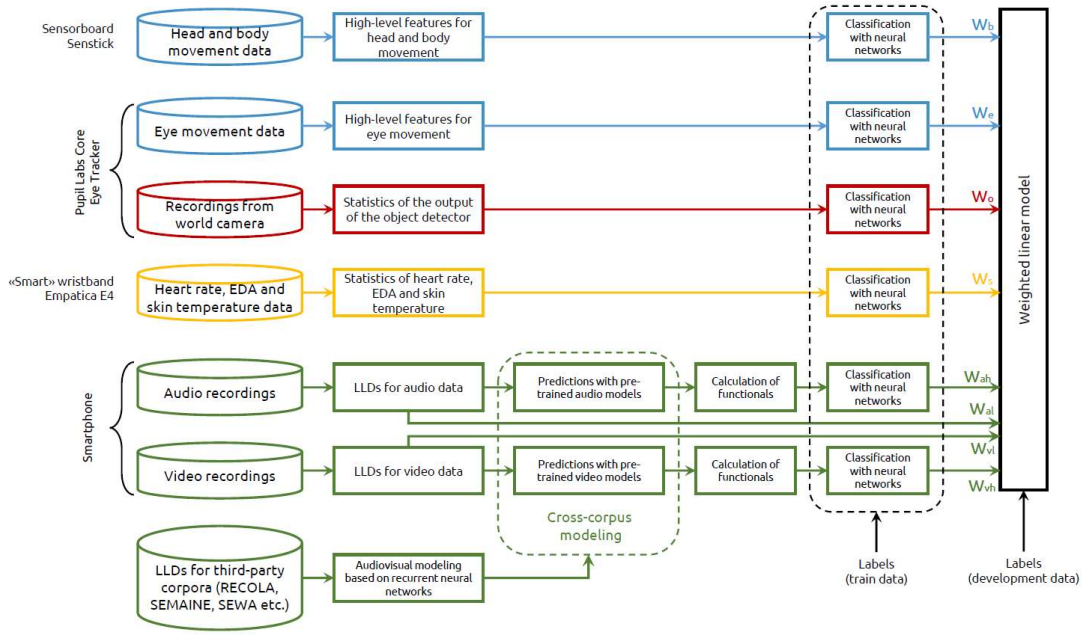


Figure 6 – Multimodal feature extraction and emotion recognition system (decision level fusion). LLDs stand for low level descriptors, EDA – electo-dermal activity, w_x is a weight corresponding to prediction of a particular unimodal system

In the **conclusions**, generalizing conclusions of the dissertation research are presented, the main results and possible directions for further research in the field of context-dependent continuous recognition of emotions are considered.

Conclusions

The main contribution of the thesis is the development of effective methods for integrating contextual information in the systems of automatic continuous emotion recognition.

Within the framework of this dissertation work, the following main theoretical and practical results were obtained:

1. Methods for flexible modeling of the context of an active user (speaker) based on recurrent models have been developed to ensure optimal loading of the model with data, which allows increasing system performance. The proposed methods make it possible to determine the dependence of the model performance on the used context and offer several ways to flexible fine-tuning.
2. Methods have been developed for integrating the context of the speaker. These methods allow for the data fusion for continuous recognition systems, both after receiving a complete set of data as well as in real-time. Moreover, it allows for variation of the context of the interlocutor and speaker independently of each other. Independent modeling of the active user and speaker context provides additional flexibility in customizing models.
3. A multimodal analysis system has been developed that serves to recognize the user's emotional state in the use case of an increased influence of the physical environment on emotions. This setup was tested in three tourist locations with 47 participants. The results showed the efficiency of such a system and the possibility of its application

for solving problems of determining the influence of the physical environment on the emotional state of the user.

List of publications

Publications indexed by the Scopus and Web of Science databases.

1. Fedotov D., Ivanko D., Sidorov M., Minker W. Contextual Dependencies in Time-Continuous Multidimensional Affect Recognition // Proceedings of 11th International Conference on Language Resources and Evaluation, LREC 2018, pp. 1220-1224 (Scopus)
2. Fedotov D., Matsuda Y., Takahashi Y., Arakawa Y., Yasumoto K., Minker W. Towards Estimating Emotions and Satisfaction Level of Tourist based on Eye Gaze and Head Movement // Proceedings of 2018 IEEE International Conference on Smart Computing, SMARTCOMP 2018 - 2018, pp. 399-404 (Scopus, Web of Science)
3. Fedotov D., Kaya H., Karpov A. Context Modeling for Cross-Corpus Dimensional Acoustic Emotion Recognition: Challenges and Mixup // Lecture Notes in Computer Science, SPECOM 2018 - 2018, Vol. 11096, pp. 155-165 (Scopus, Web of Science)
4. Fedotov D., Matsuda Y., Minker W. From Smart to Personal Environment: Integrating Emotion Recognition into Smart Houses // IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2019 - 2019, pp. 943-948 (Scopus)
5. Fedotov D., Matsuda Y., Takahashi Y., Arakawa Y., Yasumoto K., Minker W. Towards Real-Time Contextual Touristic Emotion and Satisfaction Estimation with Wearable Devices // IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2019 - 2019, pp. 358-360 (Scopus)
6. Fedotov D., Kim B., Karpov A., Minker W. Time-Continuous Emotion Recognition Using Spectrogram Based CNN-RNN Modelling // Lecture Notes in Computer Science, SPECOM 2019 - 2019, Vol. 11658, pp. 93-102 (Scopus, Web of Science)
7. Fedotov D., Perepelkina O., Kazimirova E., Konstantinova M., Minker W. Multimodal approach to engagement and disengagement detection with highly imbalanced in-the-wild data // Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data, MCPMD 2018 - 2018, pp. 1-9 (Scopus)
8. Fedotov D., Sidorov M., Minker W. Context-Awared Models in Time-Continuous Multidimensional Affect Recognition // Lecture Notes in Computer Science, SPECOM 2017 - 2017, Vol. 10459, pp. 59-66 (Scopus, Web of Science)
9. Kaya H., Fedotov D., Dresvyanskiy D., Doyran M., Mamontov D., Markitantov M.V., Salah A., Kavcar E., Karpov A., Salah A. Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics // Proceedings of the 9th International Audio/Visual Emotion Challenge and Workshop AVEC 2019, co-located with ACM Multimedia 2019 - 2019, pp. 27-35 (Scopus, Web of Science)

10. Kaya H., Fedotov D., Yesilkanat A., Verkholyak O., Zhang Y., Karpov A. LSTM based Cross-corpus and Cross-task Acoustic Emotion Recognition // Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH - 2018, pp. 521-525 (Scopus, Web of Science)
11. Matsuda Y., Fedotov D., Takahashi Y., Arakawa Y., Yasumoto K., Minker W. EmoTour: Estimating Emotion and Satisfaction of Users Based on Behavioral Cues and Audiovisual Data // Sensors - 2018, Vol. 18, No. 11, 3978 (Scopus, Web of Science)
12. Matsuda Y., Fedotov D., Takahashi Y., Arakawa Y., Yasumoto K., Minker W. Emotour: Multimodal emotion recognition using physiological and audio-visual features // Adjunct Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the ACM International Symposium on Wearable Computers UbiComp/ISWC 2018 - 2018 , pp. 946-951 (Scopus, Web of Science)
13. Matsuda Y., Fedotov D., Takahashi Y., Arakawa Y., Yasumoto K., Minker W. Estimating User Satisfaction Impact in Cities using Physical Reaction Sensing and Multimodal Dialogue System // Lecture Notes in Electrical Engineering - 2019, Vol. 579, pp. 177-183 (Scopus, Web of Science)
14. Verkholyak O., Fedotov D., Kaya H., Zhang Y., Karpov A. Hierarchical Two-level Modelling of Emotional States in Spoken Dialog Systems // Processing of IEEE International Conference on Acoustics, Speech and Signal ICASSP 2019 - 2019, pp. 6700-6704 (Scopus, Web of Science)

1 Introduction

In this introductory chapter we will first present a brief overview of emotion recognition problem, then focus on definition of context used throughout this thesis, as well as concept of smart environment and dialogue systems. After this, we will cover our motivation of conducting research in this area and possible applications. Finally, we will conclude with a short summary of thesis contributions and an outline.

1.1 Emotion Recognition

Automatic emotion recognition (AER) is a process of identifying human emotions with some computational model using input data. As many other machine learning tasks, AER can be based on several modalities, such as speech, facial expressions, textual data, user behaviour, etc. Moreover, there are several annotation schemes widely used in research and numerous models known to perform well at this task. Emotion recognition is of interest for researchers for a long period of time already and during the last two decades it received a noticeable development due to hardware and software improvements, as well as the increasing demand on intelligent conversational agents.

One of the distinctive features of AER compared to other machine learning tasks, such as automatic speech recognition, age or gender recognition, is a high subjectivity of labels (emotions). It plays an important role on two levels: while expressing emotions and while annotating them. Earlier studies on emotion recognition used acted corpora for training their systems. Such corpora include recordings in which a person is acting a particular emotion according to a predefined script. In some cases, such recordings are based on the appropriate scenario (e.g. a script of "angry" recording includes words associated with this emotion), and in others – the same content (e.g. a phrase) is repeated with different emotions. Here, subjectivity takes place during the dataset collection, as participants may express the same emotions differently. More recent studies are often focused on spontaneous or sometimes even in-the-wild data. It is based on natural emotions of a user, without any predefined script. Here, the subjectivity problem arises when recordings are being annotated, as various raters may perceive the same emotions contained in a particular recording differently.

In spite of the fact that modern emotion recognition research aims towards spontaneous data, recorded not in laboratory conditions, and real world applications, it is often being performed in an isolated manner. That means that recognition is done without taking into consideration previous actions or emotional status of a user, information about his/her interlocutor (if any) and environment while designing the pipeline. This leads to a situation, when we don't have a comprehensive picture at the modeling stage. This thesis aims to fill this gap. In the following section, we will describe in detail three levels of context considered in this work.

1.2 Contextual Information

The term *contextual* used in this thesis covers the context at three different levels, namely speaker (or user) himself, conversation and environment. These three levels, each based on an expansion of another, represent sources of impact on current emotional status of the user. It is necessary to consider them in order to create an AER system with a high level of performance and adaptability.

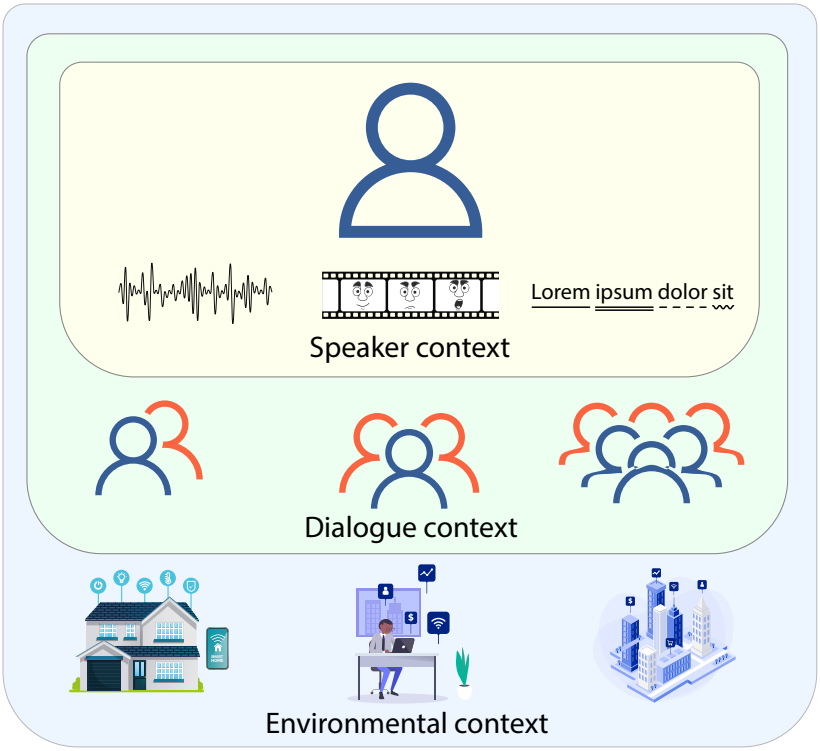


Figure 1.1: Levels of contextual information in emotion recognition used in this thesis[1]

The first level – speaker context – is related to the information produced exclusively by the user. It includes his/her speech (tone, pitch, semantic load), facial expressions, gestures, poses, body movements, etc. This information is typically used by humans to identify the current mood or emotional status of a person. Humans naturally learn to use this information in early age (Denham, 1998) in order to communicate and understand the others better. However, it is not a trivial task to train a computational model that would solve the same problem at a level of performance comparable to a human. Regarding the context of the speaker, it is important to determine if the information about his/her actions and emotional status at time points $t - n..t - 1$, is related to his/her current status (at time point t) or not; in other words, if his/her speech, facial expressions or expressed emotions of the last several seconds or minutes are important for the detection of his/her emotion at the current moment.

The second level – conversational context – covers the information that may be produced only during an interaction between the user and another person (interlocutor), e.g. a dyadic or a group conversation. Here, the same information as at the speaker level can be used, however not of the user himself, but of his/her interlocutor (-s). Instead of the direct connection between

[1]This figure was designed using icons from www.freepik.com: waveform icon is designed by Freepik; video frames icons are designed by starline / Freepik and macrovector_official / Freepik; smart house, smart office and smart city icons are designed by Freepik.

the behavior and the expressed emotion, as at the speaker level, more complex and hidden connections may exist in this case. Although people tend to share the emotions of their interlocutor (the human capacity called empathy), there is a variety of situations when it does not apply. As an example we can consider a quarrel with clear and strong domination of one participant: here, the dominating person will experience anger, while his/her interlocutor will feel guilt, shame, sadness or embarrassment. Analysis of this type of information introduces additional challenges and firmly relies on the performance of an emotion recognition system working with the data of the interlocutor. As his/her emotions are considered as features here, an error in prediction adds the unnecessary noise to the data. This may result in a lower recognition accuracy of speaker's emotions. The complexity of this task also significantly increases with the number of interlocutors.

The third level – environmental context – is connected with the information about surroundings of the user, however excluding persons that have direct contact with him/her. This primarily includes the physical environment, such as a room, an apartment or a house if the user is at home; office if the user is at work; and buildings, locations, establishments and sights if the user is outside. The physical surrounding affects our mood and thereafter our emotional states (Beilock, 2015), allowing both positive and negative changes. While dealing with the data collected in laboratory conditions, the environment is often fixed for each user, eliminating its effect on him/her. However, solving this task within natural, in-the-wild conditions, requires explicit attention to this aspect. Environmental context also covers people surrounding the user, however, they are considered not to have a direct and personal impact on the user, but to be a part of the whole environmental entity. For example, a congestion degree of some place can be related to the mood of the user – if a place (e.g. a sight) is too crowded, the user may become irritated, regardless of other factors. Nevertheless, in some cases crowdedness may not be an issue, i.e. at a music concert or another popular event.

All these aspects and sources of information play an important role in defining or affecting the current mood and emotional status of the user; hence, they should be analyzed in order to get a precise and comprehensive estimation. Contextual emotion recognition is of special importance for two related areas of research, namely, smart environments and dialogue systems. Both can significantly benefit from an integrated emotion recognition component. In the following, we will provide a brief introduction to these areas of research.

1.3 Smart Environments

Research and solutions in the area of smart environment have a firm connection with pervasive computing. Pervasive or ubiquitous computing is a concept and paradigm in information technologies, soft- and hardware engineering, that allows to spread computing to almost any device and makes it available insensible and hidden, connecting devices into the complete network, embedded into person's everyday life.

Started around 50-60 years ago, personal computing has gone a long way, evolving from large and expensive machines to something, that can be literally called personal. Several decades ago, personal computer received an explosive increase in popularity and availability, reaching the peak of sales in 2011. Later, the focus was shifted to mobile computing, allowing people to have a personal computer in the pocket.

Further development of mobile and sensing technologies introduced wearable devices to a wide range of consumers. Presence of sensors and sufficient computational power not only in a smartphone, but also on a hand of a person opened the perspectives to improve one's life quality, constantly monitoring his/her state. Nowadays even simple and inexpensive wearable

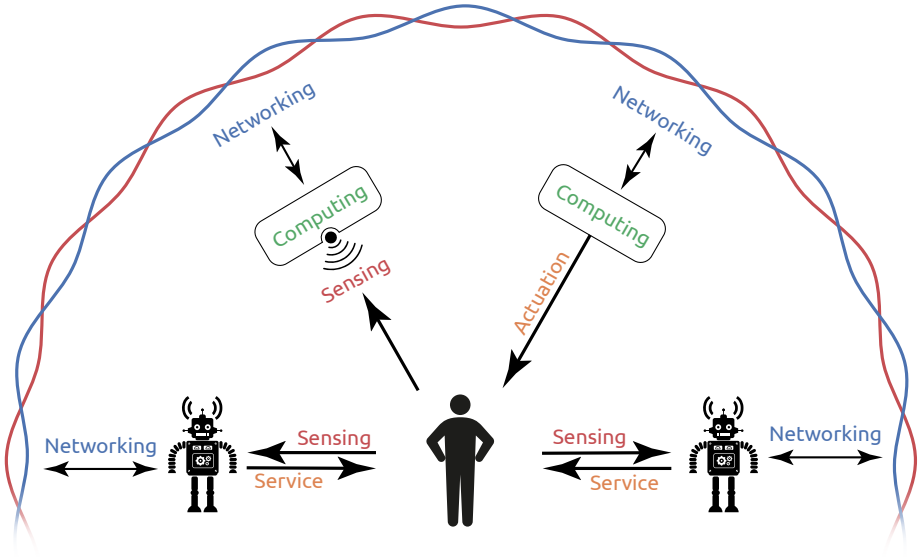


Figure 1.2: General concept of intelligent environments according to Lee and Hashimoto (2002)[2]

devices can track the amount and timing of physical activity, sleep phases, etc. and give suggestions on having a better, healthier lifestyle. Some wearables contain more sensors and provide greater possibilities of tracking one's state, including heart rate, blood pressure, and skin temperature monitoring (Garbarino et al., 2014).

Making a step further from a desktop and mobile computing, ubiquitous computing eliminates the need of a user or an operator to indeed make requests and give commands to a computing system. Such a system can be present in our lives, without any distraction or interruption. In the certain sense, following the Theory of Inventive Problem Solving (Altshuller et al., 1996), pervasive computing paradigm introduces a new functionality without explicit definition of a physical object for it.

Extrapolating the trend of wider technology implementation and increasing level of their invisibility, combined with developments in making faster, more robust, widely spread and more stable connection, pervasive computing has great prospects for the nearest future. One of the popular application areas of pervasive computing is smart or intelligent environments. According to Steventon and Wright (2010), "intelligent environments are spaces in which computation is seamlessly used to enhance ordinary activity". The main concept of intelligent space was described by Lee and Hashimoto (2002) and is depicted in Fig. 1.2

Intelligent environments are designed to be human-centered. Along with humans, there may be other participants – robots. They serve the human, help him with everyday tasks and routine assignments. Robots also sense the environment and humans in order to be aware of the surrounding. Other parts of the environment are various sensors and actuators. They scan the environment and provide feedback by performing some actions if needed, e.g. changing temperature, dimming lights, etc. All devices are connected to a network, which they use for information exchange and support of decision-making process. If equipped with an integrated emotion recognition component, such robots or environment itself may provide much better, more personal service to the user.

[2]This figure was designed using icons from www.freepik.com.

1.4 Dialogue Systems

Pervasive computing intends to make computing appear anytime and everywhere. In order to provide it, the third level of user interface is often used – the natural language interface. Let us consider historical development of the user interfaces. The first level of the user interface is the command line interface. At the beginning of the computer era, the machines were large and operated by specially trained people. In order to operate a computer through the command line, one needed to know specific commands to achieve his goals. Even though, the command line interface is not widespread among the users nowadays, it is still commonly used by engineers. The second level of the user interface is the graphical user interface. It is widely used nowadays and is the standard interface for mobile and desktop computing. The user operates the machine with the help of icons, pointers, menus and windows. The information about the current state of the system or available options and commands is shown with the help of graphical elements. Usually some experience in operating these types of machines is required, but in general, the modern user may learn it easily by reading the information available on the screen and using it to achieve his/her goals. The third level of the user interface, that has great potential for usage in pervasive computing, is the natural language interface. In this type of interface, the user speaks to the system, simulating a normal human-like conversation with the computer. It is the most natural way of interacting with the machine and requires no additional devices (such as a mouse or a keyboard) to operate, since it may be integrated into the environment. The system recognizing user's speech and responding to his commands is called dialogue system. The typical spoken dialogue system (SDS) consists of several modules. Firstly, a speech recognition module, that captures the speech of the user and extracts meaningful characteristics from it. Then, a text analysis module, which captures semantics of a user utterance. Then, a dialogue manager, which derives the exact intention of the user and seeks for the appropriate response provided by the application, to which it is connected. Then, a text generator composes a response to be said to the user, and with the help of a speech synthesis module, the dialogue system pronounces it to the user, closing one dialogue circle.

Dialogue manager may address its responses to several systems or data sources:

- to a database of interactions in order to select an appropriate answer to user's command (U: "How are you doing?" – SDS: "It is easy – you speak with me and my mood is improving!");
- to informational services to search for an appropriate answer (U: "What is the weather now?" – SDS: [checks the weather in internet] "It is +5 and raining in Ulm");
- to an application (U: "Add an appointment with Mr. Smith for tomorrow 15:00 to my calendar" – SDS: [adds an appointment to calendar] "The appointment has been added");
- to external physical systems (e.g. ones in smart house) as commands (U: "Turn on the air conditioner" – SDS: [sends the command to air conditioner, gets a reply on successful execution] "The air conditioner is on now")

The dialogue system may also be extended in order to improve the quality of interaction. Among possible extensions are recognition modules for speaker's identity, age, gender, personality traits or emotional status. The latter is especially important to be observed over time in order to evaluate the quality of interaction with the SDS. This may significantly improve the adaptivity of SDS to the user, taking naturalness, convenience and comfort of human-computer interaction (HCI) to a new level.

1.5 Motivation

Artificial intelligence makes significant steps further every year, covering additional applications and expanding its presence in our everyday lives. Aiming to enhance the quality of life and user experience, it takes care of routine assignments. The *intelligence* of a system is defined by many factors, including its ability to communicate with users at a high level, i.e. naturally and with accurate interpretation of user's intents. Emotions play a significant role here, as they can change the meaning of the words and affect user's decision-making ability, since they are strongly connected with one's mood and behavior. Emotion recognition has been a hot topic for over a decade, and many advancements in this area have been made. Every year the market for emotion recognition systems is growing and new applications emerge. Some promising applications for such systems are the following:

Human-Computer Interaction (HCI). As described previously in this chapter, the natural language interface is the third level of the user interface. It enables the most natural way for humans to interact with any computer system. Advancing from simple voice commands to SDS that is able to maintain a meaningful conversation is a crucial step of ubiquitous integration of such systems in our everyday lives. Introducing a contextual emotion recognition component into the SDS will improve its quality significantly. Without this component, it is hardly realistic to use these systems as an intelligent and human-like agent.

Human-Robot Interaction (HRI). As an extension of HCI, of special interest are smart robots. Such robots may interact with humans only on demand, e.g. a cleaning robot, or as a primary assignment, e.g. a humanoid robot used as an interactive partner or an artificial sensitive listener. A specific area of application of the latter is **elderly care**. A robot may help elderly people to cope with loneliness, stay active and exercise, prevent mental problems caused by low brain activity and be up-to-date with current technological advancements without any need of high computer literacy by means of a natural language interface. An emotion recognition module is irreplaceable in such systems in order to meet high standards of quality and intelligence.

Health monitoring in hospitals is a specific area of application for computer systems and especially for robots. Such robots may constantly monitor the physiological and psychological condition of a patient, as well as provide basic care, lowering the work load on nurses. By being assigned to a particular patient room, such robots may serve as a personal nurse, directing specific commands to a hospital personnel if necessary and providing 24/7 monitoring. Apart from lowering the work load, such robots may help to protect doctors and nurses from being infected from a patient.

Recommender systems. An integration of an emotion recognition component into a multimedia or an entertainment application may serve as an additional source of information. It may help to increase the quality of recommender system, by suggesting similar products to the user, based not only on the history of his/her views or purchase, but also on his/her emotional reactions and behaviour.

Online learning. Nowadays, many educational courses are available online and this area is constantly growing. It gives tremendous opportunities to people around the world to acquire knowledge in almost any sphere. Even local classes and exams, where people can be physically present, are sometimes held online. An introduction of an emotion recognition component may give helpful advises to tutors to structure their lessons, as well as monitor interest and engagement of students.

Mood monitoring in smart environments may be a complimentary feature for enhancing user experience. There is a potential of integrating this component to each level of environ-

ment: smart room or office, smart home, smart car or smart city. Combined through a single data processing system, information retrieved from the user may be extremely beneficial to build a comprehensive picture of his/her mood, prevent stress or depression, help to cope with emotional issues or warn a relative or a specialist if needed.

Of course, for such complicated cases as elderly care or psychological state monitoring, emotion recognition systems should be very robust, able to work time-continuously and take contextual information into account, as it affects the emotional status of the user. Despite many advances implemented in modern emotion recognition systems, most of them treat the problem in an isolated manner. Isolation is performed at data collection, data processing and modeling stages. Contextual information is often avoided and considered as noise or not relevant data. However, such type of information may be beneficial for an emotion recognition system.

The speaker level is the first point of analysis when ignoring context becomes obvious: i.e. the data of the user (speaker) at time point t is isolated from the previous data or emotional states (at time points *t-n..t-1*). In many modern systems, the emotion recognition module is often implemented for solving tasks at the utterance-level (or turn-level), i.e. one evaluation per predefined phrase or video fragment, considering previous data as to be irrelevant. This introduces an additional problem of defining turns (e.g. in a conversation), which may be not a trivial task for the system (Gunes and Schuller, 2013). It also limits flexibility of the system to work with continuously changing input and to extract valuable features from preceding data. A contextual time-continuous approach to emotion recognition seems to be more appropriate in this case. However, there are still several open questions, e.g. an observational window for emotion recognition models, i.e. how much data should be considered to make a prediction of user's emotional status at a particular time point t? On the one hand, continuous real-time predictions require a window that is relatively short in order to not introduce undesirable delay (Chanel et al., 2009). On the other hand, the observational window should be wide enough to capture important cues and to provide reliable performance (Berntson et al., 1997; Salahuddin et al., 2007). With advancements in modeling brought by recurrent models, this issue was levelled to some extent. Nevertheless, it is still an open question for real-time emotion recognition, how much data is required to achieve a proper level of performance. We study this issue in detail in Chapter 4.

The second level when context may be ignored is the conversational level, i.e. the data of the user (speaker) is isolated from the data of his/her interlocutor. Valuable information on the interlocutor's emotional status or his/her responses to the user's actions, contained in speech, facial expressions, etc., is not considered in most of the modern emotion recognition systems and is filtered at the data preprocessing stage of the recognition pipeline. In more advanced systems it is implemented at the utterance-level, it is raising the same issues as described above and introduces the necessity for an additional higher-level system to catch emotional turn-based dynamics in the conversation. A purely time-continuous system that is able to utilize the context of both the speaker and the interlocutor seems to be more preferable. We study this issue and introduce an appropriate system in Chapter 5.

The third case of ignoring context is at the environmental level. While it is the broadest aspect of context with the least explicit connections between the environmental characteristics and the user's emotional status, it is highly useful as a potential source of information to extend the emotion recognition systems. Ignoring context at this level usually happens at the data collection stage – most of the databases with emotionally rich data are collected in a strict, neutral, laboratory environment. It helps to eliminate undesired noise, that is present on the streets or in a room with other people, which do not take part in the data collection process.

However, this information may also contain useful features that can be extracted if analyzed properly. As environmental context is very broad and cannot be covered and modelled with a high precision within one model, a domain for the analysis should be defined first. After this, feature engineering within this domain is the main focus for further research – one should find the useful set of modalities and features that affect user's emotional status. We consider both of these aspects in detail in Chapter 6.

Thus, in this thesis, steps towards integration of contextual information into an emotion recognition system will be performed at each of three presented processing levels, where context has been ignored so far. This will open a room for more intelligent systems and more advanced applications. In the following section, we will briefly cover the main contributions of the thesis.

1.6 Thesis Contributions

The main objectives of this thesis are development, application and evaluation of approaches to utilization of contextual information in emotion recognition systems. As mentioned above, such information may be available at three different levels and in our work we made use of them all, defining three aims, corresponding to each level:

1. To figure out if the amount of contextual information about user, i.e. his/her previous speech or facial expressions, is related to the emotion recognition performance. If so, which amount of data is optimal and on which factor it is dependent.
2. To develop approaches to integration of interlocutor's (conversational partner's) data into an emotion recognition system for the user in order to increase its performance. This approach should be applicable to time-continuous problem statement.
3. To figure out if information about user's surroundings may help to build an emotion recognition system. If so, which modalities provide the highest performance.

For the first aim, we initially aligned features and labels using a combination of algorithms for reaction lag correction. Then, we tested models of two types: time-dependent (Recurrent Neural Network) and time-independent (Multilayer Perceptron, Linear Regression with L2 Regularization, Support Vector Regressor, Gradient Boosted Decision Trees). To prove a hypothesis of existing dependencies between the amount of context and system performance, we have developed a flexible approach to contextual modeling, considering each stage of the recognition pipeline. Our experiments were conducted on three corpora of spontaneous time-continuous audio-visual data, annotated in arousal and valence.

Based on extensive experiments with various approaches, context length, models, modalities and dimensions, we figured out that there are indeed dependencies between amount of used context and model performance. More precisely, the optimal amount is not dependent on a feature set, the amount of time steps for recurrent models and data frequency. Nevertheless, our experiments showed that the optimal context length is affected by the modality, corpus and model type. For more detailed conclusions regarding the later aspects, experiments on additional databases should be conducted. Moreover, we have conducted experiments in cross-corpus scenario, that have showed that contextual dependencies are often inherited from training and target corpora.

For the second aim, we have developed several approaches to integration interlocutor's and user's data for time-continuous problem statement. They are based on feature-level and decision-level fusion and allow context variation for the user and interlocutor in a dependent,

as well as in an independent scenario, i.e. using similar and different data amount in each sample. We have conducted our experiments on four corpora of spontaneous interactions with audio-visual data, annotated in arousal and valence. In total, we have tested four approaches to contextual emotion recognition in dyadic interaction.

Based on the performance comparison of these approaches to a speaker-only baseline, we have concluded that incorporating interlocutor's data into emotion recognition system may significantly improve its performance. Among the tested approaches, the fully independent one has showed the highest performance, while it is also the most resource demanding. The simpler approaches (partially independent or dependent ones) have showed slightly lower performances on average, but due to fewer parameters they are easier to start with.

For the third aim, we have focused on a specific use case, when environmental context has strong influence on the user's emotions, namely, a sightseeing tour. As no off-the-shelf corpora are available for this task, we have collected our own dataset of emotionally labelled touristic behaviour, using several devices and annotated on several scales. Then, we have trained several uni-, bi-, tri- and multimodal systems for emotion, satisfaction and touristic experience quality estimation using feature sets designed to extract meaningful characteristics from collected data.

Our experiments have showed that the features describing head movements (tilts and turns) provide the highest performance for emotion recognition task, these features combined with eye movements based ones – for satisfaction estimation, and the audio-visual ones – for touristic experience quality prediction. Feature-level fusion of all available modalities has showed performance gain over the best unimodal systems only for satisfaction estimation; and decision-level fusion – for each of three problem statements. The performance of decision-level fusion approach was also much higher compared to other approaches.

In the following section, we will briefly introduce a structure of this thesis.

1.7 Outline

The thesis consists of seven chapters. The current Chapter 1 has introduced the main idea and the motivation of this work. In Chapter 2 we present relevant background, different problem statements for emotion recognition, current interest to this topic in the research community and state-of-the-art machine learning methods. Then Chapter 3 introduces multimodal multidimensional time-continuously annotated corpora used in this work, as well as data preprocessing steps and evaluation metrics used throughout presented research. The following Chapter 4 proposes several methods of speaker context modeling, presents their advantages and disadvantages, as well as experimental results. Based on the conclusions made in this chapter, we extend our research to the dialogue-level context modeling in Chapter 5. There, we cover several methods of context modeling in dyadic interactions and their experimental results. The following Chapter 6 moves the focus to affective context in smart environments, where we cover it in a use case of smart cities (smart tourism). It includes the *EmoTour* project, database and concept for emotion aware smart cities, developed in cooperation with Nara Institute of Science and Technology in Ikoma, Nara, Japan. Chapter 7 wraps up this thesis by presenting its main conclusions and contributions in three groups: theoretical, practical and experimental, as well as the most promising directions for future research in the areas of time-continuous emotion recognition.

2 Background and Related Research

Emotion recognition is a growing area that attracts many researchers from different fields, including psychology, linguistics, computer science, etc. However, this area is not new and is already on the market for several decades. Over this time, emotion recognition performed a huge leap and advanced significantly, supported by developments in such other areas of computer science, as speech recognition, natural language processing and pattern recognition in general, as well as by constantly growing computational capabilities, allowing to train better, more complex and flexible models faster and on larger datasets. In this chapter, we will cover important concepts which form a basis for any emotion recognition research, as well as advancements in solving problems tackled in this thesis. We will begin with presentation of the most common approaches to emotion recognition, then consider significant research works related to *contextual* emotion recognition, followed by a short description of emotion recognition challenges, which provided a valuable contribution in the area. After that, we will cover one of the most important part of any automatic recognition system – machine learning algorithm. We will present a brief description of methods used in our thesis and conclude the chapter afterwards.

2.1 Approaches to Emotion Recognition

Studies on human emotions have a long history. Some of them claim that the ability to recognize emotions is developed in childhood and helps a child to successfully interact with other humans. Children with better understanding of human emotions have greater chances of building strong, positive relationships in their future lives (Denham, 1998). Although humans are the best emotion recognition systems ever built so far, they can understand the same expression differently due to the subjective nature of emotion. Aside from cultural and social differences, the ability to recognize emotions can vary greatly with age (Chronaki et al., 2015).

In spite of the fact that any human has an ability to recognize other's emotions with a certain degree of preciseness, various approaches were developed and applied in order to standardize this process and have a common understanding of it. For example, Facial Action Coding System was originally developed by Hjortsjö (1969) and later adopted by Friesen and Ekman (1978) and resulted in publishing a comprehensive, 527-page manual (Ekman et al., 2002). This system is used in many areas, e.g. by researchers in facial analysis and by cartoon animators. Numerous applications were developed based on it, including one used in this thesis – OpenFace – that is covered in Section 3.2.2.

How to represent emotions, how to code and standardize them – are the first questions to be answered. There are several approaches to it; two most widely used are: **categorical** and **dimensional**. The first one divides emotions into several basic, easy-to-understand categories, such as *anger* or *happiness*. There is no fixed set of emotions to be used. For example, Paul Ekman defined six basic emotions: *anger, happiness, sadness, fear, disgust* and

surprise. These emotions are independent from each other, and we cannot assign a particular order between them. Later this set was extended by *amusement, contempt, contentment, embarrassment, excitement, guilt, pride in achievement, relief, satisfaction, sensory pleasure* and *shame* (Ekman, 1999). Another researcher, Robert Plutchik, presented his concept of the *"Wheel of Emotions"*, where he defined eight basic emotions as four pairs of opposites: *anger-fear, joy-sadness, trust-distrust* and *surprise-anticipation* (Plutchik and Kellerman, 1980). Plutchik's model is often depicted as a 3-dimensional cone-like figure, where vertical dimension corresponds to intensity of emotion (see Fig. 2.1 for its 2D representation). A set of emotions to be used in the recognition system is often defined by the final task of such system: e.g. if emotion recognition module is used as a part of SDS in a call-center, the set may be limited to *{angry, not angry}* or *{satisfied, not satisfied}*.

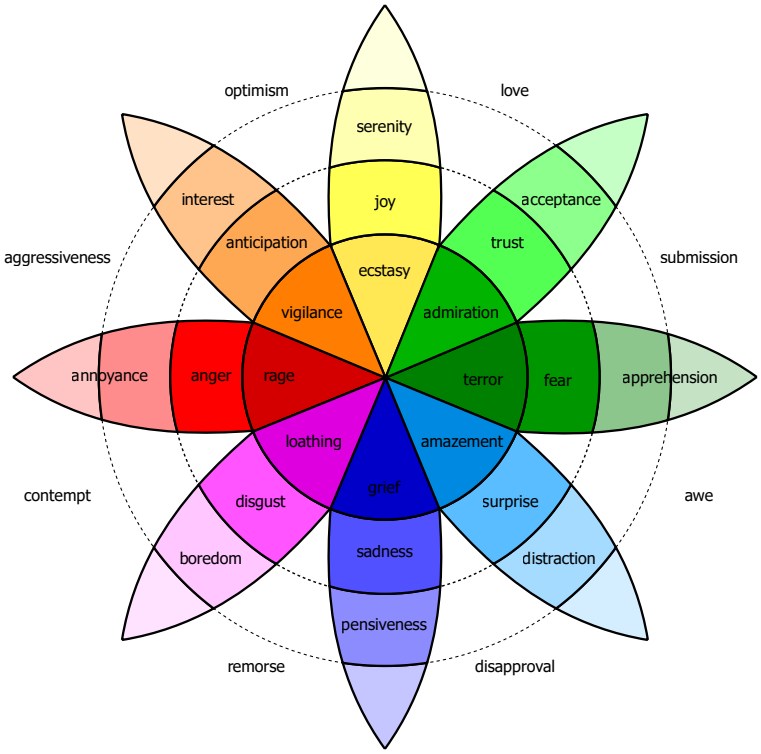


Figure 2.1: 2D representation of Robert Plutchik's wheel of emotions[1]

However, in real life, people tend to experience more subtle emotional states than represented by basic emotions. Some studies showed that cognitive mental states, such as *agreement* or *disagreement*, *concentrating*, *thinking*, etc. occur more often than basic emotions described above (Baron-Cohen, 2007). Moreover, some sets of basic emotions are too small and don't allow any transition states. Sometimes there is no *neutral* state in such sets, which is far from real-life conditions.

Taking this into account, researchers suggested another representation of emotions, where they are not independent from one another, but rather ordered in a system – a dimensional approach. Here, to each emotional state, a value on orthogonal scales of a continuum is assigned. This approach allows connections and smooth transitions between states, as well as

[1]By Machine Elf 1735 - Own work, Public Domain, `https://commons.wikimedia.org/w/index.php?curid=13285286`

intensity definition. The most widely used model is the *"Circumplex of Affect"* introduced by Russell (1980). He proposed to use the following scales: *arousal* (or activation, excitation) and *valence* (or pleasantness, pleasure, appraisal). Other authors suggest an extension of such emotional representation by introducing additional scales, e.g. *dominance* (Mehrabian, 1996). This 3-dimensional space is also popular in the field of emotion recognition, as it suggests more clear division between some states of arousal-valence space, especially in the area of negative valence (e.g. *fear* and *anger* – both are low valence, high arousal, but *fear* has low dominance, while *anger* – high dominance). Some researchers advocate for the fourth dimension – *expectation*, as a degree of anticipation (Fontaine et al., 2007) and some for the fifth – *intensity*, as a degree of rational nature of person's behavior (McKeown et al., 2010).

A transition between categorical and dimensional representations is possible, but it may lead to loss of information (Gunes and Schuller, 2013). For example, in Fig. 2.2 an arousal-valence emotional space is presented with values assigned to some basic emotions based on (Scherer, 2005) (in red) and (Cowie et al., 2000) (in green). One may notice, that in many cases the same emotions are relatively far away from each other and sometimes their positions are controversial (e.g. for *afraid*).

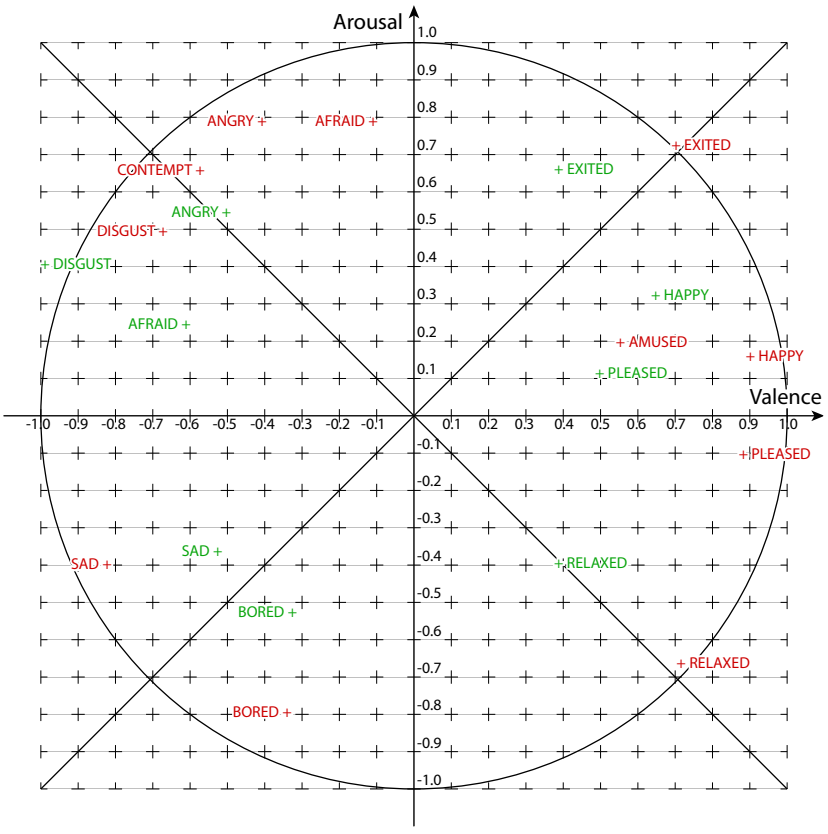


Figure 2.2: Arousal-Valence model with two emotional label sets: based on Figure 1 from (Scherer, 2005) (in red) and on screenshot of FeelTrace annotation tool (Cowie et al., 2000) (in green)

.

In spite of the fact that these scales presumed to be independent from one another and close to orthogonal, studies showed that there is positive correlation between them (Oliveira et al., 2006; Alvarado, 1997). This will be also noted in Section 5.1.

Apart from label representations, there are also two main diverse input data types: **con-**

tinuous and **non-continuous** (also often referred to as an utterance-level). While dealing with the non-continuous input, a model processes chunks of data, corresponding to one turn or one utterance of a speaker. This turn defines borders of an affective event in an original data and limits the model to perform recognition only in a certain time range. This approach implies one feature vector corresponding to one chunk, i.e. regardless of chunk duration, an input to the model describing it will always be of size $[1 \times F]$, where F is the number of features. A typical approach to obtain one feature vector per chunk is to use two-level feature extraction: (i) feature values are extracted for a fixed short time window (usually 10-30 ms); (ii) functionals, such as mean, standard deviation, minimum, maximum, etc. are calculated on data corresponding to one utterance.

Continuous input doesn't imply any turn annotation, hence the model has to deal with the original data structure and define start and finish points of an affective event by itself. As continuous input is still comprised of discrete values (due to the discrete nature of data), one may see it as a non-continuous input with a very narrow window size for utterances. However, there is one big difference between these two input types: in the non-continuous input type, utterances don't have to be consecutive, and they are usually independent from each other, while in the continuous input type, previous data frame is always connected to the one before and one after it. This provides an opportunity for more context-aware analysis of the current data frame, but also introduces additional challenges in the modeling.

As original raw data is often time-continuous (e.g. a speech signal or video footage), the type of input data is usually defined by the label's format. Many databases have continuous data, but annotated on the utterance-level.

There are no strict rules for choosing categorical or dimensional representations for a particular input type, but the general trend is to use categories for utterance-level annotation and dimensions for continuous annotation. It comes with a human's perception of the information: it is natural to assign a particular emotion (such as anger, happiness, neutral) to a short phrase or a facial expression of someone, and to register not only emotions themselves, but also changes in emotional state on a more precise level in a long run. However, there are some exceptions from this trend, e.g. the RAMAS database (Perepelkina et al., 2018), which is annotated continuously in categories. One of the disadvantages of this approach is that it does not provide transition states and the change of emotional class happens too fast, which is unrealistic. Authors use a confidence or an agreement score between annotators to mitigate this issue, but this doesn't solve the problem if we consider only one true class for each frame (no soft boundaries).

In turn, the type of the input data defines the modeling approach to be used: classification or regression. A task with categorical annotations and non-continuous data input is solved by classifiers; a task with dimensional annotations and either continuous or non-continuous data input is solved by regressors. However, from the mathematical point of view, there are only few differences between them.

For many years, most of the research on emotion recognition was concentrated on solving classification tasks with the non-continuous (utterance-level), acted input data. Numerous databases were collected, e.g. RUSsian LANguage Affective Speech (RUSLANA) (Makarova and Petrushin, 2002), Berlin Database of Emotional Speech (Emo-DB) (Burkhardt et al., 2005) and Surrey Audio-Visual Expressed Emotion (SAVEE) (Haq and Jackson, 2010). Many of them were collected in highly constrained laboratory conditions, e.g. in Emo-DB each speaker spoke one sentence with different emotions. Some corpora, however, were collected in a spontaneous scenario, e.g. Let's Go Database (LEGO) (Eskenazi et al., 2008) consists of non-acted utterances recorded by SDS used in a navigation system on a bus stop

(Raux et al., 2006).

For some databases, non-continuous data representation is used, but they are annotated dimensionally. For instance, Vera am Mittag (VAM) (Grimm et al., 2008b) database consists of phrases recorded during the popular German talk show "Vera am Mittag" and has ratings in activation (arousal), valence and dominance.

Gradually, the focus in emotion recognition research is shifting towards a **continuous** input, **dimensional** annotations and **spontaneous** interactions, as this representation not only allows more flexible modeling, but also is closer to real-life conditions, has a greater potential to be further used as a subsystem for decision support (e.g. in a SDS) as well as a standalone system. In this thesis, we focus on this type of data and describe used corpora and data specific challenges in Chapter 3.

2.2 Contextual Emotion Recognition

In this section, we will consider works related to the contextual emotion recognition in the sense of the three levels presented in Section 1.2.

2.2.1 Speaker Context

One basic question to be asked when working with continuous emotion prediction is: what is an appropriate unit for analysis (similarly to utterance in non-continuous data) in this setting? This is an issue to be addressed not only to reach an optimal prediction performance, but also to facilitate real time predictions with the shortest delay possible (Gunes and Schuller, 2013; Chanel et al., 2009).

Previously, for contextual learning in emotion recognition, utterance-level representations were used. Wöllmer et al. (2010) applied Bidirectional Long Short-Term Memory (BLSTM) Recurrent Neural Networks to this problem, using audio and visual features. They compared the performance of BLSTM to fully connected 3-state Hidden Markov Models (HMM) and Support Vector Machines (SVM) in a unimodal and multimodal (feature-level fusion) setup. They used original arousal and valence annotations to form the three-class representations: low, medium and high, as well as joint classes using three, four and five clusters. They measured performance with accuracy, recall, precision and F1 score. Authors stated that for a unidimensional setting, the highest performance was achieved with BLSTM for valence and HMM + LM (language model) for arousal. For a joint classification using arousal-valence clusters and multimodal features, LSTMs and BLSTMs were dominating in each of the three clustering setups.

Ringeval et al. (2015) experimented with different window sizes for functional extractions from audio, video, ECG (electrocardiogram) and EDA (electro-dermal activity) modality using a fully continuous representation of the input data, and LSTM models to capture temporal dependencies. Authors recalled that for some cues it is enough to have a window size of 0.5 seconds (Yan et al., 2013), while for the others it may take up to 6 seconds. Their experiments with different window sizes showed that on average for the four used modalities, valence requires a window about twice the duration of the one used for arousal in order to obtain the best performance.

Some studies were conducted to define a relation between the window size used in LSTM and the system performance. Huang et al. (2018) used overlapping windows in AVEC 2018 Cross-Cultural Sub-Challenge (Ringeval et al., 2018) on SEWA database (Kossaifi et al.,

2019) and compared predictions to the original approach provided by challenge organizers – to use a whole recording as one data sample. Authors reported performance gain for 8 out of 9 cases: three modalities (audio, video, text) used to predict three dimensions (arousal, valence, liking). They set the window size to 500 frames and an overlap of 100 frames which corresponds to 50 seconds and 10 seconds respectively, according to the frame rate of labels.

The similar approach was used by Keren et al. (2016), where authors changed original files provided for Interspeech ComParE 2016 with shorter overlapping samples extracted from them. Authors referred to it as a data augmentation and reported a significant increase in predictions while using shorter samples.

Ouyang et al. (2019) used the Autoregressive Exogenous model (ARX) to capture temporal dependencies in data without using a Recurrent Neural Network. In this case, the amount of the used context in labels and features was controlled by orders of an autoregressive sub-model and an exogenous sub-model respectively. Authors conducted experiments with three continuously annotated corpora and stated that there was a strong connection between performance and order of the autoregressive sub-model, as well as a moderate connection with order of the exogenous sub-model. They also noticed that for some of the used corpora, there was an optimal value of delay used in the model, which could correspond to a reaction lag, as it was not explicitly corrected during the preprocessing stage of the pipeline. This approach showed a performance comparable to LSTM modeling.

However, the general relation between the window size and the model performance considering model size, different modalities and dimensions was not studied comprehensively yet. This is the problem covered in Chapter 4 of this thesis.

2.2.2 Dialogue Context

Most of the current research on emotion recognition is done in accordance with a speaker-isolated scenario. However, following the trend of taking emotion recognition out of laboratory conditions and making it face real-life data and problems, the interest to dyadic emotion recognition has grown over the recent years.

Lee et al. (2009) mapped utterance-level annotated corpus IEMOCAP (Busso et al., 2008) to "turn change" data representation. Then authors analyzed four strategies utilizing the emotion evolution in dyadic interactions: (i) baseline – with no connections; (ii) individual time-dependency – with connections within the data of one speaker; (iii) cross-speaker dependency – with mutual influence between speakers; (iv) combined – with mutual influence within and between speakers data. They used Dynamic Bayesian Networks to model the mutual influence and the temporal cross-speaker dependency of emotional status and obtained the relative improvement of 3.67% in terms of classification accuracy over the defined baseline.

Chen et al. (2017) used three strategies to audio data to gain improvements from its dialogue nature: (i) mixed, when speech from both speakers are present in audio-file; (ii) purified, when speech of interlocutor was cut from the file; (iii) doubled, when feature-level fusion was used to incorporate data of both speakers in training process, while diversifying them one from another. They report "doubled" setup to be the most beneficial.

Li and Lee (2018) proposed a network architecture to obtain robust acoustic representation for an individual during dyadic interactions. It was designed to describe individual's acoustic features as a general variational deep embedding augmented with the dyad-specific representation. Their approach achieved the relative improvement of 4.48% in terms of Spearman's correlation on CreativeIT (Metallinou et al., 2010) and NNIME (Chou et al., 2017) corpora.

Zhao et al. (2018) proposed several multimodal interaction strategies to make use of the multimodal information of the interlocutor: (i) AFA – combining audio features of interlocutor with audio-visual features of speaker; (ii) AFF – combining visual features of interlocutor with audio-visual features of speaker; (iii) AFAF – combining audio-visual features of interlocutor with audio-visual features of speaker; (iv) ATFATF – same as previous, but with additional textual modality; (v) ATFAT – same as previous, but without visual features of interlocutor, as authors stated that it reduced the performance. They made an extensive analysis of feature sets for audio, video and textual modalities and suggested dyadic human-human interaction pattern under multimodal interaction scenarios. Authors used recurrent neural networks with long short-term memory (RNN-LSTM), and applying the proposed approach, they achieved the relative improvement of 34.35% and 35.70% over the baseline results for arousal and valence respectively in terms of concordance correlation coefficient on SEWA dataset (Kossaifi et al., 2019).

Extending the topic of emotions in dyadic interactions, Koutsombogera and Vogel (2018) developed the MULTISIMO corpus of collaborative group interactions in order to investigate factors that influence collaboration and group success in a multi-party setting. Group emotions recognition is also a regular topic of EmotiW emotion recognition challenge since 2016 (Dhall et al., 2016).

However, most of the strategies of utilizing interlocutor's data work with utterance-level data, but not with purely continuous.

2.2.3 Environmental Context and User State Recognition in Smart Environments

Most of the research works aiming towards emotion-aware smart environments, study the possibility of recognizing user states and conditions based on data from wearable devices. One of the most popular topics in this area is stress detection.

Referring to the lack of datasets in this field, Schmidt et al. (2018) introduced WESAD – a multimodal dataset for WEarable Stress and Affect Detection. It consists of data of 15 participants (12 male, 3 female with mean age of 27.5 years old) recorded from wrist- (Empatica E4[2]) and chest-worn (RespiBAN Professional[3]) devices. Modalities list includes blood volume pulse, electrocardiogram, electro-dermal activity, electromyogram, respiration, body temperature and acceleration. Samples of the dataset are annotated on three affective states, namely, neutral, stress and amusement. Authors created a benchmark with several classification approaches, such as decision trees, random forest, adaptive boosting on decision trees, linear discriminant analysis and k-nearest neighbors and achieved up to 93% accuracy (with 0.91 F1 score) in binary classification problem statement and 80% accuracy (with 0.72 F1 score) for three-class classification.

Another study questioned the applicability of wearable sensors to various research tasks performed in out-of-laboratory conditions. Menghini et al. (2019) conducted a series of experiments to assess the accuracy of Empatica E4 wristband under various conditions. Authors compared the data collected with this wearable device to the gold-standard – electrocardiography and the finger skin conductance sensor – the ones that are difficult to use in real-world experiments. The examined conditions included seated rest, seated activity (e.g. keyboard typing), as well as light physical exercises, such as walking. Experiments showed,

[2]https://www.empatica.com/en-int/research/e4/

[3]http://biosignalsplux.com/products/wearables/respiban-pro.html

that only heart rate measurements keep their good performance over different conditions. Other modalities, such as heart rate variability, showed relatively high performance only in resting conditions. Keyboard typing or walking caused a significant drop in accuracy.

Not only wearable devices are used to acquire data from users. Zhao et al. (2016) presented an EQ-Radio system, that transmitted radio frequency signals and used their reflections from the user's body to analyze heart rate and respiration frequency. This data was used later to extract features, the most informative of which were then selected with feature selection algorithms and emotion classification was performed. Experiments with data from 12 participants and elicited emotions showed accuracy up to 87% for person-dependent scenario of four-class classification task. The highest precision was shown for class *joy* and the lowest for *anger*. However, in the person-independent scenario this system showed 72% of accuracy, with controversial results for particular classes: the highest accuracy for *anger* and the lowest for *pleasure* (*joy* is the second lowest). However, greater developments in the area of such devices will foster emotion-aware smart environments.

Another improvement comes from increasing capabilities of mobile cloud computing (MCC). Chen et al. (2015) proposed an EMC – framework for personalized emotion-aware services by MCC and affective computing. Authors claim their system to work in several scenarios, such as elderly care (to decrease their loneliness level), people working in closed environment over a long period of time (to monitor their physical and mental state), socially autistic people (to omit sociophobia) and medical care (to help patient to recover quicker). The goal of the proposed framework is to provide personalized and intelligent emotion-aware services.

It was proven, that emotional status of a user can be measured using wearable devices with a certain level of preciseness. However, in most of the studies, an effect of environment was not considered.

2.3 Organized Challenges on Emotion Recognition

A tremendous contribution to the development in the area of emotion recognition was made by numerous challenges and competitions. By presenting the task in a competitive manner, setting baselines and benchmarks, organizers of the challenges fostered research in this field greatly and covered diverse applications of paralinguistics, not limited to only emotion recognition. Three most notable challenge series are Interspeech Computational Paralinguistics ChallengE (ComParE), ACM Multimedia Audio/Visual Emotion Challenge (AVEC) and ACM International Conference on Multimodal Interaction Emotion Recognition in the Wild (EmotiW). All of them have rather long history, organizing competitions since 2009, 2011 and 2013 respectively.

Each of the challenge series took its niche. Interspeech ComParE aimed to foster research in acoustic signal usage for various paralinguistic tasks, such as recognition of emotions (Schuller et al., 2009), gender and age (Schuller et al., 2010), alcoholic intoxication and sleepiness (Schuller et al., 2011), personality, likability and pathology (Schuller et al., 2012), autism (Schuller et al., 2013), cognitive and physical load (Schuller et al., 2014b), Parkinson's condition (Schuller et al., 2015), quality of pronunciation (Schuller et al., 2016), addressee, cold and snoring (Schuller et al., 2017), heart beat and infant crying (Schuller et al., 2018), baby sounds and even orca (toothed whale) sounds (Schuller et al., 2019).

In turn, EmotiW focused more on visual modality. Challenges on following recognition tasks were organized: facial expressions in the wild (Dhall et al., 2013, 2014), static facial expressions (Dhall et al., 2015), group-level emotions (Dhall et al., 2016, 2017), student

engagement (Dhall et al., 2018; Dhall, 2019). An approach to deep-learning based feature extraction used by one of the winner teams, was utilized for this research in Section 4.1.2 and described in detail in Section 3.2.2.

AVEC focused exclusively on conventional emotion recognition based on audio-visual signals from a user. Considering evolution of challenge tasks, baseline and winning solutions, one may easily track the changes in main trends for audio-visual emotion recognition research over the last decade. Many novelties, introduced by organizers or participants of AVEC, formed the basis of research questions for this thesis. The first challenge (Schuller et al., 2011) in the series was organized on classical emotion classification task. However, already in AVEC 2012 (Schuller et al., 2012) a transition to continuous input and labels was performed. Since then, only regression task was used in this challenge. Taking baseline correlation scores into account, it was emphasized that the task of continuous emotion recognition is indeed very challenging. A similar problem statement was used in AVEC 2013 and 2014 (Valstar et al., 2013, 2014). In AVEC 2015 (Ringeval et al., 2015), several novelties were introduced: RECOLA database (Ringeval et al., 2013), new expert-knowledge based feature set eGeMAPS (Eyben et al., 2016) and different performance measure – Concordance Correlation Coefficient (CCC) (Lawrence and Lin, 1989). All of them were used in our work, forming the basis of modeling and evaluation for Chapter 4 and Chapter 5. In the next challenge of the series (Valstar et al., 2016a), reaction lag – a delay between an actual affective event and its annotation by raters – was considered. It is also of great importance for this thesis and described in detail in Section 3.2.3. In AVEC 2017 (Ringeval et al., 2017a), another corpus for time-continuous emotion recognition was introduced – SEWA (Kossaifi et al., 2019) – which is also used in our work. One of the winner teams (Chen et al., 2017) proposed an interesting approach – to use the information of the interlocutor in order to enhance the input audio data. Together with the general problem statement for time-continuous emotion recognition, this formed the research question for Chapter 5. AVEC 2018 (Ringeval et al., 2018) used the same corpus but with extended data. One of the winner teams (Zhao et al., 2018) for this challenge, as well as for the next year's competition (Ringeval et al., 2019; Chen et al., 2019) used a deep-learning based feature extractor for audio signal, namely, *vggish*, which is utilized in this work in Section 4.1.2 and described in detail in Section 3.2.2.

An evolution in modeling approaches is also clearly visible from the challenge baselines overview. Conventional methods, such as Support Vector Machines (SVM) for classification or regression, were used at the beginning of each challenge series. However, participants introduced their deep learning based solutions, overperforming the baseline starting from 2013 (Kahou et al., 2013). Each year, the proportion of such approaches was increased, until they displaced SVM and became baseline solutions in 2015, 2016 and 2018 for AVEC, ComParE and EmotiW respectively. For more detailed overview of these methods and their applications, see the next section.

2.4 Background in Machine Learning Algorithms

The core of automatic emotion recognition is the machine learning approach. In this section, we briefly cover algorithms used throughout this thesis to build recognition systems. Most of these approaches can be used for both classification (categorical labels) and regression (dimensional labels) tasks. We will first consider Artificial Neural Networks, and their most used architectures, namely, Multilayer Perceptrons, Convolutional Neural Networks, Recurrent Neural Networks, Long Short-Term Memory blocks. After that we will briefly describe Linear Regression, Support Vector Machines and Gradient Boosting on Decision

Trees algorithms. The choice of a particular algorithm depends on a task on hand (input and output data). While several problem statements are considered in this work, all algorithms listed above are used for contextual emotion recognition.

2.4.1 Neural Networks

Artificial Neural Networks (ANNs) are the quintessential models in the area of modern machine and deep learning. Initially inspired by the structure of biological brain, they revolutionized the area by achieving great results and outperforming other methods. Nowadays, there are numerous architectures of ANNs, designed to solve specific tasks.

Feedforward Neural Networks

The most general and basic architecture is feedforward neural networks or multilayer perceptron (MLP). It defines a mapping $y = f(x;\theta)$ approximating function $y = f^*(x)$ which maps an input x to an output y. The term "feedforward" means that there are no feedback connection, therefore computations are done from x through intermediate layers to y. Networks that have feedback connections are called recurrent and used in this thesis as a basic models. Their description is provided later in this section.

ANNs are comprised of three types of layers – input, output and hidden. Input layer works directly with input data x, output layer works directly with target labels y. Hidden layers are used to learn a representation from data x, utilizing weights applied to connection between neurons of current and previous layers, summation and activation functions. First, a weighted sum of input signals of each neuron is calculated:

$$v_j = \sum_{i=0}^{m} \omega_{ji} y_i, \tag{2.1}$$

where y_i is output of neuron i from previous layer, ω_{ji} is the corresponding weight, m is a number of neurons from previous layer connected to neuron j of current one. After that, the output of neuron j is calculated by applying an activation function ϕ:

$$y_j = \phi_j(v_j). \tag{2.2}$$

Initial weights of each neuron, as well as number of hidden layers, number of neurons corresponding to each layer and type of activation functions should be set in advance; this process is called initialization. The number of neurons on input and output layers are known in advance from problem setting, and are equal to dimensionality of feature and target vectors respectively. Once initialization is completed, the output of the MLP can be calculated. First, values of input feature vector are assigned to neurons of input layer. Then, using weights ω_{ij} and activation functions ϕ, outputs of hidden layers are consecutively calculated. Finally, the output of each neuron of the output layer is calculated and the vector of these values is considered to be a response of the MLP to the input vector.

An important aspect of any machine learning algorithm is the training procedure. To facilitate training, the quality of current approximation obtained with an algorithm should be assessed first. Outputs of MLP are used to calculate an error signal e_j^n for the neuron j at iteration n as follows:

$$e_j^n = d_j - y_j^n, \tag{2.3}$$

where d_j is a target value for j^{th} component of the output vector and y_j^n – an actual output of MLP for j^{th} component at iteration n. The current energy of error at iteration n can be defined as a sum of individual errors of each neuron of the output layer:

$$E^n = \frac{1}{2}\sum_{j \in O}(e_j^n)^2 \tag{2.4}$$

where O includes all neurons of output layer. As we want our model to provide better approximation, the training procedure of MLP can be considered as an optimization (minimization) task:

$$\underset{\omega \in W}{\text{minimize}}\, E^n, \tag{2.5}$$

where W represents an overall set of possible weights ω.

One of the most widely used algorithms for solving the optimization task defined in Equation 2.5 is Stochastic Gradient Descent based on Backpropagation (Rumelhart et al., 1985). The latter allows information from the energy of error to flow backwards through the network in order to compute corresponding gradients. These gradients are used further to perform learning itself, by applying a correction element $\Delta\omega_{ji}^n$ to the corresponding weight $\omega_j i$, which is proportional to the partial derivative $\frac{\delta E^n}{\delta \omega_{ji}^n}$. Applying the chain rule of calculus, that states that for functions $y = g(x)$ and $z = f(g(x)) = f(y)$ derivative of z with respect to x can be calculated as $\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}$, we calculate it as:

$$\frac{\delta E^n}{\delta \omega_{ji}^n} = \frac{\delta E^n}{\delta e_j^n}\frac{\delta e_j^n}{\delta y_j^n}\frac{\delta y_j^n}{\delta v_j^n}\frac{\delta v_j^n}{\delta \omega_{ji}^n}. \tag{2.6}$$

This partial derivative determines search direction for weight ω_{ji}^n in continuous space. Differentiating Equation 2.4 with respect to e_j^n we obtain:

$$\frac{\delta E^n}{\delta e_j^n} = e_j^n \tag{2.7}$$

Then, by consequently differentiating Equation 2.3 with respect to y_j^n we obtain:

$$\frac{\delta e_j^n}{\delta y_j^n} = -1. \tag{2.8}$$

Then, by differentiating Equation 2.2 with respect to v_j^n we obtain:

$$\frac{\delta y_j^n}{\delta v_j^n} = \phi'(v_j^n), \tag{2.9}$$

where ϕ' is the first derivative of a function ϕ. Finally, by differentiating Equation 2.1 with respect to ω_{ji}^n we obtain:

$$\frac{\delta v_j^n}{\delta w_{ji}^n} = y_j^n. \tag{2.10}$$

Combining Equations 2.7-2.10 with Equation 2.6 we obtain:

$$\frac{\delta E^n}{\delta \omega_{ji}^n} = -e_j^n \phi'(v_j^n) y_i^n. \tag{2.11}$$

Thus, the correction term $\Delta\omega_{ji}$ for weight ω_{ji} according to delta rule is determined by:

$$\Delta\omega_{ji} = -\alpha \frac{\delta E^n}{\delta \omega_{ji}^n} \tag{2.12}$$

where α is a small constant called learning rate, which is a parameter in Backpropagation algorithm. The minus sign indicates that the minimization problem is being solved as a function decreases in the direction opposite to its gradient in the current point. Combining Equation 2.11 and Equation 2.12 we obtain:

$$\Delta\omega_{ji} = -\alpha e_j^n \phi'(v_j^n) y_i^n. \tag{2.13}$$

If the corresponding neuron is at output layer, then e_j^n can be calculated directly according to Equation 2.3. If it is at hidden layer, e_j^n can be calculated as a weighted sum of the errors from the next layer, i.e. propagating the error backwards. The procedure of weight updates is repeated until a stopping criterion is met. Such a criterion can be a number of epochs (full update of weights), reaching some value of error or no significant improvement compared to previous epochs.

The concept of MLP is essential to any other ANNs architecture. MLPs are used as classification or regression algorithms and, in spite of their black-box nature, received a lot of attention from research community. A high level of flexibility of this model fostered development of many other architectures based on MLP. Although, many of them were invented decades ago, due to low amount of computational resources, they became popular only recently. In the following, we will cover two additional ANNs architectures: convolutional neural networks (CNNs) and recurrent neural networks (RNNs).

Convolutional Neural Networks

Convolutional neural networks are a special type of ANN architecture for processing data with known, grid-like structure, such as time series (one-dimensional) or images (two-dimensional). The name *convolutional* comes from the mathematical operation *convolution* that is used instead of matrix multiplication, which, however, does not correspond precisely to the operation performed in CNN. The convolution is usually denoted with an asterisk (*) and for one-dimensional real valued case described with the following formula:

$$s(t) = (x * w)(t) = \int x(a) w(t-a) da, \tag{2.14}$$

where x is an input vector, t is a real valued time index, $w(a)$ is a weighting function and a is a position value of our data, e.g. an age of measurement.

However, in real-life setting, values at every instant are impossible to obtain. The discrete convolution operation for the same one-dimensional input can be defined as:

$$s(i) = \sum_{a=-\infty}^{\infty} x(t) w(t-a), \tag{2.15}$$

where time index t is now an integer. The second argument – w is often refereed to as the kernel. For a two-dimensional input I, e.g. an image, and two-dimensional kernel K the convolution operator for discrete values can be defined as:

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n) K(i-m, j-n). \tag{2.16}$$

however, due to the commutative property of convolution, it can be rewritten as:

$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n). \tag{2.17}$$

The operation of convolution is similar to cross-correlation, but with a flipped kernel K. However, in many machine learning libraries, cross-correlation is used. It can be defined with the following formula:

$$S(i,j) = (K * I)(i,j) = \sum_m \sum_n I(i+m, j+n)K(m,n). \tag{2.18}$$

In contrast to MLPs, CNNs use three important ideas to improve their learning abilities: sparse interactions, parameter sharing and equivariant representations. Sparse interactions or sparse connectivity is obtained by having a kernel smaller than the input. That is, for calculating consequent output, one uses not every input as in MLPs, but only a certain subset. It allows to significantly reduce number of parameters and, therefore, number of operations to compute the output, i.e. to increase computational efficiency of the algorithm. Parameter sharing refers to the usage of the same value of parameter to calculate more than one output in a model. This does not affect the runtime of forward propagation in CNNs, as the same amount of operations are to be performed, however, it significantly reduces the amount of weights, which leads to reduced storage and higher statistical efficiency. Equivariant representations refer to the situation when changes in input lead to similar changes in output. That means that we can transform an image I to I' with a function g and then apply the convolution operator to it, which will be similar to applying the convolution to I' first and then function g to an output.

A typical layer in CNN consists of three stages: convolution, detector and pooling. At the first stage, above-mentioned convolution operator is applied to input. At the second – activation function is applied to the output of convolution, similarly as described previously for MLP. At the pooling stage, the output is further modified by replacing it with a summary statistic of the nearby values. It helps to make representations invariant to small translation of the input. The most commonly used strategy for pooling is max pooling (Zhou and Chellappa, 1988), which provides the maximum value withing a rectangular neighborhood of an output. Pooling also reduces the image size and after several layers of convolution and pooling, features are small enough to be flattened and used as an input to MLP at the end of CNN, that is responsible for classification or regression itself.

The procedure of CNN learning is similar to the one described previously for MLP – error backpropagation. One should note that due to weight sharing approach, not all outputs y of layer l are connected through weights w_l as inputs in layer $l+1$. On the other hand, several pairs of output-inputs are responsible for the particular weight $w_{i,j}$. The process of learning is performed only at convolutional layers and does not affect pooling layers as they do not have any adjustable parameters (LeCun et al., 1989).

CNNs are widely used for feature extraction from the grid structured data, such as images. Trained on large datasets, they are able to solve computer vision tasks with an outstanding performance. On the tasks of face recognition or object detection, they perform at a level even close to that of humans (Russakovsky et al., 2015; Szegedy et al., 2015), which firmly established them in the field of computer vision. They also found their application for feature extraction from audio signals (Hershey et al., 2017).

Being arguably the most actively developing types of ANNs, numerous architectures of CNNs emerged in recent years and new ones keep showing the best scores on benchmark

tasks on a scale of months. Various CNN architectures are also widely applied for emotion recognition research. CNN was used as the feature extractor of the baseline end-to-end system for Interspeech ComParE 2017 (Schuller et al., 2017). Huang et al. (2017) proposed to use a wide range of features, which include deep visual features based on AlexNet (Krizhevsky et al., 2012) for video modality. These features were later fed into RNN-LSTM. Chen et al. (2017) used CNNs not only for video (Huang et al., 2017), but also for audio modality (Aytar et al., 2016). Their systems significantly outperformed baseline of AVEC 2017. Tan et al. (2017) used several architectures for their two-level (face and global image) emotion recognition system: VGG19 (Simonyan and Zisserman, 2014), BN-Inception (Ioffe and Szegedy, 2015), and ResNet101 (He et al., 2016). Similarly, Guo et al. (2017) used CNNs for their three-level (face, skeleton and global image) system: VGG-Face model (Parkhi et al., 2015), Inception-v2 (Szegedy et al., 2016) and ResNet-152 (He et al., 2016). For time-continuous emotion recognition, CNNs are often combined with RNN-LSTMs to build a complete end-to-end system.

Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are architectures of ANNs designed to process sequential data. They received a lot of attention in deep learning and form the basis for many state-of-the-art machine learning applications, such as Automatic Speech Recognition or Automatic Translation. There are several key features of RNNs in contrast to MLPs. Outputs and hidden states of RNNs are dependent on previous hidden states within a certain time series. Similarly to CNNs, RNNs also utilize a shared weight strategy, but in a different manner. Each output produced in the same manner as previous, using the same parameters (weights). A hidden state of RNN can be formulated with a following equation:

$$h^t = f(h^{t-1}, x^t, \theta), \tag{2.19}$$

where f is a function that maps the state at time $t-1$ to the state at time t, x is an input vector at time t, θ is a set of parameters. A computation graph for hidden state h^t can be unfolded:

$$\begin{aligned} h^t &= f(h^{t-1}, x^t, \theta) = f(f(h^{t-2}, x^{t-1}, \theta), x^t, \theta) \\ &= f(f(f(h^{t-3}, x^{t-2}, \theta), x^{t-1}, \theta), x^t, \theta) = \dots \end{aligned} \tag{2.20}$$

This process can be repeated until the first time step is reached. Initial hidden state at $t = 0$ is defined during initialization process. For a forward pass of RNN for time steps from $t = 1$ to $t = \tau$ (where τ is the number of time steps), the following equations are used to calculate the output:

$$a^t = Wh^{t-1} + Ux^t + b, \tag{2.21}$$

where a is an intermediate state of a hidden neuron at time t, b is a bias vector, W is a weight matrix for hidden-to-hidden connections, i.e. from previous hidden state to current one, and U is the weight matrix for input-to-hidden connections, i.e. from input value to current hidden state. Similarly to Equation 2.2, the final state of a hidden neuron is calculated by applying an activation function to a^t:

$$h^t = \phi(a^t), \tag{2.22}$$

The output for the current time step t is calculated based on hidden state according to the following equation:

$$o^t = Vh^t + c, \tag{2.23}$$

where c is another bias vector and V is a weight matrix for hidden-to-output connections, i.e. from hidden state to output. Depending on the task, an additional function can be applied to o^t, e.g. for classification task one often calculate prediction of the model using softmax activation function:

$$\hat{y}^t = softmax(o^t) \tag{2.24}$$

However, for regression tasks mostly used in this thesis, the common activation function is linear, therefore:

$$\hat{y}^t = o^t. \tag{2.25}$$

This set of equations maps an input sequence to an output sequence. Such modeling type is called sequence-to-sequence and used primarily in this thesis. For a backward pass, one applies the Backpropagation algorithm described previously. However, there are several peculiarities emerging in RNN training. As each hidden state h^t depends not only on input sequence x^t and some parameters or weight matrices, but also on previous hidden state h^{t-1}, each subsequent gradient of loss function with respect to weight matrices is connected to the previous gradient until initial hidden state h^0. That means that gradients should be calculated recursively. This process is called Backpropagation Through Time (BPTT). In general, having loss function defined in Equation 2.4, partial derivative of e^t with respect to the weight matrix V will be equal to:

$$\frac{\delta e^t}{\delta V} = \frac{\delta e^t}{\delta \hat{y}^t}\frac{\delta \hat{y}^t}{\delta o^t}\frac{\delta o^t}{\delta V}. \tag{2.26}$$

The total loss function is equal to the sum of losses at each time step t. Taking Equation 2.25 and Equation 2.8 into account for linear activation function, it can be rewritten as:

$$\frac{\delta e^t}{\delta V} = -e^t\frac{\delta o^t}{\delta V} = -e^t {h^t}^\top. \tag{2.27}$$

For the other two weight matrices W and U it is more complex, as recursive gradient is required. For example, partial derivative with respect to W is:

$$\frac{\delta e^t}{\delta W} = \frac{\delta e^t}{\delta \hat{y}^t}\frac{\delta \hat{y}^t}{\delta o^t}\frac{\delta o^t}{\delta h^t}\frac{\delta h^t}{\delta W} = -e^t V\frac{\delta h^t}{\delta W}. \tag{2.28}$$

The last term in this equation – h^t – depends on W:

$$\frac{\delta h^t}{\delta W} = \frac{\delta}{\delta W}\phi(Wh^{t-1} + Ux^t + b). \tag{2.29}$$

Let us designate the argument of function ϕ as z^t: $z^t = Wh^{t-1} + Ux^t + b$. Then we can rewrite the last equation as:

$$\frac{\delta h^t}{\delta W} = \frac{\delta \phi}{\delta z^t}\frac{\delta z^t}{\delta W} = \phi'(z^t)\left(h^{t-1} + W\frac{\delta h^{t-1}}{\delta W}\right). \tag{2.30}$$

One may notice that the last term $\frac{\delta h^{t-1}}{\delta W}$ also depends on W. In turn, to calculate it, one should know $\frac{\delta h^{t-2}}{\delta W}$ and so forth until initial state h^0. This recursiveness is the main idea of BPTT. Similar strategy is applied to the weight matrix U:

$$\frac{\delta e^t}{\delta U} = -e^t V \frac{\delta h^t}{\delta U}, \tag{2.31}$$

where the last term $\frac{\delta h^t}{\delta U}$ is defined as follows:

$$\frac{\delta h^t}{\delta U} = \phi'(z^t)\left(x^t + U\frac{\delta x^t}{\delta U} + W\frac{\delta h^{t-1}}{\delta U}\right). \tag{2.32}$$

As x^t does not depend on U, the second term in second parenthesis is zero, while the last term depends on U similarly to $\frac{\delta h^{t-1}}{\delta W}$ depending on W in previous equations. Therefore, recursiveness is required for calculation of the weight matrix U as well.

Multiplication of many gradients and outputs of activation functions leads to the problem of vanishing or exploding gradients (Hochreiter, 1991; Bengio et al., 1993, 1994). To overcome these issues, several approaches were developed. Among them is the change of RNN architecture to gated RNNs.

Long Short-Term Memory

The most widely used gated RNN is the Long Short-Term Memory (LSTM) model (Hochreiter and Schmidhuber, 1997). The main idea of it was to make a self-loop – additional recurrence – that is conditioned on context and can dynamically change the time scale of integration. Weights between various inputs in LSTM cell are controlled by three gates, namely, input, output and forget gate.

Forget gate determines which information should be drawn from previous state and current input. An equation for the output of the forget gate is as follows:

$$f_i^t = \sigma\left(b_i^f + \sum_j U_{i,j}^f x_j^t + \sum_j W_{i,j}^f h_j^{t-1}\right), \tag{2.33}$$

where x^t is an input vector for current time step t, h^t is a hidden layer vector, b^f is a bias for the forget gate, U^f and W^f are weight matrices for input and recurrent connection respectively. Similarly, the output of the input gate (also external input gate) is set as follows:

$$g_i^t = \sigma\left(b_i^g + \sum_j U_{i,j}^g x_j^t + \sum_j W_{i,j}^g h_j^{t-1}\right), \tag{2.34}$$

where b^g, U^g and W^g are parameters, similarly to forget gate. Having outputs of forget and input gates, an internal state s_i^t of the cell can be updated:

$$s_i^t = f_i^t s_i^{t+1} + g_i^t \sigma\left(b_i + \sum_j U_{i,j} x_j^t + \sum_j W_{i,j} h_j^{t-1}\right). \tag{2.35}$$

The output h_i^t of LSTM cell is calculated as:

$$h_i^t = \phi(s_i^t) q_i^t \tag{2.36}$$

and is controlled via the output q_i^t of the output gate:

$$q_i^t = \sigma\left(b_i^o + \sum_j U_{i,j}^o x_j^t + \sum_j W_{i,j}^o h_j^{t-1}\right). \tag{2.37}$$

As activation function ϕ for the output of LSTM cell, a hyperbolic tangent $tanh$ is commonly used. For gating, a sigmoid activation function σ is used in order to control the output values in range $[0, 1]$.

RNN-LSTMs are very widely used for time-continuous emotion recognition. They are dominating methodology in this area and already a standard approach. They are used as baseline systems for challenges on emotion recognition starting from 2015 (Ringeval et al., 2015) and also by most of the participants. Huang et al. (2018) used multiple feature sets with a fusion of predictions obtained with unimodal RNN-LSTM models. Zhao et al. (2018) used this RNN architecture with CNN based feature extraction subsystem (Hershey et al., 2017). Liu et al. (2018) used RNN-LSTMs in end-to-end scenario combined with VGG-16. Li et al. (2019) used a bidirectional RNN-LSTM (BLSTM) combined with CNNs to capture dynamics of information in the audio-visual emotion recognition task. All these approaches outperformed existing benchmarks in respective fields.

In this thesis, ANNs are the core modeling algorithms with a special focus to RNN-LSTM due to their ability to model long-term dependencies. However, other methods that do not take context into consideration are also used for comparison. We will further describe them.

2.4.2 Ridge Regression

Ridge Regression is a Linear Regression with L2 regularization. Linear Regression assumes a linear relationship between y (dependable variable) and x_i (independent vectors) with a noise ϵ:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \tag{2.38}$$

Where β_0 is y-intercept and β_1 to β_n are coefficients for variables x_1 to x_n. This version of the equation is used for regression problems; for classification it slightly differs, but the main ideas are the same. Coefficients β_i for $i = 0..n$ are selected by minimizing the residual sum of squares between predicted and true values:

$$RSS = \sum_{j=1}^{N} (\hat{y_j} - y_j)^2 = \sum_{j=1}^{N} \left(\hat{y_j} - \beta_0 - \sum_{i=1}^{n} x_{ji}\beta_j\right)^2, \tag{2.39}$$

where $j = 1..N$ is a number of data sample, $\hat{y}$ is a value predicted with our model. The optimization task can be solved with any optimization algorithm, such as Ordinary Least Squares or Gradient Descent. Numerical methods are preferred to solve the problem with the high number of features and/or samples. In the Ridge Regression there is a penalty term combined with RSS:

$$PRSS = \sum_{j=1}^{N} \left(\hat{y_j} - \beta_0 - \sum_{i=1}^{n} x_{ji}\beta_j\right)^2 + \lambda \sum_{i=1}^{n} {\beta_i}^2, \tag{2.40}$$

It penalizes high coefficients β_i, reducing the slope. The Ridge Regression is an effective approach to problems with many important variables, i.e. such variables x_i that make contribution to the dependent variable y. If this is not the case and only few variables x_i

are important, Lasso Regression or Linear Regression with L1 Normalization is preferred. It allows not relevant variables to be "shut down" by relatively large λ, therefore, working similarly to feature selection approaches.

Linear regression and its analog for classification tasks – logistic regression – work under several assumptions about independence of variables, which is not always the case in practice. However, due to its simplicity, it was widely used in the past, later supplanted by other machine learning algorithms, such as Support Vector Machines (SVM). Coulson (2004) used logistic regression to study attribution of emotions to body postures. Soleymani et al. (2011) used Ridge Regression for continuous emotion detection from music videos. Jenke et al. (2013) studied evaluation measures for emotion recognition and used Ridge Regression as one of the basic recognition models in their research.

2.4.3 Support Vector Machines

Support Vector Machine (SVM) is another widely used algorithm for classification and regression. Having a data set $\{x_i, y_i\}_{i=1}^{N}$, for classification, SVM seeks an optimal hyperplane that separates input values x_i with respect to output values y_i. A hyperplane can be defined as:

$$\omega^T x + b = 0, \tag{2.41}$$

that should fulfill the following conditions:

$$\begin{aligned} \omega^T x + b \geq 0, y_i = +1 \\ \omega^T x + b < 0, y_i = -1 \end{aligned} \tag{2.42}$$

The formal problem definition will then be as follows:

$$y_i(\omega^T x_i + b) \geq 1, \tag{2.43}$$

such that it minimizes the convex loss function:

$$L = \frac{1}{2}\omega^T \omega. \tag{2.44}$$

SVM works similarly for regression: it seeks a function $f(x, \omega)$ with at most ϵ-deviation of predictions $\hat{y}$ from the target vector y. The following conditions are to be fulfilled:

$$\begin{aligned} y_i - \omega_i x_i - b \leq \epsilon \\ \omega_i x_i + b - y_i \leq \epsilon \end{aligned} \tag{2.45}$$

The optimization problem defined in Equation 2.44 is solved with the method of Lagrange multipliers (Lasdon, 2002) or with another optimization method. One of the most commonly used is Sequential Minimization Optimization (SMO) (Plattet al. , 1999).

SVM and SVR (SVM for regression) are used for emotion recognition at two levels: for recognition itself and for modality fusion. The first approach was popular previously, before RNNs became widely used for this task. Thus, Grimm et al. (2007) compared SVR to other algorithms for solving the task of automatic recognition of spontaneous emotions in speech and found it to perform the best. Chang et al. (2013) used SVR for emotion recognition based on physiological features, such as electrocardiogram, galvanic skin response, blood volume pulse, etc. SVR and SVM were also used as the baseline methods for many emotion

recognition challenges, such as AVEC 2012-2014. However, Nicolaou et al. (2011) compared the performance of SVR and BLSTM, reporting higher performance for the latter algorithm. Nowadays, SVR and SVM are often used in time-continuous emotion recognition for fusion of feature sets or modalities as a meta classifier or regressor. Huang et al. (2017) proposed to use a wide range of features, train separate LSTM models feature- and dimension-independently and then fuse results with SVR. For EmotiW 2018, Liu et al. (2018) trained and fine-tuned four CNN-based models for feature extraction: Inception-v3 (Szegedy et al., 2016), DenseNet-121, DenseNet-161, DenseNet-201 (Huang et al., 2017) and then fed them to SVM for classification.

2.4.4 XGBoost

XGBoost (Chen and Guestrin, 2016) stands for eXtreme Gradient Boosting and is relatively novel. Similarly to other algorithms, it can be used for regression and classification with various types of variables. XGBoost has a slightly different procedure of tree building compared to Gradient Boosted Trees, and it is called "extreme" because of many additional optimization techniques that make XGBoost fast and efficient.

In XGBoost one starts to build a regression tree with a single leaf – the average value of input data. Then, the residuals corresponding to this leaf are split into two groups: less and more than a certain value for current parameter. The exact value among the options is decided based on gain function, which in turn based on similarity score:

$$S = \frac{\left(\sum_{i=1}^{N} r_i\right)^2}{N + \lambda}, \tag{2.46}$$

where S is a similarity score, r_i is a residual for i^{th} element of the leaf, N is the total amount of elements of the leaf, and λ is a regularization parameter. The gain of the new tree split is calculated as:

$$G = S_{left} + S_{right} - S_{root}, \tag{2.47}$$

where S_{left}, S_{right} and S_{root} are similarity scores for the left leaf of the new tree, the right leaf and its root (without splitting) respectively. After trying several splits, we select the one with the highest gain. Then, the tree is split further; usually, up to 6 levels are used. Gain values are also used for tree pruning – reducing its size. We compare gain of each branch to a defined value γ and if gain is smaller than γ, we will remove the branch. One should also note that setting $\gamma = 0$ does not cancel pruning, as gain can be negative value when regularization parameter is used.

The prediction value for the input sample is build as a sum of the output of the original tree (one leaf) and the output of each subsequent tree multiplied by learning rate. The goal is to find an optimal output value for the leaf that minimizes the following objective function:

$$L = \left(\sum_{i=1}^{n} L(y_i, p_i)\right) + \gamma T + \frac{1}{2}\lambda O^2. \tag{2.48}$$

where T is the number of terminal nodes in a tree. Prediction value p_i can be specified as the previous state plus the addition from the new tree: $p_i^t = p_t^{(t-1)} + O$. The objective function can be approximated with the Second Order Taylor Approximation:

$$L(y_i, p_i^{t-1} + O) \approx \sum_{i=1}^{n} \left(L(y_i, p_i^{t-1}) + \frac{\delta L(y_i, p_i^{t-1})}{\delta p_i^{t-1}} O + \frac{1}{2} \frac{\delta^2 L(y_i, p_i^{t-1})}{\delta (p_i^{t-1})^2} O^2 \right) + \gamma T + \frac{1}{2} \lambda O^2. \tag{2.49}$$

For simplicity, this equation is rewritten using g for the first order partial derivative (gradient) and h for the second order partial derivative (Hessian):

$$L(y_i, p_i^{t-1} + O) \approx \sum_{i=1}^{n} \left(L(y_i, p_i^{t-1}) + gO + \frac{1}{2} hO^2 \right) + \gamma T + \frac{1}{2} \lambda O^2. \tag{2.50}$$

As the first term $L(y_i, p_i^{t-1})$ does not have any effect on O, we can omit it. Then, taking the derivative with respect to O and solving the equation for it, we get:

$$O = \frac{-\sum_{i=1}^{n} g_i}{(\sum_{i=1}^{n} h_i) + \lambda} \tag{2.51}$$

Gradients g of the loss function for regression task are negative residuals $-(q_i - p_i)$ and Hessians of this loss function are derivatives of g with respect to p_i, hence, they are equal to 1. Therefore, the optimal output value for the leaf is calculated with an equation similar to the one for similarity score:

$$O = \frac{\sum_{i=1}^{N} r_i}{N + \lambda}. \tag{2.52}$$

We keep building new trees according to this procedure until the residual are smaller than a certain predefined value, or we reach the maximum number of trees. Other parts of XGBoost are related to optimization and improvement of its effectiveness. Namely, Approximate Greedy Algorithm for large datasets provides a trade-off between small number of thresholds for leaf splitting, that are not computationally expensive but may lead to poor splits, and high number that are computationally expensive but provide good splits. In Approximate Greedy Algorithm we use quantiles to decide on the positions of splits. In turn, quantiles are also calculated approximately, using Weighted Quantile Sketch Algorithm that computes histograms of subsets of data in parallel. Next technique, Sparsity-Aware Split Finding decides on how to treat missing values. While splitting, it puts all the observations with missing value of a particular variable into left and right leaf and calculated gain score for them. Based on this score, the general path through the tree for all missing values is decided. Cache-Aware Access related to computational optimizations – gradients and Hessians are stored in cache memory so that similarity scores and output values can be calculated faster. The last technique – Blocks for Out-of-Core Computation – compresses the data that is too large for cache memory and RAM, to save the time of accessing it from a hard drive.

XGBoost is relatively new approach, which, however, already became one of the most popular method for solving challenging tasks on such platforms as Kaggle[4] or KDD[5]. Winning solutions for challenges on various topic (e.g. (Mangal and Kumar, 2016)) are based on XGBoost, outnumbering deep neural networks (the second popular approach). XGBoost is also used for emotion recognition. Wang et al. (2018) used it for their entropy-based pipeline built on features from physiological signals. Xing et al. (2019) used XGBoost to train on a fusion of video and electroencephalography (EEG) features.

[4] https://www.kaggle.com/

[5] https://www.kdd.org/

2.5 Summary

In this chapter we considered general approaches to emotion recognition, important developments in utilization of the contextual information, recent advances in problem statement and applied methodology as well as relevant scientific background and state-of-the-art methods used for emotion recognition. Thus, in the first part of the chapter, we presented various approaches to emotion recognition regarding input data, namely, time-continuous and utterance-level; and output of a model, namely, categorical and dimensional. In early research on emotion recognition, utterance-level categorical approach was used. However, nowadays, focus is shifted towards time-continuous dimensional approach, which is also used in this thesis.

Then, we provided an overview of studies on each of the three levels of context considered in our work: speaker, dialogue and environment. We emphasized, that dealing with real-life conditions, an emotion recognition system faces contextual information from various sources. Previously, this information was perceived as noise and was preferably excluded from the source data prior to the recording process by using laboratory-based setups or afterwards – by filtering. However, recent studies have shown that these sources of data may be extremely useful. How person speaks, what are background sounds or surroundings, how interlocutor reacts in a dialogue setting – all these factors play an important role in better understanding the current situation. The amount of speaker level context was proven to have an impact on a system performance in several studies. It was also shown, that it depends on modality and dimension. Presence of interlocutor and his emotions were proven to affect emotions of speaker, and many interaction strategies were investigated. However, they were not time-continuous, but based on utterance-level analysis of a dialogue. A number of studies showed a possibility of recognizing emotions based on data from wearable devices. Nevertheless, the reason of emotional changes was often not considered. Thus, in spite of the fact that these issues of contextual emotion recognition were covered to some extent in existing research, a more comprehensive analysis in each area seems to be timely. This research will be carried out in this thesis at each of the three presented levels.

An additional boost to research on contextual emotion recognition was given by numerous organized challenges. Summarizing them, one may derive several conclusions. Firstly, emotion recognition and paralinguistics attracted great interest of researches over the last decade. Not only the number of challenges, but also the number of tasks and participating teams increased dramatically. Secondly a shift from classical approaches to context-aware temporal methods was made, and now these models dominate the field. Initially, such approaches were utilized by participants in their submissions and later became a state-of-the-art and were included into the baseline systems. Finally, this shift in methods allowed a gainful usage of contextual information, improving the quality of predictions. State-of-the-art and challenge-winning architectures and algorithms form a significant basis for the methodology used for experiments carried out in this thesis. They have been covered in detail in the last section of this chapter.

In the following chapter we will present a description of corpora, data preprocessing steps and model evaluation metrics used in this thesis.

3 Data and Tools

This chapter introduces corpora, methods and evaluation metrics used in this thesis. It also covers data pre-processing steps required, such as feature extraction and continuous emotion prediction specific step of reaction lag correction.

3.1 Corpora

In our work we use five corpora: RECOLA, SEMAINE, SEWA, IEMOCAP and UUDB. Our motivation for selecting these corpora among others available for research community is their: (i) recording scenario and conditions – they consist of emotionally colored interactions collected in **spontaneous** scenario or close to it; (ii) annotation approach – they are annotated **time-continuously** in at least two affective primitives: **arousal** and **valence**; (iii) available modalities – most of them have **audio** and **video** data. We will now present these corpora, providing a short description, corpus suitability for dialogue-level emotion recognition, statistics in terms of labels and recording lengths distributions, and a brief literature overview with recent or most significant studies.

For the raw signal based feature extraction (Section 3.2.2) we additionally use AffectNet database, which is briefly covered in description of the feature extraction process. In our research on smart environments (Chapter 6), we use an additional corpus. It was collected by ourselves in scope of the *EmoTour Project*. This database is rather specific and mostly does not follow standard procedures of data preprocessing and feature extraction described in this section, hence, these aspects are covered in detail later in the corresponding chapter.

3.1.1 RECOLA

The *REmote COLlaborative and Affective interactions* database (Ringeval et al., 2013) was collected for EmotiBoard project in the University of Freiburg for building real-time emotion recognition systems to augment remote collaborations with emotional feedback and measuring its impact on teamwork quality.

Corpus characteristics. The RECOLA consists of spontaneous collaborative interactions between participants divided into 23 dyadic groups (46 participants: 27 females, 19 males), recorded remotely. Participants were first asked to solve a survival task individually and then discuss it with their interlocutors. The survival task was designed by the National Aeronautics Space Administration (NASA) (Hall and Watson, 1970); one should rank items according to their importance in case of a disaster scenario, e.g. a plane crash. Mood of some participants was induced prior to discussion by short video clips with strong positive or negative content. The data was collected in four modalities: audio, video, electro-dermal activity (EDA), electrocardiogram (ECG). The mean duration of discussions within each pair was approximately 15 minutes. However, as most of the discussions occurred at the beginning of the talks, each recording was shortened to a fixed length of five minutes.

Annotators were instructed prior to providing ratings and asked to perform a trial annotation using two videos from SEMAINE database (McKeown et al., 2010). Annotation tool ANNEMO was developed and used for RECOLA database in order to facilitate remote annotation. Two affective (arousal, valence) and five social behavior scales (agreement, dominance, engagement, performance, rapport) were used. Arousal and valence were annotated continuously, with frequency of 25 Hz; social behavior was annotated as one label per recording. In total, six annotators (3 males, 3 females) participated in the process. Authors state that non-French speaking raters provide much less detailed labels, therefore, only French-speaking raters were selected. Label post-processing was performed to increase inner rater agreement and to level issues caused by re-annotation and missing data. Labels distribution is presented in Fig. 3.1.

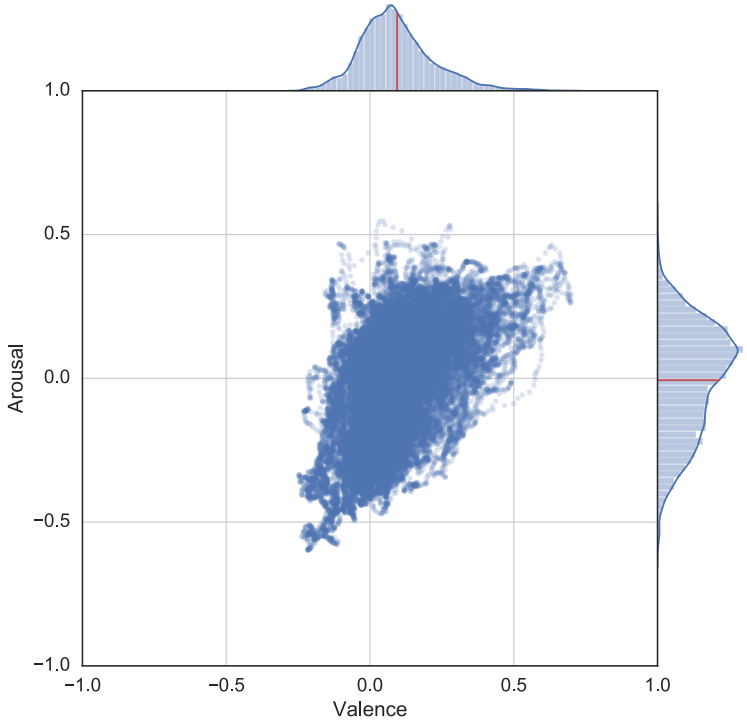


Figure 3.1: RECOLA database labels distribution. Scatter plot, histograms for valence (x-axis) and arousal (y-axis), as well as corresponding KDE-plots are in blue. Red lines show the mean value for valence and arousal calculated on the whole database.

Only a part of RECOLA database – 23 recordings from 23 speakers (10 males, 13 females, with mean age of 21.35 (σ=2.04)) – was shared with research community. In this thesis, we will refer to this part exclusively. All participants were French-speaking, but had different mother tongues: 17 French, 3 German and 3 Italian. Physiological features (ECG and EDA) are not available for 5 recordings. In spite of the fact that RECOLA was recorded in a dyadic interaction scenario, it cannot be used for dialogue based emotion recognition, as there is missing data for many interlocutors since they didn't give consent to share and publish their data.

Corpus usage. RECOLA became available to the scientific community in 2013 and since then was extensively used in many studies on emotion recognition. Trigeorgis et al. (2016) proposed a deep learning based approach to robust, context-aware feature extraction. Authors

used CNNs applied to raw data (wave audio signal) to learn the representations, and stacked BLSTM on top to perform emotion predictions. They first segmented raw signal to 6 seconds intervals, then used two 1D convolutional (with filter size 40 for both) and two pooling layers (pool size 2 and 20) to map input data to original frequency of labels. BLSTM layers have 128 cells each; however, authors reported similar performance for unidirectional LSTM. As a baseline, they used predictions obtained with SVR and BLSTM on two conventional feature sets: eGeMAPS (Eyben et al., 2016) and the one used in Interspeech ComParE 2013 (Schuller et al., 2013). Having access to the complete dataset, they used all 46 recordings and reported their approach to significantly outperform the baseline systems: 0.686 vs. 0.382 for arousal and 0.261 vs. 0.195 for valence in terms of concordance correlation coefficient (CCC).

Ringeval et al. (2015) proposed another LSTM based approach to continuous emotion prediction. They made use of two types of multi-task learning: analyzing ratings provided by each annotator as a separate target variable and training model to predict arousal and valence simultaneously. Hence, their networks could have one output (single-task, averaged labels), two outputs (multi-task, averaged labels), six outputs (single-task, original labels) or 12 outputs (multi-task, original labels). Authors experimented with each of four available modalities and reported that performance can be significantly improved by using rating of all annotators simultaneously in one model, which was, however, not the case for the audio modality. This can imply that network can deal with asynchronous dependencies and use this data beneficially. Their unimodal models showed CCC of 0.788 for arousal with audio and 0.431 for valence with video features. In addition, they implemented two fusion strategies and obtained 0.769 of CCC for arousal with audio+visual and 0.492 for valence with audio+visual+ECG features using feature-level fusion, and 0.804 for arousal with audio+visual+EDA and 0.528 for valence with all modalities combined using decision-level fusion.

Tzirakis et al. (2017) proposed another end-to-end strategy to utilize both audio and visual features. They used slightly different architecture as in (Trigeorgis et al., 2016) for audio modality and deep residual network ResNet-50 (He et al., 2016) for video modality. They fused outputs of these unimodal models on feature-level and trained 2-layer LSTM with 256 cells on both layers. Their multimodal approach showed 0.714 CCC for arousal and 0.612 for valence, outperforming contributions submitted to AVEC 2016 (Valstar et al., 2016a).

He et al. (2015) proposed a multimodal approach based on BLSTMs. They used feature sets provided by organizers of AVEC 2015 (Ringeval et al., 2015), as well as additional sets, including 158 LLDs extracted with YAAFE toolbox (Mathieu et al., 2010) for audio modality, 768 LPQ-TOP features for video modality, 52 features from ECG signal and 22 from EDA signal from various signal analysis domains. After feature extraction, authors fed them into unimodal BLSTM models, with another BLSTM model stacked afterwards for decision-level fusion. Authors reported up to 0.747 of CCC for arousal and 0.609 for valence, compared to 0.444 and 0.382 respectively as a baseline.

Architectures and performance measures used in aforementioned works form a basis for the methodology utilized in this thesis (see Chapter 4 and Chapter 5).

3.1.2 SEMAINE

The *Sustained Emotionally colored Machine-human Interaction using Nonverbal Expression* database (McKeown et al., 2011) was designed to develop systems for machine-human interactions and collected for the eponymous project by Queen's University Belfast with technical support of the iBUG group of Imperial College London.

Corpus characteristics. The SEMAINE consists of three parts, representing interaction between user and sensitive artificial listener (SAL) at three levels:

1. solid SAL – a human operator simulates an agent;
2. semi-automatic SAL – a human operator chooses phrases from a predefined list;
3. automatic SAL – the utterances and nonverbal actions are decided entirely automatically by the current version of the designed system.

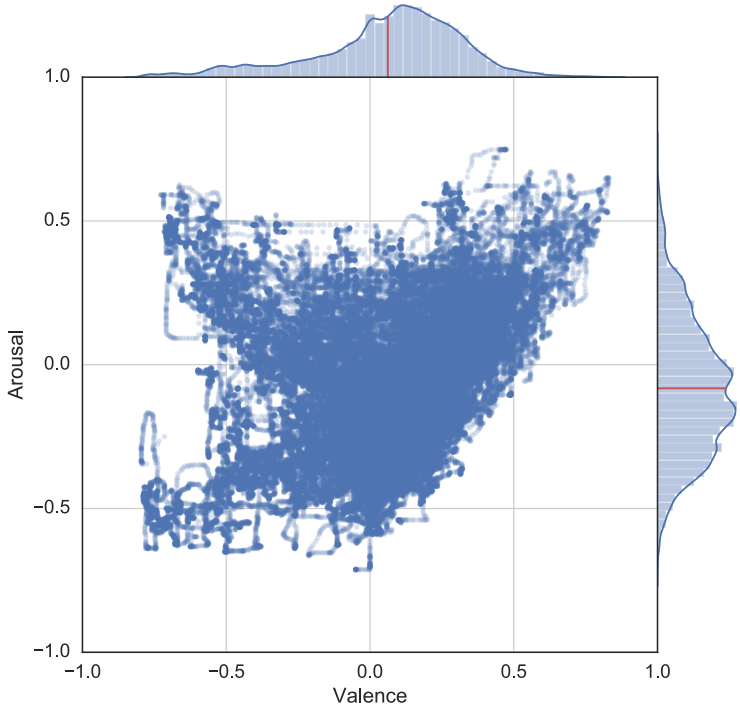


Figure 3.2: SEMAINE database labels distribution. Scatter plot, histograms for valence (x-axis) and arousal (y-axis), as well as corresponding KDE-plots are in blue. Red lines show the mean values for valence and arousal calculated on the whole database.

In our work, we use solid SAL part, which is shared with the research community on the official website of the SEMAINE database[1]. Each recording has two roles: user and operator (or agent, or artificial listener). User represents a target person, whose emotions should be predicted and whose emotions the operator tries to affect with his/her behaviour. Operator has four roles, each corresponds to a certain behaviour pattern and given a particular name:

- Spike – constitutionally angry, tries to evoke anger to something or someone in the user;
- Poppy – constitutionally happy and cheerful, tries to evoke happiness in the user;
- Prudence – constitutionally sensible, tries to evoke rational behaviour or thoughts in the user;
- Obadiah – constitutionally gloomy, tries to evoke gloom and sadness in the user.

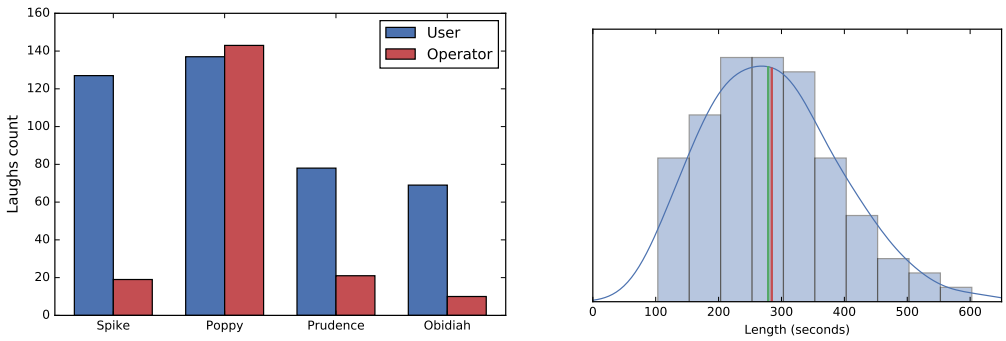
[1] `https://semaine-db.eu/`

These characters to some extent correspond to four personality types: Spike is choleric, Poppy is sanguine, Prudence is phlegmatic, and Obadiah is melancholic.

The solid SAL part used in this thesis consists of 95 multi-channel recordings with the total duration of 393.9 minutes (or 787.8 minutes of mono recordings – separately for user and operator). There are 21 participants (13 females, 8 males) of eight nationalities with average age of 31.6 years (σ = 11.7) and all speaking English.

The dataset is annotated time-continuously on several dimensions: arousal, valence, power, anticipation/expectation, intensity. In addition, there are annotations of particular emotions (e.g. fear, anger, happiness, etc.) for user, e.g happiness for Poppy, anger for Spike, etc. This was performed to check if the operator behaviour in fact affects the one of user in proper way. However, emotional annotations of operator are available only for a few sessions, hence we use the full user part for speaker-only emotion prediction (Chapter 4) and the operator part in addition for dyadic emotional context modeling (Chapter 5).

Label distribution for two basic affective primitives – arousal and valence – is presented in Fig. 3.2. It is worth to mention that the participants were easy to empathize with the positive characters, but much harder with the negative ones: they often react with laughter to anger or with ridicule and contempt to sadness, as found them inappropriate or stilted. Fig. 3.3a represents the laughs count for both roles and each character. One may notice that for Spike, Prudence and Obadiah the operator tries to behave within the predefined pattern and not laugh, but the user often reacts with laughter, especially to inappropriately angry behaviour of Spike, doing it almost as often as with cheerful Poppy.



(a) SEMAINE database laughs count for user and operator roles and each character: Spike, Poppy, Prudence and Obadiah. Data is based on word level aligned transcript, available for recordings 1 – 24 (sessions 1 – 129).

(b) SEMAINE database recordings length distribution. 10 bins histogram and corresponding KDE-plot are in blue, median value depicted as a green vertical line, mean value depicted as a red vertical line.

Figure 3.3: Laughs count and recordings length statistics for SEMAINE database.

Recordings of SEMAINE database are smoothly distributed (see Fig. 3.3b) in terms of length, with mean and median value at approximately 4.5 minutes.

Corpus usage. SEMAINE was recorded in 2008-2009 and published in several papers in 2010-2011 (McKeown et al., 2010, 2011). Being annotated time-continuously in such an extensive way, including different affective scales, basic emotions and transcripts, as well as having high-quality recordings, it became a great material for research on emotion recognition. Most studies of that time were focused on utterance-based emotion recognition and many of them dealt with laboratory-condition acted databases, but SEMAINE facilitated a gentle shift of research focus towards time-continuous emotion recognition.

Tian et al. (2016) proposed a hierarchical fusion strategy for acoustic and lexical features and compared it to classical feature-level and decision-level fusion. Authors used eGeMAPS (Eyben et al., 2016), LLDs from Interspeech ComParE 2010 (Schuller et al., 2010) and Global Prosodic Features (Bone et al., 2014) for audio and Disfluency and Non-verbal Vocalisation, Pointwise Mutual Information and Crowd-Sourced Emotion Annotation Features (Warriner et al., 2013) for text. LSTM modeling of 3-class classification problem showed that proposed approach of hierarchical fusion outperforms any unimodal model and both feature-level and decision-level strategies with accuracy of 61.7% for arousal and 51.2% for valence.

Nicolaou et al. (2011) fused audio data with facial expressions and shoulder gestures. For audio modality, MFCC and prosody features were used, for video – features extracted with 20 facial feature points, and for shoulders – with 2 points. Authors built a BLSTM based system to recognize continuous labels in terms of arousal and valence and compared its performance to the SVR based. They used (root) mean squared error and Pearson's correlation coefficient to measure the performance of approaches and utilized three fusion strategies: feature-level, decision-level and output-associative fusion. The latter incorporated cross-dimensional dependencies and was reported to achieve the highest performance of 0.642 for arousal and 0.796 for valence in terms of correlation coefficient.

SEMAINE was a subject of reaction lag analysis in several studies. Mariooryad and Busso (2013, 2015) studied the presence of a certain delay between an actual affective event and its notation by rater. As presented by authors, reaction lag correction has a significant effect on system performance and although it has greater impact on utterance-level emotion recognition, and the problem is partially vanishes when recurrent models (e.g. an RNN-LSTM) are used, correcting reaction lag clears the data from unnecessary and undesirable noise. Hence, other authors found it useful and reasonable also for recurrent models (Mencattini et al., 2017).

Reaction lag based on a combination of presented approaches is used in this thesis for the data preprocessing and covered in Section 3.2.3.

3.1.3 SEWA

The Automatic Sentiment Analysis in the Wild (SEWA) database was recently collected for the eponymous project between several academic and industrial institutions (Kossaifi et al., 2019).

Corpus characteristics. The SEWA sets the main focus to high demographic variability (it includes participants from six different countries) and strictly equal distribution of participants by age groups and gender. Participants were recorded under two different scenarios:

1. while watching an advertisement video chosen to elicit emotions and mental states, e.g. empathy, amusement, boredom or liking. The video contained no speech to be understandable regardless of participant's language knowledge;
2. while discussing the last advertisement with a partner in a video chat.

The SEWA database is rather new and for the last two years was used for several emotion recognition challenges, therefore, holders do not share it completely with research community. During AVEC 2017 (Ringeval et al., 2017a) organizers shared part of the database that contains data from German speakers divided into three subsets. Next year, at AVEC 2018 (Ringeval et al., 2018) they included data of Hungarian speakers by putting it into test set and shifted the focus to cross-cultural affect recognition. The year after, at AVEC 2019 (Ringeval et al., 2019) another speaker group was introduced – Chinese, extending cross-cultural setting to different continents.

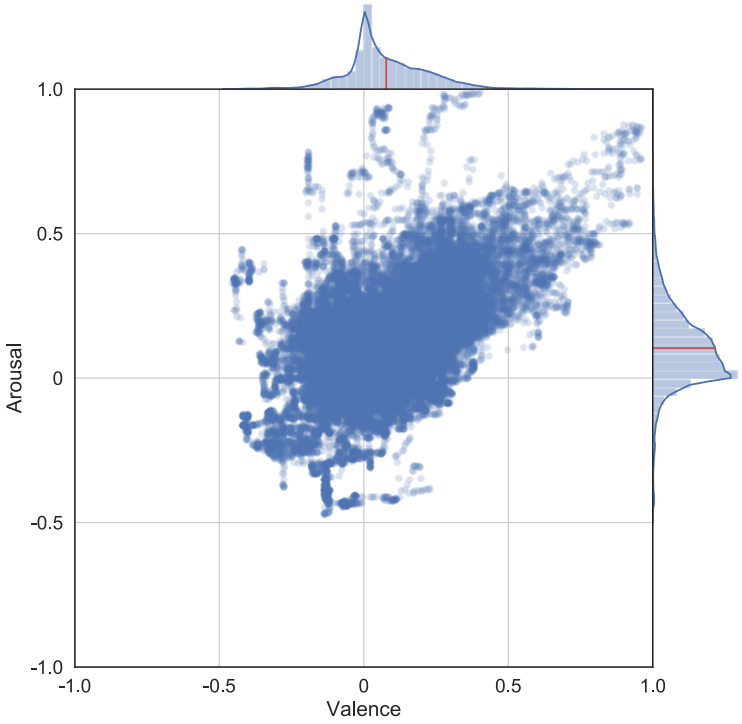


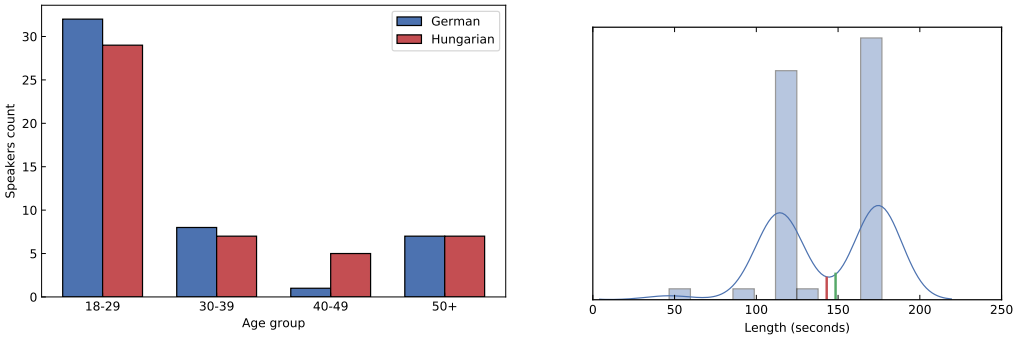
Figure 3.4: SEWA database labels distribution. Scatter plot, histograms for valence (x-axis) and arousal (y-axis), as well as corresponding KDE-plots are in blue. Red lines show the mean value for valence and arousal calculated on the whole database.

Authors claim that SEWA consists of audio-visual data from 398 people of six cultures: Chinese, English, German, Greek, Hungarian and Serbian. The participants split is almost equal in gender and covers five age groups: 18-29, 30-39, 40-49, 50-59, 60+ with most people being in the first two groups for each culture, except for Serbian. The total duration of the database is 44 hours; data is annotated by five raters using a custom-built tool on several scales, including arousal, valence and liking (in regard to the product presented in an advertisement).

However, at the moment (end of 2019) only a comparatively small part of data is available. Further in this thesis, we will refer to this part of the data exclusively as to "SEWA database". It consists of 48 pairs of interlocutors (96 participants), discussing advertisements in the German and Hungarian languages. The Chinese data was also presented in AVEC 2019, but only in test set and without any labels, therefore we do not use it in this work. Audio and close-up video data are provided for each dialogue, as well as time-continuous labelling on three affective scales: arousal, valence and liking. Label distribution for two basic affective primitives used in this thesis – arousal and valence – is presented in Fig. 3.4. The total duration of partition of SEWA database used in this research is 229 minutes, making it 114.5 minutes if we consider a dialogue as a single recording with two data channels (one for each interlocutor).

The gender ratio of participants is 1.0, which means that they are perfectly balanced. Age group distributions for both German and Hungarian cultures are presented in Fig. 3.5a. Distribution of recordings length is presented in Fig. 3.5b. One may see that it has two major peaks, covering almost each recording – approximately at 2 and 3 minutes.

Corpus usage. As SEWA database is only partially available, most of the studies used it



(a) SEWA database age groups distribution for German (blue) and Hungarian (red) speakers.

(b) SEWA database recordings length distribution. 10 bins histogram and corresponding KDE-plot are in blue, median value depicted as a green vertical line, mean value depicted as a red vertical line.

Figure 3.5: Age groups and recordings length statistics for SEWA database.

in the context of participation in emotion recognition challenges, described in Section 2.3. Several of these works were described there, and we will present only a few additional papers here.

Aspandi et al. (2020) proposed an approach based on generative adversarial networks, namely Star-GAN (Choi et al., 2018), which included two sub-networks: Auto-Encoder based Generator (AEG) and Conditional Discriminator based Affect Estimator. Their system tried to solve two tasks simultaneously: recognize fake images created by AEG and predict affective state in terms of arousal and valence. They combined audio and visual features through late fusion. Authors reported performance of 0.430 for arousal and 0.405 for valence in terms of CCC.

Atmaja and Akagi (2020) used multitask learning, combining arousal, valence and liking into one label vector. Authors used feature-level fusion of audio and video modality including such sets as eGeMAPS, DeepSpectrum, bag-of-audio words applied to MFCCs and eGeMAPS, Facial Action Units with their functionals as well as ResNet and VGG (see Section 2.3 for more details). Their three-layer LSTM network with 256-128-64 cells was optimized with RMSProp algorithm for 50 epochs. Authors proposed multistage fusion with SVR applied at several levels of the modeling pipeline. They reported this approach to significantly outperform baseline provided by organizers of AVEC 2019 (Ringeval et al., 2019), as well as unimodal, bimodal with early fusion and multimodal with late fusion systems: 0.680 for arousal and 0.656 for valence in terms of CCC.

Schmitt et al. (2019) questioned the need of using recurrent models, nowadays prevailing in the field of continuous emotion recognition. Authors used the development subset of AVEC 2018 (Ringeval et al., 2018) and tested two deep neural networks: BLSTM (4 layers) and CNN. For CNN they tried three architectures: 2-layer, 3-layer and 4-layer and chose the latter as the superior one. Filter sizes were chosen to be 5, 20, 30 and 50 respectively. This architecture allowed to capture 8 seconds of context from an audio signal through increasing receptive field provided by multilayer structure. Authors shifted labels in a range of 0 to 6 seconds to compensate delay in annotations and reported rather stable performance for CNN in contrast to BLSTM. Authors further studied the effect of the last layer's filter size on performance and reported no substantial influence. In conclusion, they listed advantages and disadvantages of both modeling approaches and stated that LSTMs are not a must in a

contextual modeling.

In contrast to the presented utterance-level approaches dealing with utilization of contextual information in dyadic interactions, we propose a purely time-continuous contextual emotion recognition pipeline (see Chapter 5). Additionally, we use one of the presented approaches to deep learning based feature extraction (see Section 4.1.2).

3.1.4 IEMOCAP

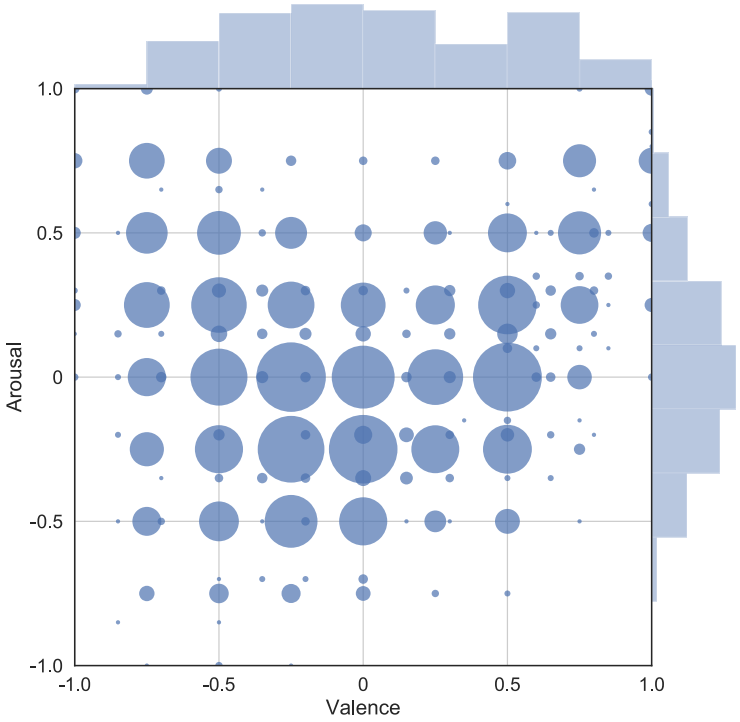


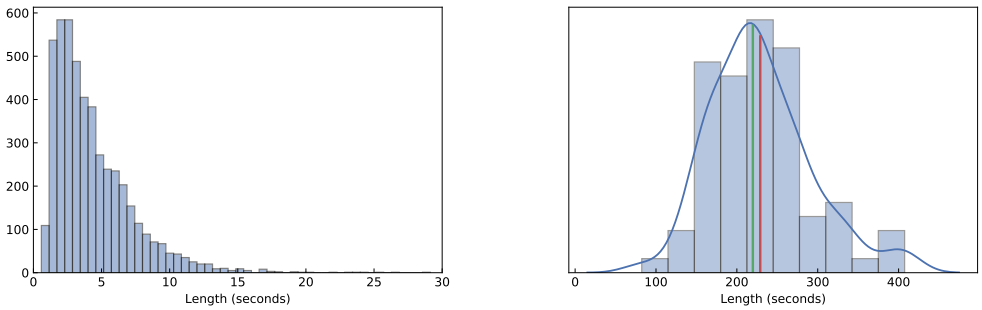
Figure 3.6: IEMOCAP database labels distribution. Scatter plot, histograms for valence (x-axis) and arousal (y-axis) are in blue.

The *Interactive Emotional dyadic MOtion CAPture* database (Busso et al., 2008) consists of recordings from ten English-speaking actors in dyadic sessions collected by the Speech Analysis and Interpretation Laboratory at the University of Southern California.

Corpus characteristics. For the IEMOCAP seven professional actors and three senior students from the Drama Department at the University of Southern California were selected. The actors performed selected emotional scripts and also improvised hypothetical scenarios designed to elicit specific types of emotions (happiness, anger, sadness, frustration and neutral state). For this study, we use only an improvised part of the database, which consists of 80 multi-channel recordings with the total duration of 305.2 minutes (resulting in 610.4 minutes of mono recordings of dialogue).

Each session has two speakers of different genders, which makes the gender ratio of the database equal to 1.0. However, no additional information, such as age, mother tongue, etc. is provided. IEMOCAP has three modalities: audio, video and motion captures, achieved by analyzing information from markers located on face, head and wrist of participants. The data representation includes annotations at turn-level with the average turn duration of 4.3 seconds. The dataset is annotated in terms of categorical emotions (neutral, happiness,

sadness, anger, surprise, fear, disgust, frustration, excited, other), as well as on arousal, valence and dominance scales. ANVIL tool (Kipp, 2001) was used to assign labels of both types to each turn. Self-Assessment Manikins (Fischer et al., 2002; Grimm et al., 2007) were used for dimensional annotations to avoid difficulties caused by individual understanding and personal bias in interpreting textual descriptions of emotions. Label distribution for two basic affective primitives used in our thesis – arousal and valence – is presented in Fig. 3.6. There are four categorical and three dimensional annotations for each turn. To obtain one value for dimensional scales, we calculate the mean of provided scores. Original labels are on a scale from one (low arousal or valence) to five (high arousal or valence). To make them consistent with other databases used in this thesis, we scaled values to a range of $[-1, 1]$. One may notice that in contrast to previously described databases, IEMOCAP has much wider distribution in each dimension, which may be caused by the discrete nature of provided annotations.



(a) IEMOCAP turn length distribution based on transcriptions provided.

(b) IEMOCAP database recordings length distribution. 10 bins histogram and corresponding KDE-plot are in blue, median value depicted as a green vertical line, mean value depicted as a red vertical line.

Figure 3.7: Turn and recordings length statistics for IEMOCAP database.

Distributions of turn and recording duration are presented in Fig. 3.7. One may notice that most of the turns (approximately 70%) are less than five seconds, with the average value of 4.33 seconds. Recordings length has moderately smooth distribution, with the mean value of 229 seconds.

Corpus usage. IEMOCAP was released in 2008 and at that time it was superior to the existing databases in many ways. It had larger duration, multiple modalities, including motion capture in addition to classical audio and video ones, annotations that are obtained with two approaches, acted and spontaneous parts as well as dyadic interaction scenario. Therefore, it opened the door for analysis and experiments in different directions. Previously, in Section 2.2.2 we reviewed several studies that used IEMOCAP in their experiments on mutual inference of speakers' emotions in a dialogue scenario.

Han et al. (2014) proposed a multistage approach to a speech emotion recognition. They used a deep neural network (DNN) for extraction of features from segments of the original audio signal; then the output of the network was used to construct features at utterance level, and finally they were fed into an extreme learning machine (ELM) architecture. Authors used MFCCs and pitch-based features as an input to DNN. In five-class (neutral, happiness, excitement, surprise, frustration) classification task, they achieved approximately 0.48 of

unweighted accuracy with EML and kernel EML, compared to 0.37 obtained with Hidden Markov Model.

Poria et al. (2016) proposed a multimodal approach with the convolutional multiple kernel learning (MKL) (Subrahmanya and Shin, 2009) classifier. This algorithm provides feature selection and effective fusion of data from several modalities. Authors extracted features from three sources: audio, video and text. For audio modality they used Interspeech ComParE feature set (6373 features), for video – features extracted with convolutional recurrent neural network, for text – CNN based on word2vec embeddings (Mikolov et al., 2013) and the part of speech tags obtained with Stanford Tagger (Toutanova et al., 2003). In the separate classification tasks for the four emotional classes (angry, happy, sad, neutral), authors achieved 72-80% of accuracy, outperforming the baseline set by Rozgić et al. (2012). In another study (Poria et al., 2017) authors used LSTM models at two levels: feature extraction and modality fusion. This approach led to the performance gain for *happy* and *sad* classes, however, also to the performance decrease for *neutral* and *angry*.

Mirsamadi et al. (2017) proposed using an attention mechanism to enhance the recurrent neural networks. Authors experimented with four architectures for applying DNNs or RNNs for speech emotion recognition: frame-wise training, final-frame training, mean pooling in time and weighted pooling with logistic regression attention model. For the latter, they also introduced a novel weighted-pooling strategy that focused on emotionally colored parts of utterances. Authors conducted experiments on audio modality with emotion LLDs and raw spectral features. In four-classes (sad, happy, neutral, angry) they achieved 58.8% of unweighted accuracy using the proposed approach.

Zadeh et al. (2018) used multi-attention recurrent network with the long short-term hybrid memory and tested it on the three multimodal tasks: sentiment analysis, speaker traits recognition, emotion recognition. On the latter, using features extracted with COVAREP (Degottex et al., 2014) for audio (MFCC, pitch, etc.), facial action units for video and GloVe embeddings (Pennington et al., 2014) for text, they achieved a correlation score of 0.65 for arousal, 0.10 for valence and 0.65 for dominance.

Being the largest dataset with data and labels available for both speaker and interlocutor, IEMOCAP seems to be extremely useful for the research on the dialogue level context described in Chapter 5.

3.1.5 UUDB

The Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UUDB) (Mori et al., 2008, 2011) is the corpus collected in Utsunomiya University, Japan.

Corpus characteristics. The UUDB consists of 27 dialogues with spontaneous interactions in Japanese. Participants were recorded, while solving a cooperative task: four cards representing a sequence of frames from a comic (manga) were shuffled and each participant had two of them; the aim was to restore the original order of frames without looking at frames of each other and only using dialogue to find it out. There is a total of 14 speakers (2 males, 12 females) with an average age of 20.35 years ($\sigma = 1.22$). Additional information of residential history of participants is available. For each recording there is audio data, a detailed transcript (orthographic and phonetic), as well as an utterance-level annotation in terms of arousal, pleasantness (valence), positivity, credibility, interest and dominance on 7-point scale provided by three annotators. We scaled labels to a range of $[-1, 1]$ in similar way as for IEMOCAP database. Label distribution for the two basic affective primitives used in our thesis – arousal and valence – is presented in Fig. 3.8.

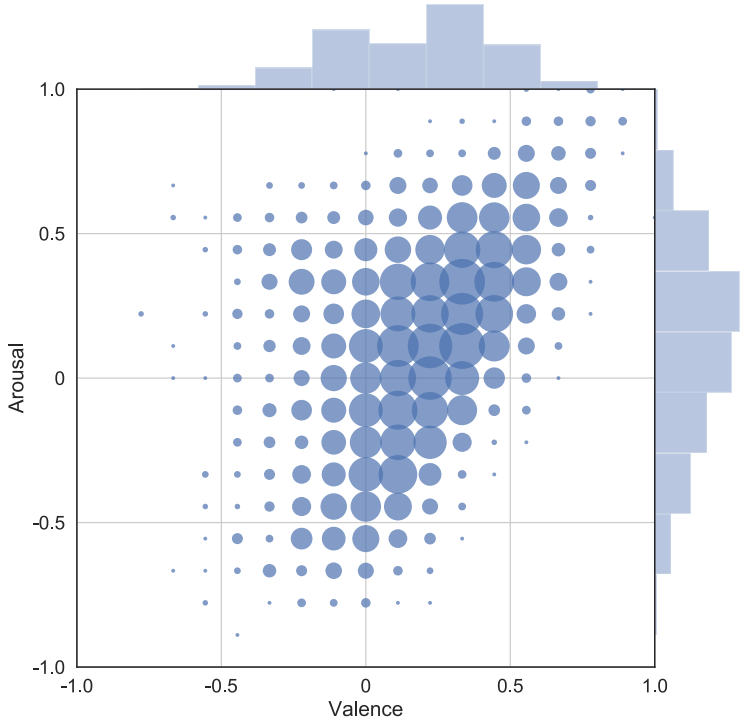


Figure 3.8: UUDB database labels distribution. Scatter plot, histograms for valence (x-axis) and arousal (y-axis) are in blue.

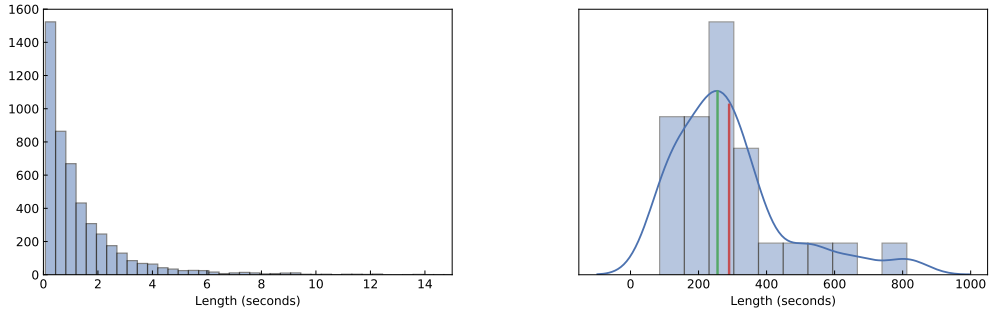
The total duration of the recordings is 260.8 minutes (130.4 minutes if recordings are considered as multi-channel) with the mean duration of utterance equal to 1.4 seconds computed from 4840 examples. Distributions of utterance and recording duration are presented in Fig. 3.9b. One may notice that utterances are shorter compared to IEMOCAP (approximately 80% are shorter than two seconds). Recordings length has less smooth distribution, with the mean value of 290 seconds.

Corpus usage. As there are very limited resources available in Japanese for emotion recognition, UUDB received much attention from Japanese researches.

The author of the database, Mori (2009), studied the effect of switching pauses on emotional labels. He reported an existing correlation between such pauses and the following emotional dimensions: pleasantness, credibility and positivity. He concluded that it can be used as a paralinguistic feature for building classifiers.

Elbarougy and Akagi (2014) suggested a three-level model for speech emotion recognition in contrast to a conventional two-level one (acoustic features – emotional labels). As an intermediate level they used semantic primitives based on adjectives (Huang and Akagi, 2008), such as bright, dark, high, low, calm, monotonous, clear, etc. The presented model is used in two ways: top-down to perform feature selection, and then bottom-up to perform emotion recognition. Authors tested this approach with two databases, in two scenarios (speaker-dependent and speaker-independent) with the Gaussian Mixture Models classifier and reported significant gain in performance compared to conventional two-layer modeling.

Sidorov et al. (2014) studied the influence of speaker information on a performance of emotion recognition models. In their work, they considered two approaches to incorporate this information: by including it as an additional feature and by building speaker-specific emotion recognition models. They evaluated these approaches in two scenarios: using



(a) UUDB utterance length distribution based on transcriptions provided.

(b) UUDB database recordings length distribution. 10 bins histogram and corresponding KDE-plot are in blue, median value depicted as a green vertical line, mean value depicted as a red vertical line.

Figure 3.9: Turn and recordings length statistics for UUDB database.

ground-truth speaker information, as well as building a separate speaker identification model based on neural network. Authors reported the performance gain for each of five used corpora with both scenarios.

Brester et al. (2016) used a two-criterion optimization model for the multi-objective feature selection approach based on evolutionary algorithms. First, authors compared several classifiers (SVM, multilayer perceptron, linear logistic regression) applied to a speech emotion recognition problem. Then, they used multiobjective genetic algorithms to perform the feature selection. As criteria they chose intra- and inter-class distances. After that, they replaced the classifier with an ensemble of models to further improve the performance. Finally, authors reported up to 13% of the relative improvement for some corpora used.

Toyama et al. (2017) studied the effect of global and acoustic features on improving the quality of automatic speech recognition system by integrating them into language model based on recurrent neural network. Their experiments with i-vectors (Dehak et al., 2010) and features extracted with openSMILE showed that this approach provides a significant reduction of perplexity and a slight reduction of word error rate.

UUDB is the only corpus used in this thesis that consists of speech data recorded in a non-European language. With lively conversations between participants, it has shown to be highly valuable for the contextual research on the dialogue level (see Chapter 5).

3.1.6 Summary

We presented the five corpora of emotionally rich interactions used further in this thesis (see Table 3.1 for an overview). Additional corpus for smart environments will be described in corresponding section of Chapter 6. Throughout Chapter 4 and Chapter 5 the presented corpora are used for all the experiments.

Corpus	Language	Speakers	Gender	Age	Modalities	Recordings	Paired
RECOLA	French	23	10 m 13 f	21.35 (2.04)	audio video ECG EDA	23	6
SEMAINE	English	21	8 m 13 f	30.33 (10.38)	audio video	95	95[2]
SEWA	German Hungarian	96	48 m 48 f	-	audio video	96	96
IEMOCAP	English	10	5 m 5 f	-	audio video mocap	160	160
UUDB	Japanese	14	2 m 12 f	20.35 (1.22)	audio video	54	54

Table 3.1: Corpora summary

For all the experiments, we use predefined, carefully sampled partitions to ensure reproducibility and conformity of the data in order to exclude the possibility of results corruption by sampling partitions. We use each available aspect of meta information to perform partitioning (age, gender, nationality, mother tongue) to provide equal division into train and test subsets, yet diverse from one partition to another.

3.2 Data Preprocessing

In this section, we will cover the procedure of obtaining ready-to-use features and labels from the original data. It includes data cleaning, feature extraction and gold standard calculation. All of these three steps are essential for the time-continuous contextual emotion recognition. Thus, data cleaning ensures correct separation of the **contextual** information from different sources, feature extraction turns a raw input signal into meaningful characteristics, and gold standard calculation, that includes reaction lag correction and labels alignment, is critically important for the **time-continuous** emotion recognition model to eliminate undesired bias.

3.2.1 Data Cleaning

In real life and even in an experimental setup in laboratory conditions, additional audio signals may be present in an audio file, containing data of emotional speech. While in some cases this data may be useful for emotion analysis (e.g. laughs, sighs, etc.), it often introduces

[2]Recordings for SEMAINE database are not paired with each other – interlocutors' data is available as additional 95 recordings without annotations.

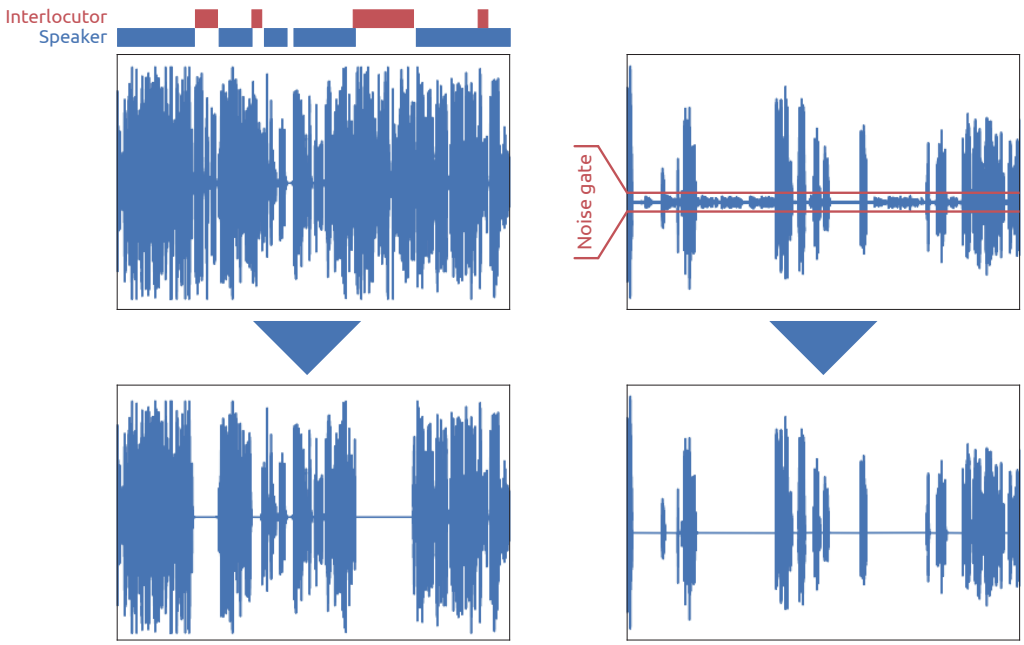


Figure 3.10: Approaches to data cleaning. Turn based (left) and noise gate based (right).

undesired noise, which may corrupt the data drastically, e.g. when speech of the interlocutor is loud enough to be interpreted by an emotion recognition system as data of the speaker.

Nevertheless, the speech data of interlocutor may influence emotions of speaker, and it may be beneficial for the contextual emotion recognition to analyze them in a joint manner. However, data from these two sources should not be mixed together in a single input, as annotation is usually performed separately, considering the data of one participant only. A better approach is to have two separate inputs: for the user and the interlocutor.

In order to clean the data from interlocutor's speech, we perform the following procedures. For some databases, transcriptions or turn-level annotations are provided along with audio files. In this case, we mute the parts of audio file that meet two criteria: (i) are turns of interlocutor; (ii) are **not** turns of speaker, i.e. only interlocutor is speaking. There may be situations, when speaker and interlocutor are speaking simultaneously – here the audio signal will not be muted. If neither transcriptions nor annotations are provided and audio signal of interlocutor has a significant difference in volume compared to the one of speaker, we use noise gate with the threshold chosen empirically. An example for both approaches can be found in Fig. 3.10.

3.2.2 Feature Extraction

Two basic modalities widely used for emotion recognition are audio and video. In this section, we will cover feature extraction methods for these modalities used in this work. More specific features as well as their extraction procedures will be described in corresponding sections later, e.g. head movement or object detection for smart environments.

Feature extraction methods can be divided into two main categories: (i) conventional, when logical rules and/or direct analysis are used to extract numerical representations from data; (ii) deep learning based, when a pre-trained deep learning model (usually artificial

neural network) is used to extract high-level representations from raw data.

Conventional features in turn can be divided into low-level descriptors (LLDs) and high-level descriptors. LLDs are closely related to the signal itself and usually are hard to interpret by a human. High-level descriptors are extracted from LLDs, using functionals (e.g. min, max, mean, standard deviation, etc. over some period of time) and are intended to increase the level of interpretability of data or level of feature complexity, which can be useful when using simpler models. Deep learning based features are also considered to be high-level ones, but they lack any interpretability due to the black-box nature of most models.

Audio. Conventional Features.

As a third-party tool for conventional audio features extraction, we use *OpenSMILE* software (Eyben et al., 2010). It was developed by *audEERING GmbH*[3] and widely used to extract features for various paralinguistic tasks and challenges, including ComParE and AVEC, discussed in Section 2.3. Firstly, *openSMILE* calculates a set of LLDs from the input signal, then functionals can be applied to these LLDs. It is also a common practice to use LLDs directly, especially for such models as recurrent neural networks (Ringeval et al., 2019, 2018). *openSMILE* can be used in both offline and online modes. Offline approach implies feature extraction as a separate and isolated step of preprocessing pipeline, while with online feature extraction, *openSMILE* can be used as a module of a larger real-time recognition system.

A feature set comprised of a group of 4 energy related LLDs, 55 spectral related LLDs, and 6 voicing related LLDs along with their first order derivatives (130 features in total) was a popular starting point for many emotion recognition research since Interspeech ComParE 2013 challenge (Schuller et al., 2013), where it was used. Total feature set (LLDs $\times$ functionals) contains 6 373 features, which provides a heavy computational load on a recognition model. Eyben et al. (2016) developed a Geneva Minimalistic Acoustic Parameter Set (GeMAPS) and its extended version (eGeMAPS), which consists of much smaller amount of features carefully selected based on experts' opinion on their usability and impact for paralinguistic tasks.

eGeMAPS contains LLDs in three groups (taken from original paper of Eyben et al. (2016)):

1. Frequency related parameters:
    a) *Pitch* – logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz (semitone 0);
    b) *Jitter* – deviations in individual consecutive F0 period lengths;
    c) *Formant 1, 2, and 3 frequency* – centre frequency of first, second, and third formant;
    d) *Formant 1* – bandwidth of first formant.
2. Energy/Amplitude related parameters:
    a) *Shimmer* – difference of the peak amplitudes of consecutive F0 periods;
    b) *Loudness* – estimate of perceived signal intensity from an auditory spectrum;
    c) *Harmonics-to-Noise Ratio* – relation of energy in harmonic components to energy in noiselike components.
3. Spectral (balance/shape/dynamics) parameters:
    a) *Alpha Ratio* – ratio of the summed energy from 50–1000 Hz and 1–5 kHz;

[3] `https://www.audeering.com/opensmile/`

b) *Hammarberg Index* – ratio of the strongest energy peak in the 0–2 kHz region to the strongest peak in the 2–5 kHz region;
c) *Spectral Slope 0–500 Hz and 500–1500 Hz* – linear regression slope of the logarithmic power spectrum within the two given bands;
d) *Formant 1, 2, and 3 relative energy* – ratio of the energy of the spectral harmonic peak at the first, second, third formant's centre frequency to the energy of the spectral peak at F0;
e) *Harmonic difference H1–H2* – ratio of energy of the first F0 harmonic (H1) to the energy of the second F0 harmonic (H2);
f) *Harmonic difference H1–A3* – ratio of energy of the first F0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3);
g) *MFCC 1–4* – Mel-Frequency Cepstral Coefficients 1–4;
h) *Spectral flux* – difference of the spectra of two consecutive frames.

The complete eGeMAPS feature set includes additional descriptors and functionals, resulting in 88 features. We use it to correct reaction lag of annotators (see Section 3.2.3). For modeling, in this work we extract two functionals from eGeMAPS LLDs mentioned above: mean and standard deviation over a certain time window (see Fig. 4.9 in Section 4.1.3 for additional details).

Audio. Deep Learning Based Features.

As suggested by several challenge-winning works (Zhao et al., 2018; Chen et al., 2019), deep learning based features may provide meaningful representations for any data type, including audio, video and text. Most of the deep learning models used for the feature extraction are based on convolutional neural networks (CNNs). In our work, we use VGGish[4] (Hershey et al., 2017) embeddings for audio modality in addition to conventional features.

VGGish is based on VGG model (Simonyan and Zisserman, 2014) and consists of 11 hidden layers. It takes an audio file and performs the following steps: (i) resamples it to 16kHz mono audio; (ii) computes a spectrogram using Short-Time Fourier Transform and a periodic Hann window; (iii) maps computed spectrogram onto mel spectrogram to 64 mel bins for a range of $[125Hz, 7500Hz]$; (iv) applies a logarithmic function to mel spectrogram; (v) frames features into non-overlapping examples of 0.96 seconds; (vi) feeds these features into adjusted VGG model; (vii) postprocesses obtained embeddings by whitening and feature transformation using Principal Component Analysis (PCA).

Video. Conventional Features.

Similarly to audio features extraction with *openSMILE*, we extract features from video with the open-source software *OpenFace*[5]. As LLDs in case of video files one may consider facial landmarks – points on the face of a person used to identify significant parts of it, e.g. eyes, nose, mouth, etc. Landmarks are then used to extract Action Units (AUs). AUs represent the movement of facial muscles in accordance with *Facial Action Coding System (FACS)* introduced in Section 2.1. They are widely used in research community to analyze emotions based on facial expressions (Tarnowski et al., 2017; Chu et al., 2017) and in many emotion recognition challenges as standard feature set derived from facial data (Ringeval et al., 2017a, 2018, 2019).

OpenFace extracts AUs in two different form:

[4]https://github.com/tensorflow/models/tree/master/research/audioset/vggish
[5]https://github.com/TadasBaltrusaitis/OpenFace

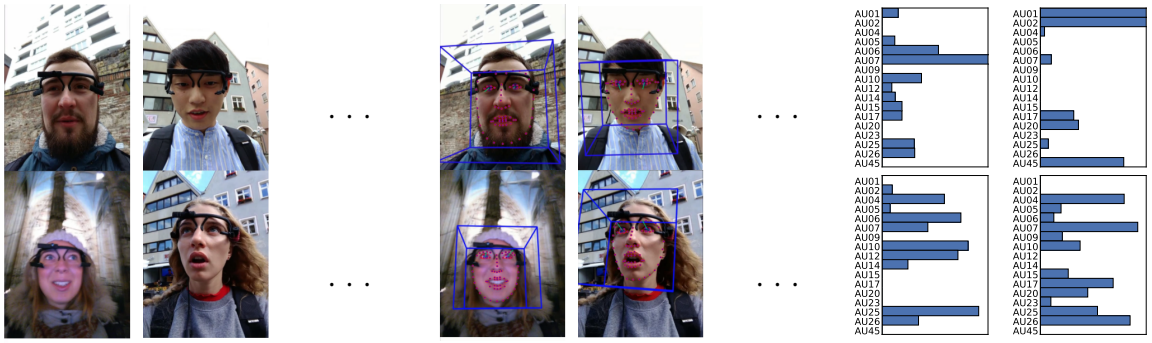


Figure 3.11: Action Units extraction pipeline with OpenFace. Facial landmarks are being extracted from an original image, then they are used for detecting Action Units and real values assignment.

1. binary – with 0 or 1 output representing presence or absence of particular AU respectively;
2. real – with output greater than 0 (but not limited to 1), representing a measure of confidence of the system that a particular AU is present in the current frame.

In this thesis, we use real values provided by *OpenFace* as video features. There are 17 of them:

- AU01 – Inner Brow Raiser;
- AU02 – Outer Brow Raiser;
- AU04 – Brow Lowerer;
- AU05 – Upper Lid Raiser;
- AU06 – Cheek Raiser;
- AU07 – Lid Tightener
- AU09 – Nose Wrinkler;
- AU10 – Upper Lip Raiser;
- AU12 – Lip Corner Puller;
- AU14 – Dimpler;
- AU15 – Lip Corner Depressor;
- AU17 – Chin Raiser;
- AU20 – Lip Stretcher;
- AU23 – Lip Tightener;
- AU25 – Lips Part;
- AU26 – Jaw Drop;
- AU45 – Blink.

Originally, there were more AUs (Ekman and Friesen, 1978) but not all of them are included into *OpenFace*. An example of AUs extraction pipeline is presented in Fig. 3.11

Video. Deep Learning Based Features.

Similarly to VGGish for audio, we use CNN to extract deep learning based features for the video modality. We take ResNet-50 (He et al., 2016) pretrained on VGGFace2 dataset as a basic architecture. VGGFace2 (Cao et al., 2018) consists of more than 3 million images of 9 131 persons (celebrities of different professions) with an average of 362.2 images per person. Hence, this architecture is pretrained on a face recognition task and extracts meaningful features from a person's face.

We then cut the last layer of this network and replace it with two additional dense layers: (i) of 100 neurons with SELU (scaled exponential linear units) activation function and both L1 and L2 weight regularization; (ii) of two neurons with linear activation function, corresponding to arousal and valence dimensions. The resulting network is then fine-tuned on manually annotated images of AffectNet dataset (Mollahosseini et al., 2017). It consists of 450 000 images with respective arousal and valence values. Prior to fine-tuning, for each image we detect face coordinates with *dlib*[6] library and resize it to 224×224 pixels. All layers of the network, except for the last two (with 100 and two neurons) are frozen during the fine-tuning, which means that the training process does not affect their weights.

3.2.3 Gold Standard and Annotations Shifting – Concept and General Approaches

As mentioned in Section 2.1, time-continuous dimensional annotation of affective states has several advantages over utterance-level one, the most important of them are closer emotional perception to the real life scenario and higher flexibility of ratings. Moreover, the process of continuous data annotation reduces required time resources, as it is performed in real-time, and it is unnecessary to map the beginning and end of each utterance. However, this approach introduces challenges, not inherent in utterance-level annotation. One of them is reaction lag (RL) of annotator. RL is the value of a delay between an actual affective behaviour exposure by a speaker and its annotation by a rater. This topic was previously discussed in the research community. In this section, we will cover significant studies on RL and present our manual analysis of RL. In the subsequent section we will propose our method of automatic RL correction, based on the combination of several previously developed approaches, in order to profit the most.

In the comprehensive study, conducted by Mariooryad and Busso (2015), authors used the SEMAINE database (McKeown et al., 2011) to show the effect of RL on the emotion recognition system performance. As a recognizer they used Sequential Minimal Optimization (SMO) implementation of the linear kernel Support Vector Machine (SVM) provided by an open source *Java* based platform, *WEKA* (Hall et al., 2009). Their analysis, based on a Maximum Mutual Information (MMI) criterion, showed that RLs differ across annotators and modalities (audio or face). Audio features tend to introduce longer RLs compared to face features.

In the study, conducted by Mencattini et al. (2017), authors reported longer RLs for valence than for arousal, using audio modality and correlation-based feature selection (CFS) measure. In order to estimate it, they used quadrant-based temporal division (QBTD) – an approach to separate RL calculation for each quadrant of the arousal-valence space in respect to the overall length of such fragments.

RL was also noticed by He et al. (2015) and handled as offsets between features and labels. Authors reported approximately 2-3 seconds of the offset for audio and video (LPQ-TOP) modalities, and shorter for other video representations.

In our work, we assume that RL is composed of two components:

- the affective behaviour notation – the time between exposure of some affective behaviour (e.g. raising voice, smiling, etc.) and its first notation by a rater;

[6] https://github.com/ageitgey/face_recognition

- the annotator's response – the time between notation of affective behaviour and completion of appropriate actions (e.g. moving mouse pointer from one position to another).

The first component relies on abilities of the rater to perceive emotions, and may vary from one affective event to another. The second one relies more on a general (physical) reaction of a person and has much weaker dependence on an annotation object.

Manual Reaction Lag Analysis. In order to prove the existence of RL and study its dependence on annotators or subjects, we have selected six affective expressions from three male and three female speakers of RECOLA database, and manually calculated RL in terms of its two components (see Table 3.2). Affective expressions were selected to be clear and explosive, i.e. change from neutral state is rather quick. An example of manual RL detection is presented in Fig. 3.12

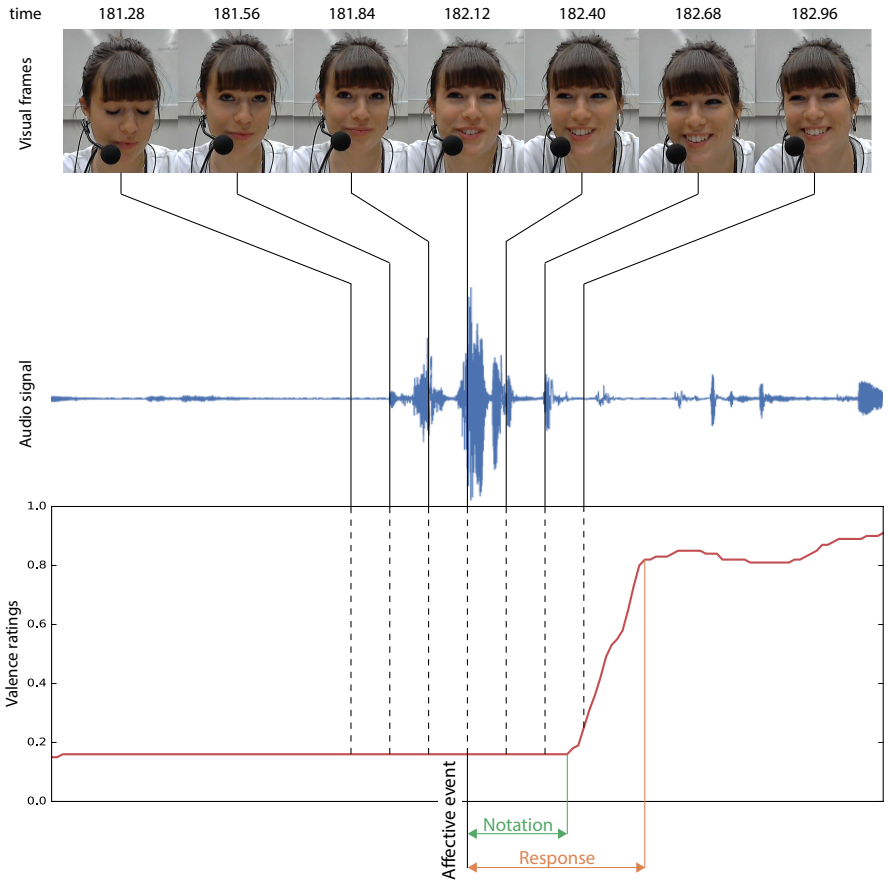


Figure 3.12: Manual reaction lag detection for recording *P25* and valence ratings of annotator *FF1* from RECOLA database. An actual affective event is marker by the middle vertical line. Three frames preceding and three following the affective event (with a step of 0.28 seconds) are presented to show the explosiveness of the emotion. Audio signal and valence ratings are representing 5.88 seconds of data. For this case, notation delay is 0.72 seconds and response delay is 1.28 seconds.

It may be noticed that RL differs across annotators, speakers and dimensions. It is hard to derive any reliable connection between RL and the gender of a speaker and/or an annotator. On average, valence dimension requires longer time to response to an emotional behaviour compared to arousal, which goes in a consensus with (Mencattini et al., 2017). The RL

Rec (F/M) (time)	D	FM1		FM2		FM3		FF1		FF2		FF3		mean	
		N	R	N	R	N	R	N	R	N	R	N	R	N	R
P16 (M)	A	0.66	2.28	0.16	0.92	0.56	1.24	0.48	0.6	0.52	2.12	0.72	0.96	0.52	1.35
(127.68)	V	0.16	1.08	0.16	2.00	0.52	1.52	0.44	1.6	3.76	4.40	0.36	0.96	0.90	1.92
P17 (M)	A	0.44	0.80	0.44	1.36	0.08	1.20	1.28	2.08	1.12	2.00	-	-	0.67	1.49
(108.00)	V	0.08	1.04	0.12	1.04	0.08	0.84	0.16	1.68	1.72	2.08	0.04	1.92	0.37	1.43
P26 (M)	A	0.40	0.56	0.80	1.56	0.64	1.52	0.32	0.88	1.04	1.28	0.00	1.16	0.53	1.16
(62.56)	V	0.12	1.48	0.16	0.64	0.20	0.68	0.60	2.40	0.52	2.52	0.28	0.76	0.31	1.41
P19 (F)	A	2.52	2.92	0.28	0.56	0.04	0.80	1.92	2.68	0.40	0.64	-	-	1.03	1.52
(281.60)	V	0.12	0.72	0.72	1.36	0.56	0.88	0.60	1.76	0.64	0.84	0.56	1.36	0.53	1.15
P25 (F)	A	0.16	1.16	0.00	0.76	-	0.80	0.92	1.72	-	1.32	0.12	0.80	0.30	1.09
(182.12)	V	0.40	1.20	0.20	0.44	0.20	0.68	0.72	1.28	1.92	2.60	0.12	0.68	0.59	1.15
P30 (F)	A	0.36	0.80	0.40	0.72	0.20	0.88	0.40	0.72	0.36	1.68	0.72	1.12	0.41	0.99
(37.88)	V	-	1.24	0.16	0.84	0.04	1.76	0.08	1.08	1.24	2.12	0.24	0.76	0.35	1.30
mean	A	0.76	1.42	0.35	0.98	0.30	1.07	0.89	1.45	0.69	1.51	0.39	1.01	0.58	1.27
	V	0.18	1.13	0.25	1.05	0.27	1.06	0.43	1.63	1.63	2.43	0.27	1.07	0.51	1.40

Table 3.2: Manual analysis of reaction lag for RECOLA database. N – notation delay, R – response delay; A – arousal, V – valence. Gender of a speaker (F – female, M – male) is indicated in the first column; gender of an annotator is coded in its name by the second letter (as in original RECOLA annotation data), e.g. *FM1* – male, *FF1* – female. The last row and column represent mean values for the corresponding annotator and particular affective event, respectively.

calculation may be performed on several levels, e.g. for each emotional event, each pair annotator-speaker, each annotator or across the whole data.

Representations and Metrics. In order to calculate RL automatically, we have to define criteria for estimating its effect on data. During the annotation process, raters usually focus on basic changes in observable human behaviour, such as voice pitch and loudness or facial expressions. We find it to be well-describable by high-level audio and visual feature sets, such as *eGeMAPS* and *AUs*. As a proper RL value should provide better alignment between behaviour and emotions, we measure the similarity between their representations – features and labels – using Pearson's correlation coefficient and Mutual Information (MI) as metrics.

Calculating MI measure between two continuous variables is a resource demanding task; therefore, we preprocess the data (both features and labels) prior to this procedure by dividing it into seven clusters, based on their mean and standard deviation values.

Automatic Annotator-related Reaction Lag Analysis. Using data and metrics described above, we have calculated the similarities between features and labels for each annotator of time-continuously annotated databases (SEWA database has only one annotation series, representing the gold standard; therefore no annotators-dependent analysis is available). Graphical representation of results is provided in Fig. 3.13.

Although in all cases for RECOLA database and many cases for SEMAINE database, a trend of increase in correlation up to some point and decrease afterwards (with the peak approximately at the position of 2 seconds) is noticeable, we cannot state it for each annotator-dimension pair of each database. Hence, completely annotator-dependent RL correction – i.e. separate feature-label based calculation of RL for each annotator – is not possible.

Nevertheless, differences in RLs of different annotators are obvious and cannot be ignored by combining the ratings into a single gold standard using plain averaging. Marioryad and Busso (2015) used pre-aligning of ratings from several annotators within each recording session, but this approach introduces additional complexity without any increase in accuracy. Mencattini et al. (2017) used weighting procedure that favors annotators, that agree more

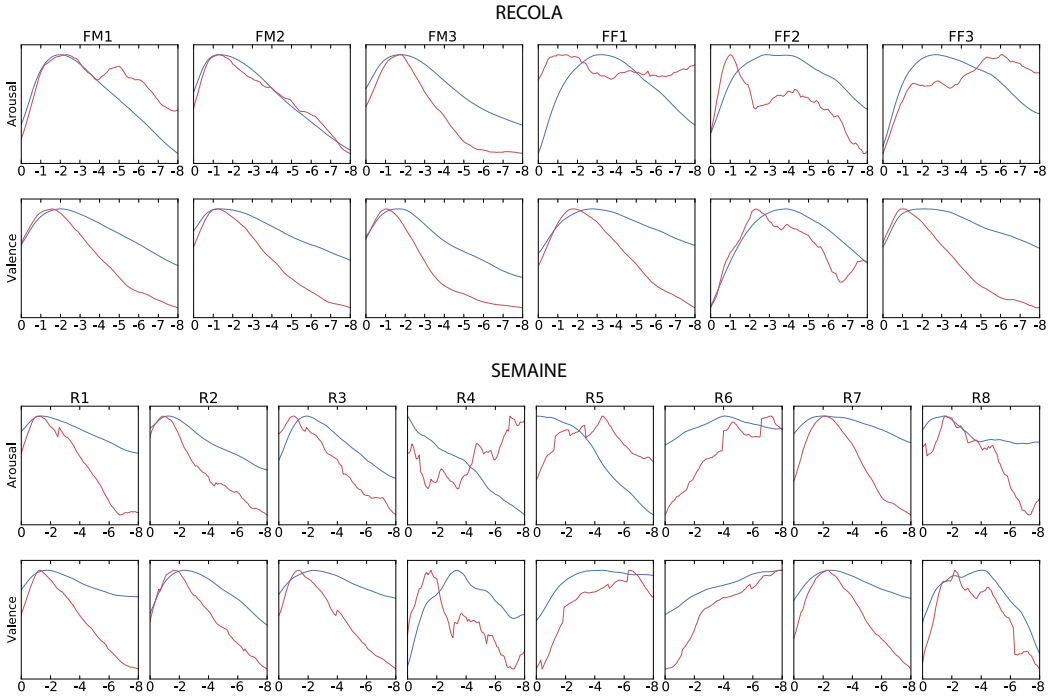


Figure 3.13: Normalized correlation (blue) and mutual information score (red) between audio-visual features and annotations of RECOLA and SEMAINE databases for each rater. X-axis represents shift of labels backwards (in seconds).

with the pool of others. However, the agreement is measured only at a starting point (without alignment) and could be corrupted if annotations are in perfect agreement, but have a delay between each other.

3.2.4 Gold Standard and Annotations Shifting – Combination of Approaches

Rater	R1	R2	R3	R4	R5	R6	R7	R8
Coverage	62.98%	69.73%	77.91%	67.14%	**99.26%**	90.42	46.82%	28.39%

Table 3.3: Annotation coverage in SEMAINE database. Percentage of data annotated by each rater.

Combining the approaches described above, we use annotation pre-alignment over the whole database prior to calculation of the weighted gold standard. For each database, one annotator was selected as a reference, then ratings from other annotators were shifted in the interval [-4, 4] seconds and Pearson's correlation coefficient was calculated. For SEMAINE, the amount of annotations differs from one session to another, therefore we chose one annotator, covering the most sessions as the reference (see Table 3.3) – *R5*. RECOLA database has consistent rating, i.e. each annotator provided data for each whole recording. We chose *FM2* as a reference, because he has the shortest RL, according to our both manual (see Table 3.2) and automatic (see Fig. 3.13) analyses.

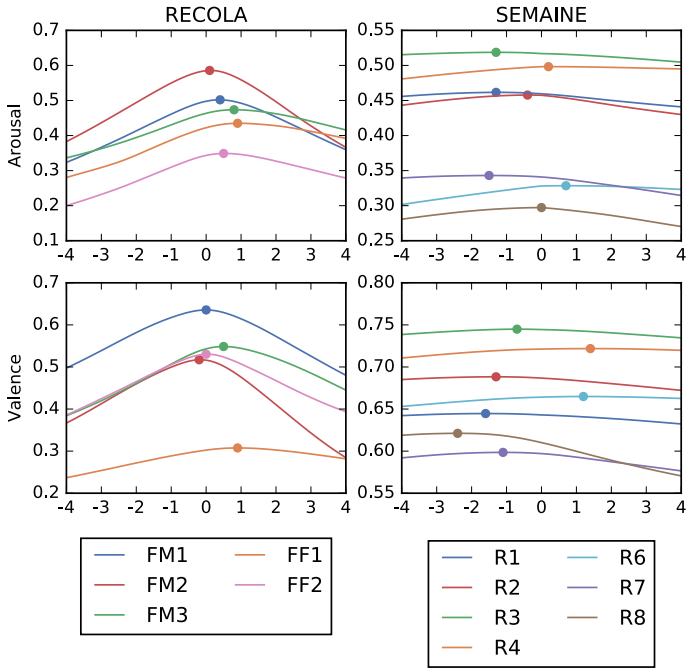


Figure 3.14: Ratings alignment for RECOLA and SEMAINE databases. Round points show the highest correlation between ratings of current and reference annotators

We use *argmax* of correlation between ratings from each annotator and the reference to select the appropriate shifting value for the alignment. The correlation graphs with different shifting values are presented in Fig. 3.14. Now the labels are aligned to the fixed reference withing each database.

As a next step, we perform the evaluator weighted estimation (EWE) procedure (Grimm et al., 2008a; Ringeval et al., 2015; Mencattini et al., 2017) in order to favor ratings of the annotators that agree more with others. To do so, we shift evaluations $e_d^{a_i}(t)$ of each dimension $d \in \{arousal, valence\}$ provided by each annotator a_i, $i = 1, \ldots, N$ to the same value s_d calculated as follows:

$$s_d(t) = \frac{1}{\sum_{i=1}^{N} \rho_d(i)} \sum_{i=1}^{N} \frac{1}{T} \sum_{t} e_d^{a_i}(t) \rho_d(i), \tag{3.1}$$

where $\rho_d(i)$ is a mean pair-wise Pearson's correlation coefficient between evaluations of annotator a_i and the rest $N-1$ annotators:

$$p_d(i) = \frac{1}{N-1} \sum_{j=1 i=j}^{N} max(0, p_d(i,j)), \tag{3.2}$$

Thus, the final label series (gold standard) are calculated according to the following formula:

$$e_d(t) = \frac{1}{N} \sum_{i=1}^{N} (e_d^{a_i}(t) - s_d). \tag{3.3}$$

After the gold standard is obtained using EWE procedure, we perform RL analysis one more time with the single time series for each affective dimension of databases similarly to

Algorithm 1: Gold-standard calculation with pre-alignment based on annotator-dependent reaction lag correction

Result: Corrected gold-standard

$e_d^r(t) \leftarrow$ evaluations of reference annotator a_r for dimension d

for *all* $e_d^i(t)$ *except* $e_d^r(t)$ **do**

  for $t_s = +4\ sec$ ***to*** $-4\ sec$ *with* ***step*** $-0.01\ sec$ **do**

    $c_i(t_s) \leftarrow \rho\left(e_d^i(t - t_s), e_d^r(t)\right)$

  end

  $s_i \leftarrow argmax(c_i(t_s))$

  $\widehat{e_d^i(t)} \leftarrow e_d^i(t - s_i)$

end

$g_d(t) \leftarrow EWE\left(\widehat{e_d(t)}\right)$

for $t_s = 8\ sec$ ***to*** $0\ sec$ *with* ***step*** $0.01\ sec$ **do**

  $c_d^m(t_s) \leftarrow \rho\left(g_d(t - t_s), f_m(t)\right)$

end

$s_d \leftarrow argmax\left(\sum_m c_d^m(t_s)\right)$

$g_d^s(t) \leftarrow g_d(t - s_d)$

the one depicted in Fig. 3.13. We wrap this three-steps procedure up into a single algorithm (Algorithm 1). An illustration of it is presented in Fig. 3.15.

Notation for Algorithm 1 is as follows: $e_d^i(t)$ is evaluation of annotator a_i for dimension

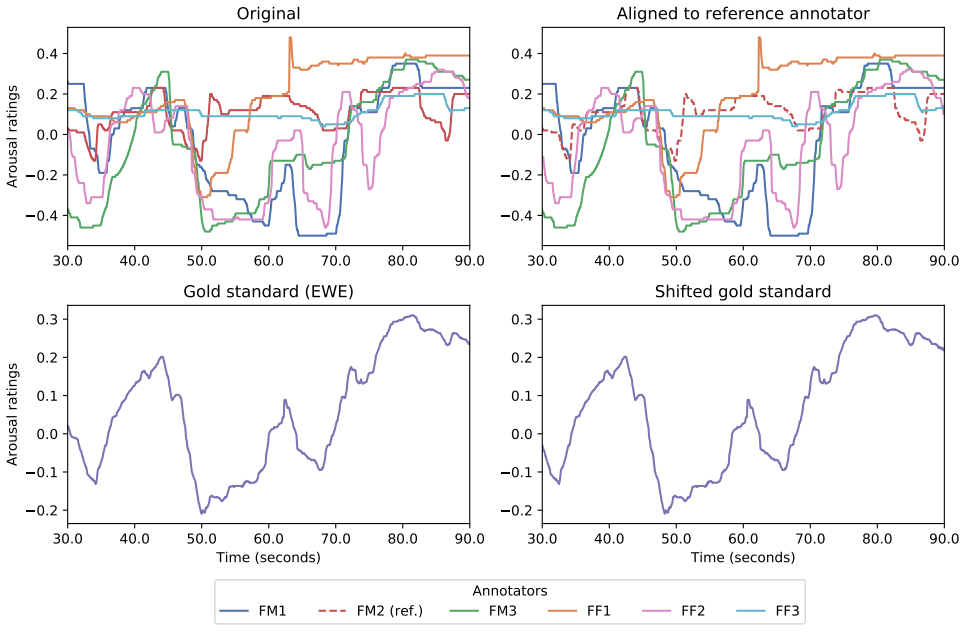


Figure 3.15: Example of gold standard calculation pipeline for RECOLA database. Left-top: original ratings from six annotators; right-top: ratings, aligned to reference annotator (see Fig. 3.14); left-bottom: gold standard obtained according to EWE procedure; right-bottom: gold standard shifted with accordance to correlation analysis between labels and features

d; $c_i(t_s)$ is correlation between evaluations of the annotator a_i shifted to t_s and the reference annotator a_r; $\rho(x, y)$ is Pearson's r between x and y; s_i is an optimal shift for the annotator a_i; $\widehat{e_d^i(t)}$ is evaluation of the annotator a_i shifted to the optimal value; $g_d(t)$ is a gold standard obtained with EWE procedure for dimension d; $c_d^m(t_s)$ is a correlation between gold standard for dimension d and features for modality m; s_d is an optimal shift for gold standard for dimension d; and g_d^s is a corrected (shifted) gold standard for dimension d.

Gold standard calculation and shifting concludes the preprocessing pipeline for the data. Later, at the stage of modeling, features and labels are normalized with Z-transform using information about the train subset (which is defined by sampling partitions), and predicted labels are denormalized in order to fit the original data range.

3.3 Evaluation Metrics

In the area of practical machine learning and data analysis, there is a variety of ways to evaluate model's performance. Since a metric is used to evaluate the ability of the model to solve the particular task, as well as to compare approaches and derive conclusions, one should choose it carefully. If the chosen metric does not appropriately represent proximity of predictions and original labels, it will lead to corrupted decisions.

The degree of suitability of a particular metric to a particular problem is highly dependent on the data and task itself – there are different metrics for classification and regression tasks, for balanced and imbalanced data, etc.

For classification, one estimates the degree of the class match between predictions and original labels. The simplest approach to perform it is to calculate *accuracy*:

$$A = \frac{C}{T}, \tag{3.4}$$

where A is accuracy, C is a number of correctly classified examples and T is a total number of examples. This metric shows straightforward behavior and is very easy to understand. However, one should note that it is less representative for imbalanced classification problems, i.e. such where one class dominates over other(-s) in number of examples. For instance, if we have a problem of medical diagnostic, it is often the case that the overwhelming majority (let us say 98%) of samples are labelled as *negative*, i.e. the patient does not have a disease. If our model will always predict *negative* label, without any data analysis, it will show 98% of accuracy, which may be wrongly considered to be a high performance.

Thereby, two additional aspects should be mentioned: lower performance limit and confusion matrix. Lower performance limit can be defined in two ways:

- random (or chance-level) performance, i.e. a value of metric that can be produced by chance (with randomly generated labels);
- the best performance that can be achieved with fortuitous system initialization but without any training.

If the labels are randomly generated, then chance-level performance will be on average equal to $\frac{1}{C}$, where C is the number of classes. If we choose the second definition, the lower performance level can be set as $\frac{N_{c\_max}}{N}$, where N_{c_max} is an amount of examples with labels of majority class and N is a total number of examples.

The confusion matrix represents the relation of number of examples, assigned to each class by the system, to the real distribution of data among these classes. For a binary classification

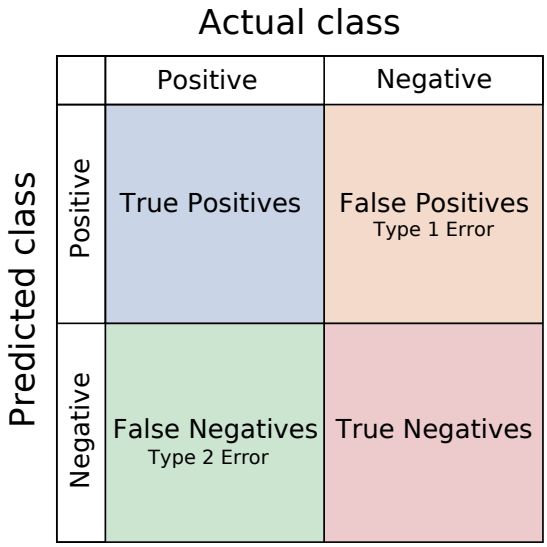


Figure 3.16: Confusion matrix for binary classification problem.

problem (when there are two classes) the confusion matrix is presented in Fig. 3.16. Using it, we can rewrite Equation 3.4 as:

$$A = \frac{TP + TN}{TP + FP + TN + FN}, \tag{3.5}$$

where TP are true positives, TN – true negatives, FP – false positives, FN – false negatives.

Usually, one should not go below the lower performance limit. However, if the chosen metric does not suit the task well, it is not always so. Getting back to the example with medical diagnostic: if our system will train and always predict the *positive* labels (patient has a disease) correctly, but will lose 3% of correctly predicted *negative* labels due to adjustments to *positive* ones, our overall performance will drop and be close to 97%, which is lower than the performance limit. Nevertheless, now our system performs much better in regard to assigned task, and we can see it with the confusion matrix.

Confusion matrix provides us with a greater flexibility for precise results analysis. For example, we can decide, whether each error type is equally bad for a particular application of our system or not. In many cases it is not clear, depends on the decision making person or cannot be determined. In our medical example, type 1 error stands for falsely diagnosing a person to have a disease, while he does not have it; and type 2 error stands for falsely diagnosing a person to not have a disease, while he has it. In this case, type 2 error seems to bear worse consequences compared to type 1 error, as the patient will be probably analyzed additionally to confirm the positive diagnosis, while it is not the usual practice for originally negative results, which can lead to a delayed treatment and a disease progression. However, one can not use these advancements while working with *accuracy*. Therefore, additional metrics are suggested for the classification task: *precision*, *recall* and *F1 score*.

Precision (P) represents the ratio between the number of examples, correctly labelled as class C, and the total number of all the examples assigned to class C by the system:

$$P = \frac{TP}{TP + FP}. \tag{3.6}$$

In its turn, *recall (R)* represents the ratio between the number of examples, correctly

labelled as class C, and the total number of examples of class C in the data:

$$R = \frac{TP}{TP + FN}. \tag{3.7}$$

F1 score is a harmonic mean of both precision and recall, and it is defined as follows:

$$F1 = 2 \times \frac{P \times R}{P + R}. \tag{3.8}$$

There are several approaches to calculate *F1 score*:

- micro – calculates true positives, false negatives and false positives globally by counting their total numbers;
- macro – calculates true positives, false negatives and false positives separately for each class, then finds their unweighted mean. This approach does not take imbalances in labels into an account;
- weighted – calculates true positives, false negatives and false positives separately for each class, then finds their mean, weighted by the number of real examples of each label. This takes imbalance of labels into an account and can result in metric value, that lays not between P and R.

While *F1 score* is an advanced and widely used metric, it is not always trivial to interpret it. Let us consider three examples with imbalanced data (see Table 3.4):

1. We have nine objects in our dataset, six of them are of the first class and three are of the second class. Our system has not trained at all and assigns the first class to any object;
2. Now eight of them are of the first class and only one of the second class. Our system performs the same;
3. We have nine objects in our dataset, six of them are of the first class, two are of the second class, and one is of the third class. Our system has trained to distinguish the first and the third classes, but still assigns objects of the second class to the first one;

Example 1	True labels	1	1	1	1	1	1	2	2	2
	Predictions	1	1	1	1	1	1	1	1	1
Example 2	True labels	1	1	1	1	1	1	1	1	2
	Predictions	1	1	1	1	1	1	1	1	1
Example 3	True labels	1	1	1	1	1	1	2	2	3
	Predictions	1	1	1	1	1	1	1	1	3

Table 3.4: Three examples of imbalanced data.

Three approaches to calculate *F1 score* described above will provide the following results (see Table 3.5).

F1 score (micro) provides the same value as *accuracy* and does not represent performance of the system properly. *F1 (macro)* is hard to interpret and *F1 (weighted)* is biased to majority class. For better interpretability and understanding of the model performance, we decided to use *unweighted average recall* (*UAR*, also *macro recall*) instead of *F1 score*. An important advantage of *UAR* over *F1 score* is an easily set lower performance limit – it is always equal to $\frac{1}{C}$, where C is the number of classes. For the three examples of imbalanced data described above, *UAR* is presented in Table 3.6.

	F1 score (micro)	F1 score (macro)	F1 score (weighted)
Example 1	0.667	0.400	0.533
Example 2	0.889	0.471	0.837
Example 3	0.778	0.619	0.683

Table 3.5: F1 scores (micro, macro and weighted) for imbalanced data examples from Table 3.4.

	UAR
Example 1	0.500
Example 2	0.500
Example 3	0.666

Table 3.6: UAR for imbalanced data examples from Table 3.4.

In contrast to any definition of *F1 score*, *UAR* did not increase while changing from Example 1 to Example 2, as the system was still not trained and did not show any positive adjustments. Considering our medical example mentioned previously, *UAR* will be 0.500 for not trained system – at the lower performance limit; and 0.985 for the system that classifies each example of *positive* class correctly, but misclassifies 3% of *negatives* – a significant difference.

All the metrics described above provide values in range $[0, 1]$ and are positively-oriented, i.e. the greater the metric value is, the better solution was provided by the system.

Even though there are some databases annotated time-continuously in terms of basic emotions (anger, sadness, neutral, happiness, etc.) (Perepelkina et al., 2018), most of the problem statements for time-continuous emotion recognition imply solving a regression task in regard to affective primitives (arousal and valence). Metrics mentioned beforehand provide a solid basis for the classification task evaluation, but do not suit the regression task well. We will now cover metrics widely used for regression tasks, with a special focus on time-continuous emotion recognition.

One of the most simple, in terms of understanding, metrics is the *mean absolute error (MAE)*. As the name implies, it provides an estimation of how far away the predictions are from original labels on average. It can be calculated using the following formula:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} \mid y_i - \hat{y}_i \mid, \tag{3.9}$$

where N is a total number of samples, y_i is a ground-truth for i^{th} element of the sample, $\hat{y}_i$ – a prediction of the system for this element.

In practice another metric is widely used – *root mean square error (RMSE)* in order to penalize greater differences between ground truth and predictions harder. *RMSE* is similar to *MAE*, but the errors are firstly squared, then a square root of their average is taken:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}. \tag{3.10}$$

Both *MAE* and *RMSE* do not consider the direction of errors due to absolute or squared values respectively, provide the estimation in the same units as true and predicted values,

moreover, they are negatively-oriented (the less the value is, the better) in range of $[0, \infty)$. These metrics are often used for classical regression tasks, e.g. prediction of house prices, where samples are not connected with each other on a time scale.

On the other hand, if samples are connected, such metrics as Pearson's correlation coefficient (Pearson's **r** or PCC), concordance correlation coefficient (CCC) or mutual information (MI, also information gain) are widely used.

PCC applied to a sample is defined with the following formula:

$$PCC = \frac{\sum_{i=1}^{N}(y_i - \mu_y)(\hat{y}_i - \mu_{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \mu_y)^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \mu_{\hat{y}})^2}}. \tag{3.11}$$

where μ_y and $\mu_{\hat{y}}$ are sample means of ground truth and predictions respectively and defined as:

$$\mu_y = \frac{1}{N}\sum_{i=1}^{N}(y_i). \tag{3.12}$$

PCC ranges in $[-1, 1]$ with 1 and -1 meaning relationship perfectly described with linear equation and 0 meaning no existing linear correlation between samples.

PCC plays an important role in CCC, which measures an agreement between samples. The general formula of CCC is as follows:

$$CCC = \frac{2\rho\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2}. \tag{3.13}$$

where ρ is PCC, σ_y and $\sigma_{\hat{y}}$ are sample variances of ground truth and predictions respectively and defined as:

$$\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_y)^2. \tag{3.14}$$

CCC can also be defined through covariance by changing the numerator of Equation 3.13 accordingly:

$$CCC = \frac{2cov(y, \hat{y})}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2}. \tag{3.15}$$

where covariance $cov(y, \hat{y})$ is defined with:

$$cov(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_y)(\hat{y}_i - \mu_{\hat{y}}). \tag{3.16}$$

MI measures mutual dependencies between two variables and, in contrast to PCC, is not limited to only linear dependencies. It can be measured using the following equation:

$$MI(y, \hat{y}) = \sum_{i=1}^{|y|}\sum_{i=1}^{|\hat{y}|} P(i, j) log\left(\frac{P(i, j)}{P(i)\hat{P}(j)}\right). \tag{3.17}$$

where $P(i)$ and $\hat{P}(j)$ define probabilities that a randomly selected object is of class y^i or $\hat{y}^j$:

$$P(i) = \frac{|y^i|}{N} \tag{3.18}$$

and $P(i, j)$ is a probability that a randomly selected object is of both classes y^i and $\hat{y}^j$:

$$P(i, j) = \frac{|y^i \cap \hat{y}^j|}{N} \tag{3.19}$$

therefore, Equation 3.17 can be rewritten as follows:

$$MI(y, \hat{y}) = \sum_{i=1}^{|y|} \sum_{i=1}^{|\hat{y}|} \frac{|y^i \cap \hat{y}^j|}{N} log\left(\frac{N|y^i \cap \hat{y}^j|}{|y^i||\hat{y}^j|}\right). \tag{3.20}$$

MI introduces the more comprehensive comparison between two variables as it covers also non-linear dependencies. However, while working with the discrete representation of continuous variables (regression task), it requires enormous computational and temporal resources to compute the metric value, compared to e.g. PCC.

Therefore, one often uses discretization into limited (low) number of classes prior to MI calculation. In our work we use seven classes, five of which are defined in range, $[\mu - 2.5\sigma, \mu + 2.5\sigma]$ each representing a range of one σ. Two additional classes are defined for values above and below $\mu + 2.5\sigma$ and $\mu - 2.5\sigma$ respectively. This approach allows to drastically speed up the calculation of MI, without significant loss of its precision.

3.4 Summary

In this chapter we covered the data, the basic preprocessing steps and metrics used further in this work.

In Section 3.1 we provided a short description, basic statistics and a brief literature review related to data used in this thesis. At the moment, there are only a few corpora suitable for performing contextual analysis available to the research community. Many corpora used previously for emotion recognition tasks consist of acted data and often annotated categorically at utterance-level. In our work we aim to study contextual dependencies in real-world, spontaneous, time-continuous emotional behaviour, therefore, we limited the scope of appropriate data sources.

Further, in Section 3.2 we covered data preprocessing pipeline from raw data to ready-to-use features or representations. First, we described the data cleaning procedure used in our work to remove unnecessary noises and speech of interlocutor from audio recordings. Then, we reviewed feature extraction methods for audio and visual data, including classical expert knowledge based approaches, as well as recently developed deep learning based ones. The latter provides embeddings, obtained with black-box machine learning models, proven to extract useful data representations for further analysis. Afterwards, we tackled the problem arising for time-continuous emotion recognition – labels shifting and reaction lags. We reviewed existing methods to cope with this issue and suggested a combined methodology for the feature-labels alignment and the gold-standard calculation.

Finally, in Section 3.3 we covered evaluation approaches to measure the quality of recognition, model training and system performance.

4 Modeling Speaker Context in Time-continuous Emotion Recognition

The first level at which the contextual information may be considered is the user data itself. This information is related to his/her actions and behaviour (e.g. speech, facial expressions, gestures, pose, etc.) as well emotional status at time points *t-n..t-1* preceding the current time point *t*. In other words, the amount of speaker context defines, how much previous data of the user should be analyzed by the system to make a reliable prediction of his/her current emotional status. Already in the early age, humans naturally learn to determine the relevant amount of such data and use it for their decision making on person's emotions (Denham, 1998). However, for artificial systems still remains an open question. In this chapter we will consider this issue in detail, propose approaches to speaker context modeling and test those in various scenarios.

There are several approaches to model the different amount of context using time-continuous data, and they correspond to the stages of the emotion recognition pipeline. Contextual modeling can be performed at a feature extraction stage, a data preprocessing stage or at a modeling stage. In this work, we consider approaches covering context modeling at all these stages as well as their combinations (see Fig. 4.1). In the following sections, we will describe them in detail, considering their advantages and disadvantages. We utilize two types of models: time-dependent, i.e. such models that consider time-continuous nature of data (e.g. recurrent neural networks), and time-independent, i.e. such models that do not consider it (e.g. linear regression).

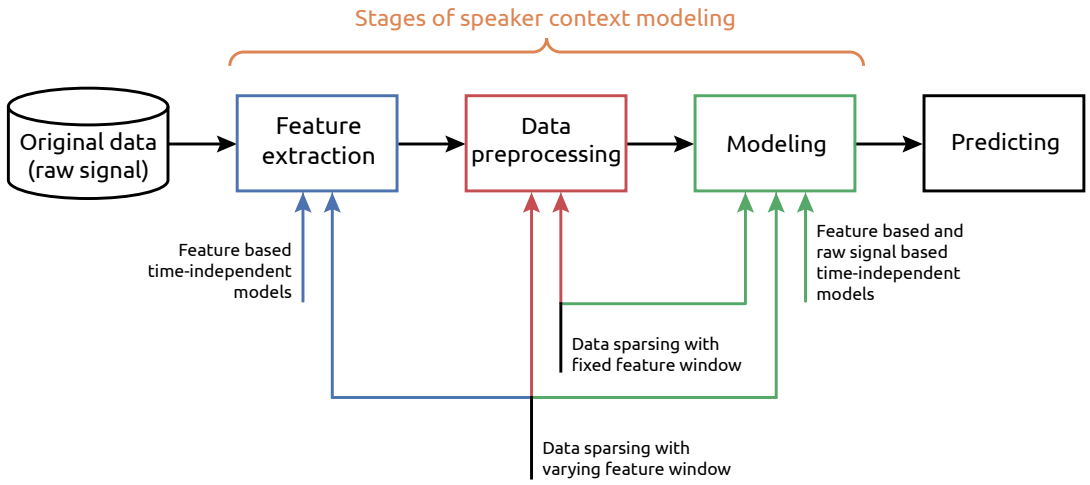


Figure 4.1: Speaker context modeling in general pipeline for emotion recognition system. Approaches to context modeling corresponding to each stage of the pipeline are indicated at bottom part of the figure.

First, in Section 4.1 we consider several approaches to the straightforward context modeling, such as performed with feature based and raw data based time-dependent models (modeling stage) or feature based time-independent models (feature extraction stage). Then, in Section 4.2 we introduce a simple yet effective approach to context modeling at the data preprocessing stage, called *data sparsing*, which is combined with the straightforward approach from the previous section. Further, in Section 4.2.3 we extend it and make it more flexible, combining with the context modeling at the feature extraction step, thus using all the three levels of context modeling simultaneously. In Section 4.3 we study the introduced approach in cross-corpus setting. Finally, in Section 4.4 we analyze the results obtained with each of the presented approach.

All results presented in this chapter are averaged and obtained with ten fixed speaker-independent sampling partitions, carefully selected in order to preserve the original distribution of the available speaker information (e.g. age, gender, mother tongue, etc.) in corpus, as well as to keep the train-to-test ratio close to 0.7/0.3 without splitting any recording. The performance measure is the concordance correlation coefficient (CCC) described in Section 3.3. Throughout this chapter we will use various data frequencies. In order to make comparison of performance reasonable, we interpolate prediction vectors to 10Hz prior to metric calculation. CCC is calculated on each recording separately and the final value of metric obtained by summation of CCC measures for each recording of the test subset, weighted by their length.

$$CCC_f = \sum_{r=1}^{N} (w_r \times CCC(true_r, pred_r)). \tag{4.1}$$

where N is a number of recordings in database, $true_r$ are labels for recording r, $pred_r$ are predictions for recording r, w_r is a weight corresponding to the recording r and calculated as follows:

$$w_r = \frac{l_r}{\sum_{i=1}^{N} l_i} \tag{4.2}$$

where l_r is a length of the recording r. The data frequency of 10Hz for time-dependent models was selected as the lowest one available for original annotations of corpora used in this chapter (SEWA).

For RNNs we use RMSprop optimizer with adaptive learning rate that is optimized in range of $[0.00015625, 0.08]$, where each consecutive rate is twice as high as the previous (LeCun et al., 1998). We use two LSTM layers with 20 and 10 cells respectively, each followed by a dropout layer with $p = 0.3$. Our previous experiments showed that the similar architecture with 80 and 60 cells as in (Ringeval et al., 2015) or 64 and 32 cells as in (Ringeval et al., 2019, 2018) achieves slightly higher results, but shows completely similar behaviour in terms of dependencies between performance and context length. Therefore, we decided to use the less complex architecture in order to reduce computational time and resources for the extensive experiments presented in this chapter. The activation function for LSTM cells is the hyperbolic tangent, the last layer of the network is the fully-connected time-distributed dense layer with the linear activation function. In order to optimize weights of the network, we use CCC-based loss function (Weninger et al., 2016), which is equal to $1 - CCC$.

4.1 Straightforward Approach

The first approach to model context is simple and straightforward – one should use more time-continuous data to form each sample, e.g. more time steps for RNNs, wider windows for the feature extraction or larger filters size for convolutional layers of CNNs at the feature mapping stage.

While this approach suggests a high level of simplicity in the modeling and understanding, it has several substantial disadvantages corresponding to each type of data processing or modeling:

- if we use RNNs for modeling emotions time-continuously and have a hop size of one frame between data samples, the memory load of data increases considerably with the increasing time window, which requires additional computational time and resources;
- if we use RNNs for modeling emotions time-continuously and change a hop size in accordance with size of the time window (e.g. hop is equal to $\frac{2}{3}$ of the time window), the number of data samples decreases significantly toward the situation, when each sample represents a complete recording. Moreover, starting from the particular window size (the value is dependent on the data and model), model may start to lose information about earlier frames of data;
- if we use deep learning approaches as a feature extraction subsystem (e.g. CNN dealing with raw audio or images) at the input stage of the time-continuous emotion recognition system pipeline (e.g. RNN) and we model the context through changing the size of the filter for the feature extraction, it introduces additional computational load for training this system;
- if we use high-level features and model the context through changing the width of window for the feature extraction, it will cause loss of data dynamics, as data within this window is averaged.

In the following, we will cover the effect of straightforward context modeling in the three different types of models: feature based time-dependent, raw signal based time-dependent and feature based time-independent.

4.1.1 Feature Based Time-Dependent Models

In this subsection, we will discuss an effect of the context amount on the model performance with the feature based time-dependent models. As *time-dependent* we define any model that considers the input data as a structured array, where the state (i.e. features and labels) at the time step t is connected to the state at the time step $t-1$ and in turn to the one at $t+1$. In this type of model, the data is represented as a three-dimensional array $[samples \times features \times timesteps]$ and it cannot be shuffled across $timesteps$ dimension nor comprised into this dimension with inconsistent order. One should also note that the data from different recordings may not appear in one sample and should be filtered, e.g. replaced with zeros.

As the *feature based* model, we consider any model that uses high-level features as an input. The model itself nor its parts should not serve as the feature extractor from the raw signal. Some examples of such features are eGeMAPS for audio modality or AUs for video modality (see Section 3.2.2). This type of the context modeling enables the context regulation at the *modeling* stage of the general emotion recognition pipeline (see Fig. 4.2 for more details).

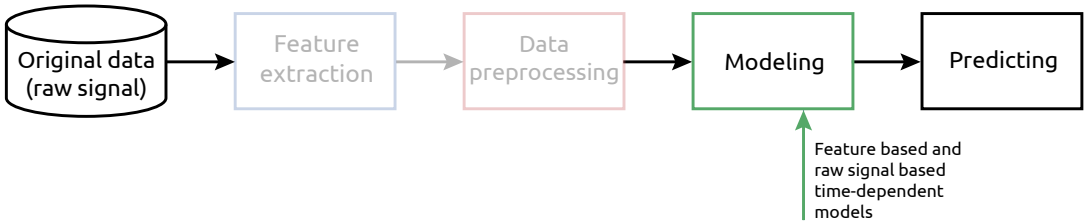


Figure 4.2: Straightforward approach to the speaker context modeling with feature based time-dependent models in the general pipeline of the emotion recognition system

Let us consider the time-continuous emotion recognition model (RNN-LSTM) trained on audio data (eGeMAPS feature set) of RECOLA database at its original sampling rate – 25 Hz. It is easy to see that in order to model *one second* of context at this rate, we need *25 time steps* for the model. We will gradually increase the amount of context using from 2 to 8192 time steps in our model, which corresponds to context length from 0.1 to 328 seconds (see Table 4.1 in the Section 4.2 for additional details and explanation). The maximal number of time steps is connected to peculiarities of data: each recording of RECOLA database has the length of 300 seconds, therefore it is not rational to consider any window lengths beyond this value. We choose 328 and not exactly 300 seconds to have a smooth increment in number of time steps: $window\_size = \{2, 3, 2 \cdot 2^1, 3 \cdot 2^1, 2 \cdot 2^2, 3 \cdot 2^2, ..., 2 \cdot 2^n, 3 \cdot 2^n\}$. Values corresponding to the data after the 300th second are replaced with zeros and then masked during the training procedure to eliminate the negative effects of empty time steps.

Another question to ask is which length should we choose if our data is potentially limitless. For example, we have a database of 100 recordings, each containing 24 hours of continuously recorded and annotated data. Which length in terms of time steps should we choose to train a model? In this chapter, we address this question and study it from different perspectives.

During model training, we shift the sampling window to $\frac{2}{3}$ of window size. Hence, the window size is connected to the number of samples for training, while the total amount of data (memory) used by the model is kept at approximately the same level: we can either have a lot of samples with the window size of two, many of them with the window size of 64, few of them with the with window size of 1024 or merely the same number as we have recordings with the window size of 8192 (for RECOLA database). The latter approach was used as the baseline model for AVEC2019 (Ringeval et al., 2019). The authors used a full recording as one sample and trained the RNN-LSTM model based on this data. This approach showed rather high performance on both modalities and emotional dimensions, but was outperformed by one of the team participating in this challenge that split original samples into shorter ones. Moreover, this approach has a significant drawback: assume that we trained our model with recordings of 300 seconds each, as we have for RECOLA. Now in order to get the prediction for new, unseen data we should provide an input of 300 seconds. This introduces additional limitations:

- if our new recording is shorter than 300 seconds, we can handle the inconsistency in the time sequence lengths by zero-padding the input during the data preprocessing stage or masking it at the modeling stage;
- if our new recording is longer than 300 seconds, e.g. 400 seconds, we can model emotions only for the first 300 seconds, then we need to shift the window to 100 seconds and model emotions for period of 100-400 seconds, which is similar to the idea of splitting recordings into shorter steps.

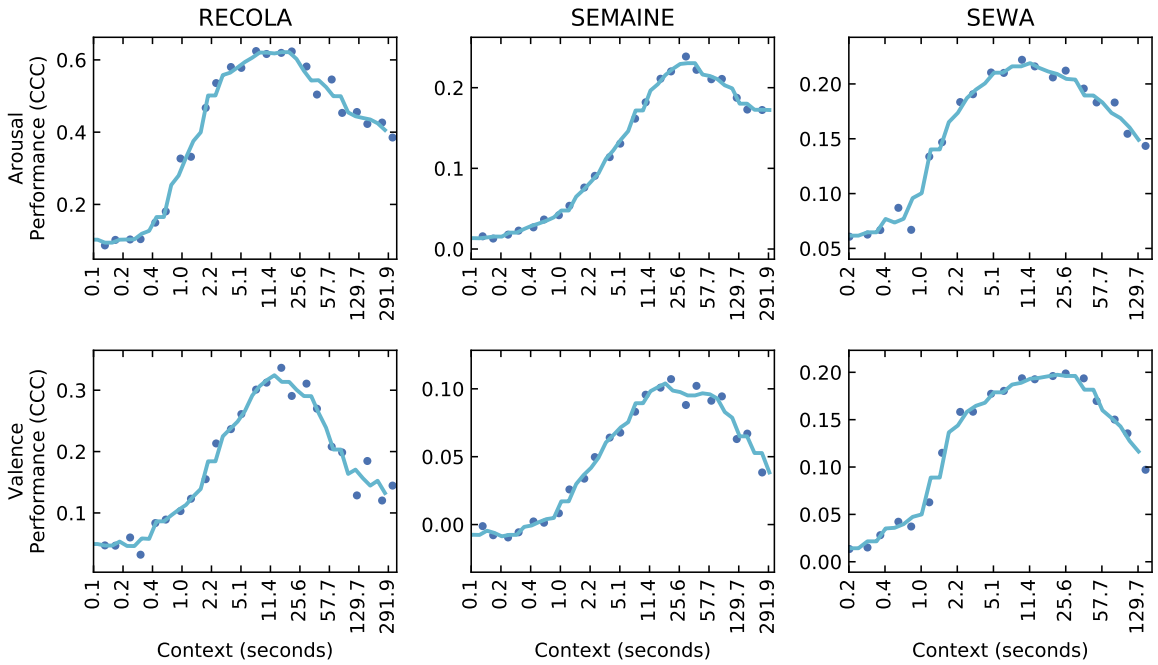


Figure 4.3: Results for straightforward context modeling on audio modality (eGeMAPS) with feature based time-dependent models. Hop size for samples is equal to $\frac{2}{3}$ of modelled context (values of x-axis), which is in turn equal to $\frac{timesteps}{data\,frequency}$. Arousal (top row) and valence (bottom row) dimensions for three time-continuously annotated corpora: RECOLA, SEMAINE and SEWA. The performance is measured in terms of CCC and is dependent on amount of context represented by number of time steps used in RNN and data frequency.

We model the context by simply increasing the length of time window as described previously, taking into account their original sampling rate and the median length of recordings as the maximal number of time steps (see corpora description in Section 3.1). As audio features we use eGeMAPS, as video features – AUs. The results for audio modality are presented in Fig. 4.3 and for video modality in Fig. 4.4.

The resulting graphs show strong dependencies between the modelled context and the model performance across each database, modality and dimension. There are two general patterns. With gradually increasing context coverage, the performance can reach some area of "high" values and stay there with minor variability. We will refer to this type of situation as to a "performance plateau" in order to distinguish it from the second pattern – a "performance peak" – where the performance increases at first, reaches its peak value or region and then has a clear and distinguishable decrease.

One can notice the performance plateaus for the video-valence pair of RECOLA and SEMAINE, as well as for each dimension on the video modality for SEWA. Peformance peaks are shown for the audio modality of each corpus, as well as for video-arousal pair for RECOLA and SEMAINE.

Results on RECOLA database demonstrate the steepest increasing trend and often reach the peak value or region earlier compared to the other databases, e.g. the performance of the audio-arousal pair reaches its first local maximum at 3.8 seconds and global maximum near ten seconds. The same values of local maxima at 3.8 seconds are to see for both dimensions on the video modality, and the wider context window length (around eight seconds) for the

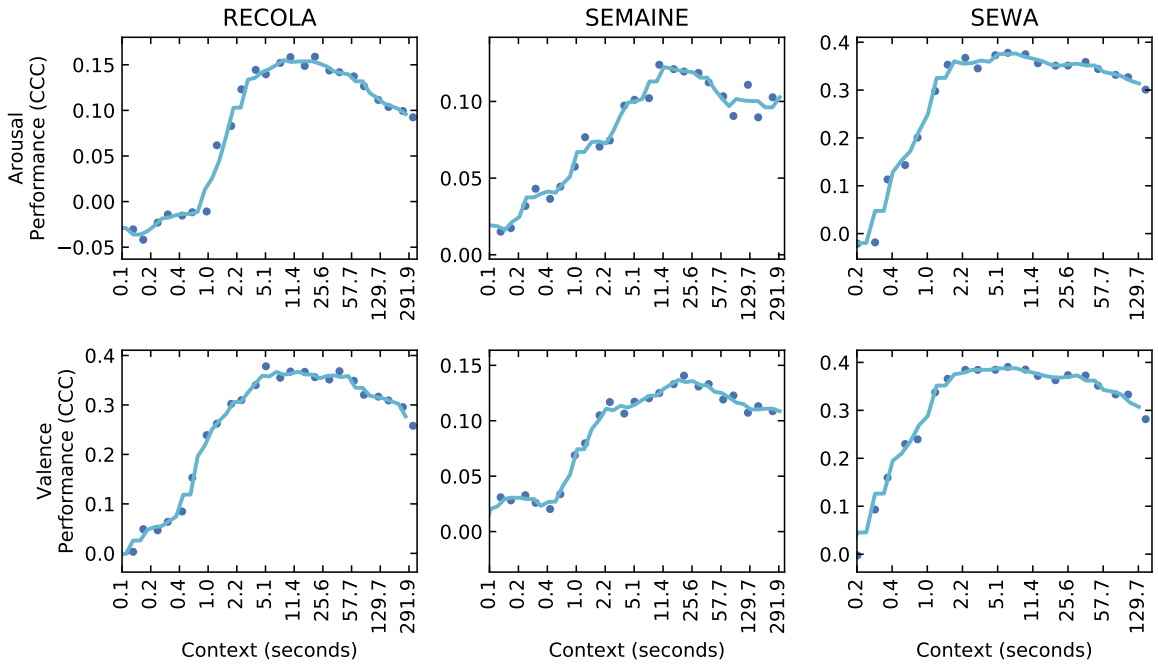


Figure 4.4: Results for straightforward context modeling on video modality (AUs) with feature based time-dependent models. Hop size for samples is equal to $\frac{2}{3}$ of modelled context (values of x-axis), which is in turn equal to $\frac{timesteps}{data\,frequency}$. Arousal (top row) and valence (bottom row) dimensions for three time-continuously annotated corpora: RECOLA, SEMAINE and SEWA. The performance is measured in terms of CCC and is dependent on amount of context represented by number of time steps used in RNN and data frequency.

audio-valence pair.

Then, results on SEMAINE database demonstrate the longest context window lengths required to reach the area of top performance. For audio-arousal, audio-valence and video-valence it is after approximately 30 seconds; for video-arousal – above ten seconds. The first second of increasing context provides little to no improvement and after reaching the top performance region, performance decreases at approximately 30 seconds for audio-arousal, 60 seconds for audio-valence and 40 seconds for video-arousal.

Finally, results on SEWA database demonstrate the most smooth pattern, considering both the increase tempo and the fluctuations. Performance on the video modality reaches its top region already at approximately two seconds, thus, competing with RECOLA to have the shortest context window. Performance on the audio modality requires a slightly wider context window – approximately eight seconds. However, similar to RECOLA, SEWA has a decrease in performance with wider context windows, which is especially significant for the audio modality.

Summarizing the results shown in this subsection, we can conclude that the largest differences in the performance dependence on context may be noticed between the audio and video modalities and also between databases. The video modality usually requires shorter context windows. In general, patters are similar for all modality-dimension pairs within each database, which indicates strong connection between the performance patterns and the data itself.

In order to check the relation between the amount of context and the performance of the

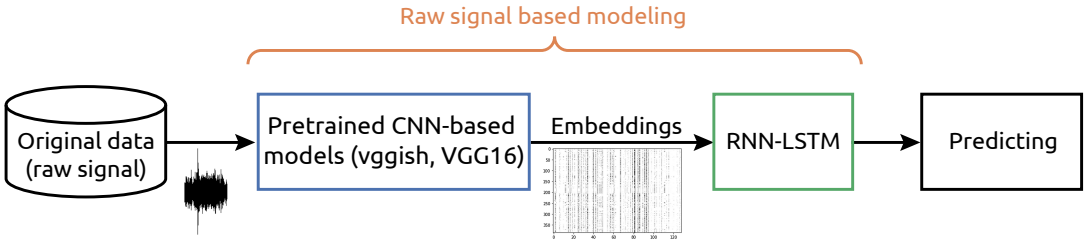


Figure 4.5: Pipeline of raw signal based time-dependent modeling. The first block is used to extract embeddings, those further used directly as an input to recognition model.

system, and find out, whether they depend on the input and/or model type, we repeat these experiments with the raw data based recurrent models (time-dependent) and with the feature based conventional (time-independent) models.

4.1.2 Raw Data Based Time-Dependent Models

First, we change the input of the model from features to raw data such as audio signal or video frames. The process of the feature extraction should be performed within the model. It can be implemented as a subsystem at front-end part of the model or trained from scratch. One of the most commonly used feature extraction architecture, successfully applied for many different machine learning tasks and able to work with different kinds of input, is based on CNNs. These networks usually play the role of finding such mapping between an input and an output signals, that a simple system applied to the output data of the CNN could solve the original classification or regression task easily, in contrast to the complex and computationally expensive problems, that the system would be facing while working with the original input signal directly. Also, CNNs often allows using less trainable parameters, as they are shared for the different input parts within the model.

Training models for the feature representations from scratch using raw data is time-, data- and resource-demanding task, therefore, in this work we use systems that are pre-trained on large datasets and described in Section 3.2.2. These are *vggish* for audio signal and *VGG16* fine-tuned with the *AffectNet* database for video frames. The embeddings extracted with these systems are used further as an input to RNN, similarly to the feature-based approach described in the previous section. The pipeline of this approach is shown in Fig. 4.5.

Although this process does not significantly differ from the feature based one, it implies the use of the embeddings that extracted with black-box models and are hardly explainable in contrast to the expert knowledge based feature sets used in the previous subsection.

Due to the specifics of the *vggish* embeddings, we are constrained to use a fixed low sampling rate of data, namely approximately 1.5 Hz. *vggish* were trained on data with the window length of 0.96 sec which cannot be changed while calculating the embeddings. We use hop size of 0.64 sec (1.5625 Hz) as it corresponds to exactly $\frac{2}{3}$ of the window length, therefore, it coheres to our approach to modeling used in the previous session. *VGG16* extract features from each data frame, therefore, we do not have to change anything in the modeling process for video modality.

We use these raw signal based features extracted with deep-learning models for audio and video modalities in the same fashion as in the previous section and increase the number of time steps used in the recurrent model. As we have a fixed window length for audio features and cannot change it, a particular amount of context may be modelled with much fewer time

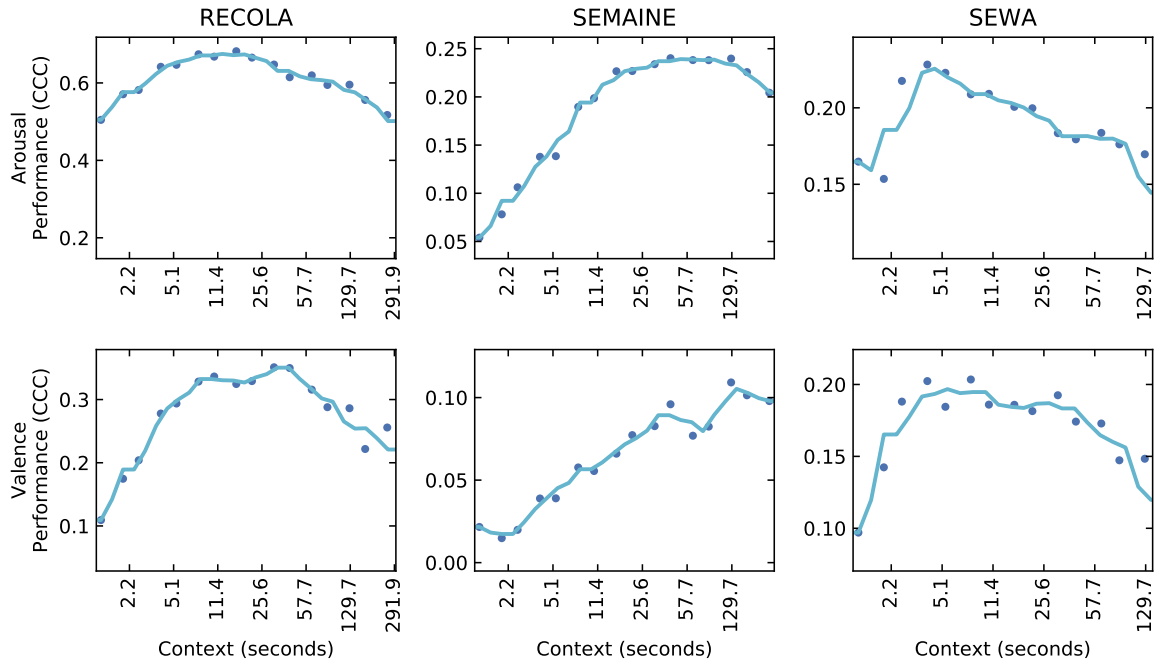


Figure 4.6: Results for straightforward context modeling on audio modality (vggish) with raw data based time-dependent models. Arousal (top row) and valence (bottom row) dimensions for three time-continuously annotated corpora: RECOLA, SEMAINE and SEWA. The performance is measured in terms of CCC and is dependent on amount of context represented by number of time steps used in RNN and data frequency.

steps. For example, in the feature based approach we used 25 time steps to model one second of context for RECOLA database, or 8192 time steps to model 328 seconds. Here, we only need one or two timesteps to model one second of context, or 512 time steps to model the same 328 seconds of context – 16 times less.

The results for context modeling with the raw data based recurrent models are presented in Fig. 4.6 for *vggish* (audio) and in Fig. 4.7 for *VGG16* (video). Note that for *vggish* we start the performance graph at 1.6 seconds in contrast to 0.1 seconds used in the previous section, as we are constrained with the low frequency.

Analyzing the results, one may notice similar patterns for all presented corpora. Audio-arousal pair of RECOLA demonstrates the ascending performance, with the peak at approximately ten seconds. Both dimensions with video modality demonstrate rather steep increase, reaching the top performance region at four seconds with slight fluctuations afterwards and the decrease of the same magnitude at the wider context windows (closer to 300 seconds).

Models for audio-arousal of SEMAINE database reach the top performance at similar values of the context length with less significant decrease at larger window sizes. Video-arousal pair demonstrates the most unstable behaviour among other combinations, also similarly to the feature based models.

Models for audio modality of SEWA database reaching the peak value earlier compared to the feature based ones, which is not the case for the video modality, where these values are approximately the same as for models from the previous subsection. Additionally, considering the results of the raw signal based models, one may notice similar performance decrease at the large window size compared to the feature based models.

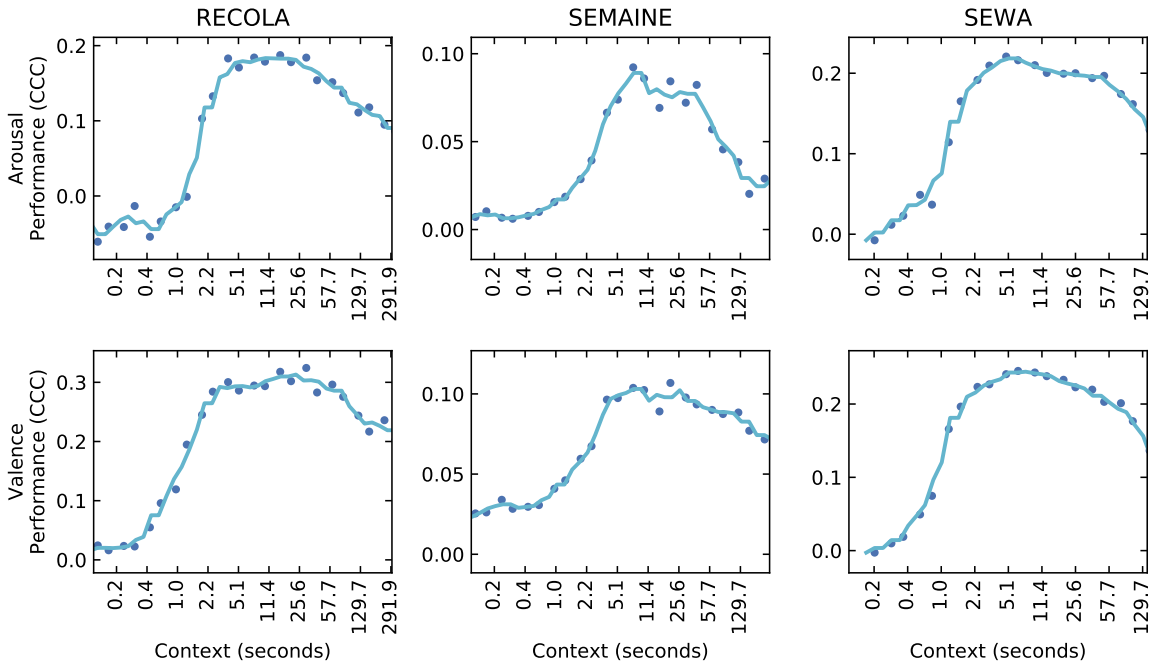


Figure 4.7: Results for straightforward context modeling on video modality (embeddings of VGG16 fine-tuned with AffectNet) with feature based time-dependent models. Arousal (top row) and valence (bottom row) dimensions for three time-continuously annotated corpora: RECOLA, SEMAINE and SEWA. The performance is measured in terms of CCC and is dependent on amount of context represented by number of time steps used in RNN and data frequency.

While these patterns have slight differences, they are in general similar and demonstrate a strong dependency between the amount of used data and the system performance. This proofs that this dependence is not related to a feature set. In the following subsection, we will compare them to the ones obtained with non-recurrent models.

4.1.3 Feature Based Time-Independent Models

Another way of the straightforward context modeling is to use conventional, non-recurrent, time-independent models, those have no information about the values of features and labels for previous time steps. Examples of such classification and/or regression algorithms are Linear Regression, Logistic Regression, k-Nearest Neighbors, Support Vector Machines, Decision Trees, Feedforward Neural Networks and many others. This type of the context modeling enables the context regulation at the *feature extraction* stage of the general emotion recognition pipeline (see Fig. 4.8 for more details).

As in our work we consider time-continuous emotion recognition, we are facing the regression problem. For our further analysis, we have selected four algorithms, those are widely used for various machine learning tasks nowadays, and are simple and effective:

- Linear Regression with L2 Regularization (Ridge Regression)[1];
- Epsilon-Support Vector Regression (SVR)[2];

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html`

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html`

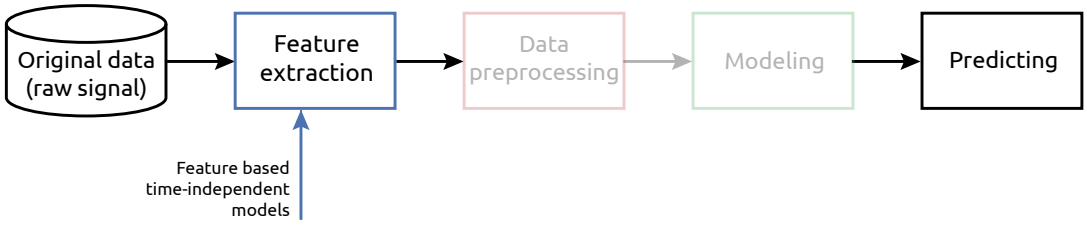


Figure 4.8: Straightforward approach to the speaker context modeling with feature based time-independent models in the general pipeline of the emotion recognition system

- Gradient Boosted Decision Trees (XGBoost)[3];
- Feedforward Neural Network (NN).

These models were described in detail in Section 2.4. Even though most of the current research on the *time-continuous* emotion recognition is based on the models designed to take time relation within the data into account, we consider the following prerequisites in order to be sure that our time-independent models are able to perform well at this task:

- annotations alignment and reaction lag correction (Section 3.2.3) – to ensure that the features and the labels are aligned;
- varying feature extraction window width – to allow flexibility in the context modeling.

By varying the feature extraction window width, we consider different amount of time steps used for calculation of functionals. In our work, the window width is strictly related to the hop size for functionals (i.e. the final data frequency) and is 1.5 times greater, providing an overlap between neighbouring feature vectors in the size of one-third of the window width. This approach is presented in Fig. 4.9, uses straightforward (in contrast to deep learning based) features as a basis, and is used further in our work with a special importance to Section 4.2.3.

In order to test the performance of simple models described above with regard to the amount of considered context, we use feature window width in a range from 0.06 to 30 seconds (i.e. data frequency rate from 25Hz to 0.05Hz) for the functionals based feature extraction. Here we use mean and standard deviation as functionals as they represent the most significant data changes. It is worth noticing that increasing the window width we significantly decrease the amount of training samples, as well as (in contrast to recurrent models) we lose the dynamics of data. After averaging over the feature window, any fluctuations of the original signal are lost. This is of the special importance for the larger window sizes.

Performance of the four time-independent models mentioned above in terms of CCC is presented in Fig. 4.10 for audio modality and in Fig. 4.11 for video modality.

Analyzing the figures, we can distinguish several patterns of dependence:

1. an increase of performance at the beginning with significant decrease afterwards, e.g. audio-arousal of RECOLA;
2. constant increase of performance with performance plateau at larger window widths, e.g. audio-arousal of SEMAINE;
3. performance plateau at shorter window lengths with decrease of performance after certain length, e.g. video-valence of SEWA.

[3] `https://xgboost.readthedocs.io/en/latest/index.html`

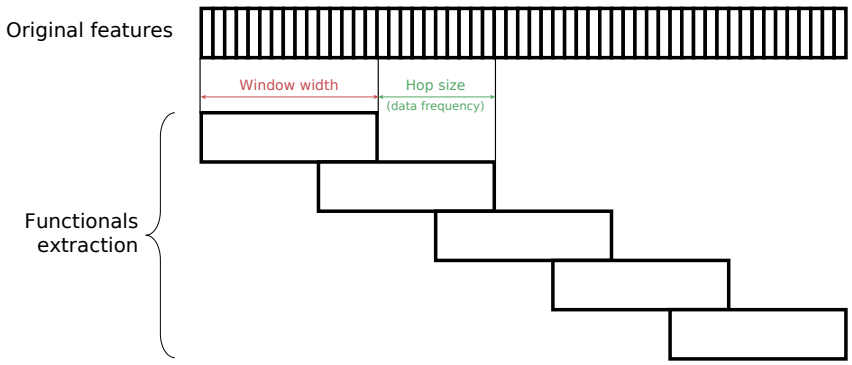


Figure 4.9: Extracting functionals from original LLDs. A part of original time-continuous feature vector, corresponding to particular amount of time steps (window length), is taken, then a value of functional (e.g. mean, standard deviation, max, min, etc.) is calculated on this data, after this the starting position for functional extraction is moved forward to hop size, which is equal to $\frac{2}{3}$ of window length, and the next value of functional is calculated.

The first pattern is inherent in the audio based models for RECOLA and SEWA databases. The second one – in the audio based models for SEMAINE. The video based models for RECOLA and SEMAINE demonstrate the first pattern and for SEWA – the third one.

Comparing the performance graphs of the time-independent models presented in this subsection and the RNNs discussed earlier in Section 4.1.1 (cf. Fig. 4.3 and Fig. 4.4), one may notice some similarities in patterns. Databases with the "earlier" peak – RECOLA and SEWA – also show similar behavior with nearly the same values of the optimal context for audio modality – approximately seven to ten seconds. Performance for the video modality starts to decrease already at one to two seconds, much earlier than for audio. However, a significant difference between the results of the time-independent and the time-dependent models for RECOLA and SEWA is performance behavior after the optimal value. In case of time-dependent models, the performance sometimes stays at similar level after the context reached the peak up to the moment when the context window represents the average recording length or decreases gradually. Nevertheless, it is not the case for time-independent models: performance drops drastically after reaching the peak. It may be caused by the loss of data dynamics, while the structure of RNNs allows to preserve it even with wider context windows.

Summary. In this section, we have tested contextual dependencies with two feature representations – feature based and raw signal based, as well as with two model types – recurrent and conventional. Comparing all obtained results, we can conclude that there are strong patterns between the amount of context represented by the input data and the model performance. Behaviour of the performance curve as well as the value of the performance metric itself may differ depending on the model. However, the general pattern remains the same – there is a particular amount of context needed to represent the underlying cues with a certain degree of precision. According to the results presented in this section, the existence of these dependencies is not connected to the particular model or feature set.

All the three approaches presented in this section have drawbacks. The feature based time-dependent models described in Section 4.1.1 may require wide time window, which can be the cause for a performance reduction with large length of the input sequences. The raw data based recurrent models described in Section 4.1.2 are highly resource-demanding at a

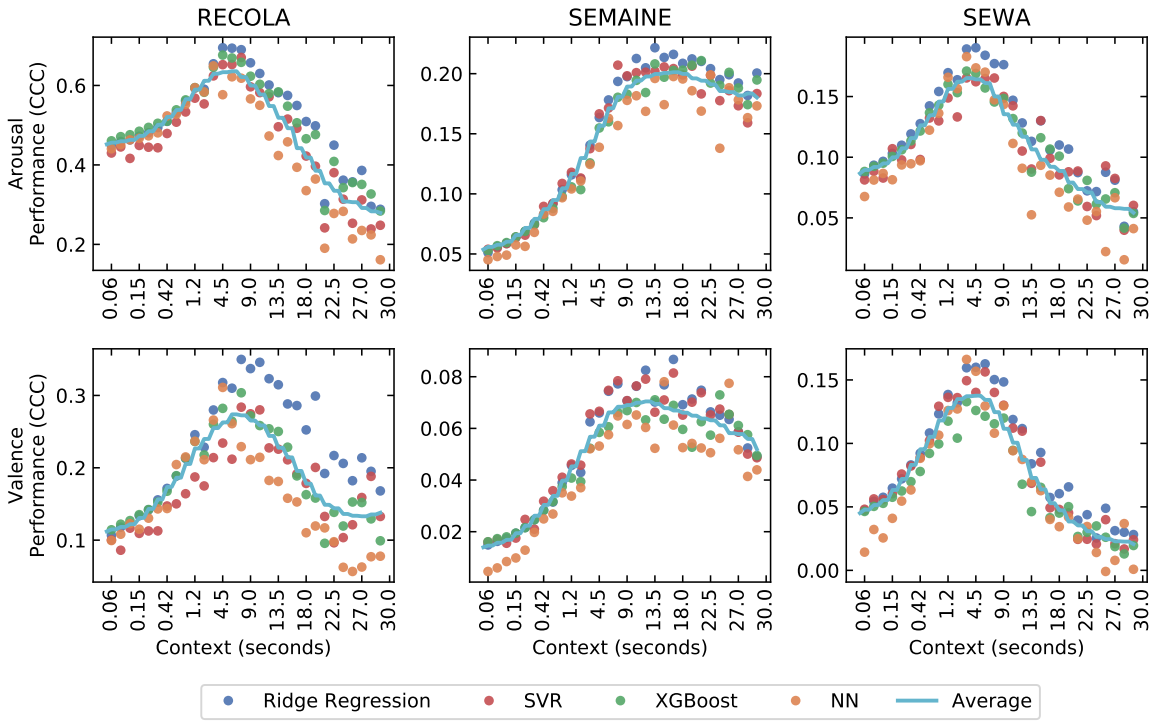


Figure 4.10: Results for straightforward context modeling on audio modality (eGeMAPS) with classical time-independent models: Ridge Regression (blue), SVR (red), XGBoost (green), NN (orange). The average values over four models are in cyan line. Arousal (top row) and valence (bottom row) dimensions for three time-continuously annotated corpora: RECOLA, SEMAINE and SEWA. The performance is measured in terms of CCC and is dependent on amount of context represented by one data sample.

step of feature extraction without any performance gain. The feature based time-independent models described in Section 4.1.3 lose data dynamics with the high lengths of context, which leads to a decrease in performance. To level these negative features of straightforward context modeling, we developed a simple and flexible data sparsing approach. In the following section we will explain it in detail, introduce its effect on the modeling and the system performance and consider its weaknesses.

4.2 Data Sparsing

In order to discover if the model performance is dependent on the context coverage and not peculiarities of the model or features, as well as to increase the level of flexibility in the context modeling, we developed the contextual data sparsing approach, which can be easily applied to any feature-based recurrent model.

4.2.1 General Concept

Having feature-based data, extracted in accordance with Section 3.2.2 and Fig. 4.9, the amount of covered context can be described with the following formula:

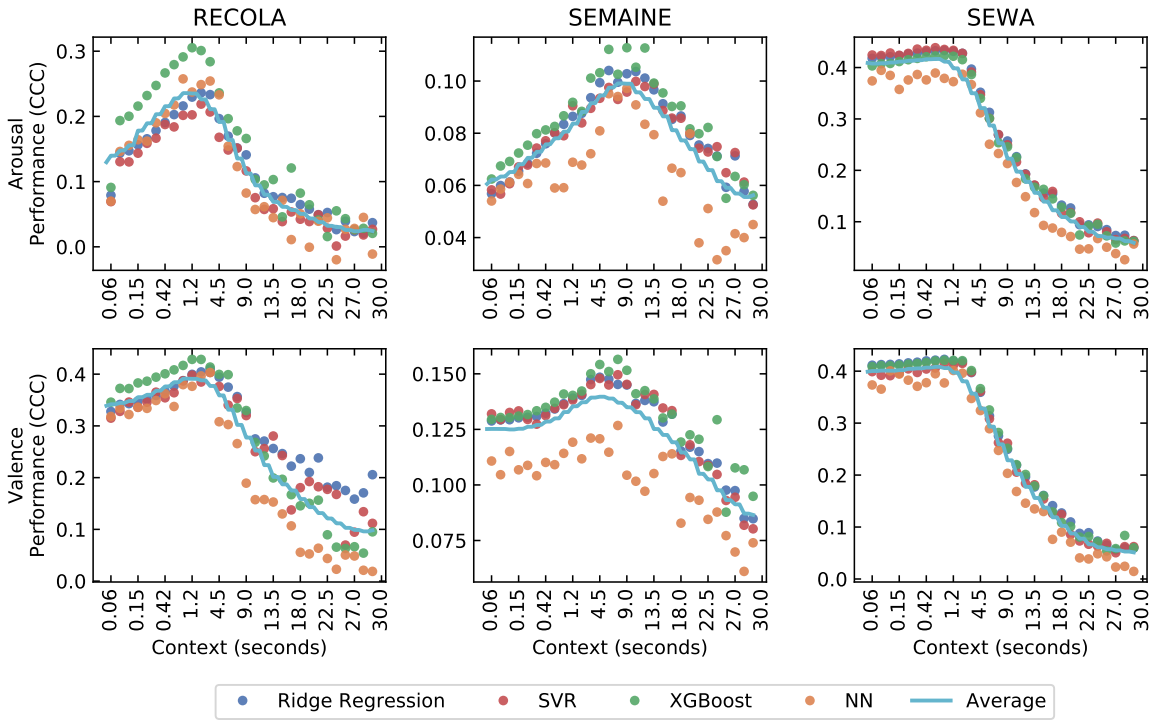


Figure 4.11: Results for straightforward context modeling on video modality (AUs) with classical time-independent models: Ridge Regression (blue), SVR (red), XGBoost (green), NN (orange). The average values over four models are in cyan line. Arousal (top row) and valence (bottom row) dimensions for three time-continuously annotated corpora: RECOLA, SEMAINE and SEWA. The performance is measured in terms of CCC and is dependent on amount of context represented by one data sample.

$$C = H \cdot (T - 1) + W, \tag{4.3}$$

where C is a context coverage of one data sample in seconds, W is a window size for feature extraction in seconds, T is a number of time steps, H is a hop size. In our case, the hop size H is strictly connected to the window size and $H = \frac{1}{3}W$, therefore, (4.3) can be rewritten as:

$$C = \frac{1}{3}W(2 \cdot T + 1), \tag{4.4}$$

in turn, the window size W used for the feature extraction in our work is strictly connected to the data frequency (F) as $W = \frac{3}{2F}$. Hence, the hop size is an inverse value of the frequency, which is set by labels according to the original annotation of the databases. Taking this into account, we can rewrite (4.4) using F:

$$C = \frac{T + 0.5}{F} \tag{4.5}$$

Let us go back to the example with RECOLA database. It has the data frequency of 25Hz, i.e. $W = 0.06$ and by using from 2 to, for example, 64 time steps, we model from 0.1 to 2.58 seconds of context (see first row of Table 4.1).

One may see, that extending our model from 2 to 64 time steps, allows the context coverage of rather short time period of 2.58 seconds. Fixed data frequency applies additional

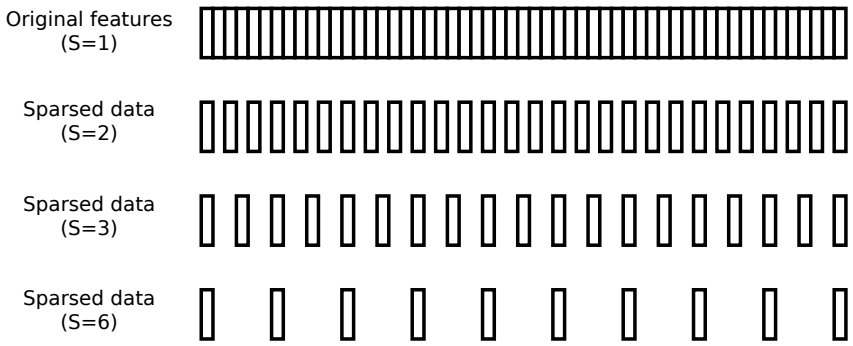


Figure 4.12: Concept of data sparsing. Each *S-th* frame is kept as a result of data sparsing, other frames are omitted. S is a sparsing coefficient, which can be set to any positive integer value. With $S = 1$ the features are kept in original form.

constraints on the context modelling, whose can be partially overcome by the data sparsing. The main idea of this approach is to sparse the time-continuous input data, i.e. to use not each time step while forming data samples, but only each *n-th* time step. The degree of data sparsity is controlled by a *sparsing coefficient (S)*, which defines the hop between two time steps being considered for modeling within one sample (see Fig. 4.12).

The sparsing coefficient introduces an additional factor in (4.3):

$$C = H \cdot S(T-1) + W, \tag{4.6}$$

where S is a sparsing coefficient. Or rewritten with the frequency and according to values of the hop size and the window size used in our work:

$$C = \frac{S(T-1) + 1.5}{F}, \tag{4.7}$$

Then the amount of modelled context can be greatly expanded (see Table 4.1).

		Time steps										
		2	3	4	6	8	12	16	24	32	48	64
Sparsing coefficient	1	0.10	0.14	0.18	0.26	0.34	0.50	0.66	0.98	1.30	1.94	2.58
	2	0.14	0.22	0.30	0.46	0.62	0.94	1.26	1.90	2.54	3.82	5.10
	3	0.18	0.30	0.42	0.66	0.90	1.38	1.86	2.82	3.78	5.70	7.62
	4	0.22	0.38	0.54	0.86	1.18	1.82	2.46	3.74	5.02	7.58	10.14
	6	0.30	0.54	0.78	1.26	1.74	2.70	3.66	5.58	7.50	11.34	15.18
	8	0.38	0.70	1.02	1.66	2.30	3.58	4.86	7.42	9.98	15.10	20.22
	12	0.54	1.02	1.50	2.46	3.42	5.34	7.26	11.10	14.94	22.62	30.30
	16	0.70	1.34	1.98	3.26	4.54	7.10	9.66	14.78	19.90	30.14	40.38
	24	1.02	1.98	2.94	4.86	6.78	10.62	14.46	22.14	29.82	45.18	60.54
	32	1.34	2.62	3.90	6.46	9.02	14.14	19.26	29.50	39.74	60.22	80.70
	48	1.98	3.90	5.82	9.66	13.50	21.18	28.86	44.22	59.58	90.30	121.02
	64	2.62	5.18	7.74	12.86	17.98	28.22	38.46	58.94	79.42	120.38	161.34

Table 4.1: Context coverage (in seconds) for RECOLA using data sparsing

One may see that values on diagonals of the table often correspond to similar amount of context with a certain moderate bias. For example, 5.82 seconds can be modelled with {S=48, T=4}, and almost the same context (5.70 seconds) can be modelled with {S=3, T=48}. This introduces additional flexibility in the context modeling and allows one to use models that are different in size and depth in order to work with the same amount of context. Using this

approach, we can model a wide range of context length, while simultaneously being flexible in model complexity.

The context modeling with data sparsing enables the context regulation simultaneously at the *data preprocessing* and the *modeling* stages of the general emotion recognition pipeline (see Fig. 4.13 for more details).

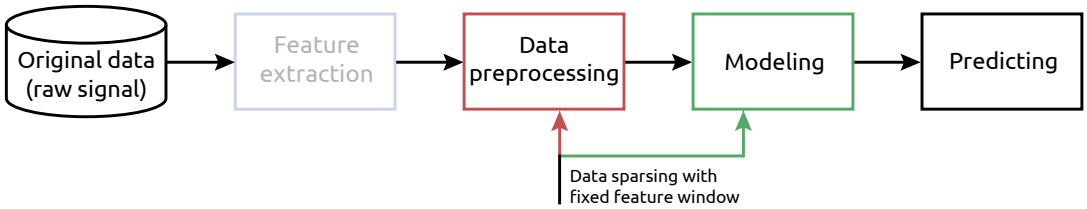


Figure 4.13: Data sparsing approach to the speaker context modeling with feature based time-independent models in the general pipeline of the emotion recognition system

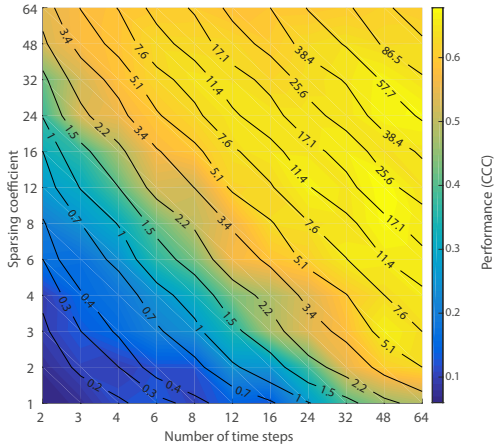
4.2.2 Data Sparsing for Feature Based Time-Dependent Models

A recurrent model shows a behavior of performance depicted as heat map in Fig. 4.14a on the audio-arousal pair of RECOLA database, considered as an example previously, for one run of the system (not statistics). X-axis represents the number of used time steps, y-axis represents the sparsing coefficient, color represents the performance in terms of CCC, black contour lines shows the amount of modelled context in seconds. One may notice the strong dependence between the model performance and the amount of context used. Performance values are not dependent on the amount of time steps and also do not decrease significantly with the high values of sparsing coefficient, i.e. when we sparse the data greatly. For this particular example we can see that the performance of the model starts with low values, then increases gradually, reaching the peak at approximately eight seconds and stays at this performance plateau similarly to the previous experiments (cf. Fig. 4.3).

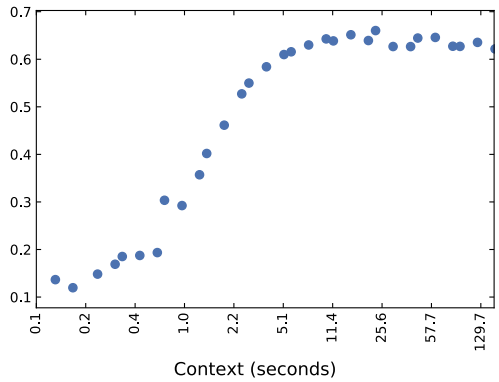
Although this graph provides a clear and vivid picture of dependencies between the amount of context and the model performance, it is hardly suitable for comparison of these dependencies across modalities, dimensions, corpora, etc. as it is three-dimensional (performance depicted with color is the third dimension). Thereby, we reconstruct it into a simple scatter plot with x-axis representing the amount of used context in seconds (similar to black contour lines) and y-axis – the performance in terms of CCC, similarly to the graphs used previously. As in some cases values of context are almost identical, we use clustering to reduce the amount of data points and averaging performance over certain time window, e.g. each value that falls into the range of $[5.0, 5.5]$ seconds is assigned to 5.25 seconds. The simplified version of Fig. 4.14a is presented in Fig. 4.14b.

We will follow the data representation of Fig. 4.14b for further experiments with the data sparsing throughout this work. Additional graphs represented by heat maps for the other databases and modality-dimension pairs are presented in Appendix A.

Studying the results further, we plot graphs for remaining databases in the same fashion as Fig. 4.14b with the original data frequency – Fig. 4.15 for the audio modality and Fig. 4.16 for the video modality.



(a) Performance heatmap for data sparsing approach using RECOLA database for audio-arousal pair. X-axis represents amount of time steps (T) used in model, y-axis – sparsing coefficient (S); color is used to show the model performance in terms of CCC (from blue (low) to yellow(high)); black contour lines shows amount of modelled context in seconds.



(b) Scatter plot version of the performance heatmap for data sparsing approach using RECOLA database for audio-arousal pair. The performance is measured in terms of CCC and is dependent on amount of context represented by number of time steps used in RNN, data frequency and sparsing coefficient.

Figure 4.14: Heatmap and scatter plot demonstrating results of data sparsing approach.

Analyzing the results, one may notice patterns almost identical to the ones presented in Section 4.1.1 for both audio and video modalities. The only differences are in the video-valence pair of RECOLA – with the data sparsing its performance decreases with larger windows, and with the video-arousal pair of SEMAINE – with data sparsing the graph is much smoother and demonstrates more stable results. In contrast to Section 4.1.1, here we achieve the higher values of context not by increasing the number of time steps, but by using the high sparsing coefficients, and the patterns remain the same. This fact proofs existing dependencies between the amount of context itself and the system performance, independent of model peculiarities (such as amount of time steps used).

Moreover, considering the performance values, one may notice the same level for both approaches. This means that on average, with increasing steps between the time steps within one data sample, performance of the system does not drop. This introduces an additional degree of freedom to the context modeling procedure, as now it can be controlled not only through the number of time steps, but also through the sparsing coefficient. More details on the effect of data sparsing on system performance are presented in Section 4.4.

We have increased the level of flexibility for contextual modeling by introducing the time-continuous data sparsing approach. However, we are still constrained with the original data frequency. In the next section, we will overcome it by using the variable data frequency.

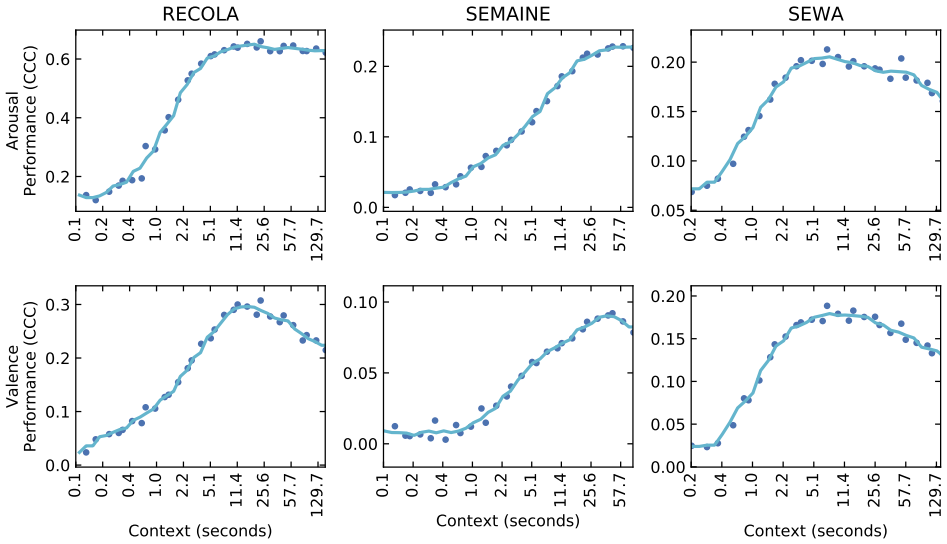


Figure 4.15: Results for data sparsing approach to context modeling on audio modality (eGeMAPS) with feature based time-dependent models. Audio modality, arousal (top row) and valence (bottom row) dimensions for three time-continuously annotated corpora: RECOLA, SEMAINE and SEWA. The performance is measured in terms of CCC and is dependent on amount of context represented by number of time steps used in RNN, data frequency and sparsing coefficient.

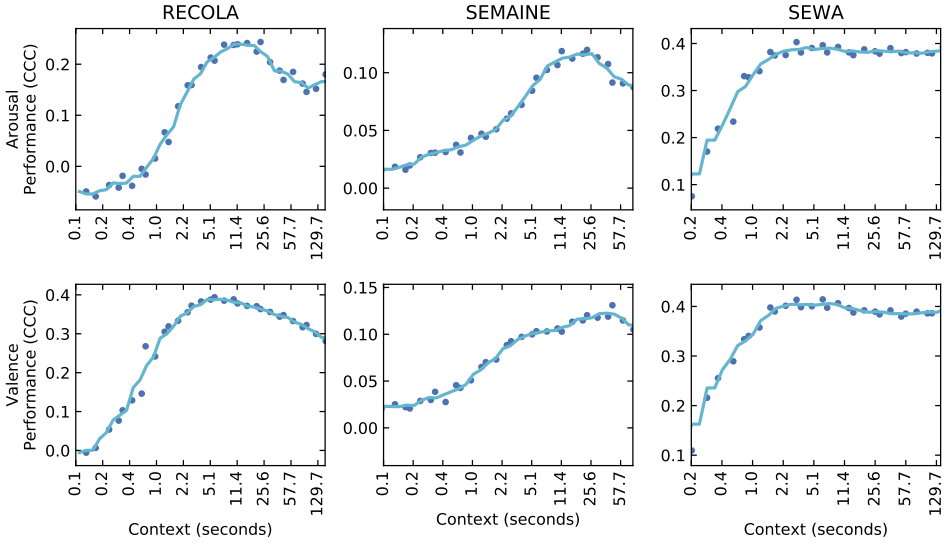
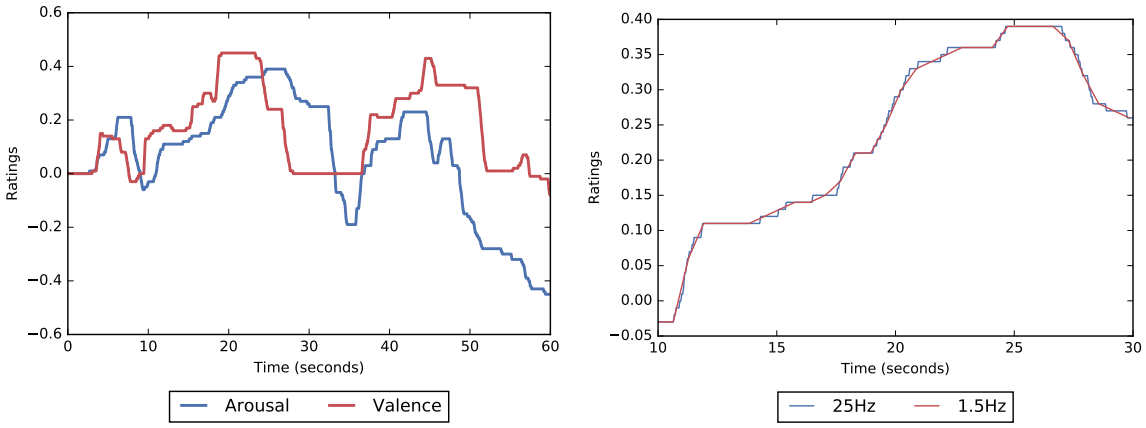


Figure 4.16: Results for data sparsing approach to context modeling on video modality (AUs) with feature based time-dependent models. Audio modality, arousal (top row) and valence (bottom row) dimensions for three time-continuously annotated corpora: RECOLA, SEMAINE and SEWA. The performance is measured in terms of CCC and is dependent on amount of context represented by number of time steps used in RNN, data frequency and sparsing coefficient.

4.2.3 Data Sparsing with Varying Feature Window



(a) Arousal (blue) and valence (red) ratings for recording *P16* of RECOLA database in time interval of $[0, 60]$ seconds.

(b) Arousal ratings for recording *P16* of RECOLA database in time interval of $[10, 30]$ seconds with frequencies 25Hz (blue) and 1.5Hz (red).

Figure 4.17: Examples of ratings (true labels) for RECOLA database.

Time-continuous data (including both features and labels) is presented with different sampling rates in different databases. It is not clear, which sampling rate is sufficient for the modeling with the high level of performance and not overcomplicates the model with unnecessary data at the same time. One way to find it out is upsampling or downsampling of data. By upsampling one often considers the interpolation of the data with the increased sampling rate. While upsampling may not work well for features, it is rather a rational solution for the time-continuous labels of emotions, because they do not fluctuate rapidly. An example of ratings for recording *P16* of RECOLA database is presented in Fig. 4.17a. The original frequency of 25Hz used in this graph provides us with 1500 label datapoints for one minute of the data. Of course, emotions cannot change so many states in one minute, and annotation of databases, collected by a human-expert while watching a video clip in real time, also cannot contain many different states which are distinguishable from one another. Therefore, downsampled ratings does not differ much from the original ones (see Fig. 4.17b) and allow using them for the modeling without any significant loss of information.

The frequency of audio features may be changed easily. As they derived from the raw wave signal, features of audio modality with an increased frequency emerge not artificially, but from real data. On the other hand, the frequency of video features is limited to frame rate

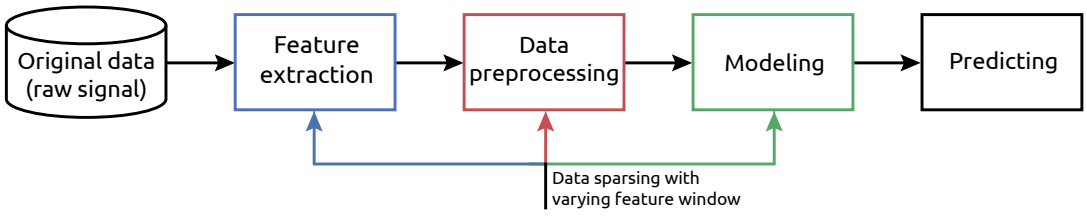


Figure 4.18: Data sparsing approach to the speaker context modeling with feature based time-independent models and varying feature window in the general pipeline of the emotion recognition system

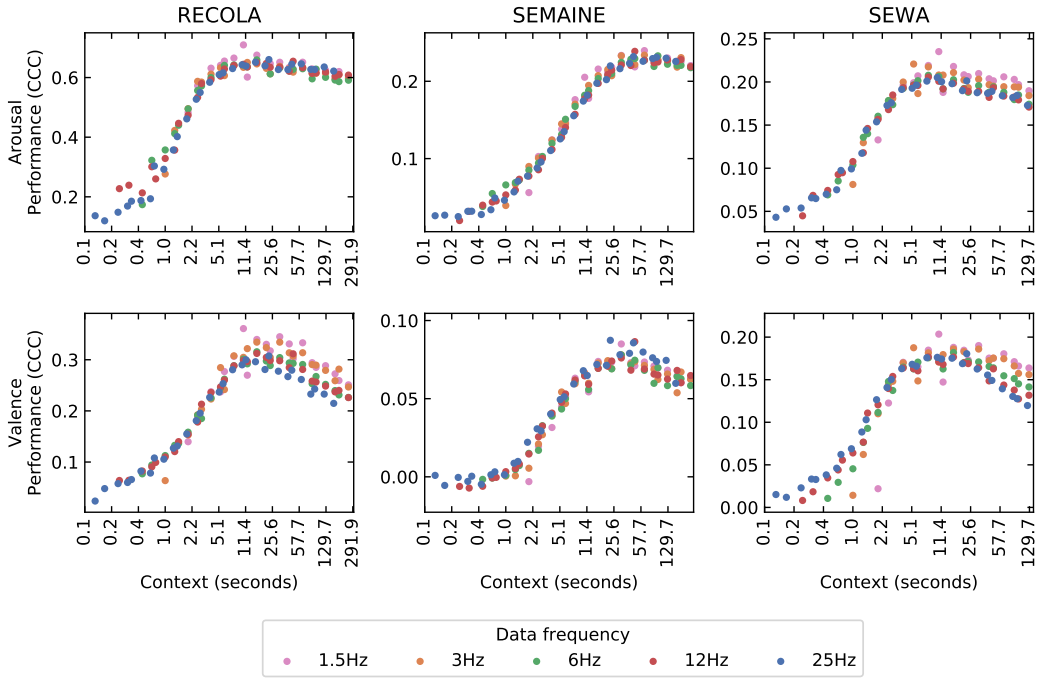


Figure 4.19: Results for data sparsing approach with varying feature window to context modeling on audio modality (eGeMAPS) with feature based time-dependent models. Arousal (top row) and valence (bottom row) dimensions for three time-continuously annotated corpora: RECOLA, SEMAINE and SEWA. The performance is measured in terms of CCC and is dependent on amount of context represented by number of time steps used in RNN, data frequency and sparsing coefficient.

of source video file. Usually it is equal to 25 or 50 frames per second.

In this section, we will adapt data sparsing approach from previous section, but will change the frequency of the data in order to increase flexibility of the context modeling, which now covers all the three parts of the general emotion recognition pipeline (see Fig. 4.18 for more details).

We select the following data frequencies to work with: 25Hz, 12.5Hz, 6.25Hz, 3.125Hz, 1.5625Hz. We start with 25Hz as it is the highest frame rate of corpora used in our work, and limit it to 1.5625Hz. We round these values in the further notation for simplicity and more reader-friendly appearance to [25Hz, 12Hz, 6Hz, 3Hz, 1.5Hz], but keep them unchanged in the experiments.

We upsample and downsample ratings for arousal and valence to the corresponding frequencies and use the feature extraction method described in Section 4.1.3 to change the frequency of data respectively. This introduces an additional degree of freedom to (4.7) and, therefore, increases flexibility of context modeling.

Results for the context modeling using the data sparsing with varying feature window are presented in Fig. 4.19 for audio and Fig. 4.20 for video. Comparing them to graphs from the previous section, one may notice identical patterns and, in many cases, also identical averaged performance. For some cases, e.g. audio-valence and both dimensions with video of RECOLA, one may notice that lower frequencies, such as 1.5Hz provide higher performance on larger context length. The lower the frequency, the fewer time steps or lower sparsing coefficient with a fixed number of time steps is needed to provide the same amount

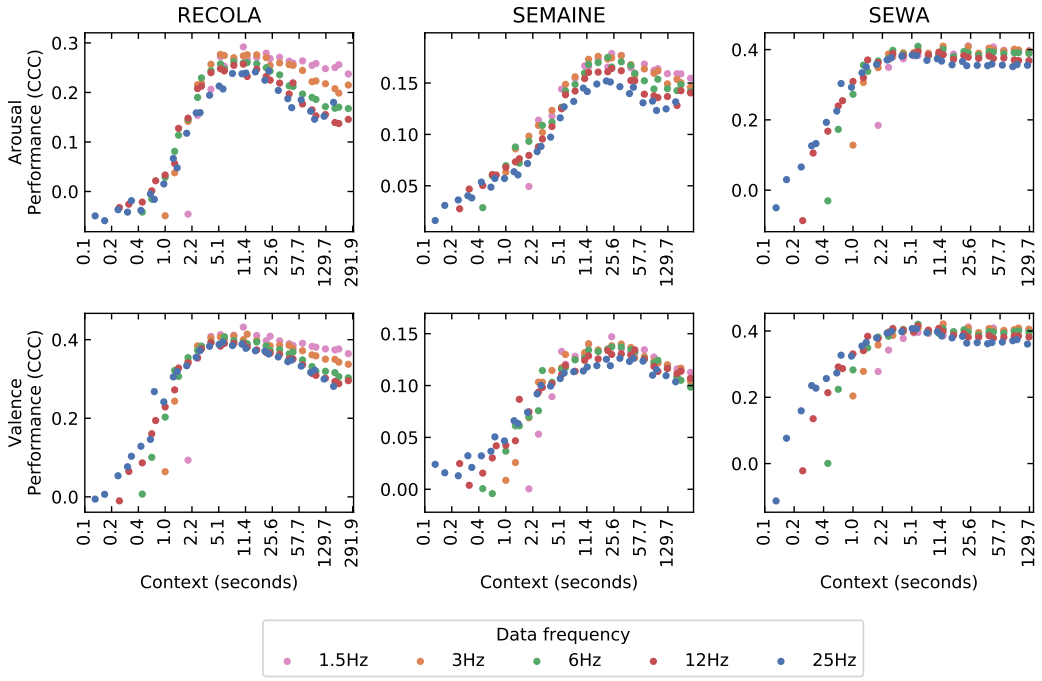


Figure 4.20: Results for data sparsing approach with varying feature window to context modeling on video modality (AUs) with feature based time-dependent models. Arousal (top row) and valence (bottom row) dimensions for three time-continuously annotated corpora: RECOLA, SEMAINE and SEWA. The performance is measured in terms of CCC and is dependent on amount of context represented by number of time steps used in RNN, data frequency and sparsing coefficient.

of context. We will cover these differences in greater details in Section 4.4.

Summary. The data frequency, introduced as the third degree of freedom in the context modeling, allowed us to further increase a flexibility and derive more trustworthy conclusions about dependencies between the amount of context and the model performance: independent of the frequency of data used, patterns remain the same. This also proofs that the optimal amount of context is corpus-dependent, as even with the same conditions (data frequency, number of time steps, sparsing coefficient), one may notice a significant difference between performance curves for e.g. RECOLA and SEMAINE.

In the following section, we will study behavior of these patterns in a cross-corpus scenario to see if the optimal amount of context is inherited from the test or train corpus.

4.3 Transferability to Cross-corpus Setting

Patterns presented in the previous sections show strong dependencies on the database. Optimal value of context stays invariant to changes in the data representation, and in most cases also does not change significantly with emotional dimensions within one database. Hence, it is interesting to see, how these patterns will behave in a scenario of mixed data – cross-corpus modeling.

The cross-corpus modeling implies that the system is either trained on one database (corpus) and tested on another, or trained on a mix of the databases and tested on another (data from target corpus may also be included into training subset). Thus, the testing and training data do not origin from the same data distribution. The cross-corpus modeling

improves many aspects and qualities of models, such as their applicability for working with a wide range of data, including real-world or in-the-wild data if it was present in the training corpus. In addition, it expands the amount of data that can be used for training significantly, which in turn has a potential to boost the performance of the model and its ability to generalize the data. Data-driven approach is used for training most of the models at the moment.

Nevertheless, this approach introduces many additional challenges, as the data used for training and/or testing comes from different sources. Apart from differences in the recording conditions, it may be annotated differently: categorically or dimensionally, with the different annotation tools, by annotators with different level of expertise. Some of these inconsistencies may be handled by the model itself, but some of them should be addressed at the stage of data preprocessing. For example, in our previous work we found out that it is likely for model trained on the data from one corpus to perform within its distribution for the target corpus as well (Fedotov et al., 2018b) despite the fact that the models were able to operate within the range of $[-1, +1]$ and labels scaling was used. Source data restricts the model to predict close to its distribution, which significantly hinders the model in terms of performance.

In order to overcome the issue of model's strong dependence on the input data conditions, we apply a corpus domain adaptation approach based on combination of Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA) suggested by Sagha et al. (2016). The main idea is to consequently use these two approaches to transform features of the two data sources into a common distribution and eliminate the ground differences in data. First, we fit PCA on both databases separately to define the most important components from original data. Then we apply each trained PCA model to both source and target data. Two matrices obtained with each model are concatenated and later used as an input to CCA model. The resulting data representation of CCA model is split back to correspond to the original split of data to the source and target corpus. For better understanding, this procedure is presented in Fig. 4.21 and as Algorithm 2.

Notations for Algorithm 2 are as follows: X_s represents the data of source corpus, X_t – the data of target corpus; X_{s_norm} and X_{t_norm} – normalized data of the source and target corpora respectively; T_j^i – features of j data transformed with PCA algorithm trained on i data; M_i – concatenated T_i^i and T_j^i feature matrices; W_i – resulting matrix of CCA algorithm; $\hat{X}_s$ and $\hat{X}_t$ – transformed and adapted feature sets for the source and target corpus respectively.

We keep 99% of the original data variance during the PCA procedure to have a sufficient amount of components after feature transformation to perform analysis onto, and avoid singularity during CCA. According to our previous experiments (Kaya et al., 2019), we keep

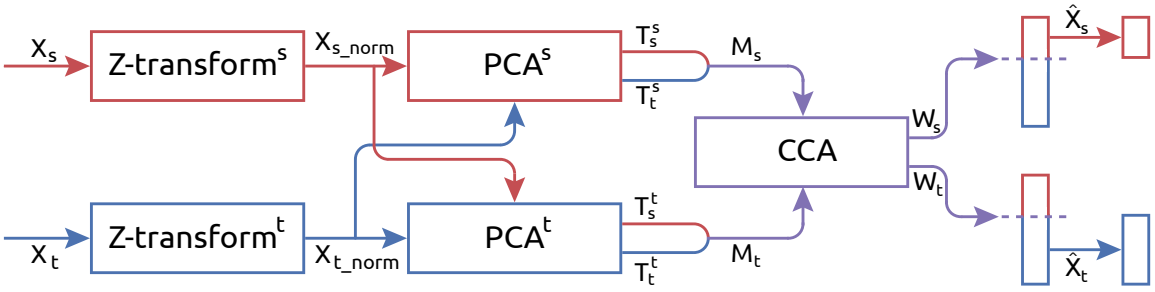


Figure 4.21: PCA-CCA approach to cross-corpus domain adaptation adopted from (Sagha et al., 2016). Red lines represent data flow of source corpus, blue lines – of target corpus, violet lines – of combined data. For more details on notation, see Algorithm 2.

Algorithm 2: PCA-CCA approach to cross-corpus domain adaptation

Result: Transformed and adapted feature sets: $\hat{X_s}$ and $\hat{X_t}$
for i *in* $\{s, t\}$ **do**
  $X_{i\_norm} \leftarrow Z$-transform applied to X_i
end
for i *in* $\{s, t\}$ **do**
  $PCA^i \leftarrow PCA$ algorithm fitted with X_{i_norm} to keep 99% of corpus variance
end
for i *in* $\{s, t\}$ **do**
  for j *in* $\{s, t\}$ **do**
    $T_j^i \leftarrow PCA^i$ applied to X_{j_norm}
  end
end
for i *in* $\{s, t\}$ **do**
  $M_i \leftarrow concatenate(T_s^i, T_t^i)$
end
$[W_s, W_t] \leftarrow CCA(M_s, M_t)$
$\hat{X_s} \leftarrow W_s[: length(X_s)]$
$\hat{X_t} \leftarrow W_t[-length(X_t) :]$

the number of components for CCA close to the number of principal components.

Similarly to methodology described in the previous sections, we perform the context modeling in cross-corpus setting. We use RNN-LSTM models and the data sparsing approach on data with the frequency of 6Hz (as middle value among frequencies used in Section 4.2.3). Data of each pair of corpora (source and target) is preprocessed according to Algorithm 2 prior to context modeling with data sparsing.

As we use the domain adaptation approach, we build, train and test separate models for each combination of corpora, modality and dimension, as well as the sparsing coefficient and the number of time steps. In the following part of this thesis, we will present the results for the contextual cross-corpus modeling, using similar graphical representation as in the previous sections of this chapter. As we test each pair of corpora separately, the number of graphs necessary to present the results increases significantly. Therefore, we will present here the results of cross-corpus context modeling for audio-arousal (Fig. 4.22) only. Similar graphs for the other modality-dimension pairs can be found in Appendix B.

One may notice that for most of the cases, the optimal value of context for the cross-corpus model (blue dotted line) lies near one for the train corpus (green dot-dashed line) or for the test corpus (red dashed line). The results for other modality-dimension pairs, presented in Appendix B demonstrate similar behavior. However, based on these graphs, we cannot explicitly define which corpus (train or test) dominates the optimal amount of context for the cross-corpus scenario. Out of total 24 cases, in 11 it is closer to the optimal amount of the train corpus, in 13 – of the test corpus. In 11 cases it lies between them and in 13 – not.

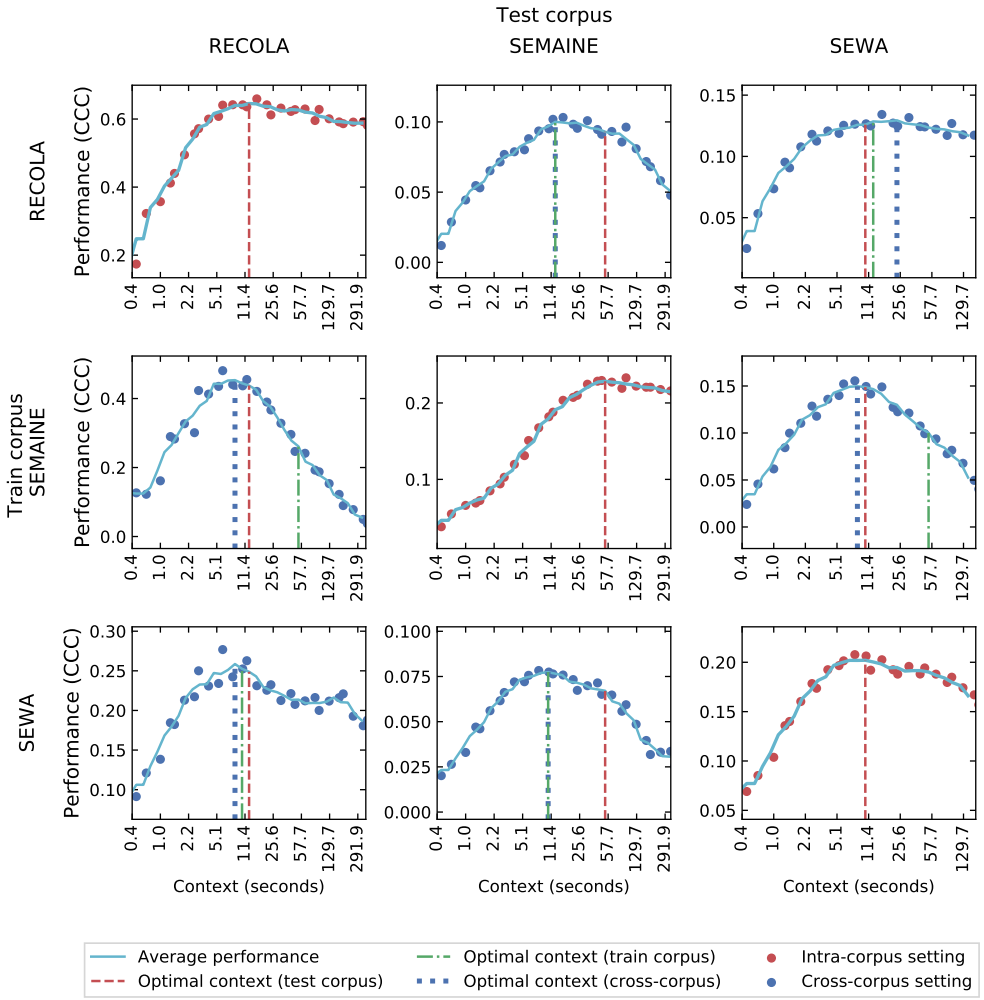


Figure 4.22: Results for data sparsing approach with varying feature window to cross-corpus context modeling on audio modality (eGeMAPS), arousal dimension with feature based time-dependent models. Red dashed line represents optimal context amount for test corpus, green dot-dashed line – for train corpus, blue dotted line – for cross corpus setting. The performance is measured in terms of CCC and is dependent on amount of context represented by number of time steps used in RNN, data frequency and sparsing coefficient.

4.4 Analysis and Discussion

Results of the previous sections proved that the amount of used context affect the system performance and the optimal value is neither feature set nor the amount of time steps nor the data frequency dependent. However, there are dependencies on modality and corpora. In this section, we will consider it in greater detail and discuss possible reasons for it.

Foremost, we will define the optimal amount of context for each corpus, modality and dimension. As the approach presented in Section 4.2.3 provides the highest flexibility in context modeling, we will use it for this purpose. First, we calculate the average performance across the five used data frequencies f (from 1.5Hz to 25Hz). Then, we find the lowest amount of context that provides such a performance, that there is no significantly higher value for longer context. We will check if the difference is significant with paired sample t-test (p=0.01).

The optimal values, obtained with this criterion, are presented in Table 4.2.

		RECOLA	SEMAINE	SEWA
Audio	Arousal	15.4	23.0	7.7
	Valence	11.5	24.0	5.8
Video	Arousal	5.3	12.0	6.0
	Valence	4.0	11.5	2.9

Table 4.2: Optimal values of context length for RECOLA, SEMAINE and SEWA according to results from Section 4.2.3

One may see that the optimal value for the video modality is significantly lower that for the audio modality in all six cases. We cannot see similar differences between arousal and valence. Regarding corpora, RECOLA and SEWA demonstrate the shorter optimal context coverage than SEMAINE.

A possible reason for these differences in the optimal context across the corpora might be the duration of speaker's turns and pauses. As we are dealing with the time-continuous emotion recognition problem, we need to provide a curve of emotional changes over time in terms of arousal or valence. The data has been annotated by watching video clips and

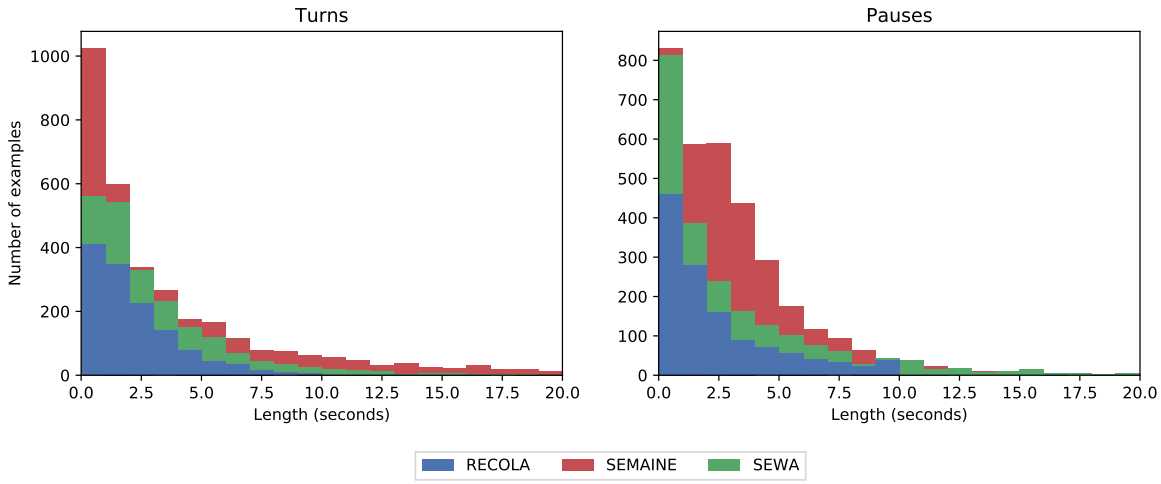


Figure 4.23: Turn/pause length distributions for RECOLA, SEMAINE and SEWA databases.

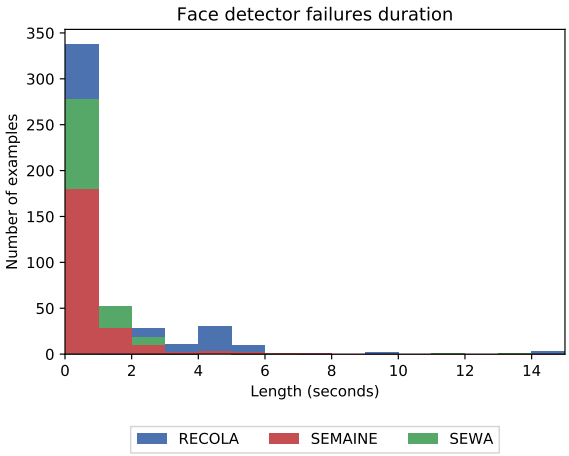


Figure 4.24: Face detector failure duration distributions for RECOLA, SEMAINE and SEWA databases.

capturing ratings simultaneously, hence, both audio and video modalities were available. However, we consider the models trained with the audio and video data separately, therefore the audio model has to learn how to cope with pauses – moments in time when nothing is being said (essentially no data is present), but arousal/valence rating are changing. We can isolate this type of the data at the moment of training by masking it (ignoring examples with no data while calculating the loss function), but the model nevertheless has to provide predictions at the testing stage. In Fig. 4.23 we present histograms of turns and pauses distributions for corpora used in this chapter.

		RECOLA	SEMAINE	SEWA
Turns	Mean	2.19	4.78	2.96
	Median	1.67	2.07	1.95
Pauses	Mean	3.09	3.27	3.54
	Median	1.67	2.51	2.02

Table 4.3: Mean and median value of turn and pause duration for RECOLA, SEMAINE and SEWA.

All three corpora have rather short duration of both turns and pauses. We present the mean and median values for them in Table 4.3. SEMAINE has the highest median value for turns and pauses, but this difference is not large.

On the other hand, the video modality is always present but the optimal amount of context for it is not close to zero, especially for SEMAINE. Similarly to a pause in speech, an analog for video modality would be frames, from which the system cannot detect face and, therefore, extract features. This may be caused by several reasons, including:

- person's face is not in the field of view of the camera;
- there are difficult conditions for the camera, e.g. a bad light;
- some objects (e.g. a hand) covering the face of the person.

All these situations may lead to a failure in feature detection pipeline, resulting in corrupted or missing data. *OpenFace* software used in our work to extract AUs from videos, provide additional variables in an output file – *confidence* and *success*. The first one corresponds

to the confidence of the system that the face was recognized correctly; the second – binary version of the confidence. We will use *success* variable to track the facial feature detection failures by measuring the length of time periods when the *success* variable is equal to 0. Distribution of the duration distribution is presented in Fig. 4.24.

One may notice, that the duration of failures for the video modality is shorter than the pauses for audio. Moreover, most of them (57% for RECOLA, 69% for SEMAINE and 61% for SEWA) are shorter than 0.1 seconds and approximately one-third of the failures for each database correspond to merely one frame of video recording.

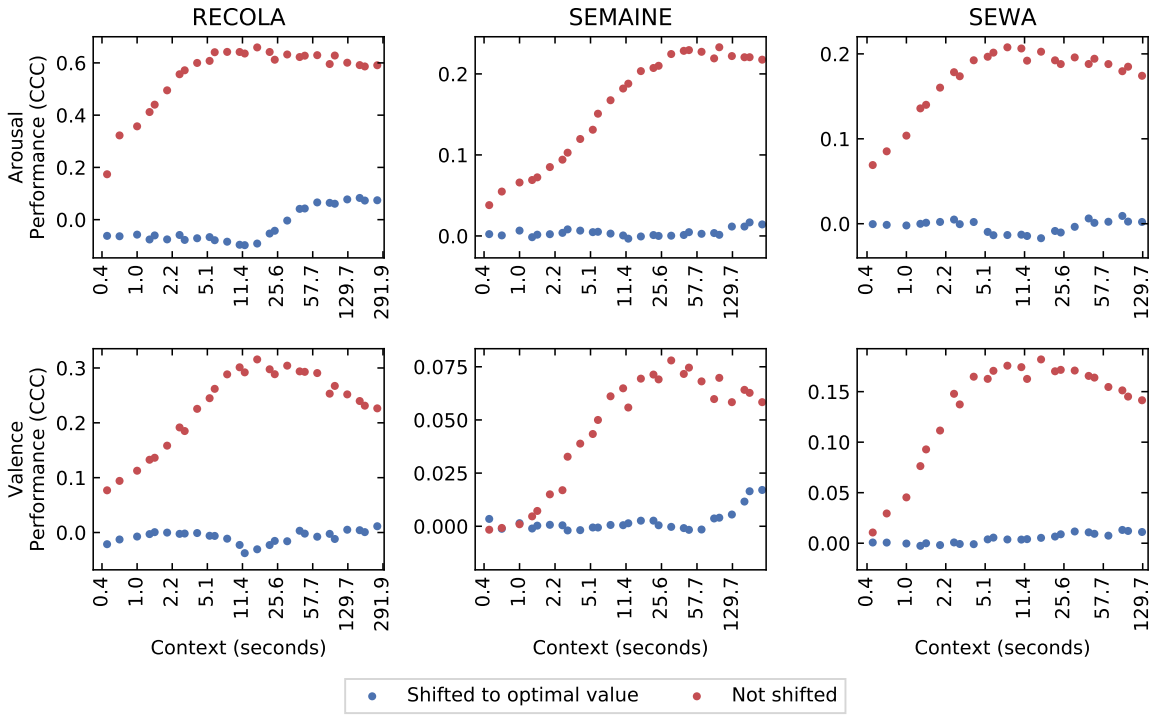


Figure 4.25: Comparison of model performance trained with gold standard data and shifted to optimal context value.

Duration of the turns, pauses and video failures do not correlate with the differing optimal context value. Another issue that may cause these differences is the shift of labels. Previously, in Section 3.2.3 we described procedure of correcting the reaction lag of annotators and aligning features and labels. However, there we considered shift of [-4,4] seconds for the annotators' ratings alignment and [-8,0] seconds for the features and labels alignment. It is possible, that the optimal value of shift lies beyond this range and by finding it we will achieve the perfect alignment and the emotion recognition could be performed without any context at all, only using the current data. Here we will check this hypothesis: first, we shift gold standard labels used previously in this chapter to the optimal values from Table 4.2; then, we will train and evaluate the same model used in Section 4.2.3 for data with the frequency of 6Hz; finally, we will compare it to the performance curves obtained previously. Results presented in Fig. 4.25 demonstrate that this approach does not work, and the gold standard extracted in the previous chapter provide reasonable evaluation of user's emotions. For all the cases showed in this figure, training models with the labels shifted to the optimal values, i.e. eliminating this gap, leads to almost zero performance. This proves an importance features-labels alignment in the contextual time-continuous emotion recognition.

Thus, the differences in the optimal context values may rely on the other aspects of data,

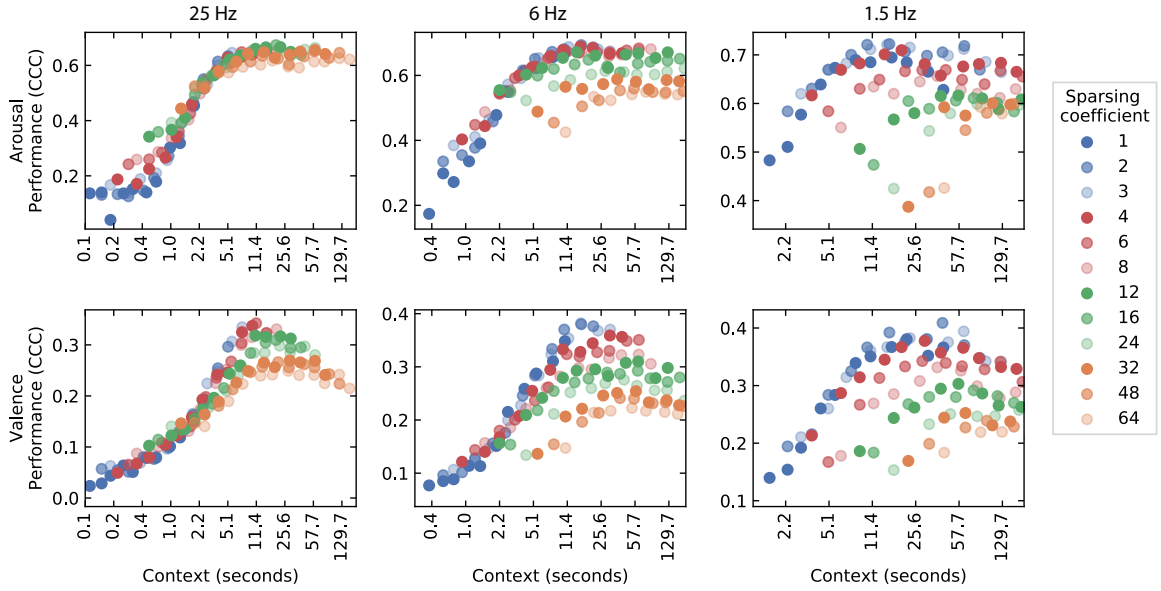


Figure 4.26: Full graph for results of context modeling for RECOLA on audio modality presented in Fig. 4.19 (audio-eGeMAPS), separated by sparsing coefficient used (depicted with color) for three frequencies: 25Hz, 6Hz and 1.5Hz.

such as language, surroundings or even annotation tools (Kessler et al., 2015), which is the direction for future research in this area.

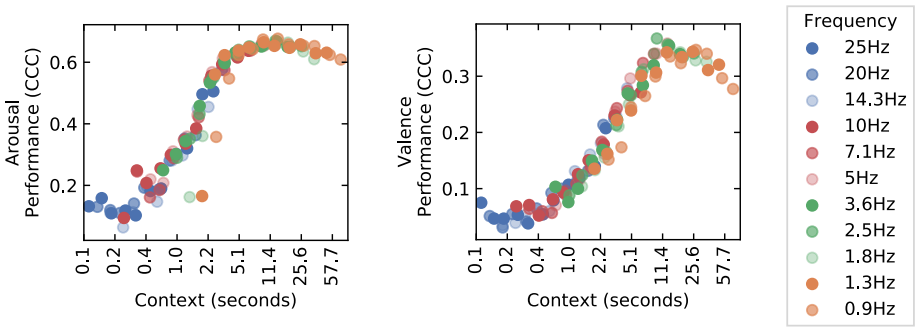


Figure 4.27: Results for context modeling on RECOLA, audio (eGeMAPS) by changing data frequency (depicted with color).

Furthermore, we want to study the effect of sparsing coefficient on performance. Previously, in Section 4.2, we used clustering to reduce the amount of data points on a performance graph to make it more readable. Here, we will consider the full graph including all the data point obtained. We will plot data points corresponding to the different sparsing coefficients with different color to discover dependencies.

In Fig. 4.26 we present the full graphs for context modeling with the data sparsing approach for the audio modality of RECOLA database and 25Hz, 6Hz and 1.5Hz data frequency. One may notice that for valence dimension, there is a significant performance loss for the frequency of 25Hz with the sparsing coefficients above eight. This difference is growing as we lower the data frequency. For example, for 6Hz we can notice a wider span between the performance obtained with the low sparsing coefficient and the high one, which is even wider for 1.5Hz. We present the remaining graphs to cover both modalities and data frequencies of 25Hz, 6Hz

and 1.5Hz in Appendix C to save place here. Analyzing these graphs, we may notice that the approach to sparse the data affects audio modality much stronger. Models trained on the visual features do not lose the performance even with such high values of the sparsing coefficients, such as 48 or 64. However, if we use low frequency of data, high sparsing coefficients quickly increase the amount of covered context, which may cause instability of performance curve, especially with low number of time steps (2, 3, 4) – for most of them at 1.5Hz we can notice much lower performance compared to the one for the number of time steps above 6. Moreover, it is not the case for the higher frequencies – the low number of time steps does not cause such instabilities, as data points are located close to each other. Nevertheless, the sparsing coefficients up to 10 provide reasonably stable results for the most of used data frequencies.

One way of designing a more robust context modeling approach is to use not the sparsing coefficient to flexibly control the amount of covered context, but the data frequency. That is, to fix S in (4.7) to 1 and change F instead. In order to provide similar context values, we should use the following frequencies: $[25, 20, 14.29, 10, 7.14, 5, 3.57, 2.5, 1.79, 1.25, 0.89]$ Hz. We present the results for this approach for RECOLA on audio modality in Fig. 4.27. One may notice the much narrower span in the performance compared to Fig. 4.26 and graphs in Appendix C, while reaching the highest values obtained with the previous approach. With the value of the sparsing coefficient fixed to one and with 64 time steps used throughout this chapter, we can model only approximately 70 seconds of context, which, however, can be corrected by a slight increase of the sparsing coefficient.

4.5 Summary

In this chapter, we presented several approaches to the speaker context modeling for the time-continuous emotion recognition systems based on the audio and the video data. We used these systems and modelled speaker context at three levels to see if there are dependencies between the amount of used context and the performance of emotion recognition system.

Thus, at the beginning, we considered the straightforward approaches to the speaker context modeling. More precisely, we used a higher number of time steps for recurrent neural models and a wider window of functional extraction for conventional time-independent models. Results on the expert knowledge based and the raw signal based feature sets demonstrated an existence of dependencies between the amount of context and the system performance. Conventional time-independent models cope with larger context worse, than recurrent models. Presumably, it is caused by averaging functionals for conventional models over a wide window and, therefore, lost feature dynamics. Raw signal based features didn't provide any improvement of the performance and showed similar patterns as expert knowledge based feature set.

Then, we introduced an approach to increase flexibility of the context modeling – the data sparsing. The main idea of it is to use not each available time step of data, but each n-th time step. We showed, how this approach may be utilized to reduce an amount of time steps, but keep the context length constant. Furthermore, we extended it by introducing varying data frequency. Results showed remaining dependencies between the amount of context and system performance, regardless of the amount of time steps, the sparsing coefficient and the data frequencies used.

Further, we studied the behavior of these dependencies in a cross-corpus setting. We used PCA-CCA approach for domain adaptation and reduction of differences in train and target corpora. Results, obtained with similar models, showed that the system performance

in a cross-corpus setting is also dependent on amount of used context and the optimal value usually lies near ones of the train and the target corpora. However, we couldn't state that it is dominated by one of them in most cases.

Finally, we concluded the chapter with some discussion about the optimal values of context for different corpora, modalities and dimensions and possible reasons for their differences. We found out that the video modality requires less data than the audio, but there are no consistent differences between arousal and valence. Analysis of length of turns and pauses, as well as failures of video capture (face recognition) did not answer the question on these differences as well. Our deeper study on sparsing coefficient showed that the large values (ten and higher) may cause lost of the performance, especially with low data frequency (6Hz and lower). A possible direction for further investigation in the topic of the speaker context is to study an effect of the language or surroundings of the users, which, however, requires much larger datasets.

5 Utilizing Contextual Information in Dyadic Interactions

In real life, at least two people take part in a conversation. Besides the information flow, there is also an emotional exchange, as humans are social beings and tend to empathize with their interlocutors. Likewise, in an argument, one person may provoke anger in the interlocutor or feel angry himself, which will most probably have an effect on the interlocutor's emotions.

As discussed earlier in Section 2, most of the current methods are using utterance-level data and design their strategies of interlocutors mutual emotional effects based on turns. In our work, we exploit the time-continuous emotional modeling and aim to study these relations on the continuous basis, without any connection to turns. The general pipeline used in this chapter is presented in Fig. 5.1.

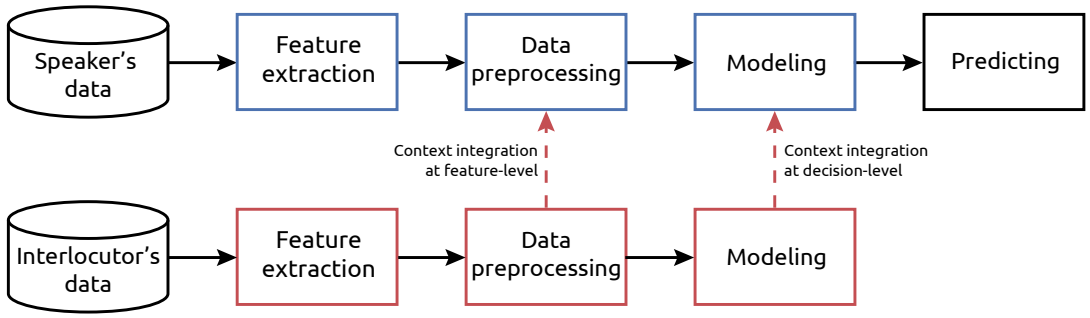


Figure 5.1: The general pipeline of the time-continuous dialogue-level contextual modeling

There are several approaches to the emotion modeling in dyadic interactions. Similarly to the other emotion recognition tasks, it can be categorical or dimensional, unimodal or multimodal. Audio and video modalities are often considered to be complementary for the emotion recognition, as usually during the conversation, subjects do not speak at the same time and listen to each other, i.e. there are periods of speech and silence. During the silence periods, the audio modality can not be representative and the natural way to deal with it is to use the facial expressions or behavioral cues (Wagner et al., 2015). However, these modalities are not always available, e.g. during a phone call. In order to not lose the information about the emotional flow of the conversation when only the audio modality is present, it is important to consider contextual data of both interlocutors over time. Similarly, for the video modality, the data may be missing if the person does not look straight into a camera or something is blocking the view, e.g. he/she has a hand in front of the face. In our study, we conduct the emotion modeling in dyadic interactions and use two approaches to fuse the data of interlocutors: decision-level based and feature-level based ones.

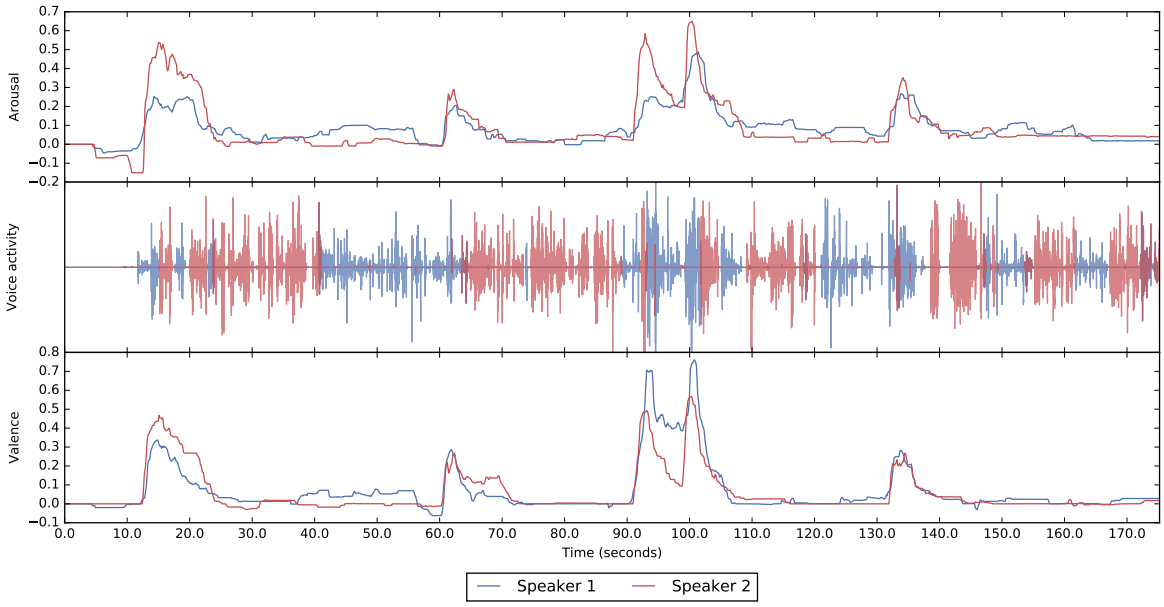


Figure 5.2: Example of empathetic changes of emotions. Blue lines represent data of speaker 1, blue lines – data of speaker 2. Recordings "Devel_DE_04" and "Devel_DE_06" of SEWA database are used.

5.1 Discovering Mutual Effects in Emotional Dynamics of Interaction

Emotional status of the speaker and the interlocutor, appearing, having an impact on each other and changing over time, creates an emotional flow of interaction. There are many different patterns that can occur in a dyadic conversation. Emotional expressions may induce not only similar reactions in the interlocutor, but in some cases different or even polar. Further, we will introduce some examples of the mutual emotional effects occurring during a dyadic interaction, using the available databases.

Changes in emotions may be caused by several reasons. Namely, if the person is speaking at the time point t – he is active and his emotions indicate his current status and support his words, if he is silent (listening) – he is passive and processes information produced by his interlocutor, reacting to it. In general, three main patterns may exist in the data: (i) interlocutor's emotions are changing similarly to those of the speaker; (ii) interlocutor's emotions are changing contrary to those of the speaker; (iii) interlocutor's emotions are indifferent to those of the speaker. Sometimes data does not fall within any of those patterns, but it is rarely the case for the corpora analyzed in this work.

If emotions of interlocutor change similarly to the ones of speaker, it may refer to empathy. An example of it is presented in Fig. 5.2. Empathy is a dominant pattern in interactions between two persons that get along well with each other.

The second pattern, when emotions of the interlocutor change contrary to the speaker's, usually is not persistent, and emerges for a rather short period of time. The situations that may cause such a behavior include positively-colored, e.g. when the speaker is angry (negative valence) at some third-party person and his interlocutor is supporting him in his anger, agreeing with him and cheering him up (positive valence), as well as negatively-colored, e.g. when the speaker is angry (negative valence) at the interlocutor, but this makes the interlocutor laugh (positive valence) and mock the speaker. In the first situation both speakers are on the

same side, in the second one – they are opposed to each other. An example of the contrary emotion changes is presented in Fig. 5.3. In this particular case, the speakers were discussing a product and both found it not good. The speaker 2 (blue line) asked the speaker 1 (red line) in jest (positive valence), if he wants to buy this product, to which the speaker 1 made a facial expression of disgust (negative valence) and answered "rather not".

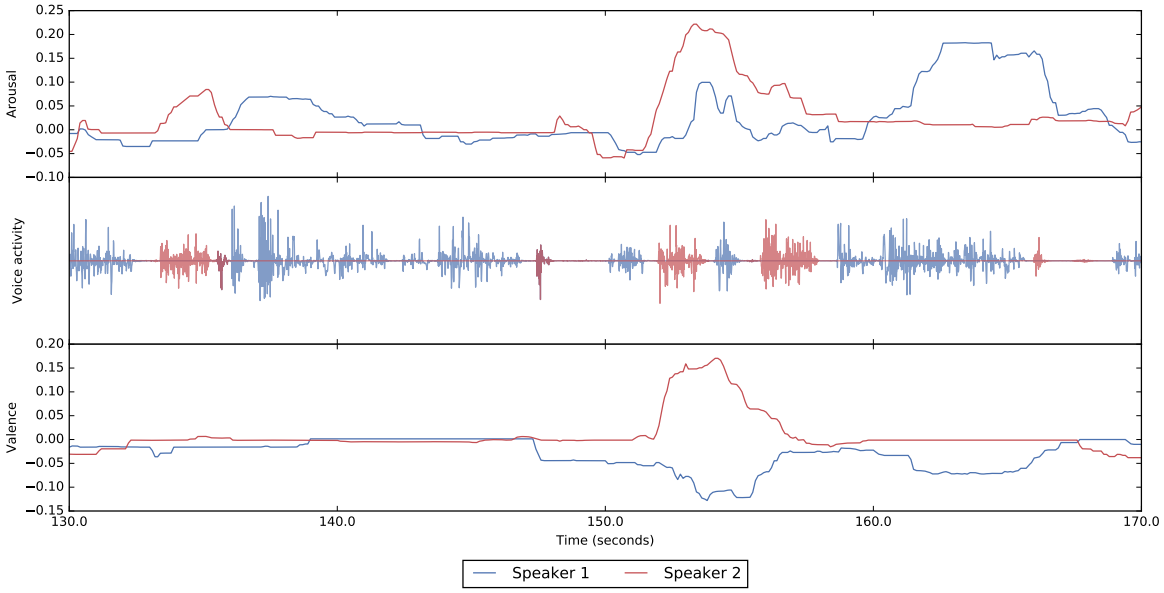


Figure 5.3: Example of contrary changes of emotions (see valence for range of 150-160 seconds). Blue lines represent data of speaker 1, red lines – data of speaker 2. Beforehand (130-150 seconds) and afterwards (160-170 seconds) valence rating of both speakers are almost identical. Between 150 and 160 seconds, valence ratings are diverging to opposite directions. Recordings "Train_DE_05" and "Train_DE_20" of SEWA database are used.

Finally, the third pattern may occur when interacting with a phlegmatic person that does not show emotions much. An example is presented in Fig. 5.4.

In this chapter, we use the databases described in Section 3.1 those are suitable for dyadic emotion recognition, i.e. have the data and labels for both speakers of recordings. RECOLA database was collected in a dialogue scenario, but only for a few recordings contains the data of both speakers. Therefore, we do not use it in this chapter. For SEMAINE database data of both speakers is always available, but for most of the recordings, only the "user" role is annotated and the "operator" role is not. Therefore, we use it here for testing only, but not when working with labels exclusively.

Emotion corpora with dyadic interactions and dimensional annotations are still very rare, and not many of them are available for the research community. Two corpora used in this thesis – IEMOCAP and UUDB – are annotated not time-continuously, in contrast to the other three. However, they consist of continuous dyadic interactions, i.e. turns have consistent sequential structure and follow each other without breaks or long pauses. In the most parts of these recordings there is at least one person speaking – the speaker or the interlocutor, and each of these turns is annotated.

To expand the data collections and therefore make them suitable for our research, we make the following assumption: independent annotations of consequent turns of one speaker reflect continuous changes in his emotional status to a certain degree. This allows us to use

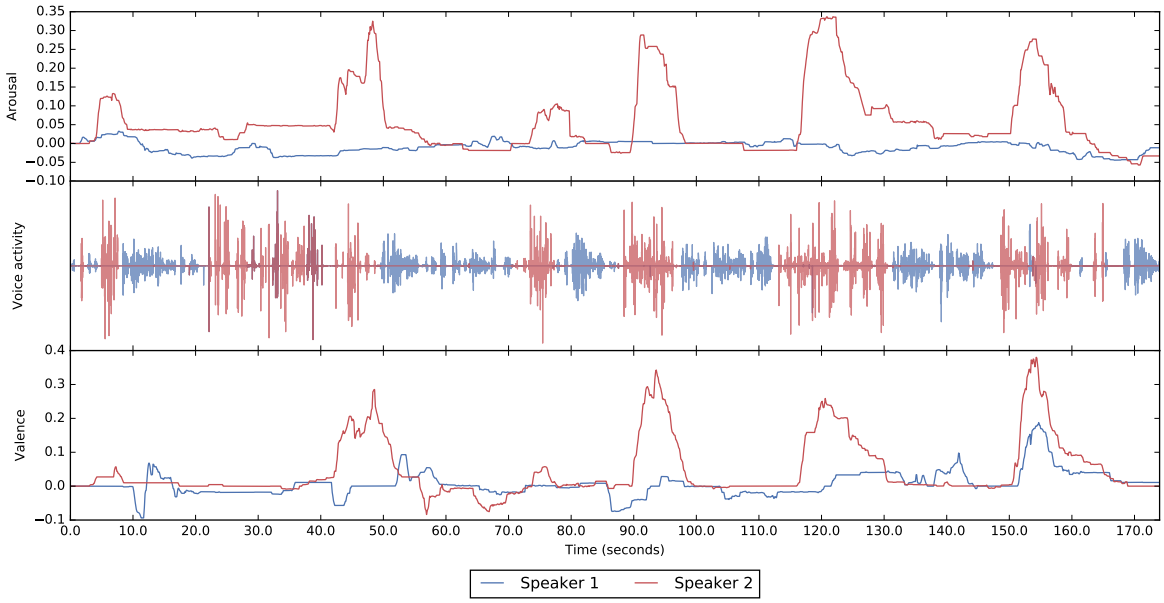


Figure 5.4: Example of interlocutor indifferent to emotions of speaker. Blue lines represent data of speaker 1, red lines – data of speaker 2. An example of empathetic changes is presented between 150 and 160 seconds. Apart from that, speaker 1 does not show strong emotions, regardless of his interlocutor. Recordings "Devel_DE_09" and "Devel_DE_14" of SEWA database are used.

original turn-level annotations and approximate their continuous representation. As we don't have any information about the speaker's emotion state between turns, we set it to neutral values. Taking the average turn duration into account (see Fig. 3.7a and Fig. 3.9a), this leads to frequent and intense fluctuations in the annotation curve. To reduce the impact of this issue, we smooth the obtained time series with a Hamming window, corresponding to one second of data. The new form of an annotation curve is used in this thesis for IEMOCAP and UUDB as with continuous databases.

While this approach allows a fairly reasonable transition from turn-level to continuous annotations, we should note that in contrast to proper continuous ratings, this representation yields a much stronger correlation with audio features (and the audio signal itself), as audio files are cleaned using similar concept and identical time frames for turns (see Section 3.2.1).

A correlation analysis of the labels of the used corpora provides an additional proof of the frequent occurrence of the first pattern (empathetic emotions). We use Pearson's correlation coefficient and compare labels of speaker 1 to labels of speaker 2. The results are presented in Table 5.1. We compare labels of the same dimension, e.g. arousal to arousal, as well as of the different dimensions, e.g. arousal to valence.

One may see that for SEWA database there is rather strong correlation between the labels of speaker 1 and 2 for both scenarios: same dimensions and different ones. UUDB and IEMOCAP both demonstrate weak correlation, as the labels for them are mapped from the utterance level annotation, which often do not intersect, as speakers do not interrupt each other much.

The results of correlation analysis show that the data of the interlocutor may be gainfully used to increase the performance of emotion recognition for the speaker. It may be introduced to the model by fusing it with the data of the speaker on two levels: feature-level and decision-level. Feature-level (early) fusion (FLF) implies concatenation of the feature matrices before

D1-D2	SEWA	UUDD	IEMOCAP
A-A	0.429	-0.137	0.059
V-V	0.542	0.012	0.258
A-V	0.490	-0.067	0.119
V-A	0.523	-0.087	0.122

Table 5.1: Pearson's correlation between different label dimensions of speakers in dyadic interaction. D1 – emotional dimension of speaker 1, D2 – emotional dimension of speaker 2, A – arousal, V – valence.

the model training, i.e. the model may use the data of both speakers from the beginning to adjust its parameters for solving the emotion recognition task in a dyadic interaction scenario. Decision-level (late) fusion (DLF) implies the separate emotion predictions for two speakers, followed by a simple meta-model that takes these predictions as inputs and provides the final result. As a meta-model for decision-level fusion, one may use simple classifier/regressor or linear combination of predictions. However, to tune the parameters of the meta-model, an additional partition of the data is required (development subset), which may not be previously used for training. Feature-level fusion techniques do not introduce this requirement, but models have to be trained on larger feature vectors.

In the following sections, we will use these methods to model the dyadic context with the two approaches: dependent and independent. By the *dependent* dyadic context modeling, we will understand the case when the amount of context modelled for the speaker is exactly the same as for his interlocutor. By the *independent* dyadic context modeling – when the amount of context modelled for the speaker may be smaller than, equal to or larger than for his interlocutor. Note that throughout this chapter we exclusively predict emotions of the speaker and use the data of the interlocutor only as an additional helpful source of information. Nevertheless, for most of the used databases (except for SEMAINE, where no emotional ratings for the operators are available), each participant plays two roles: he is a "speaker" for one recording and an "interlocutor" for another.

To compare the methods and be able to judge if they provide improvements, we set a baseline for non-dyadic modeling with ten independent runs using the fixed data partitions. We measure performance in terms of CCC and use model architecture and hyperparameters similar to Chapter 4. As databases used in our work have the different optimal context length, we select the one that provides the good performance for databases with both early and late performance peaks – 24 seconds. We model it with the data frequency of 6Hz, the time window size of 36 frames and the sparsing coefficient of 4. We will use these baseline performances throughout this chapter to study an effect of the dialogue context modeling compared to the single speaker modeling. They are presented in Table 5.2.

Modality	Dimension	SEMAINE	SEWA	IEMOCAP	UUDB
Audio	Arousal	0.181	0.222	0.273	0.665
	Valence	0.070	0.197	0.227	0.428
Video	Arousal	0.207	0.398	0.096	-
	Valence	0.172	0.392	0.098	-

Table 5.2: Baseline performance of non-dyadic modeling (in CCC)

5.2 Dependent Dyadic Context Modeling

The general concept of the dependent context modeling is presented in Fig. 5.5. In case of FLF we concatenate feature vectors at the very beginning of the pipeline and perform the context modeling thereafter. This way we feed feature vectors of double size to our model, compared to non-dyadic modeling (e.g. models from Chapter 4). In case of DLF, we use the same preprocessing steps and models as in Chapter 4, but we split the original train subset into two: new train and development. The development subset is used further in the pipeline to train the parameters of the meta-model. Therefore, we don't affect the test subset to keep the results of different fusion approaches comparable to one another and to the baseline.

In subsequent subsections, we will present the results for FLF and DLF approaches and compare them to the baseline values presented in Table 5.2.

5.2.1 Feature-level fusion

By using the dependent approach with FLF, we are dealing with similar data as in Chapter 4, but of doubled size due to the merged interlocutor's data. In this subsection we use concatenated train and development subsets to train an RNN-LSTM based model with two layers of 20 and 10 neurons respectively, the dropout layers with $p = 0.3$, the RMSprop optimizer and the CCC-based loss function, similar to the previous chapter. In spite of the increased size of the input vector for each sample, we do not increase the model size in order to ensure plausible comparison between the models based on the speaker-only data and ones with the data from both the speaker and the interlocutor.

With respect to the results obtained and conclusions derived in the previous chapter, as well as to simplify the process of the flexible context modeling, we use the data frequency

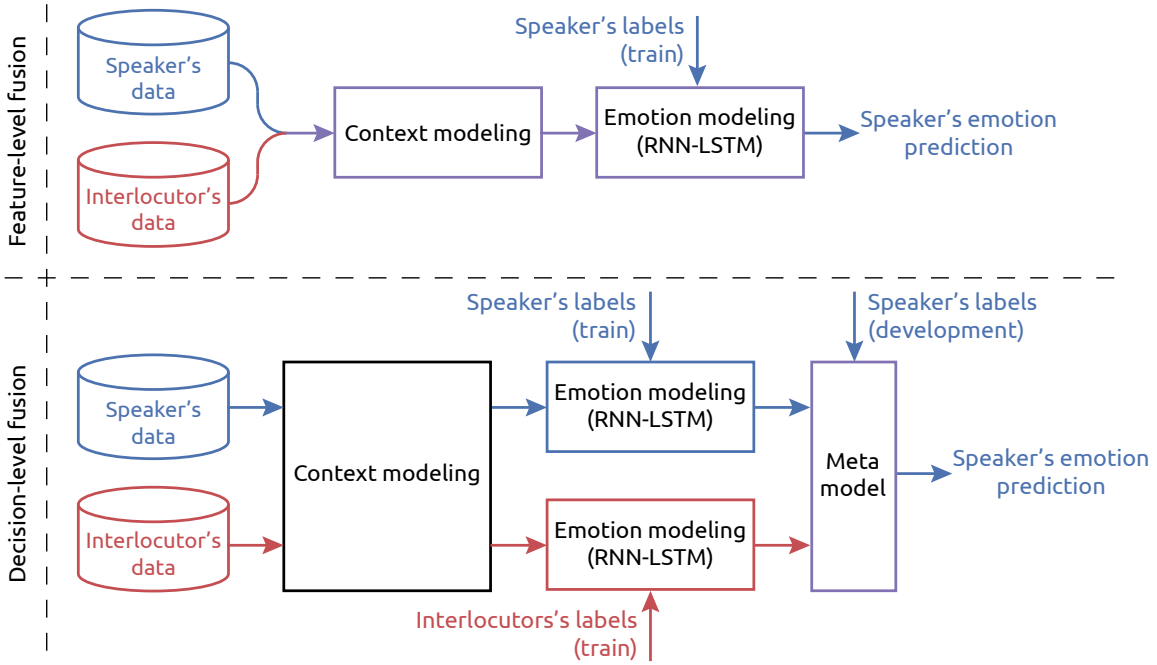


Figure 5.5: Pipeline of dependent dyadic context modeling. Blue lines represent data flow of speaker's features and labels, red lines – data flow of interlocutor's features and labels.

of 10Hz and the time window size of 10 frames. Similar to the previous chapter, we use the sparsing coefficients of {1, 2, 3, 4, 6, 8, 12, 16, 24, 32, 48, 64}. Therefore, according to (4.7) the amount of modelled context is equal to {1.05, 1.95, 2.85, 3.75, 5.55, 7.35, 10.95, 14.55, 21.75, 28.95, 43.35, 57.75} seconds respectively. We round these values to the closest integer to facilitate better visualization. The results for audio modality are presented in Fig. 5.6 and for video – in Fig. 5.7.

One may notice similar trends to ones presented in the previous chapter for SEMAINE and SEWA corpora. Not only the trends are similar, but so is the optimal context length and the performance curve behavior with some minor exceptions (e.g. a slight instability in the video-arousal pair for SEWA database). For the two corpora, that were not used previously in this thesis, there are also similar patterns to notice. For IEMOCAP there is an optimal context length of three to four seconds for each modality-dimension pair, except for audio-arousal (it is approximately 11 seconds in this case). For UUDB only the audio data is available, and the optimal context length lays between three and six seconds.

These results support the theory of existing dependencies between the recognition performance and the amount of information used by the system, which also works in the dyadic data setup. Concerning the results compared to the speaker-only based models, one may notice that with correctly selected context data, one may achieve the higher performance by utilizing interlocutor's data even with a simple FLF strategy. This approach provided the performance gain for five out of eight cases for audio modality. However, it is different for video modality – only one solid performance gain out of six cases. A reason for this may be in the more active effect of the audio modality in terms of influence on the interlocutor's emotions, as data was recorded during a spoken conversation. Therefore, the video data serves more as an indicator of one's emotions, than as a source of influence to the interlocutor. However, it does not cancel an effect of the video modality in dyadic interactions completely, as such signs as smiles may induce emphatic reactions of the interlocutor and their annotations.

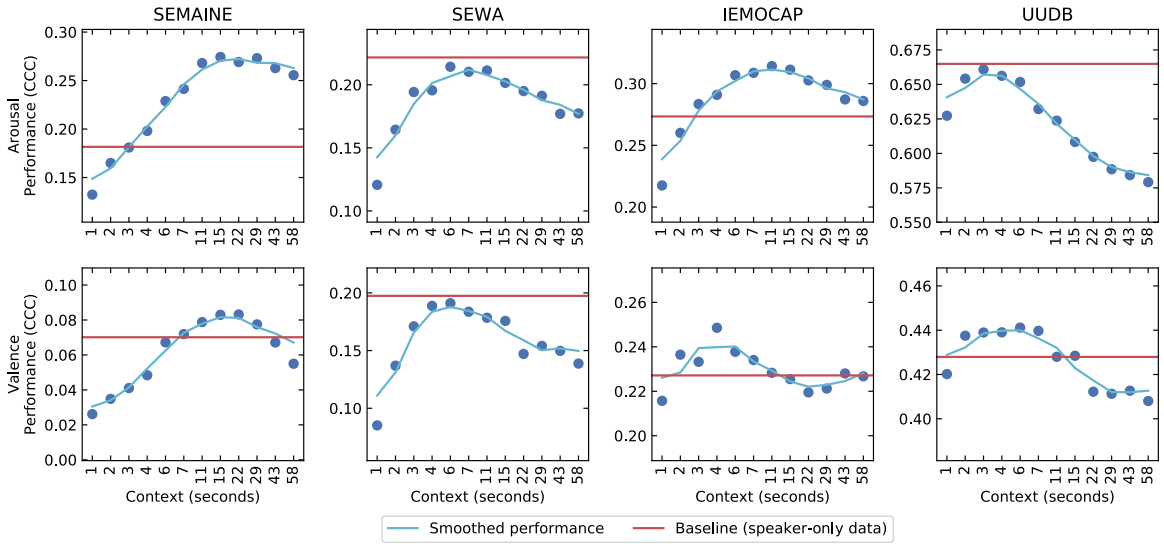


Figure 5.6: Results for dependent dyadic context modeling with FLF on audio modality (eGeMAPS). Arousal (top row) and valence (bottom row) dimensions for four corpora: SEMAINE, SEWA, IEMOCAP and UUDB. The performance is measured in terms of CCC and is dependent on amount of context represented by sparsing coefficient used in RNN

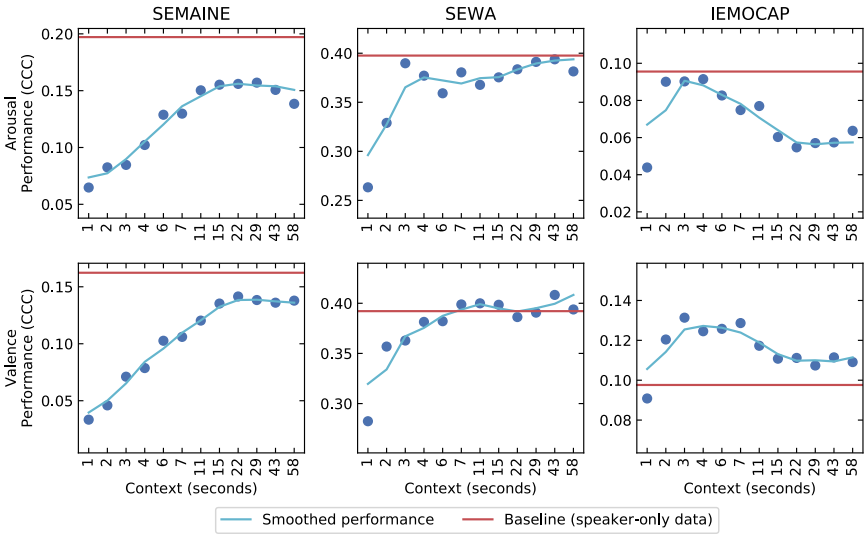


Figure 5.7: Results for dependent dyadic context modeling with FLF on video modality (AUs). Arousal (top row) and valence (bottom row) dimensions for three corpora: SEMAINE, SEWA and IEMOCAP. The performance is measured in terms of CCC and is dependent on amount of context represented by sparsing coefficient used in RNN

In some cases, for example for SEMAINE database with the audio-arousal pair, the performance gain is remarkable. Note that for this database there is no annotated data for the interlocutor available; therefore, in contrast to other databases, features extracted for him/her always appear at the same positions of the merged feature vector.

5.2.2 Decision-level fusion

In the DLF setup, we train the models as usual, on the speaker-only data and combine the speaker's and interlocutor's data later, after predictions for both of them are obtained. The resulting vector is as small as two features for single task learning (predicting one label at a time, i.e. arousal **or** valence) or four features for multi-task learning (predicting both labels simultaneously, i.e. arousal **and** valence). These features represent predictions for the speaker and the interlocutor corresponding to a particular frame or time step. One can work with them directly or extract functionals to have a better statistical description of data. In our work, to test the pure DLF setup, we use them directly and only apply smoothing over one second of predictions with a Hamming window in order to eliminate unnecessary frequent fluctuations. Solving the single-task learning problem, vectors of two features are fed into the meta-model to provide the final prediction for the speaker's emotions.

Models and parameters for the first level of prediction – separate for the speaker and the interlocutor – are the same as for the previous subsection and for the baseline. As a meta-model, we use another RNN-LSTM to be able to catch contextual dependencies in the data. As it has to deal with only two features, we select its architecture to be one layer with five neurons. As mentioned previously, we split the original train set into two subsets – new train and development; we use the former to train the single-speaker model and the latter to train the meta-model. We use a proportion close to 0.7 to 0.3 for new train/development split and keep recordings of a particular speaker and his/her interlocutor not separated (in the

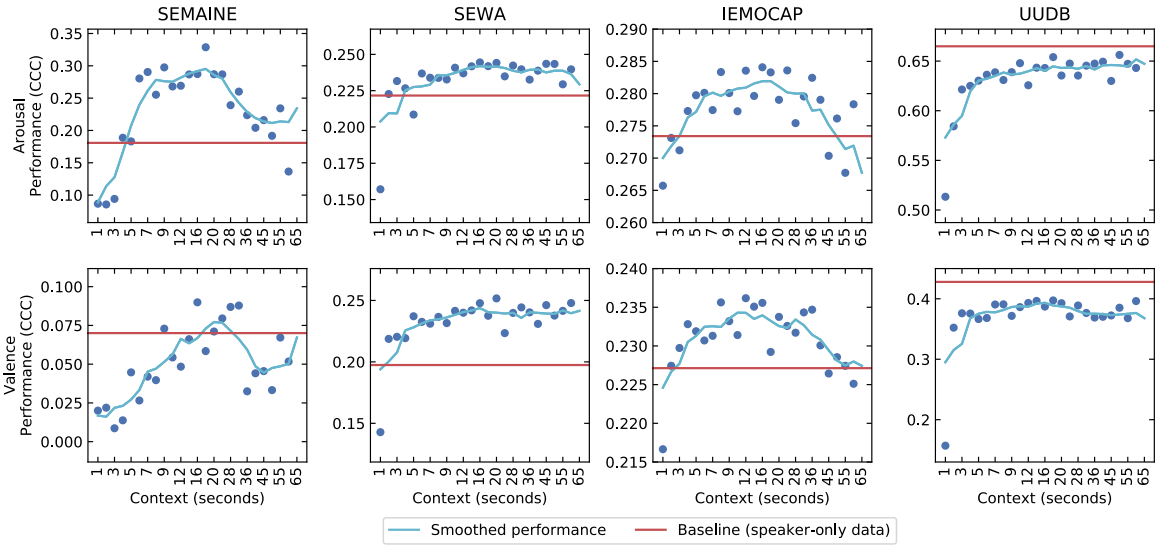


Figure 5.8: Results for dependent dyadic context modeling with DLF on audio modality (eGeMAPS). Arousal (top row) and valence (bottom row) dimensions for four corpora: SEMAINE, SEWA, IEMOCAP and UUDB. The performance is measured in terms of CCC and is dependent on amount of context represented by window size used in meta-modal (RNN)

same subset).

To study contextual dependencies in data with the DLF setup, we use the varying window size for the meta-model, corresponding to the following amount of seconds: {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, 20, 24, 28, 32, 36, 40, 45, 50, 55, 60}. This set is chosen empirically to represent smooth increase of the context in the range from one second to one minute of the data. Based on the conclusions derived in the previous chapter, we find values below one second and above one minute not reasonable. Single-speaker models trained with the same parameters as the baseline for this chapter. Results for the audio modality are presented in Fig. 5.8 and for the video – in Fig. 5.9.

Analyzing the results, one may notice some similarities between them. Starting with one second of the data – the shortest option, that corresponds to almost one-to-one (in contrast to sequence-to-sequence) analysis – models do not provide the best performance and in many cases it is much inferior to e.g. a context at the middle of a range. Models on SEMAINE database reach the best performance with context of approximately 20 seconds for each modality-dimension pair, except for the video-arousal (ten seconds).

However, with wider context window, the performance drops significantly. Similar to single-speaker scenario, models on SEWA database reach the high performance relatively fast and keep it with wider windows. For video-valence it is rather unstable, but at the same level throughout the experiments with the different window sizes. Models on IEMOCAP database also reach the performance plateau fast for the video modality, but have middle context range oriented curve for the audio modality (16-20 seconds). Models on UUDB demonstrate similar behavior to ones on SEWA.

Comparing results of DLF to the baselines defined earlier, we can notice the performance gain for three out of four databases (SEMAINE, SEWA and IEMOCAP) on the audio modality. However, models trained with dyadic data of UUDB do not show any improvements. Similar to results obtained in Section 5.2.1, models trained on the video modality seems to profit from dyadic data much less, than the ones trained on the audio modality. We can see improvements

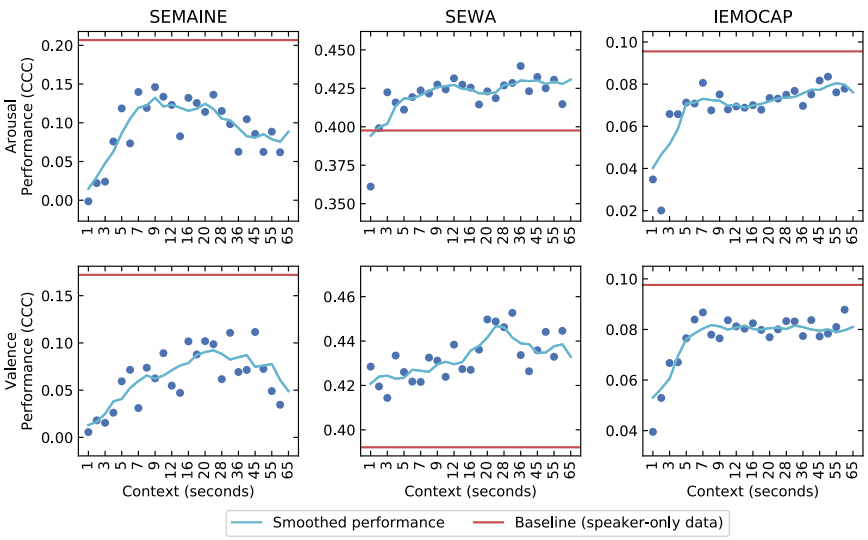


Figure 5.9: Results for dependent dyadic context modeling with DLF on video modality (AUs). Arousal (top row) and valence (bottom row) dimensions for three corpora: SEMAINE, SEWA and IEMOCAP. The performance is measured in terms of CCC and is dependent on amount of context represented by window size used in meta-modal (RNN)

only for SEWA database; nevertheless, they are notable.

Further in this chapter, we will consider the independent dyadic context modeling. We could implement this approach with DLF strategy, but in this case the independence would have been achieved at the first stage of modeling (speaker-only data). This influences not the final predictions directly, but only the intermediate results. As the amount of context was chosen as the optimal on a basis of conclusions derived earlier, a different value would cause a decrease in performance of intermediate predictions and accumulate in the error of the meta-model. Therefore, we will consider an independent approach applied only to the FLF strategy.

5.3 Independent Dyadic Context Modeling

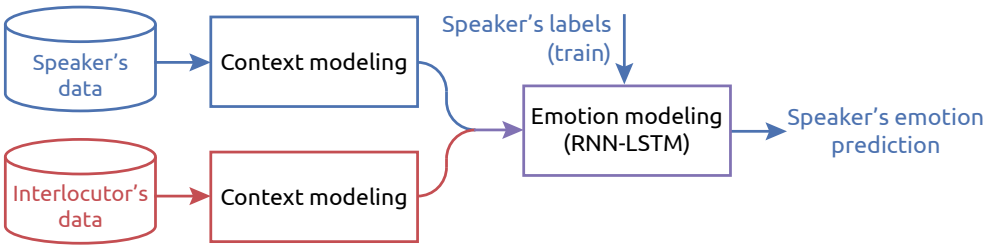


Figure 5.10: Pipeline of independent dyadic context modeling. Blue lines represent data flow of speaker's features and labels, red lines – data flow of interlocutor's features and labels.

The general concept of the independent context modeling is presented in Fig. 5.10. Comparing it to the pipeline for the dependent context modeling (cf. Fig. 5.5), one may notice

that for FLF, feature concatenation is shifted one step further and now is directly after the context modeling. This is usually not possible with time-dependent models, as it will raise an issue of feature vectors' shape mismatch in the time step dimension. However, using the data sparsing approach presented in Section 4.2, we can fix the amount of time steps and still change the amount of context by varying the sparsing coefficient or even the data frequency as in Section 4.2.3. This approach increases our flexibility in the context modeling, allowing to use FLF for emotion recognition in dyadic interactions with independently varying context. This concept is presented in more details with an example of audio data in Fig. 5.11.

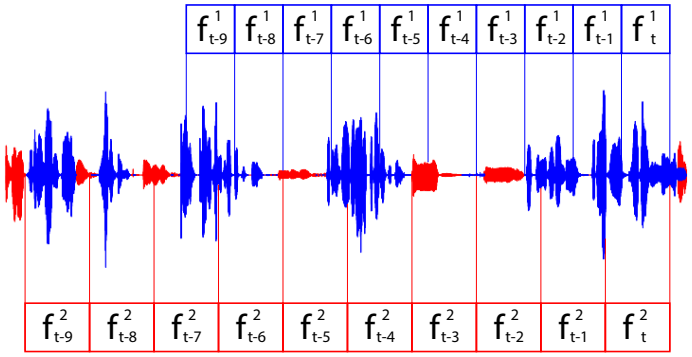


Figure 5.11: Independent context modeling with FLF for speech signal of speaker (blue) and interlocutor (red).

In our work we use an approach similar to the one presented in Section 5.2.1, but with context of speaker fixed to the optimal value, obtained with dependent FLF, and context of interlocutor changing by varying the sparsing coefficient in the same range of $[1, 64]$ as previously. This allows us to use the different observation windows for the speaker and the interlocutor and consider more context with fewer details or less context with more details. As feature positions are strictly defined and do not change during the experiment, the model has a possibility to catch and process the dependencies from both data sources effectively, profiting from various context windows.

The results for such an approach are presented in Fig. 5.12 for the audio modality and in Fig. 5.13 for the video modality. In addition to the system performance and the baseline themselves, we mark (green dashed line) the optimal value of context based on Fig. 5.6 and Fig. 5.7 respectively.

One may notice that for the audio modality, the amount of context of the interlocutor which provides the highest performance in this setting is often close or equal to the amount of context used for data of the speaker. It is, however, not the case for SEMAINE database on arousal dimension – here the performance is rather stable regardless of the interlocutor's context and, nevertheless, notably higher than the baseline. Application of this approach to SEWA database does not lead to any improvements, but shows the aforementioned pattern. The results on IEMOCAP show the performance gain and the same pattern, while for UUDB we may see a slight increase in the performance for valence dimension and none for arousal.

For the video modality, it is much more unclear. There is no strict and obvious pattern persistent at each presented graph. The optimal context value provides the worst performance for SEMAINE on arousal, and unremarkable compared to the other values on other databases and dimensions. The only exception is IEMOCAP on valence dimension, which, however, may not be considered as a rule.

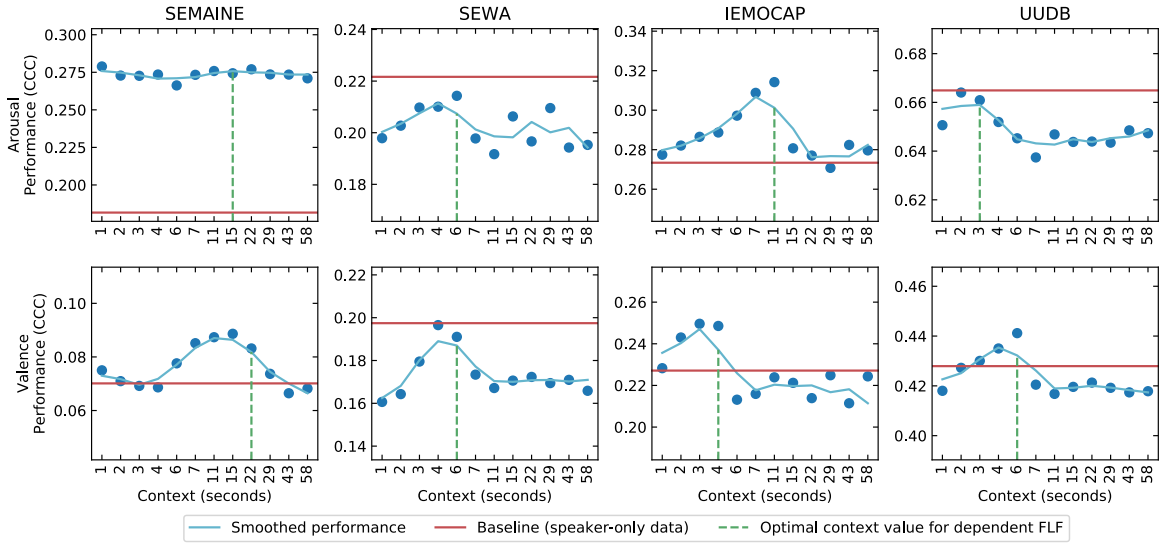


Figure 5.12: Results for independent dyadic context modeling (fixed for speaker) with FLF on audio modality (eGeMAPS). Arousal (top row) and valence (bottom row) dimensions for four corpora: SEMAINE, SEWA, IEMOCAP and UUDB. The performance is measured in terms of CCC and is dependent on amount of context of interlocutor represented by sparsing coefficient used in RNN

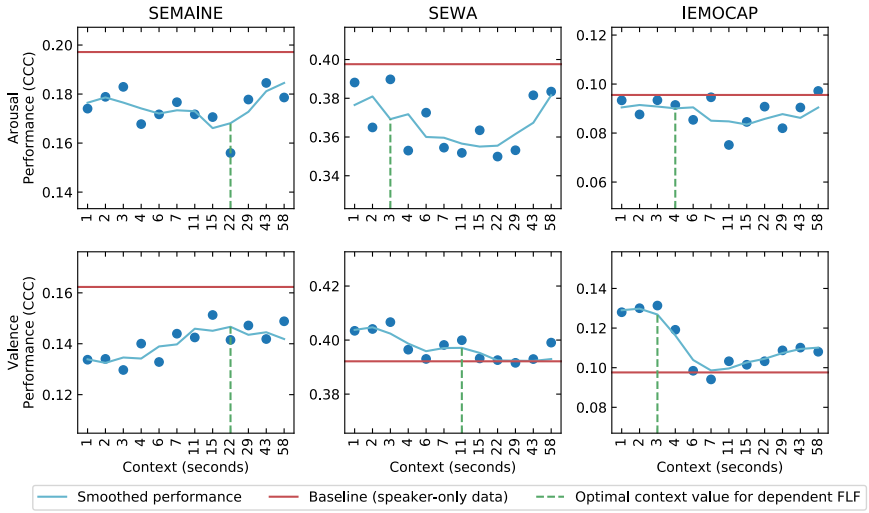


Figure 5.13: Results for independent dyadic context modeling (fixed for speaker) with FLF on video modality (AUs). Arousal (top row) and valence (bottom row) dimensions for three corpora: SEMAINE, SEWA and IEMOCAP. The performance is measured in terms of CCC and is dependent on amount of context of interlocutor represented by sparsing coefficient used in RNN

In order to increase the depth of the interlocutor's context effect analysis, we further extend our experiments and do not keep the speaker's context at its optimal value according to our previous experiments, but variate it independently of the interlocutor's data, covering all possible combinations with the same options in a range of $[1, 64]$, as used previously in this chapter.

To present these results, we use heat maps similar to the ones used in Section 4.2, but more compact. The color of each graph represents the performance, with dark blue being the lowest value and yellow – the highest. Note that color range is based on the minimal and the maximal performance and is independent for each graph. Results are presented in Fig. 5.14 for the audio modality and in Fig. 5.15 for the video modality.

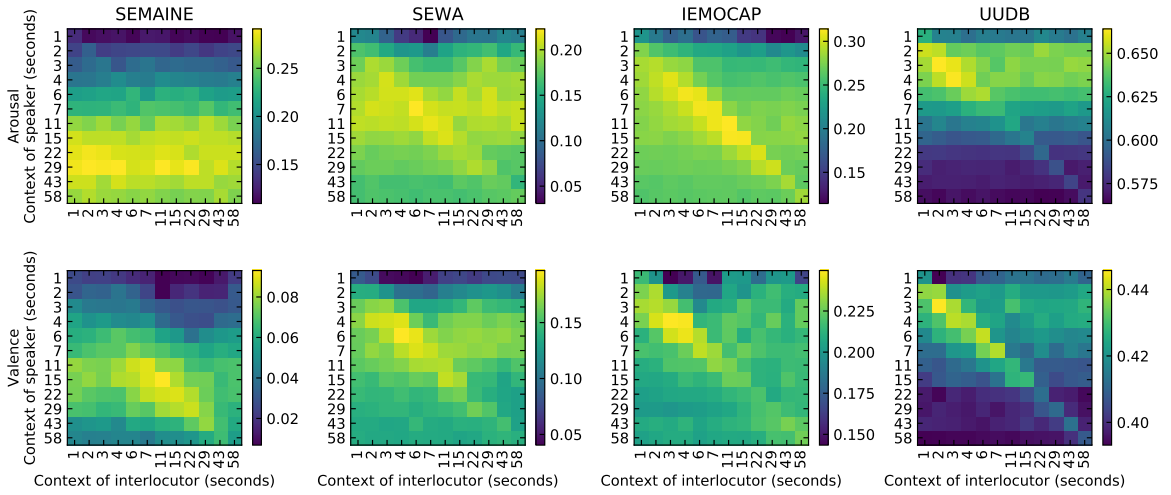


Figure 5.14: Results for independent dyadic context modeling with FLF on audio modality (eGeMAPS). Arousal (top row) and valence (bottom row) dimensions for four corpora: SEMAINE, SEWA, IEMOCAP and UUDB. The performance is measured in terms of CCC and is dependent on amount of context of interlocutor represented by sparsing coefficient used in RNN

This type of results representation allows us to see the big picture of the context effect on the performance. One may notice several patterns in these graphs. Firstly, similar to results of Chapter 4, the performance is highly dependent on the amount of context used in the model. It is seen by gradient-like changing color along y-axis for many of the presented graphs (e.g. SEMAINE audio-arousal or UUDB both dimensions). Secondly, for the audio modality in 7 out of 8 cases, the equal context length for the speaker and the interlocutor provide the highest results and in spite of the fact that it is changing with the sparsing coefficient, regardless of the coefficient value used, is higher than other options for the interlocutor's context. It is seen by clear diagonal line for SEWA, IEMOCAP and UUDB and its shortened version for SEMAINE on valence. Finally, similarly to the results with the fixed amount of speaker's context, patterns are more explicit for the audio modality, than for the video.

For the video modality, there is a similar pattern for IEMOCAP valence, but other graphs show either opposite (e.g. SEMAINE, SEWA and IEMOCAP on arousal) or no patterns at all regarding mutual effect of the speaker's and the interlocutor's context.

We could explain the diagonal pattern for IEMOCAP and UUDB on the audio modality through the original nature of the data presented in these databases – it is utterance-based and annotations between utterances are equal to zero, therefore, with equal context values for the speaker and the interlocutor, data is fused in the "correct" manner and there are no artificial

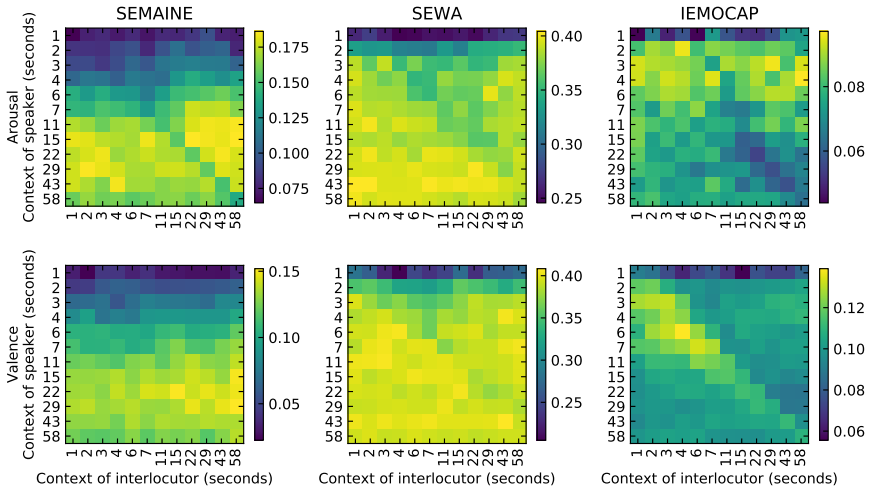


Figure 5.15: Results for independent dyadic context modeling with FLF on video modality (AUs). Arousal (top row) and valence (bottom row) dimensions for three corpora: SEMAINE, SEWA and IEMOCAP. The performance is measured in terms of CCC and is dependent on amount of context of interlocutor represented by sparsing coefficient used in RNN

interruptions caused by disparity of the sparsing coefficients and shrinkage or enlargement of the contextual window. However, we can see the same pattern for databases that are originally annotated continuously (SEMAINE and SEWA) and do not suffer from this issue.

5.4 Analysis and Discussion

The results presented in this chapter cover various strategies of integrating interlocutor's data in speaker emotion recognition models in a pure time-continuous fashion, i.e. without usage of any interaction units such as turns or utterances. In many cases, this data provided the performance gain compared to the single-speaker based model, which is however different for used approaches. We will further present the table, showing the best results achieved with each strategy in comparison to the baseline presented at Table 5.2. BL represents the baseline performance from Table 5.2 and obtained with the speaker-only (non-dyadic) data, FLF(d) stands for the dependent context modeling with the feature-level fusion described in Section 5.2.1, DLF(d) – for the dependent context modeling with the decision-level fusion described in Section 5.2.2, FLF(i-f) means the independent context modeling with the feature-level fusion and the fixed context of the speaker described in the first part of Section 5.3, finally, FLF(i) stands for the fully independent context modeling with the feature-level fusion described in the second part of Section 5.3. Numbers in **bold** show the performance gain over the baseline, underlined numbers show the significant differences compared to the baseline in terms of the paired sample t-test.

Wrapping the results up, we achieved the performance gain with applied approaches in 33 out of 56 cases, where one case is one of the four presented approaches applied to certain modality-dimension pair of the particular database. For significance check, we use paired sample t-test and consider differences compared to the speaker-only baseline significant if $p < 0.01$. Our approach resulted in 12 statistically significant cases of positive performance gain. If we consider them modality-wise, most of them (22 cases, 9 significant) were obtained with audio features; dimension-wise – 19 cases (5 significant) for valence and 14

	Dimension	Approach	SEMAINE	SEWA	IEMOCAP	UUDB
Audio	Arousal	BL	0.182	0.222	0.273	0.665
		FLF(d)	**0.274**	0.214	**0.314**	0.661
		DLF(d)	**0.329**	**0.244**	**0.284**	0.656
		FLF(i-f)	**0.279**	0.214	**0.314**	0.664
		FLF (i)	**0.291**	**0.223**	**0.314**	0.664
	Valence	BL	0.070	0.197	0.227	0.428
		FLF(d)	**0.083**	0.191	**0.249**	**0.441**
		DLF(d)	**0.090**	**0.252**	**0.236**	0.397
		FLF(i-f)	**0.089**	0.197	**0.250**	**0.441**
		FLF (i)	**0.093**	0.197	**0.250**	**0.446**
Video	Arousal	BL	0.207	0.398	0.096	-
		FLF(d)	0.157	0.394	0.091	-
		DLF(d)	0.146	**0.439**	0.084	-
		FLF(i-f)	0.183	0.390	**0.097**	-
		FLF (i)	0.185	**0.404**	**0.097**	-
	Valence	BL	0.172	0.392	0.098	-
		FLF(d)	0.146	**0.408**	**0.131**	-
		DLF(d)	0.112	**0.453**	0.088	-
		FLF(i-f)	0.149	**0.407**	**0.131**	-
		FLF (i)	0.150	**0.408**	**0.139**	-

Table 5.3: Performance overview of applied approaches to the continuous dyadic modeling compared to the baseline from Table 5.2. BL – baseline, FLF(d) – dependent context modeling with feature-level fusion, DLF(d) – dependent context modeling with decision-level fusion, FLF(i-f) – independent context modeling with feature-level fusion and fixed context of speaker, FLF(i) – fully independent context modeling with feature-level fusion. Numbers in bold show performance gain over baseline.

(7 significant) for arousal. The highest percentage of improvement achieved on IEMOCAP database – in more than 80% of the cases applying dyadic context modeling provided an improvement over the baseline, on continuously annotated SEWA and SEMAINE it was in approximately 50% of the cases and the worst results are for UUDB with 37.5% of the cases. Considering approaches used, the highest number of improvements (10, 3 significant) was obtained with the fully independent context modeling with FLF, 8 (3 significant) – with the independent context modeling with FLF and the fixed context for of speaker, 8 (4 significant) – with the dependent context modeling with DLF and, and finally 7 (2 significant) with the dependent context modeling with FLF. There are no significant cases for performance decrease, therefore, utilization of the proposed approach either increases the quality of the emotion recognition system or does not affect it negatively.

Performance level for used approaches is highly dependent on the database and the modality-dimension combination in the first place and on the train/test partitions in the second place. Modality-dimension combination sets an approximate level that is possible to achieve while using this data. Each of the presented databases consists of 2-10 hours of data, which is a low duration compared to the corpora for the other machine learning tasks, such as speech recognition. Moreover, due to the subjective nature of emotions, one would require even more data to train models robustly. These factors lead to a situation where a

certain recording can change the system performance drastically if selected for a train or a test subset. An analogous situation, where two presumably similar subsets provide completely different performance results, is observable for the emotion recognition challenges covered in Section 2.3. According to the baselines provided by organizers themselves or to results of participants on the development and test sets, the same approaches may have completely different behavior and often the conclusions derived with the results obtained on the development subset are not transferable to the test subset. This may be a consequence of overfitting, which partially derives from the small amount of data.

On the other hand, our experiments showed that with a particular sampling partition, the performance of the system does not change much between several independent runs. Therefore, we fix all sampling partitions at the very beginning of our experiments and keep them throughout, to ensure plausibility of comparison. We used stratified subsampling to have the train and test sets as close to the original dataset as possible in terms of any available information on participants, such as age, gender, nationality, etc.

We used a dependent context modeling with FLF described in Section 5.2.1 during AVEC 2019 (Ringeval et al., 2019) as a part of our final recognition system based on the multimodal DLF of several approaches. The results of this system compared to the baseline provided by organizers are presented in Table 5.4.

Culture	Dimension	Baseline	Proposed approach
German	Arousal	0.562	**0.621**
	Valence	0.646	**0.750**
Hungarian	Arousal	0.527	0.501
	Valence	0.548	0.462
Chinese	Arousal	0.355	**0.391**
	Valence	0.468	**0.499**

Table 5.4: Performance of system that includes proposed approach to dyadic modeling compared to baseline of AVEC 2019 (Ringeval et al., 2019).

With the proposed approach, we achieved performance gain for both arousal and valence dimension on German and Chinese parts of the challenge dataset (SEWA corpus), taking third place in the competition. As reported in Section 2.3, both winner and runner-up also used dyadic data in their prediction systems.

5.5 Summary

In this chapter, we presented several approaches to utilize the data of interlocutor to increase the speaker emotion recognition performance in a dyadic conversation. More precisely, we applied feature-level fusion and decision-level fusion strategies and modelled the context using methods from previous chapter dependently, i.e. context of speaker and interlocutor is the same, or independently, i.e. context of speaker and interlocutor are not connected to one another and can change freely.

Thus, at the beginning of the chapter we provided examples of several patterns of emotional flow in a dyadic conversation: (i) empathetic, when emotions of the speaker and the interlocutor have similar trends, (ii) contrary, when they have the opposite trends (e.g. one becomes angry and other happy), (iii) indifferent, when emotions of the speaker do not influence emotions of the interlocutor and/or vice versa. However, more than one pattern often happen to be present in a recording.

Then, we set the baseline for this chapter with the emotion recognition performance in terms of CCC obtained with the models that utilize speaker-only data. For the baseline and other experiments in this chapter, we used RNN-LSTM. After this, we presented the concept of dependent contextual modeling in dyadic interactions and tested it with two strategies: feature-level fusion, where we fused data of the speaker and the interlocutor prior to modeling and decision-level, where we used an additional meta-model based on a smaller RNN-LSTM to combine predictions obtained separately for the speaker and the interlocutor. Feature-level fusion approach showed patterns similar to the ones obtained in the previous chapter, i.e. the dependence between amount of used context and model performance. With decision-level fusion strategy, we noticed that models perform well on all used corpora in a range of approximately 20 seconds for meta-model. We achieve improvements over the baseline defined earlier for three out of four databases on audio modality and one out of three on video.

Further, we presented the concept of the independent contextual modeling in dyadic interactions. We used the data sparsing approach defined in the previous chapter to independently change the amount of modelled context for the speaker and/or the interlocutor. For the audio modality, we noticed the significantly higher model performance when the context window for the speaker and the interlocutor are the same or close to each other. However, it is not the case for the video modality. Overall, the independent contextual modeling with the speaker's context window fixed to the optimal value according to the previous experiments, provided improvements in five out of eight cases for the audio modality and in three out of six cases for the video; the fully independent approach – for six out of eight for the audio and for four out of six for the video.

Finally, we concluded that dyadic emotion modeling can provide more improvements with audio features, then with video. This approach can be successfully applied to databases with originally utterance-level annotations without any knowledge about the interaction units, such as turns, as well as purely continuously annotated data. Proposed methods were tested during the Audio Visual Emotion Challenge 2019 and also showed significant improvements over the baseline results.

6 Towards Contextual Emotion Recognition in Smart Environments

The third and highest level of the contextual information considered in this thesis is the environmental context. It covers many aspects of user's surrounding and in real, in-the-wild conditions it is almost impossible to model with a high precision. The reason for it is the tremendously high amount of components that comprises an environment. Yet, by recognizing user's emotional responses to his environment we open the room for possible actuations and mood regulation, significantly improving interaction systems and/or smart systems embedded into environments. A distinctive feature of the context integration at this level, compared to the ones considered in the previous chapters, is a wide range of possible modalities that are useful for information extraction. Previously, we considered only audio and visual features; another useful modality for the emotion recognition is the textual one. For the environmental context modeling, the list of modalities as well as their connections with the emotional status of the user are not explicitly defined. Almost any source of information, from the physical activity of the user to the outside temperature, may be useful. Therefore, the search of useful modalities is an important stage of the recognition pipeline (see Fig. 6.1).

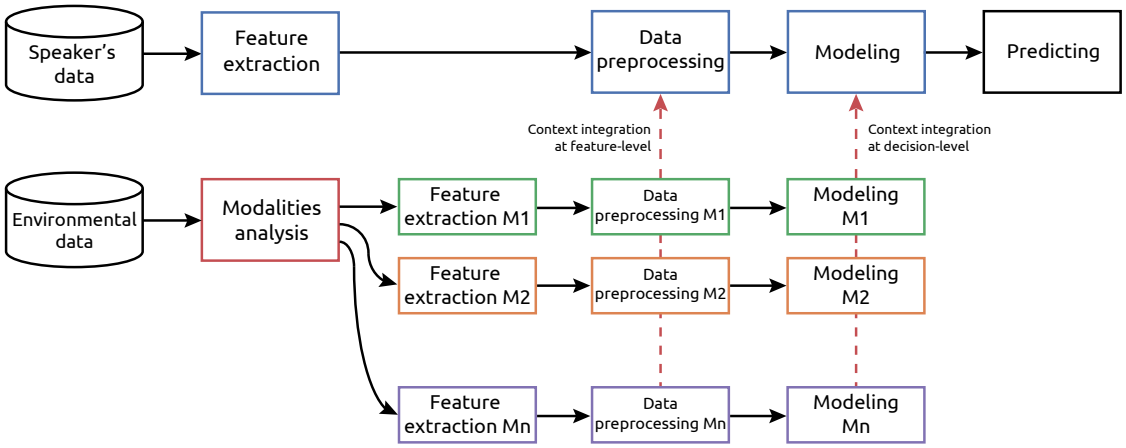


Figure 6.1: The general pipeline of the environmental-level contextual modeling

In order to study this level of context to some extent, we narrowed the number of considered aspects of the surrounding and their effect on the user's emotional status. As a particular use case, we have chosen the scenario where emotions are especially strongly affected by environment – a sightseeing tour. Here, our aim is to estimate emotions and satisfaction level of tourist during the tour using audio-visual recordings, behaviour cues and some additional modalities. Work presented in this chapter was completed in the cooperation with Nara Institute of Science and Technology (Ikoma, Nara, Japan), especially, with Dr. Yuki Matsuda from Ubiquitous Computing Lab. Researchers affiliated with this laboratory have extensive

experience in applications of sensing technologies (Otoda et al., 2018; Holatka et al., 2019; Mizumoto et al., 2020), context awareness (Nishigaki et al., 2005; Yamamoto et al., 2011; Amornpashara et al., 2015), smart environments (Ueda et al., 2015; Arakawa, 2019; Tani et al., 2020), etc. Moreover, their work also includes the development of the modern sensor-based devices, that are easy-to-use in research purposes for data collection (Nakamura et al., 2016, 2017, 2019). Combined with other products on the market of smart devices, we use them in this chapter to collect data at several different modalities, including head and body movements.

In the following sections we will describe our approach to this use case in detail, considering various aspects of data collection, pre- and postprocessing, modeling, etc. In Section 6.1 we provide a short overview on Smart Tourism, as this research field is the most connected to the considered use case. In Section 6.2 we describe the collected database – EmoTourDB, namely, in Section 6.2.1 we describe the data collection process itself, in Section 6.2.2 we provide a detailed description of the feature extraction used in this chapter, in Section 6.2.3 and Section 6.2.4 we present the set of annotations and post-experiment surveys used in our work, in Section 6.2.5 we briefly cover an aspect of synchronization between devices and in Section 6.2.6 – missing data of EmoTourDB. Further, in Section 6.3 we cover the modeling procedure and experimental results, in Section 6.4 we discuss limitations of the implemented system, use case and the concept in general. Finally, Section 6.5 summarizes this chapter.

6.1 Smart Tourism

The smart devices and wide mobile network connection allow one to provide touristic information to people based on various data sources, this paradigm is called "smart tourism". From the point of information usage or production, the current smart tourism systems can be categorized into: (i) context understanding, (ii) content generation, and (iii) information presentation. Context understanding mainly focus on *environmental context recognition* and *human activity estimation and prediction*. The mobile devices, equipped with many kinds of sensors, enable to recognize various environmental contexts, e.g. a congestion degree in the city (Nishimura et al., 2014). Some studies have proposed the tourists' behavior prediction method for recommending the next tourist spot based on machine learning modeling of the nationality, gender, and age of tourists (Lim et al., 2015; Lu et al., 2010; Popescu and Grefenstette, 2011). To generate contents from various data sources, *touristic content summarization* is often discussed as a research topic (Hidaka et al., 2017). Some studies have proposed the method for extracting important scenes from the whole video of the sightseeing tour (Okamoto and Yanai, 2013; Zhang et al., 2013; Kanaya et al., 2019). As the part visible to the tourists, various *tourist route and spot recommendation* approaches are proposed (Borràs et al., 2014; Xu et al., 2016). Huang and Bian (2009) proposed the recommendation system based on the Bayesian network, analytic hierarchy process (AHP), and tourist's behavior.

On the other hand, according to Gretzel et al. (2015), the smart tourism is comprised of three layers: (i) smart destinations, (ii) smart business and (iii) smart experience. The *smart destinations* mostly cover a physical infrastructure, integration of various devices and state-of-the-art technological advancements into urban area, increasing not only the quality of tourism, but also the quality of life for residents in this area (Lopez de Avila, 2015; Jovicic, 2019; Jeong and Shin, 2019). The *smart business* often considered to be an ecosystem, supporting and creating the tourists experiences and resources, connecting the stakeholders of tourism domain, and assuming that tourists are not only the consumers, but also play a regulatory role by creating or changing the value of and interest to certain places (Buhalis

and Amaranggana, 2013; Gretzel et al., 2015). However, in some cases it is not completely clear, how the business can gain financial profits directly from smart tourism (Gretzel et al., 2015). The third layer – the *smart experience* – focuses on personalization, context-awareness and usage of technologies in order to advance the quality of tourism (Neuhofer et al., 2015; Buhalis and Amaranggana, 2015). In this case, personal information about the user, his/her feelings during past trips and current mood play a significant role.

Most of the currently used approaches are often not human-centered, and they do not consider the emotional reactions and feelings of the users, therefore lacking the information to create personalized recommendation or analysis. In our studies on smart tourism we focus on the layer of smart experience and work on multiple informational levels, opening a room for creation, analysis, understanding and practical use of personal features of tourists.

In the past years, the psychological context of tourists (e.g., preferences, personality, emotional status, satisfaction levels) have been attracting researchers' interests as the data source, because such information might be an effective cue for summarizing and providing tourist information, especially in recommendation systems. For example, recommendation based on preferences of the tourist can provide more appropriate information to them, and the curating system can summarize it by omitting uninteresting topics. To collect psychological contexts, questionnaire-based surveys and online user reviews such as TripAdvisor[1] are still widely used. The questionnaire-based survey can be collected by asking the tourists directly. It is suitable for hearing tourist's preferences/personality for future recommendations, however, difficult to collect continuous psychological feedbacks. In contrast, online user reviews are suitable to collect such feedbacks, but it is difficult to keep the users motivated to write the reviews, especially with the medium rating values.

Following the demand on the data that is suitable to design human-centered smart tourism systems, we collected the multimodal quantitative dataset, relying on user's natural behavior and collected in real time in contrast to a questionnaire-based survey and online user reviews. We develop feature extraction approaches and test this data with several uni- and multimodal systems.

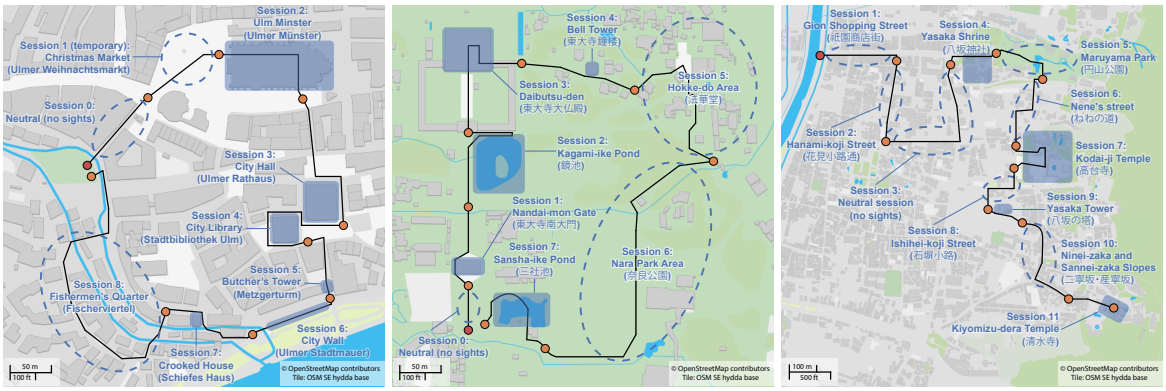
6.2 EmoTourDB

For our research on effect of environmental context on user's emotional state, in cooperation with Nara Institute of Science and Technology (Ikoma, Nara, Japan), we have collected a database of emotionally labelled multimodal touristic behavior – EmoTourDB. In this section we provide details on data collection, feature extraction and labelling.

6.2.1 Data collection

To collect the data, we chose three touristic routes located in historical parts of three different cities: Ulm (Germany), Nara (Japan) and Kyoto (Japan). The total lengths of the routes are approximately 1.5, 2.1 and 3.5 km respectively, and it took on average from 50 to 110 minutes for tourists to complete them. The routes were divided into parts (sessions) of two kinds: (i) containing one particular sight, such as a building, church or temple; (ii) representing an area, such as a street, park, etc.

[1] `https://www.tripadvisor.com/`



(a) Ulm (Germany) (b) Nara (Japan) (c) Kyoto (Japan)

Figure 6.2: Touristic routes in three data collection locations

Touristic route in Ulm, Germany (Fig. 6.2a) is located in the city center of Ulm and consists of nine sessions, including one neutral session (with no sights) at the beginning. Most sessions cover a particular historical building, such as Ulm Minster (Ulmer Münster), Ulm City Hall (Ulmer Rathaus) or Butcher's Tower (Metzgerturm), but some of them also represent a touristic area, e.g. Fisherman's Quarter (Fischerviertel) or a seasonal sightseeing place – Christmas Market (Ulmer Weihnachtsmarkt), which was available to participants only in December.

The route in Nara, Japan (Fig. 6.2b) covers eight sessions, also including one neutral session, but is located in Nara Park – the historical outskirts of Nara. Most buildings have scenic background, e.g. Nandai-mon Gate (南大門), Daibutsu-den (大仏殿), Bell Tower (鐘楼), and Hokke-do (法華堂), which are parts of Todai-ji Temple (東大寺). Additionally, there are many natural spots, e.g., Kagami-ike Pond (鏡池) and Sansha-ike Pond (三社池). The whole route is located in nature, and has no distraction from the sights included in the sessions.

The route in Kyoto, Japan (Fig. 6.2c) is located in Gion district of Kyoto, that was originally developed around Yasaka Shrine (八坂神社) in 15-16th centuries. The route consists of 11 sessions (one neutral) and apart from Yasaka Shrine itself it includes Kodai-ji Temple (高台寺) and Kiyomizu-dera Temple (清水寺) and different sightseeing areas, such as Hanami-koji Street (花見小路通), Ninei-zaka (二寧坂) and Sannei-zaka (産寧坂) Slopes, and Ishihei-koji Street (石塀小路), all famous for their traditional Japanese architecture. This route contains the most area-based sessions (7 out of 11).

Up to date, 47 participants (37 males, 10 females) from 12 countries have participated in our experiment. Most of them were visiting or short-term students. The average age of participants is 25.8 years (σ=2.9). Combining touristic experiences of the participants together, they have visited more than 50 countries around the globe.

Before each session, all the participants were provided with the short information on the sight they are about to visit. During the sessions participants were accompanied by two assistants, showing them the route, providing additional information about sight if requested, and controlling the devices if necessary. The participants were not restricted in time and actions during the experiment. They could spend as much time as they need to explore the sight, take photos, use their smartphones, chat with assistants or another participants, etc.,

[1] © OpenStreetMap contributors. The data is available under the Open Database License: `https://www.openstreetmap.org/copyright/en`

[1] Hydda map tile: `https://github.com/karlwettin/carto-style-hydda`

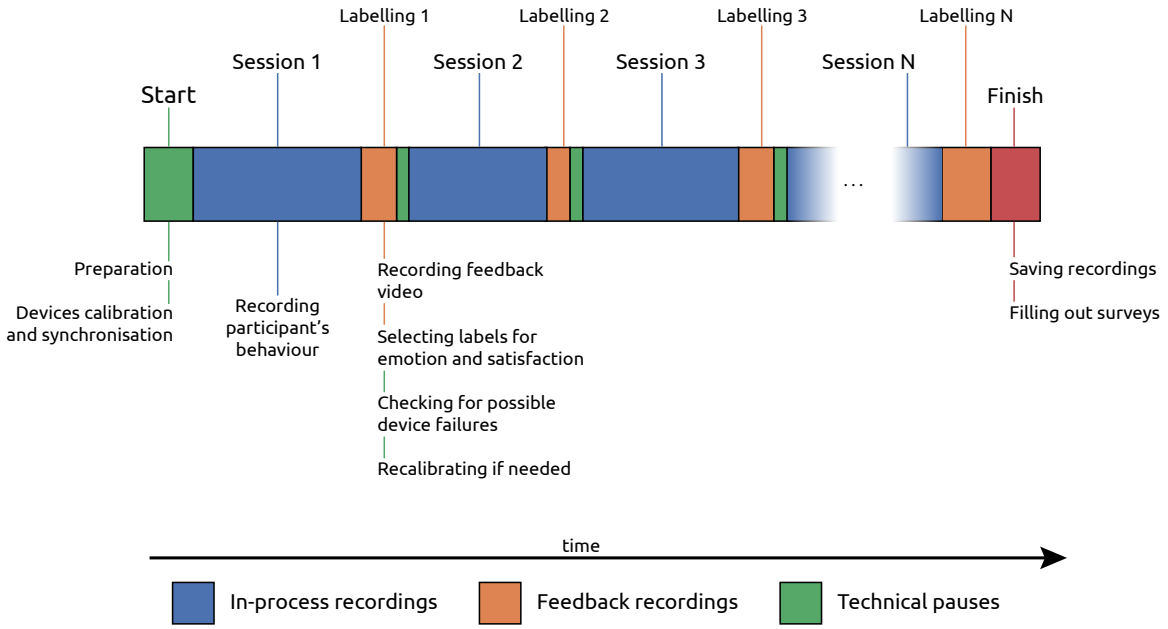


Figure 6.3: Data collection pipeline for EmoTourDB

simulating a natural touristic behavior in-the-wild.

At the end of each session, participants were asked to provide a short feedback in the form of a "selfie-video", expressing their feelings and impressions of the latest touristic session. They were not restricted to use any language and were encouraged to use the mother tongue, as they may express the emotions in the most naturalistic manner, without constraints from language skills. After taking a short movie, participants were asked to rate the session (sight) in terms of satisfaction and emotions (see Section 6.2.3 for more details). After the last session, participants were asked to fill out surveys on their personality and touristic experience. A schematic overview of the data collection pipeline is presented in Fig. 6.3.

Devices

To collect the data from participants and analyze their behavior, we use the set of devices to capture conscious and unconscious cues. The general description is presented below, for more detailed specifications see Table 6.1 and Fig. 6.4.

Android Smartphone for tracking GPS data and recording feedback from participants. Our developed application records selfie-video using the frontal camera with resolution of 960x720 or 1024x768, depending on smartphone model.

***Pupil Labs* Core Eye Tracker**[2] for tracking eye gaze and extracting higher-level features (Kassner et al., 2014). It is equipped with two infrared eye cameras and one world camera. Eye tracker was constantly connected to Apple MacBook Pro via USB-C cable to record the data.

***SenStick* Sensor Board** (Nakamura et al., 2017) mounted on the ear of the eye tracker for collecting the data on head/body movement (through data from accelerometer and gyroscope) as well as magnetic field, illumination, ultraviolet (UV), environmental temperature, humidity, and atmospheric pressure. It was connected to Apple iPod Touch via Bluetooth

[2] `https://pupil-labs.com/products/core/`

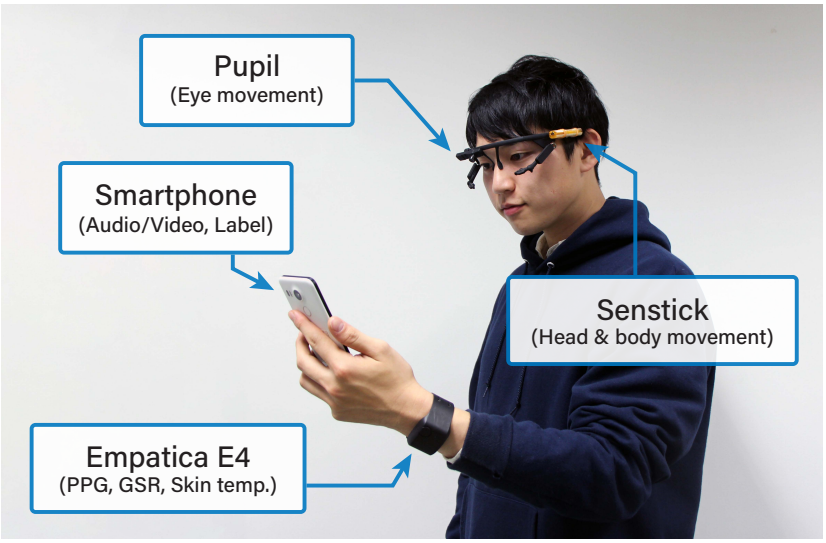


Figure 6.4: Device setup for EmoTourDB

Device	Sensor	Specification
Android Smartphone	Frontal camera	1024x768@30fps 960x720@30fps
	GPS	1 Hz
Pupil Core Eye Tracker	Binocular infrared eye cameras	192x192@124fps
	Optical world camera	1280x720@30fps
SenStick Sensor Board	Motion sensors (accelerometer, gyroscope, magnetometer)	50 Hz
	Environmental sensors (Illuminometer, UV sensor, Temperature meter, Humidity meter, Barometer)	5 Hz
Empatica E4	Photoplethysmography (PPG) sensor	64 Hz
	Electrodermal activity (EDA) sensor	4 Hz
	Skin temperature sensor	4 Hz
	Accelerometer	32 Hz

Table 6.1: Device set for EmoTourDB

and *SenStickViewer* application[3] to control SenStick logging state.

***Empatica E4* Wristband**[4] for measuring physiological signals, such as photoplethysmography (PPG), heart rate, electro-dermal activity (EDA) and body skin temperature. To stream data continuously, it was connected to Apple iPhone XS via Bluetooth and *E4 realtime* application[5].

6.2.2 Features

For our experiments, we developed and extracted several feature sets from the collected data. In this thesis, we differentiate between two types of data recordings: (i) feedback – recorded

[3] https://itunes.apple.com/jp/app/senstickviewer/id1195315881
[4] https://www.empatica.com/en-int/research/e4/
[5] https://itunes.apple.com/us/app/empatica-e4-realtime/id702791633

right after the end of the session along with the corresponding label; and (ii) in-process – recorded during the session (after finishing recording feedback for previous session, until the start of recording feedback for the current session). These two types of recordings are disjoint sets and complement each other.

Audio-visual features

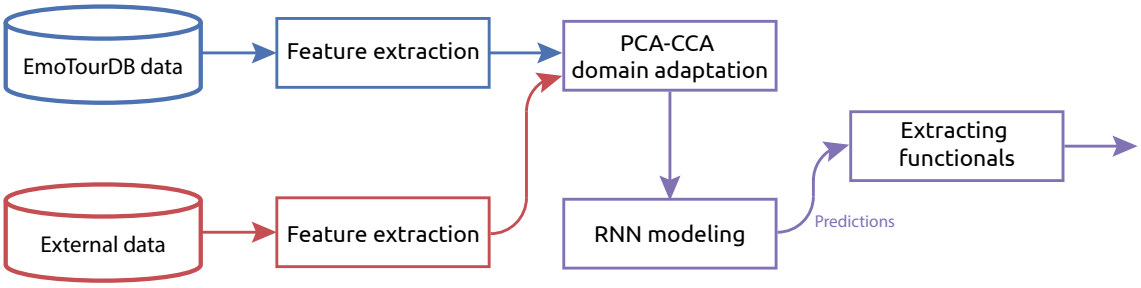


Figure 6.5: Pipeline of extracting higher level features for audio and video modality of EmoTourDB. This schema shows functional extraction for one modality-dimension pair of one external database (RECOLA, SEMAINE or SEWA). Values of functionals with all possible combinations of external database and dimension are merged afterwards.

Similarly to the previous chapters, we use LLDs of eGeMAPS feature set extracted with *openSMILE* toolkit for audio modality and Action Units extracted with *OpenFace* for video modality. However, in contrast to the other modalities described below, we use these feature sets on two levels: low and high.

At the low level, we use features, averaged over each recording, directly in our models to make predictions for emotion or satisfaction of the user. At high level, we first obtain predictions with pretrained models based on time-continuous representation of these features and then extract functionals from the prediction curves. As data for pretrained models we use RECOLA, SEMAINE and SEWA corpora and PCA-CCA approach of domain adaptation described in Section 4.3. Pipeline of the high-level feature extraction is presented in Fig. 6.5. We use models, pretrained for both modalities (audio and video) and both dimensions (arousal and valence) separately and merge the high-level features arrays corresponding to each modality afterwards, obtaining two feature sets at the final stage: audio and video.

We use the following six functionals: mean value, standard deviation, minimal value, maximal value, relative position of minimal value (where zero is the very beginning of recording and one is the last frame of recording), and relative position of maximal value. Calculating these functionals over each of the three databases and two dimensions, we obtain 36 values for each recording and each modality.

Object detection

On the data of in-process recordings (world camera of *Pupil Core* Eye Tracker) we perform object detection with the *TensorFlow Object Detection API*[6] (Huang et al., 2017) to have a general idea about the scene observed by the participant. The information on objects detected

[6] `https://github.com/tensorflow/models/tree/master/research/object_detection`

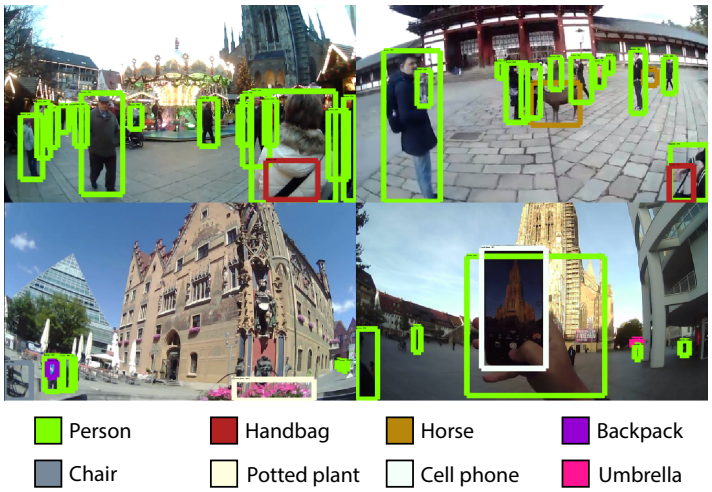


Figure 6.6: *TensorFlow Object Detection API* output for four frames of in-process recordings captured with world camera of *Pupil Labs* Eye Tracker

from images of the world camera, including object class, confidence, count and borders in terms of frame's coordinates system is available (Fig. 6.6).

We use object class and confidence score to calculate higher level features. First, we check the confidence score S for each prediction and keep only values with $S > 0.97$. This value of threshold was chosen empirically as a trade-off between the number of misclassified cases and the ability of the algorithm to extract sufficient number of recognized objects. Then, we analyze classes within each recording and compute a score for each class K (if at least one object is present) according to the following formula:

$$X_r^K = \frac{\sum_{C_i=K} S_i}{N_r} \tag{6.1}$$

where X is a resulting score of a particular class K for recording r, C_i is a class of the object i, S_i is a confidence score of the object i and N_r is a number of frames in the recording r. Hence, this score roughly represents the number of occurrences for objects of the class K in the recording r.

Videos from world camera of *Pupil Core* Eye Tracker have a frame rate of 30 frames per second, resulting in thousands of images for each session. As in most cases they are similar within several seconds, we extract one frame from video recordings every five seconds. In some conditions, e.g. at evening with not enough lighting, frame rate of videos decreases significantly (up to five frames per second), therefore, we calculate five seconds intervals based on the timestamp data, which is available for each frame of recording, and not based on position of frame in video.

Distribution of detected classes is presented in Fig. 6.7. As data shows great imbalance in classes, we use logarithmic scale for y-axis. For example, the class "person" is most represented with almost 70 000 detected objects for 28 622 frames in total. Other often detected classes are "car" (4663 cases), "cell phone" (828 cases), "backpack" (573 cases), "bicycle" (480 cases), "bus" (378 cases) and "truck" (337 cases).

It is worth to mention, that in spite of the relatively high precision on many challenging object detection tasks (see Fig. 6.8a), there are also some typical misclassification examples, such as classifying sequential objects of similar rectangular structure into trains (Fig. 6.8e). Touristic routes were not close to train stations, therefore it is doubtful that relative high

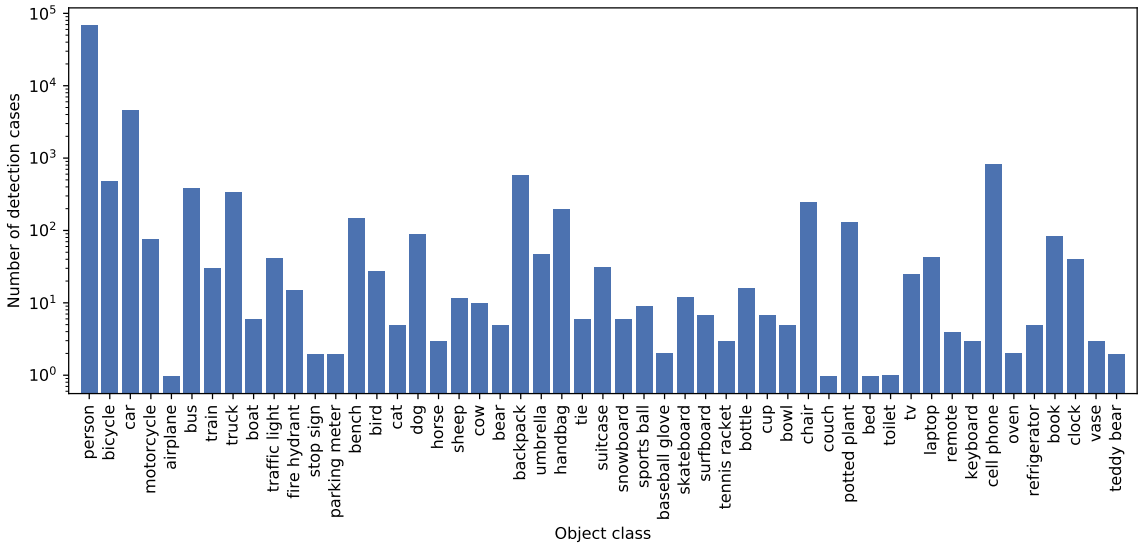


Figure 6.7: Distribution of classes detected with *TensorFlow Object Detection API*. Confidence threshold is 0.97, y-axis represents values on logarithmic scale

number of detected cases (several hundreds) represent objects correctly, even considering such high threshold value. Other examples include finding objects in the background with complex structure (e.g. Fig. 6.8b,c,d), finding objects with respect to surroundings (e.g. Fig. 6.8f,g,h) and misclassification due to lack of the categories (e.g. Fig. 6.8i,j,k,l). In the last four examples one may see deer of Nara Park, recognized either as cow, dog, sheep or cat. All these categories represent animals and some of them (cow and sheep) are close to the target object, but their selection by the recognition system (partly) caused by lack of class "deer".

Results of object detection may be used to analyze the crowdedness level (e.g. using classes "person" and "car"/"bus"/"bicycle") or diversity of objects in a particular sightseeing area. However, the latter may not work robust enough, due to the limited predefined amount of classes and misclassification cases.

Physiological signals

Among data provided by Empatica E4 Wristband, we use three sources: heart rate, electro-dermal activity and skin temperature. One should mention, that the data should be analyzed according to the participant-dependent scenario, as such values as electro-dermal activity or resting heart rate (heart rate in the calm state, without any physical activity; to some extent serves as a zero-point for heart rate analysis) differ greatly from one participant to another.

From raw time series we extract the following functionals to get the high-level features: minimum, maximum, mean, standard deviation, range (difference between minimal and maximal value), relative minimum and maximum (ratio to reference value) and main coefficient of first order polynomial fit to time series corresponding to one recording (represents general trend of data, i.e. increasing or decreasing). As the reference value, we chose data of first second of first recording for each participant.

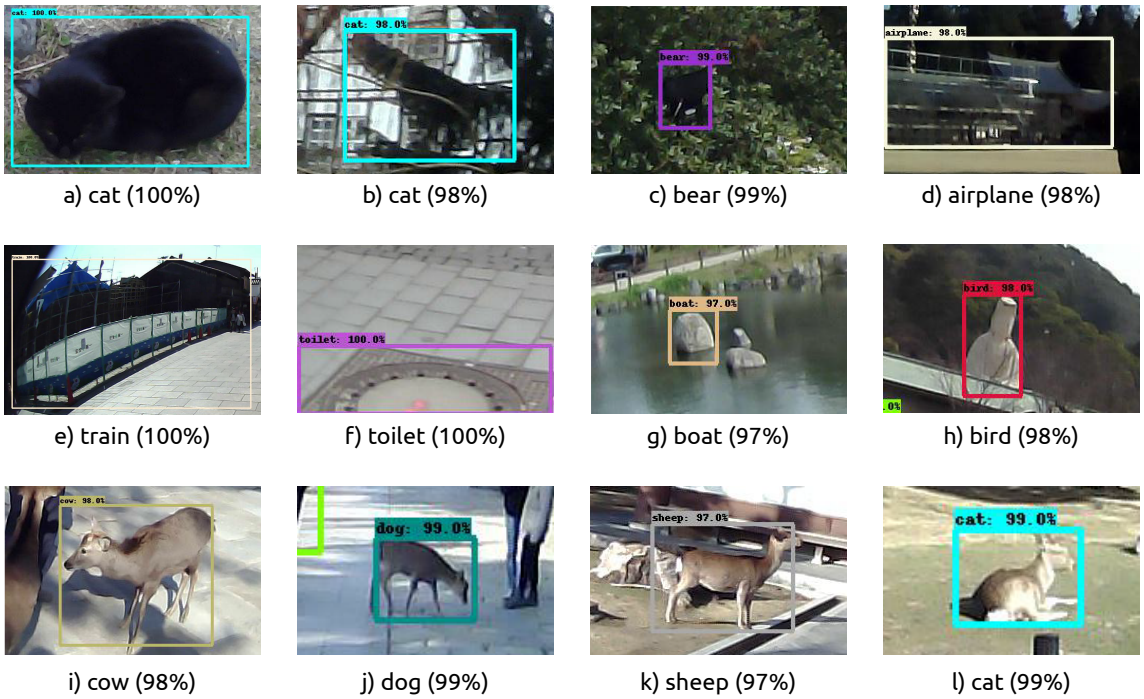


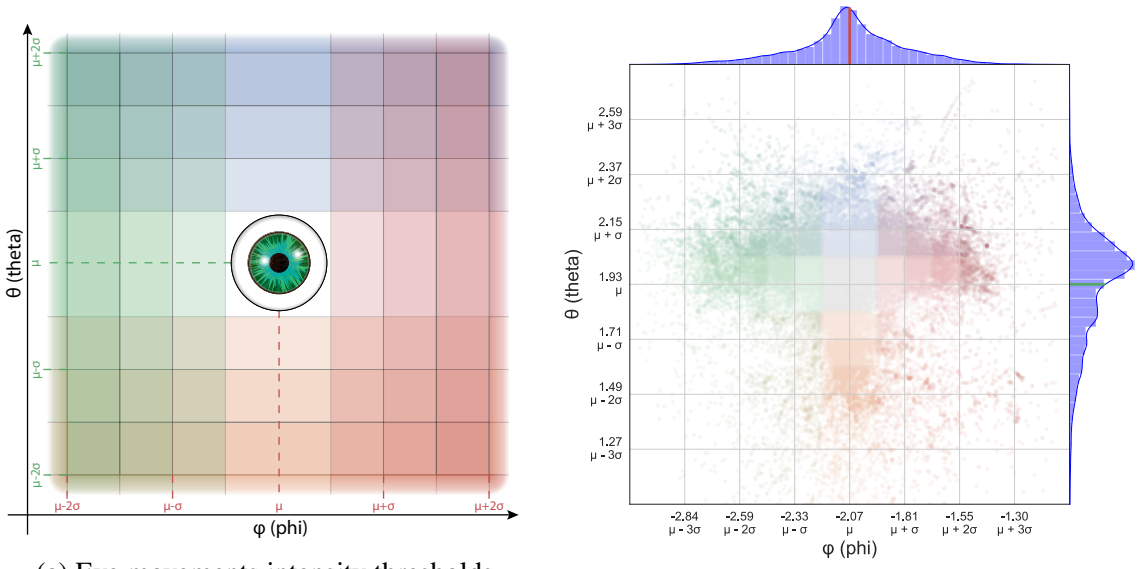
Figure 6.8: Examples of misclassification with *TensorFlow Object Detection API*

Eye movement features

We use data from eye cameras of *Pupil Core* Eye Tracker to analyze eye movements of participants. We use *theta* and *phi* values, which represent the normal of the pupil as 3D circle in spherical coordinates. Note that the raw values of eye movement data differ across users and depend on the physical setting of a camera and eye peculiarities, therefore it should be processed participant-dependently within each experiment. Using spherical coordinates of eye gaze, we calculated two sets of features described below.

Eye movement intensity. Mean (μ) and standard deviation (σ) values for *theta* and *phi* were calculated for each participant; eight thresholds (with a step of ($\frac{1}{2}\sigma$) were set and then used to count the percentage of time outside each threshold per session (see Fig. 6.9a). It is worth noticing that the significant number of participants have similar T-shaped distribution of pupil position coordinates (see Fig. 6.9b as an example). According to this figure, the participants use the top half of eye gaze space with much wider horizontal range compared to the bottom half. The latter is used almost exclusively to look directly down (presumably, to watch one's steps), while the top part is often being used to look around and view sights. This also has a physiological explanation, as one's nose blocking part of the view in the bottom range, and it is physically more comfortable to slightly turn the head, rather than to move the eyes to extreme positions.

Time-window eye movement statistics. Mean and standard deviation of *theta* and *phi* were calculated for a time-window with a particular length, and the values corresponding to the same session with data of one participant were averaged. The following window lengths were used: 1, 5, 10, 20, 60, 120, 180, 240 seconds with the offset of $\frac{1}{3}$ of the window length.



(a) Eye movements intensity thresholds

(b) Pupil coordinates distribution of one participant

Figure 6.9: Eyes and head movement features

Body movement features

Head tilts and turns. As stated above, eye movements are strongly connected to head turns. One can consider head turns as fast change of person's sight and eye movements as more precise "fine-tuning". We use the gyroscope data of SenStick during sessions to model head movements. First, the mean value μ and the standard deviation σ of the gyroscope values are calculated for each participant. Then, the upper and lower thresholds $\psi = \{\psi_{upper,a}, \psi_{lower,a}\}$ are set with the following equations:

$$\psi_{upper,a} = \mu_a + 2\sigma_a \tag{6.2}$$

$$\psi_{lower,a} = \mu_a - 2\sigma_a \tag{6.3}$$

where a is the axis of gyroscope (x, y or z). The head tilt or turn (looking up/down, right/left) is detected, when the position of the head crosses these thresholds ψ (see Fig. 6.10). In our case, y-axis indicates an action of looking up or down and z-axis indicates an action of looking left or right. Since the duration of each session is different, we extract several higher-level features based on counting, and statistics of the mean and standard deviation values of the head tilts data. We used four *basic* actions (tilt from neutral position to up, down, left, right) and two *complex* (tilt from right to left position or vice versa; the same for up to down) and apply the following functionals to them:

- To each type of actions:
  - *count*: average count of head tilts per second;
  - *span_mean*: average time interval between repetition of the same actions;
  - *span_std*: standard deviation time interval between repetition of the same actions;
- To *complex* actions only:
  - *val_mean*: average angular acceleration of head tilt;
  - *span_std*: standard deviation of angular acceleration of head tilt.

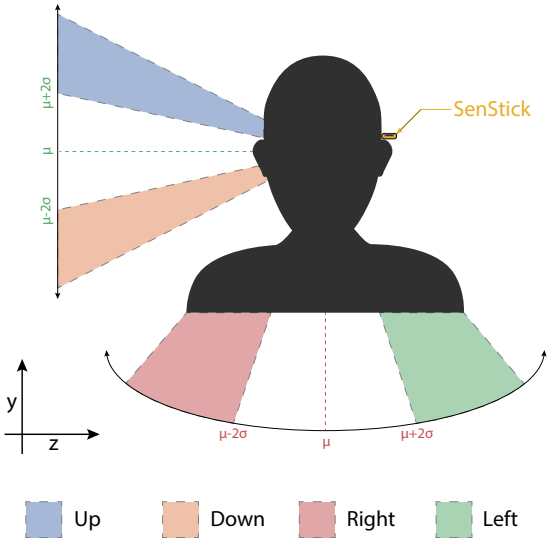


Figure 6.10: Head movements thresholds

We also count the average count of head tilts (of any type) per second. This results in 23 head movement related features in total.

Footsteps are counted based on the approach of Ying et al. (2007). First, the noise of accelerometer values is removed by applying Butterworth filter with 5 Hz cutoff frequency. Then, high-frequency components are emphasized through differential processing as follows:

$$y(n) = \frac{1}{8}\{2x(n) + x(n-1) - x(n-3) - 2x(n-4)\} \tag{6.4}$$

where $x(n)$ is an accelerometer value at the time index n. Furthermore, we use an integration process to smooth accelerometer values and remove small peaks. Since the sensor position in our case is different from the original method, we use the modified integration parameter equal to *5* chosen empirically (in contrast to *20* in original paper (Ying et al., 2007)).

$$y(n) = \frac{1}{N}\{x(n-(N-1)) + x(n-(N-2)) + \cdots + x(n)\} \tag{6.5}$$

Finally, footsteps are extracted by counting local maximum points. We analyze them and extract the following five features: (i) average footstep counts per second; (ii) average time interval between two consecutive steps; (iii) standard deviation of time interval between two consecutive steps; (iv) average acceleration of person, while walking; (v) standard deviation of acceleration of person, while walking.

6.2.3 Labels

In contrast to most of the datasets designed for emotion recognition, the third-party data annotation may not be applied, when dealing with touristic data. Only the person himself is able to correctly evaluate his feeling and status during a sightseeing tour, hence, flexibility and depth of ratings are crucial in this case. For this database, we used several types of data labelling and surveys. During the experiment, labels are assigned by the participant and collected right after the session end, using the developed labelling smartphone application.

After completion of the last session, participants were asked to fill out several survey forms.

Satisfaction estimation

Participants rated their touristic experience in terms of their satisfaction after visiting the sight using 7 point Likert-scale from 1 (completely unsatisfied) to 7 (completely satisfied).

In addition to satisfaction level, participants were asked to answer two questions related to tourist satisfaction using same 7-point scaling system: (i) *Do you want to introduce this sight to another person?*; and (ii) *Do you want to visit this sight again?*.

Analyzing different levels of satisfaction (see Fig. 6.11a) one may notice, that on general satisfaction scale (first distribution), female participants tend to provide more extreme labels, i.e. "1" and "7" are selected more often than other values on negative or positive side respectively; having "2" or "3" almost never chosen. Male participants' labels have smooth distribution, with "6" as the most selected value. Other satisfaction scales demonstrate the same situation for women, but to a certain extent different for men, as they are less smooth. However, for both men and women, the positive part of each satisfaction scale was preferred over the negative one with a strong domination. While tackling slightly different aspects of satisfaction, the values of these scales have rather strong correlation to each other ($PCC = 0.81 \pm 0.01$ between each pair).

Emotional status

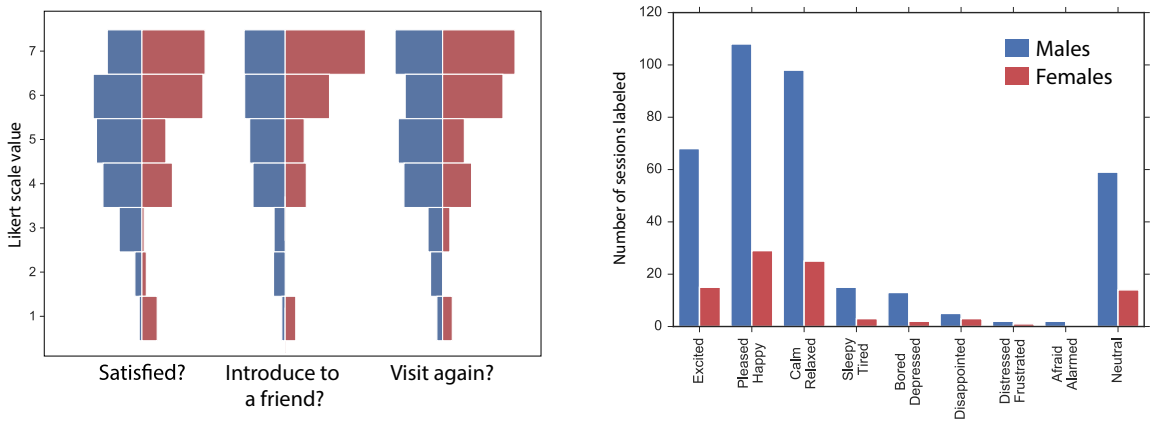
Upon completion of satisfaction-related ratings, participants were asked to provide some information about their emotional status on two different levels: general and precise. The general level represents a division of emotions into three simplest valence-related groups: negative, neutral and positive. It is intended to facilitate an easy emotion recall and to prepare the participants for more detailed answers.

The second level of emotional rating is more precise and participants were asked to make a choice among the following options: excited (0), pleased/happy (1), calm/relaxed (2), sleepy/tired (3), bored/depressed (4), disappointed (5), distressed/frustrated (6), afraid/alarmed (7), neutral (8). These emotional classes were generalized from Russell's circumplex model of emotion (see Section 2.1).

Having an a priori positive experimental setup and naturalistic data, distributions of both satisfaction levels (see Fig. 6.11a) and emotional classes (see Fig. 6.11b) are skewed to positive values. The classes *exited*, *pleased/happy* and *calm/relaxed* comprise 74% of the labels. Among the negative classes, *alarmed/afraid* is the least selected one. Also, in spite of similar patterns of men's and women's label distributions, one may notice slight imbalance between *disappointed* men and women, what makes it the most selected negative emotion class among women on a par with *sleepy/tired*.

Touristic experience quality

However, emotional classes adopted from the Russell's circumplex model do not always describe the situation in the best and most suitable way for the tourism domain. For example, the *neutral* class might be rather negative in terms of touristic experience, if selected after a sight that was expected to make a strong impression on the person. Likewise, the *tired* class may be ambiguous and reflect two different states: tired of being in this (not interesting) place or tired physically, which is not related to the place itself.



(a) Three-sided satisfaction label distribution

(b) Emotional classes distribution

Figure 6.11: Satisfaction and emotion dimensions

To have more relevant system for touristic experience rating, we have developed the two-dimensional evaluation space called Touristic Experience Quality (TEQ). Dimensions of the TEQ plot represent *interest* and *meeting expectations* and contain emotional states listed in Section 6.2.3, as well as provide more flexibility in selecting intermediate states or intensity of a particular emotion. Although the TEQ space is represented by two axes, namely interest and meeting expectation, we avoid using these names explicitly during the labelling process to omit the presence of the space origin (point of axes intersection) on the graph, which might be difficult to define and to interpret by the participants, leading to a confusion in ratings. Instead, we limit the space to boundary states: from *not interesting / nothing to see* to *interesting* on x-axis and from *didn't get what expected* to *got what expected* on y-axis (see Fig. 6.12a for more details). However, values were later scaled to range *[-1,1]* for both axes.

Analyzing the relation between the TEQ values and emotional classes (Fig. 6.12b), it may be noticed, that some classes, adopted from the Russell's circumplex model have a strong correlation to interest or meeting expectation level, e.g. *"excited"* have well-defined cluster, centered to the right-top part of the TEQ-space; some classes have a weak correlation, e.g. *"calm/relaxed"* and *"sleepy/tired"* classes are spread over the TEQ-space and may not always represent an emotion corresponding to the sight itself, but a general state of the user at that moment.

6.2.4 Additional information

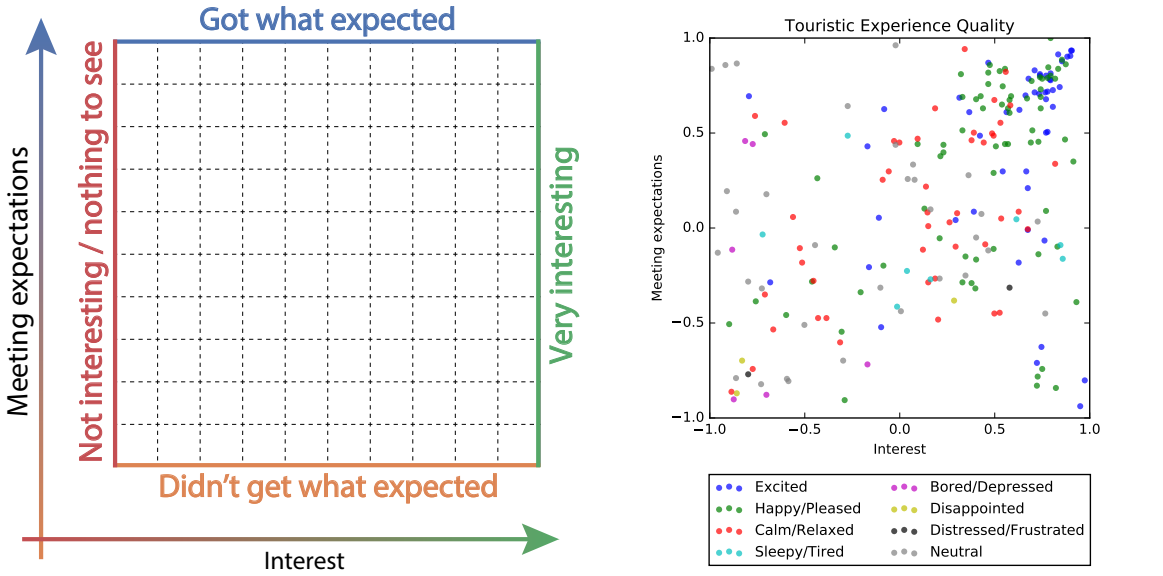
Apart from the labelling process, performed upon completion of each session, we asked participants to provide additional information that can be used to increase adaptability of the recognition system.

Sightseeing profile

To better understand participant's attitude to sightseeing, we asked them to provide a sightseeing profile, consisting of two main parts: personal information and trip experience.

Personal information covers the following aspects:

- home country;
- mother tongue (or main language if bilingual);



(a) TEQ space used in labelling application

(b) Relation between TEQ and emotional classes

Figure 6.12: Touristic experience quality

- language skills of the country of experiment;
- duration of stay in the country of experiment;
- residence place type (city center, suburb, countryside).

Trip experience is being measured through the following direct questions:

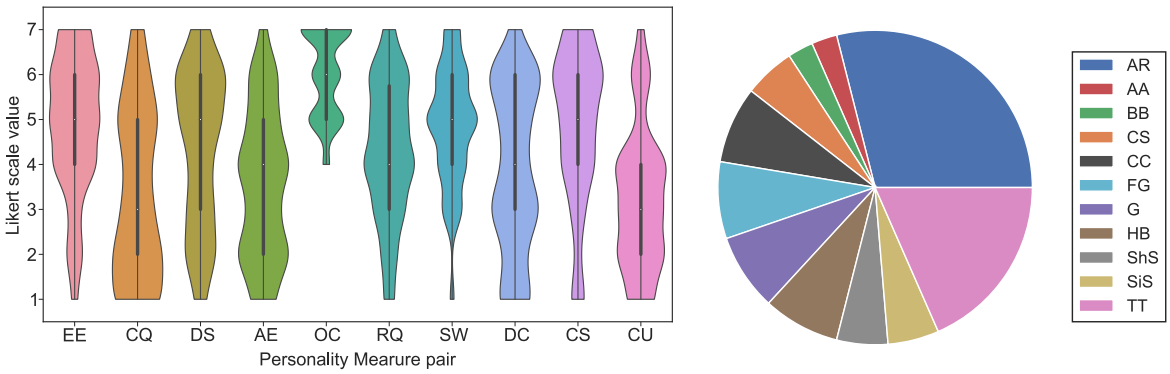
1. Do you like sightseeing? (on a scale from one to five);
2. How many times do you go sightseeing per year?
3. Which countries have you already visited for sightseeing?
4. Is it the first sightseeing trip on the continent of experiment (Europe or Asia)?
5. Is it the first sightseeing trip in the country of experiment?
6. Is it the first sightseeing trip in the city of experiment?
    a) If not: how many times did you do sightseeing in the city of experiment?
    b) If not: which particular sights of the current experiment have you already visited?

This questionnaire is designed to compose a picture of past touristic experience of the participant in order to help correcting the bias of ratings.

Personality survey

Another way of better interpretation of the ratings is having an understanding about participant's personality traits. To estimate them, we asked participants to fill out the form of Ten Item Personality Measure (Gosling et al., 2003) and rate each pair of the following traits using a seven-point Likert scale:

- Extroverted, enthusiastic (EE);
- Critical, quarrelsome (CQ);
- Dependable, self-disciplined (DS);
- Anxious, easily upset (AE);
- Open to new experiences, complex (OC);



(a) Ten Item Personality Measure distribution (according to (Gosling et al., 2003))

(b) Travel personality distribution (according to (Gretzel et al., 2004))

Figure 6.13: Personalities distribution

- Reserved, quite (RQ);
- Sympathetic, warm (SW);
- Disorganized, careless (DC);
- Calm, emotionally stable (CS);
- Conventional, uncreative (CU).

Analysis of distributions (Fig. 6.13a) shows that most of the participants are extroverted, open to new experiences, sympathetic, and neither critical nor conventional.

In addition, we asked them to choose the most suitable way to spend their leisure time from 12 travel personalities according to Gretzel et al. (2004). The options and corresponding type of preferred activity are the following (taken from original paper of Gretzel):

- All rounder (AR): go where there is lot to do and see;
- Avid athlete (AA): games of any type;
- Beach bum (BB): laying around the beach;
- Boater (B): water activities or attractions;
- City slicker (CS): clubs, meeting people, in need of pulse of the city;
- Cultural creature (CC): theaters, shows, museums, festivals, local culture;
- Family guy (FG): spend time with family during vacation;
- Gamer (G): gaming, fantastic fares and night entertainment;
- History buff (HB): historical facts and sites;
- Shopping shark (ShS): like shopping;
- Sight seeker (SiS): always stop for landmark, event or attraction;
- Trail trekker (TT): hiking, parks, mountains, forests and bird watching.

Similarly to Jani (2014), *all rounders* make up the significant number of participants (28%). However, in his research, *sight seeker* was the most preferred option among the respondents. In our case, in spite of the touristic context of the experiment, only 5% of participants marked themselves as *sight seekers*. Another large group is *trail trekker* (21%). Other participants are split into the rest categories without significant patterns. The only category that was not selected at all is *boater*. This might be caused by the home or the current living place of the participants – not many of them live directly near the sea.

These questionnaires open a room for psychological analysis of rating diversity, preferable emotions and existing natural bias across participants. However, due to rather low amount of data collected so far, they are not used in our models.

6.2.5 Synchronization and Calibration

As we use many recording devices simultaneously to collect the data for the EmoTourDB, synchronization of this data is a special challenge. In order to ensure the correct mapping between features and labels, as well as between various data sources, we follow several procedures during the process of data collection as well as data preprocessing.

Before the beginning of each experiment, we ensure that devices used for data capture (e.g. a smartphone or a laptop) are synchronized to current time via time servers on the Internet. Data recorded with each of the used device uses timestamps either with offset to local time zone or without it (UTC timestamps), which can be corrected later, at the data preprocessing step.

We use timestamps with precision of one second, e.g. *1553611283* for *2019-03-26 15:41:23*. Some devices provide timestamps with precision of one millisecond, it this case they are used in this format for data preprocessing, but rounded to one second for synchronization.

After the data was collected, session intervals are defined manually using data from world camera of *Pupil Labs* Eye Tracker. As the beginning of the session, we use the timestamp corresponding to the moment when the participant starts to walk following the route; as the end of the session – when the participant stops for the labelling. This information is used to split the data into session for each device. Such an approach is time-consuming, nevertheless, it allows precise division of data, especially for such devices as *SenStick* or *Empatica*, when no other cues can be used to distinguish between the sessions and filter technical pauses and labelling periods.

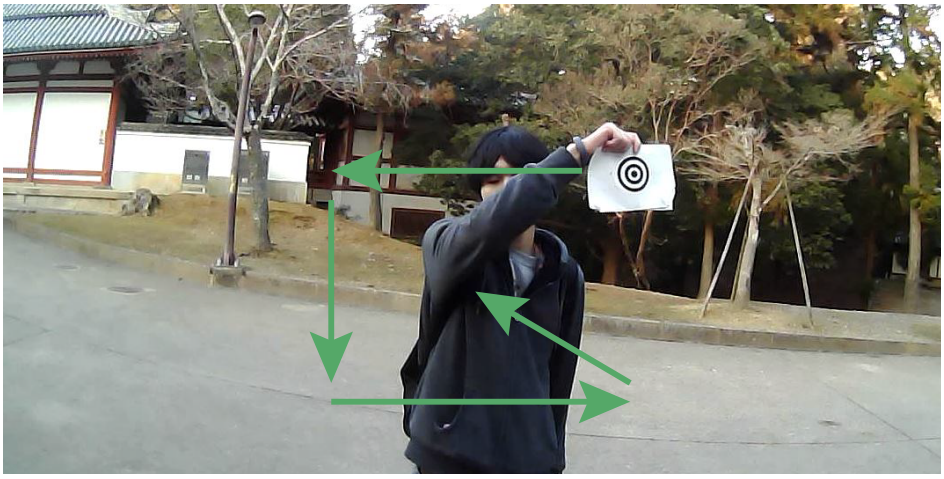


Figure 6.14: Prerequisite actions for offline calibration with *Pupil Labs Player*. Participant follows the point printed on a paper with eyes, without any head movements.

Among used devices, *Pupil Labs* Eye Tracker requires the most extensive calibration. Before the beginning of each experiment, infrared eye cameras are set to match eyeballs and the camera's field of view. After the data collection, we manually calibrated the parameters of eye gaze detection algorithm with *Pupil Labs Player* software. In order to enable an offline calibration, we asked participants to perform a simple sequence of actions prior to the session beginning, in which they should've followed an object with their eyes without any head movements (see Fig. 6.14).

6.2.6 Missing Data

Due to the real life recording setting, strict limitations of tourist data collection pipeline (no person can be recorded twice on the same route if some technical problem occurs) and continuous development of experimental setup, data may be partially missing for some participants. It may be caused by a lost connection during an experiment, a data transfer issue or hard-/software problems, e.g. the experiments in Nara were recorded under the direct sun, which caused problems with exposure for infrared cameras of *Pupil Labs* Eye Tracker (this issue was fixed in later software updates) and without using Empatica E4 Wristband.

At early stages of data collection, we used only the precise level of emotion labelling (nine classes) and one satisfaction scale ("Are you satisfied?"). Later, we introduced an additional emotional scale (three classes) and two additional satisfaction scales. From some participants we have received feedback, that emotional scales are not always precisely correspond to the effect of a sight. Therefore, we developed the TEQ labelling approach described in detail in Section 6.2.3, introducing two more relevant scales. However, it is not available for the first and the second data collection locations (Ulm and Nara).

In Table 6.2 we present percentage of missing data and labeling coverage for each database part and device.

6.3 Modeling

In this section we discuss the modeling pipeline used in this chapter. In order to perform prediction of emotional state or satisfaction level of a user, we used features described above in unimodal and multimodal scenarios. As the basic model for this chapter, we use a feed-forward neural network with a single hidden layer of 30 neurons and number of output layers dependent on the task. For emotion prediction, we use three output neurons with the softmax activation function; for satisfaction estimation, we use one output neuron with the linear activation function; for the TEQ values estimation, we use two output neurons with the linear activation function.

For model optimization, we use the *Adam* optimizer with the learning rate of 0.01. The loss function is selected according to a task: for classification we use the categorical cross-entropy, for regression – the root mean squared error (RMSE).

Due to the high imbalance of emotional classes, we do not perform nine class classification, as for several of them it would be hardly possible to split the data samples in such a way, that examples of particular class were present in both subsets. We perform three class classification

Part	Number of Participants	Number of Sessions	Missing data			Labels[7]
			Pupil Labs	Empatica	SenStick	
Ulm	18	132	12.1%	41.7%	11.3%	S(1) + E(1)
Nara	5	35	66.7%	100.0%	0.0%	S(1) + E(1)
Kyoto	24	264	0.4%	0.0%	0.0%	S(3) + E(2) + TEQ
Total	**47**	**431**	**9.5%**	**20.9%**	**3.5%**	

Table 6.2: Data statistics of EmoTourDB

[7]Number in parentheses shows the amount of scales used (see Sec. 6.2.3 and Sec. 6.2.3 for additional explanations). *E(1)* implies using only *precise level* of emotional labelling, S(1) implies using only *"Are you satisfied?"* level of satisfaction labelling.

instead, splitting labels into *positive* (excited, pleased/happy and calm/relaxed), *neutral* and *negative* (sleepy/tired, bored/depressed, disappointed, distressed/frustrated, afraid/alarmed) classes. While still highly skewed for positive classes, such label representation provides better input data. To decrease the effect of class imbalance, we additionally set weights for each class, adjusting the way that loss function is computed in order to favor samples from the negative or the neutral class. As weights are empirical values, we used nested K-fold cross-validation with $K = 10$ for the first level and $K = 9$ for the second one, to select them. Keeping the test fold held out, we perform another cross-validation (development) within this cross-validation cycle (test) in order to assess an ability of model to generalize with selected weights. Averaged values of performance, obtained within development cross-validation, were used to set weights for test model and compute test metrics. Satisfaction and TEQ estimation are regression tasks, therefore we don't use weights and perform K-fold cross-validation ($K = 10$) directly.

As the performance measure for the emotion classification we use unweighted average recall (UAR); for the satisfaction estimation we use mean absolute error (MAE), as it provides us with a clear understanding of how far away (on the 7-value scale) are our predictions from the original labels; for the TEQ prediction we use RMSE, as it shows the shortest distance between the prediction and the original labels in the TEQ space. Note that we aim for the highest metric value for the emotion recognition and the lowest for the satisfaction and the TEQ estimation. The chance-level performance for a random system would be 0.33 for the emotion recognition, 1.608 for the satisfaction and 0.610 for the TEQ estimation. In case of a "smarter" random system, the chance-level performances are 0.33 for the emotion recognition, 1.171 for the satisfaction and 0.564 for the TEQ estimation. It is a system that (i) for the emotion recognition always selects the majority class of the train data, (ii) for the satisfaction estimation always provides the mean value of the train data, (iii) for the TEQ – also provides the mean values of the train data on both scales. It is worth noticing, that models for the TEQ estimation are trained on much smaller samples, as most of the data doesn't have the TEQ annotations.

For a multimodal fusion, we use feature- and decision-level. For the feature-level fusion, we merge feature matrices of each available modality and perform the recognition task with this new matrix. For the decision-level fusion, we use the weighted linear combination of systems with weights selection on the development set. For the emotion recognition we use class scores of the output from the uni-modal models obtained previously, for the satisfaction and the TEQ estimation – the output values directly.

The results for the uni-, bi-, tri- and multimodal systems are presented in Table 6.3.

In the following, we will take a closer look at the results, considering each modality group and task. Head tilt features showed relatively high performance for the emotion recognition, compared to other modalities in the group of behavioral features, as well as their combinations. Similarly, it shows good results compared to other uni-modal systems on the task of satisfaction recognition, but performs slightly poorer compared to modality combinations. On the task of TEQ recognition, we may notice a different picture: head tilt features perform worse and the best uni-modal system is built with the footstep features.

Considering audio-visual features, one may notice that the low-level features (averaged from recordings) perform better than the high-level ones (obtained with cross-corpus learning with PCA-CCA domain adaptation) in most of the cases. A possible reason for it may be in the difference between tasks of train and target corpora. While for train databases (RECOLA, SEMAINE and SEWA) labels represent an emotional status of the speaker at *current* moment, in EmoTourDB they represent participant's feeling and attitude to the sight visited during

Modality	Emotions (UAR)	Satisfaction (MAE)	TEQ (RMSE)
Head tilt	0.458	1.105	0.508
Eye movements	0.417	1.148	0.496
Footsteps	0.429	1.260	0.489
Head tilt + footsteps	0.440	1.101	0.507
Eye movements + head tilts	0.419	1.101	0.498
Eye movements + head tilts + footsteps	0.452	1.104	0.496
Audio (low level)	0.484	1.197	0.473
Video (low level)	0.413	1.202	0.470
Audio (high level)	0.431	1.236	0.494
Video (high level)	0.438	1.235	0.501
Audio + video (low level)	0.434	1.160	0.484
Audio + video (high level)	0.412	1.323	0.503
Object detection	0.401	1.355	0.542
Empatica	0.390	1.293	0.525
Feature-level fusion (all modalities)	0.451	1.083	0.506
Decision-level fusion (weighted)	**0.513**	**1.033**	**0.458**

Table 6.3: Experimental results of uni-, bi-, tri- and multimodal systems on EmoTourDB. Values are obtained with K-fold cross validation ($K = 10$).

the latest session, e.g. at *previous* moments. Nevertheless, the high-level features perform reasonably well compared to some other modality groups. Feature-level fusion of audio and visual features results in the performance gain only in one case – for satisfaction recognition with low-level features. In other cases, uni-modal systems show higher values.

Object detection features show low performance for each task and even lowest among all modalities for the regression tasks – satisfaction and TEQ estimation. These features may be less informative for several reasons: (i) participants may be indifferent to crowdness level, amount of cars, etc., and concentrate on the sights directly; (ii) most of the target objects (sights) were not recognized by the system, as only generic classes are covered with the used API; (iii) the task of deriving dependencies between these features and emotional response of the participant cannot be solved with these amount of data.

Features extracted from the data of Empatica E4 Wristband showed the second-worst performance for the regression tasks and the worst for the emotion recognition. Three time series used from the tracker represent the skin temperature, the heart rate and the electro-dermal activity. While these features are proven to be suitable for emotion recognition tasks, one should note that most of the previous studies were performed in the laboratory conditions. All these parameters have strong connection with physical activity of a person at current moment in time (Menghini et al., 2019). Changes in emotions may cause heart rate change, but it is insignificant compared to such change caused by a simple walk. Similar applies to the temperature and the electro-dermal activity.

Feature-level fusion of all available modalities shows higher performance compared to most of the uni-modal systems, but not outperforms all of them for emotion recognition and TEQ estimation. For satisfaction – one may notice the performance gain over each uni-modal system. It may be caused by controversies in features of various modalities and increased size of input vector, with that the model has not enough samples to train onto. The situation is different with the decision-level fusion. For each task, it shows the highest results among all uni- and multimodal systems. The weighting procedure selects the optimal combination of weights to incorporate confidence scores of each unimodal system into a comprehensive

picture and profit from all of them. An example of weights selected for emotion recognition are: 0.131 for audio low-level, 0.102 for audio high-level, 0.060 for Empatica, 0.099 for eyes movements, 0.003 for object detection, 0.172 for head tilts, 0.110 for video low-level, 0.115 for video high-level and 0.208 for footsteps. One may notice, that highest priority was given to the footsteps, the head tilts and the audio low-level features. Each of these modalities provided relatively high performance in unimodal setup. Lowest priority was given to the object detection and the Empatica features – which is also coherent to the unimodal results.

In general, one may notice that behavioral and audio-visual features provide a good foundation to perform analysis of emotions, satisfaction and TEQ of a user. While some modalities may provide relatively high performance in a unimodal setup, decision-level fusion is a preferable approach to profit from all the available data.

6.4 Discussions and Limitations

Having the experimental setup, where general mood assumed to be positive by default and only self-annotation is possible, we have a great diversity in rating ranges of participants. Some of them were quite critical and used almost the whole emotions list, some of them have chosen either "happy" or "pleased" even if they didn't express emotions on the sight in feedback videos. The surveys were specially aimed to distinguish between the different types of personalities and correct this bias. However, current set of data does not allow to perform reliable distinctions and need to be extended in the future. While imbalance in emotional classes is nevertheless still likely to be present in an extended data collection, additional "negative" and "neutral" samples may help to significantly increase the quality of the recognition system and allow adjustments with the survey data.

In spite of the fact, that we planned our experiments to be as close to real life conditions as possible, one may not ignore that some bias in behavior of participants and their labeling is still present. Eye tracker wired to a laptop in a backpack is a relatively uncomfortable thing to wear with unusual sensation for those who don't wear glasses in everyday life and even less comfortable for those who do, as tracker and glasses hinder each other's appropriate fit. It is easier with other devices, as smartphones and smart watches or fitness trackers are very common nowadays and don't cause much inconvenience. Regarding the labels, participants were advised to be critical in their evaluations, however, in some cases they reported rather not their impressions of visited sight, but the general emotional status or mood. One may notice that by observing inconsistencies in TEQ graph with respect to emotional classes (Fig. 6.12b). Some samples are annotated as "excited" or "happy/pleased", but have very low value of "interest". This supports the necessity of the specific annotation space for the studied use case.

Being the most technically demanding part of the designed data collection and processing system, developed raw data acquisition setup is hardly imaginable to operate in real life. However, the amount of connection devices used in this experiment is exaggerated in order to increase the robustness of data collection and not to lose the data from several modalities at the same time in case of technical problems. In the future, it can be significantly reduced and combined to one device running the applications simultaneously. Furthermore, recent developments of devices allow us to use them in more natural conditions, e.g. *Pupil Labs* released *Pupil Invisible* – eye tracker that looks similar to a normal pair of glasses and does not require setup and calibration, in contrast to *Pupil Core* used in this experiment.

6.5 Summary

In this chapter, we presented the use case of environmental context usage for the emotion recognition or satisfaction estimation task – a sightseeing tour. We designed the data acquisition setup, developed feature extraction algorithms in addition to existing ones and tested them with several uni- and multimodal systems.

Thus, at the beginning of the chapter, we provided a brief introduction to smart tourism – a concept that is closely connected to our problem statement. We emphasized that most of current approaches to smart tourism are not human-centered, and lack the information required for their personalization to the needs and preferences of a particular user. This may be overcome by introducing emotions as an input feature.

Then, we presented the collected database – EmoTourDB – and covered such aspects as the experimental location, data collection pipeline, used devices, extracted features, labels, additional surveys of participants, data synchronization and missing data. We used five sources of data: (i) eye movements based on analysis of the eye tracker data; (ii) head tilts and footsteps based on analysis of the gyroscope and accelerometer; (iii) audio-visual data collected with the smartphone by recording of short videos; (iv) object detection features based on analysis of video frames from the world camera of the eye tracker; and (v) physiological features based on analysis of the heart rate, temperature and electro-dermal activity. For labelling we used (i) emotion classes in three and nine categories; (ii) satisfaction level on 3 scales; (iii) touristic experience quality – a specially designed coordinate system, representing interest and meeting expectations of the user.

Further, we trained systems on each available modality, logically combining them into bi- and trimodal models using feature-level fusion, as well as multimodal system with all available modalities in feature-level and decision-level fusion setup. Results showed that the performance significantly increases over the chance-level, on each task. Unimodal systems trained on the head tilt and the audio features, showed the highest performance for the emotion recognition; trained on the head tilt and the eye movements – for the satisfaction estimation; and trained on the audio-visual features – for the touristic experience quality estimation. Feature-level fusion of all modalities showed the performance gain over the best unimodal system only for the satisfaction estimation. However, weighted decision level fusion showed much higher results, especially for the emotion recognition. Weights of built linear combination system cohere with unimodal results, favoring the top performing feature sets and models.

Finally, we discussed the limitations of the applied approach and possible future improvements. With the data collected, we hope to make a step towards continuous automatic emotion estimation of tourists. The provided feature sets and labels present a good starting point for researchers in the area of human-centered smart tourism, covering different aspects of a person's behavior and personality. Wide range of nationalities and several experimental fields make the dataset closer to real-world data representation and open a room for the analysis of cultural differences.

7 Conclusion and Future Directions

In this chapter, we sum up the thesis by concluding the main achievements of this work and listing the most promising directions for further research on the topic of contextual emotion recognition. Firstly, we summarize the general results and highlight the essential findings. Secondly, we categorize thesis contributions into three groups: theoretical, practical and experimental. Finally, we present our vision of future research.

7.1 Overall Summary

In this work, we addressed the problem of contextual time-continuous emotion recognition. Highlighting the fact that in real life emotions are changing continuously over time, we focused on this type of data and made use of five corpora with audio and video data, annotated in arousal and valence. Corpora were collected in different languages (English, French, German, Hungarian, Japanese) and conditions, but consist of recordings of spontaneous interaction between users. Time-continuously annotated data requires additional preprocessing of labels, which was performed using a combination of several existing approaches. We evaluated developed systems in uni- and multimodal, single- and cross-corpus scenarios, as well as proposed several approaches to modeling and data processing.

We mentioned that context of the user may be analyzed at three levels:

- Speaker level: when previous words, facial expressions or behaviour of the user (speaker) are considered;
- Dialogue level: when data from the conversational partner of the user (i.e. his/her interlocutor) is considered;
- Environmental level: when effect of surroundings on user's emotional status is considered.

In this thesis we made use of all these levels. First, to ensure correct processing of time-continuous data, in Section 3.2.3 we proposed combined approach to labels and features alignment based on automatic detection and correction of annotator's reaction lag. User's data was cleaned from data of his/her interlocutor using provided timings or by applying a noise gate filter. Then, in Chapter 4 we proposed several approaches to model context using time-continuous data at each level of recognition pipeline, namely (i) feature extraction, (ii) data preprocessing, and (iii) modeling. One can adjust the amount of context at the feature extraction stage by varying a width of a feature (or functional) extraction window. This approach allows taking control over used context length for any method, including classical machine learning algorithms, such as linear regression. However, this approach has limitation of reasonable amount of context due to a loss of data dynamics. Using recurrent models, e.g. recurrent neural network, one may adjust a context length at the modeling stage by changing the number of time steps (frames or feature vectors) fed into the model as one data sample. For adjustments at the data preprocessing stage, we developed a simple yet flexible

and effective approach – data sparsing – which introduces an additional degree of freedom by controlling hop size between time steps within one sample. We tested these approaches to context modeling in single- and cross-corpus scenario on three time-continuous corpora of audio-visual data.

Subsequently, in Chapter 5 we proposed several approaches to introduction of conversational partner's (interlocutor's) data into emotion recognition system. By incorporating this information, we assumed an increase in accuracy of the user's (speaker's) emotion recognition. Proposed approaches were based on the feature-level and decision-level fusion and separated into two groups: dependent and independent. More precisely, for dependent approaches one regulates the amount of context for speaker and interlocutor simultaneously, setting them to exactly the same value of a particular range; for independent – these values may differ and do not have to be the same for speaker and interlocutor. The independent approach was based on data sparsing concept introduced in the previous chapter. We tested these approaches on four corpora of audio-visual data.

Finally, in Chapter 6 we studied the effect of environmental context on user's emotions. As this is a very broad term that covers various aspects, we focused on a specific use case, namely, a sightseeing tour. First, we designed a data collection pipeline, then collected a database using several recording devices, such as camera, eye tracker, sensor board and smart watch. Data from each source was analyzed, and specific feature sets were designed. We tested emotion recognition systems in uni- and multimodal scenarios, using several approaches to data fusion. As target values, we used user's emotions in three groups, satisfaction level and touristic experience quality – a novel annotation space, developed specifically for a touristic domain and aimed to eliminate some drawbacks of classical categorical annotation of emotions.

In the following section, we will present the main contributions of this thesis.

7.2 Thesis Contributions

All contributions achieved in context of this thesis fall within one of the following three groups: theoretical, practical or experimental.

7.2.1 Theoretical

1. A flexible approach to speaker context modeling – data sparsing – was proposed. By increasing the hop between time steps within sample and reducing an amount of data required to cover a certain context length, a step towards a flexible context modeling is made. Combined with varying time steps of recurrent models and data frequency at preprocessing step, it offers even higher flexibility in context selection, allowing many alternative combinations to represent the same amount of input data. This approach was first described in (Fedotov et al., 2018) and then used in several developed emotion recognition system for different problem statements (Fedotov et al., 2018b; Kaya et al., 2018; Fedotov et al., 2018; Matsuda et al., 2018a; Kaya et al., 2019; Verkholyak et al., 2019).
2. Existence of strong dependencies between the amount of modelled context and the system performance was shown in single- and cross-corpus scenario, for time-dependent and time-independent models, for speaker-only data and for dialogue data (Fedotov et al., 2018). Moreover, we came to conclusions that (i) the optimal amount of data is

individual and dependent on model type (recurrent models are more tolerant to a larger context coverage), corpus and modality (video modality requires lower amount of the context than audio), (ii) however, this amount is not dependent on the feature set, the number of time steps for recurrent neural network and the data frequency. When using cross-corpus setting, contextual dependencies are usually inherited from the train or test corpus.

3. Using the data sparsing concept, we developed several approaches to a time-continuous emotion recognition in dyadic interactions. They allow dependent and independent context modeling. The advantage of these approaches is their ability to be used not only with the decision-level but also with the feature-level fusion of data from both interlocutors, which eliminates the need of additional meta system, combining the single-speaker predictions, increasing flexibility and reducing the amount of sources for possible errors. A model based on one of the described approaches, was used as a subsystem for (Kaya et al., 2019). Independence of context modeling may be represented on two levels: with the fixed amount of context for speaker and varying for interlocutor and when both context lengths are a subject to change (fully independent dyadic context modeling).
4. It was demonstrated that the incorporation of the interlocutor's data into the emotion recognition system, may significantly improve its performance over a single-speaker baseline. The most improvements were obtained by fully independent dyadic context modeling, which also opens a room for more precise and flexible adjustment. However, the fixed context for speaker and even dependent approaches offer a good start and are significantly less resource-demanding.
5. A novel domain-specific labelling space – Touristic Experience Quality (TEQ) – was proposed. Represented by two axes, namely, "interest" and "meeting expectations", it makes a room for precise annotations and, in contrast to classical labelling approach with emotion categories, allows representation of many different user states, while, in contrast to arousal-valence space, allows easier understanding and provides higher suitability for the considered use case. For example, TEQ enables to distinguish between "good" and "bad" surprise, and also is not affected by physical state of a user (e.g. tired), eliminating this bias.

7.2.2 Practical

1. A system for flexible context modeling based on recurrent neural networks was implemented (Fedotov et al., 2018). It uses three steps of modeling pipeline to define the length of the contextual window for each data sample. As a result, one may adjust these parameters to fit any requirements on the input data size, the data frequency or the feature extraction window size.
2. Combining this approach with PCA-CCA domain adaptation, we implemented the system for flexible context modeling that works also in a cross-corpus scenario. All three levels of context adjustment are supported in this system as well.
3. Several systems for emotion recognition in dyadic interactions based on time-continuous feature-level and decision-level fusion with dependent and independent context regulation were implemented (Kaya et al., 2019). Similarly to speaker-only systems, these allow flexible context modeling on three stages of the recognition pipeline. Independent approach based on the feature-level fusion is possible due to the data sparsing.
4. A first of its kind database of emotionally labelled touristic behaviour – EmoTourDB

– was collected (Fedotov et al., 2018; Matsuda et al., 2018c). The database consists of recordings from 47 participants from 12 countries. Behavior was collected with a camera, an eye tracker, a sensor board and a smart watch. Sessions are annotated in emotions on two scales, in satisfaction on three scales, as well as in touristic experience quality.

5. Several uni-, bi-, tri- and multimodal systems for emotion recognition based on behavioral data were proposed. Systems for audio-visual data were combined with contextual cross-corpus approaches. Modality fusion was performed on feature-level, as well as on decision-level with weighted linear model (Matsuda et al., 2018b).

7.2.3 Experimental

1. Speaker-level context modeling systems were evaluated on three corpora of time-continuous spontaneous interactions: RECOLA, SEMAINE and SEWA. Each performance score was obtained on a test set using concordance correlation coefficient as a measure. Performances for each combination of sparsing coefficient, number of time steps, data frequency, etc. were averaged over 10 sampling partitions kept throughout this thesis and selected in a stratified manner, i.e. data statistics of train and test set regarding age, gender and other available information is as close as possible to original dataset (whole data). Models were built for two modalities (audio and video) and two annotation dimensions (arousal and valence) (Fedotov et al., 2018).
2. Speaker-level context modeling systems were additionally evaluated in a cross-corpus scenario on each combination of three corpora: RECOLA, SEMAINE and SEWA with the same testing methodology as for single-corpus models (Kaya et al., 2018).
3. Proposed systems for dyadic emotion modelling were evaluated on four corpora of time-continuous spontaneous interactions with the dialogue structure and data available for both conversational partners: SEMAINE, SEWA; IEMOCAP and UUDB. We used the same data partitions as for the speaker-only models, while for some systems an additional subset – development – was required. In this case, it was selected from the train set in order to keep the test part the same for all experiments. Each of four proposed systems was tested on two modalities (audio and video if available) and two annotation dimensions (arousal and valence) (Kaya et al., 2019).
4. Emotion recognition system for smart tourism was evaluated on the collected database – EmoTourDB (Matsuda et al., 2018b). Each modality or modality combination was tested on each of three annotation dimension (emotions, satisfaction, TEQ) using K-Fold cross-validation with $K = 10$.
5. Reaction lag of each annotator for RECOLA and SEMAINE corpora was automatically analyzed, and an optimal correction value based on the results on both audio and video modality for each annotation scale was found.

7.3 Future Directions

This thesis focused on the problem of utilizing contextual information at different levels to improve the quality of an emotion recognition system. However, since this area is broad-ranging, many aspects remain unstudied. In the following, we propose several future directions for further improvements of contextual emotion recognition systems.

The performance of any *time-continuous* emotion recognition system strongly relies on

the quality of the annotations and their alignment with input data or features. We emphasized this issue in Section 3.2.3 by considering the reaction lag. However, approaches to feature-label alignment and gold standard calculation based on the inter-rater agreement are still far from perfect. In turn, without a highly reliable mapping between input and output data one cannot train a time-continuous emotion recognition system with low bias, which hinders such systems to achieve and outperform human-level performance. While the rater's subjectivity will always be the case for emotions, novel approaches to annotation, with shorter reaction lags and high level of agreement between raters is a promising direction for future studies.

The second issue that impedes the research in the area of emotion recognition is the lack of high-quality data of time-continuously annotated emotionally-rich spontaneous interactions. As mentioned above, emotion recognition is an extremely complex task due to the high subjectivity of annotations. In order to be able to extract meaningful dependencies between features and labels, one require large amount of data. However, most of the available databases are not longer than 6-8 hours and many of them are as short as 1-2 hours, which does not seem to be enough to train a well-performing model from scratch. Many databases collected so far, used a laboratory setup with the highly constrained scripts and scenarios. While such databases allowed to build the first low-level systems, they are not suitable for models designed to work in the real world conditions.

In the scope of this thesis, in Chapter 4 we studied the dependencies between the amount of data used by the model and its performance. However, due to the poor variety of databases, we could not test the hypotheses and derive conclusions on the relation between the optimal amount of context and such aspects as language, culture or other corpus-specific features. The language and its grammatical (e.g. a sentence structure) and vocal (e.g. an intonation) peculiarities are likely to have an impact on the audio-based emotion recognition systems. On the other hand, cultural differences affect the range and intensity of facial expressions, which is crucial for the video-based systems. Nowadays, the number of time-continuous databases annotated dimensionally is low and most of them are recorded in European languages, such as English, French or German. New databases with high diversity in characteristics will enable further analysis of dependencies at the level of speaker context.

The range of used methods can also be extended. In this thesis, we worked with recurrent neural networks to analyze the contextual data. However, there are other approaches that take sequential nature of the data into account. Temporal Convolution Neural Networks (TCCNs) have been successfully applied to various machine learning tasks based on time series (Gao et al., 2019; Pelletier et al., 2019; Pandey and Wang, 2019). Apart from deep learning approach, some classical models also allow considering temporal structure of the data. Ouyang et al. (2019) used the Autoregressive Exogenous model (ARX) and achieved performance level comparable to the RNN-LSTM modeling. Orders of an autoregressive and an exogenous sub-model can be used to control the amount of data analyzed by the system.

In Chapter 5 we considered context at the level of conversations. However, only the dyadic interactions (dialogue) was studied in this thesis. The complexity of the emotion recognition task rises significantly with the increase in number of interlocutors. This also offers a wide range of research questions in the area of multi-party conversations. Such questions tackle interaction strategies, group dynamics, individual engagement in group assignments, emotional flow in the group, etc. Multi-party interactions have been the subject of research for interaction strategies (Strauss and Minker, 2010) or addressee detection (Akhtiamov et al., 2017); however, multi-party emotion recognition remains underdeveloped. The Emotiw emotion recognition challenge (Dhall et al., 2016, 2017) introduced a sub-challenge for group emotion recognition, but it is based on static photos. Moreover, there are databases for

group emotion recognition (Chen et al., 2018; Poria et al., 2018; Li et al., 2017), but most of them are not multimodal and annotated categorically. Hence, they are less suitable for the contextual group emotion recognition.

Furthermore, since the environmental context is the widest level of contextual information, it yields the highest potential of improvement. Future directions of study on the use case of smart tourism considered in Chapter 6 include the extension of EmoTourDB with a special focus on "neutral" and "negative" cases. It can be achieved by participation of people that are less impressed by particular sights, e.g. local residents. The data collected in this scenario may be used as a baseline behaviour, in contrast to a behaviour of the actual tourists, presented in this work. Additional studies on feature sets or suitable modalities may help to improve the recognition system as well. Finally, new use cases will enable to look at the problem of the contextual emotion recognition from different angles. Some examples include stress of office workers, emotions in sport competitions, emotions while driving, etc.

In conclusion, a change in view on user's surroundings is required: the data that was considered as "noise" in earlier research should be used for further information extraction and improvements of recognition systems. This will allow artificial systems to reach and surpass the human-level performance in emotion recognition tasks.

A Heat maps representation of performance graphs

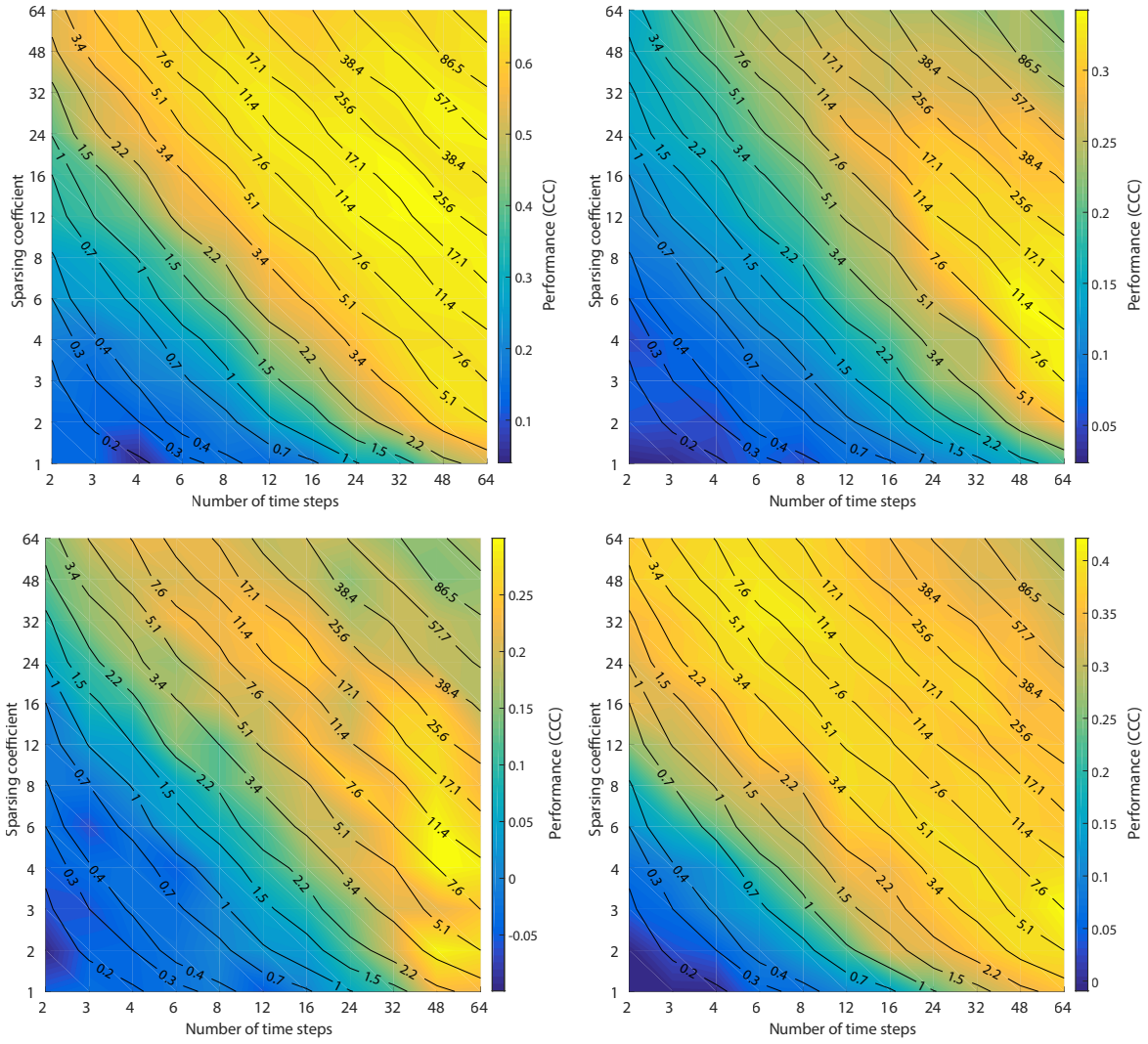


Figure A.1: Performance heat maps for data sparsing approach with RNN-LSTM model and RECOLA database for the original data frequency of 25Hz. Top-left: audio-arousal; top-right: audio-valence; bottom-left: video-arousal; bottom-right: video-valence. Performance is measures in CCC and depicted with color.

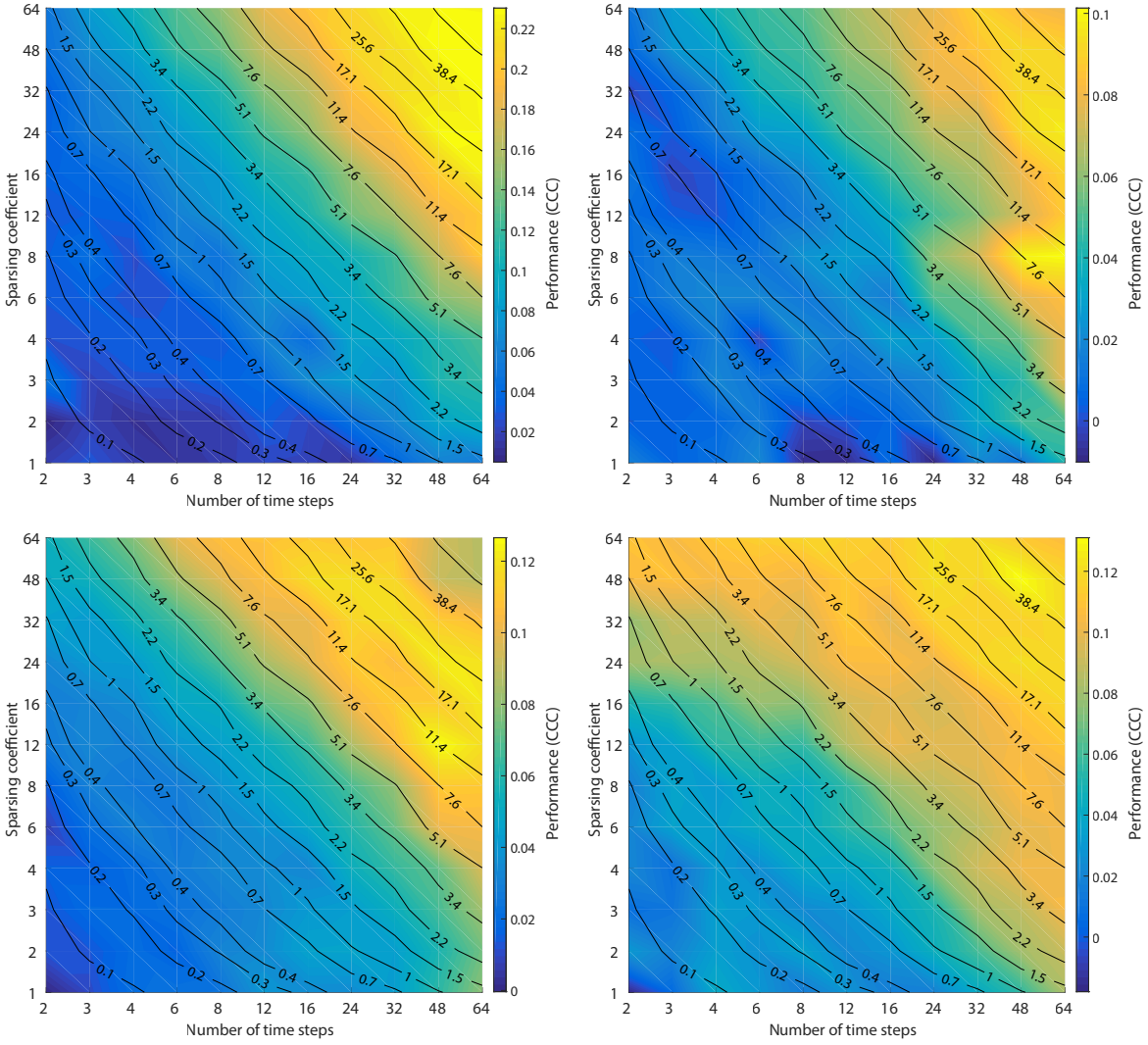


Figure A.2: Performance heat maps for data sparsing approach with RNN-LSTM model and SEMAINE database for the original data frequency of 50Hz. Top-left: audio-arousal; top-right: audio-valence; bottom-left: video-arousal; bottom-right: video-valence. Performance is measures in CCC and depicted with color.

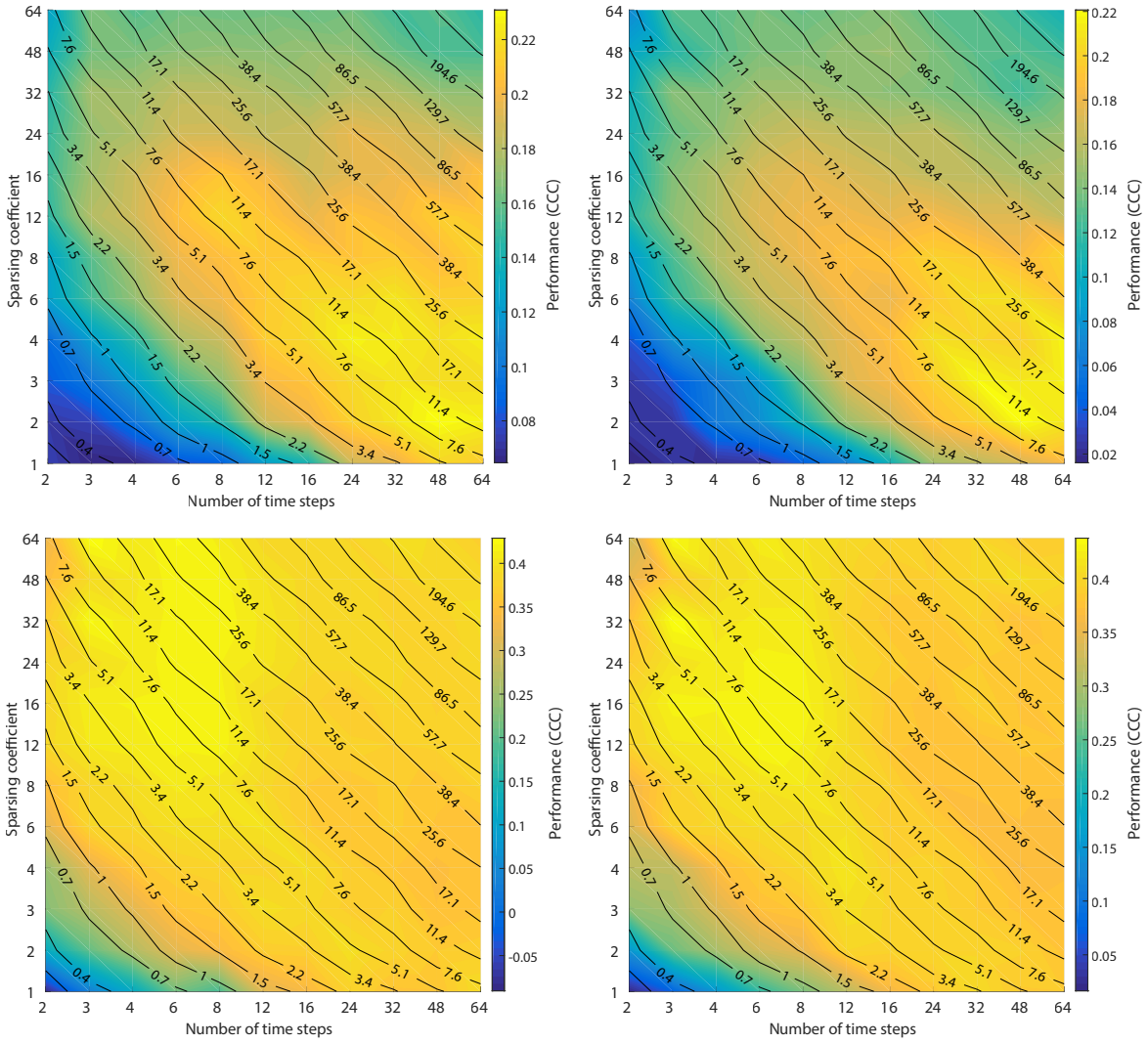


Figure A.3: Performance heat maps for data sparsing approach with RNN-LSTM model and SEWA database for the original data frequency of 10Hz. Top-left: audio-arousal; top-right: audio-valence; bottom-left: video-arousal; bottom-right: video-valence. Performance is measures in CCC and depicted with color.

B Additional results for speaker context modeling in cross-corpus scenario

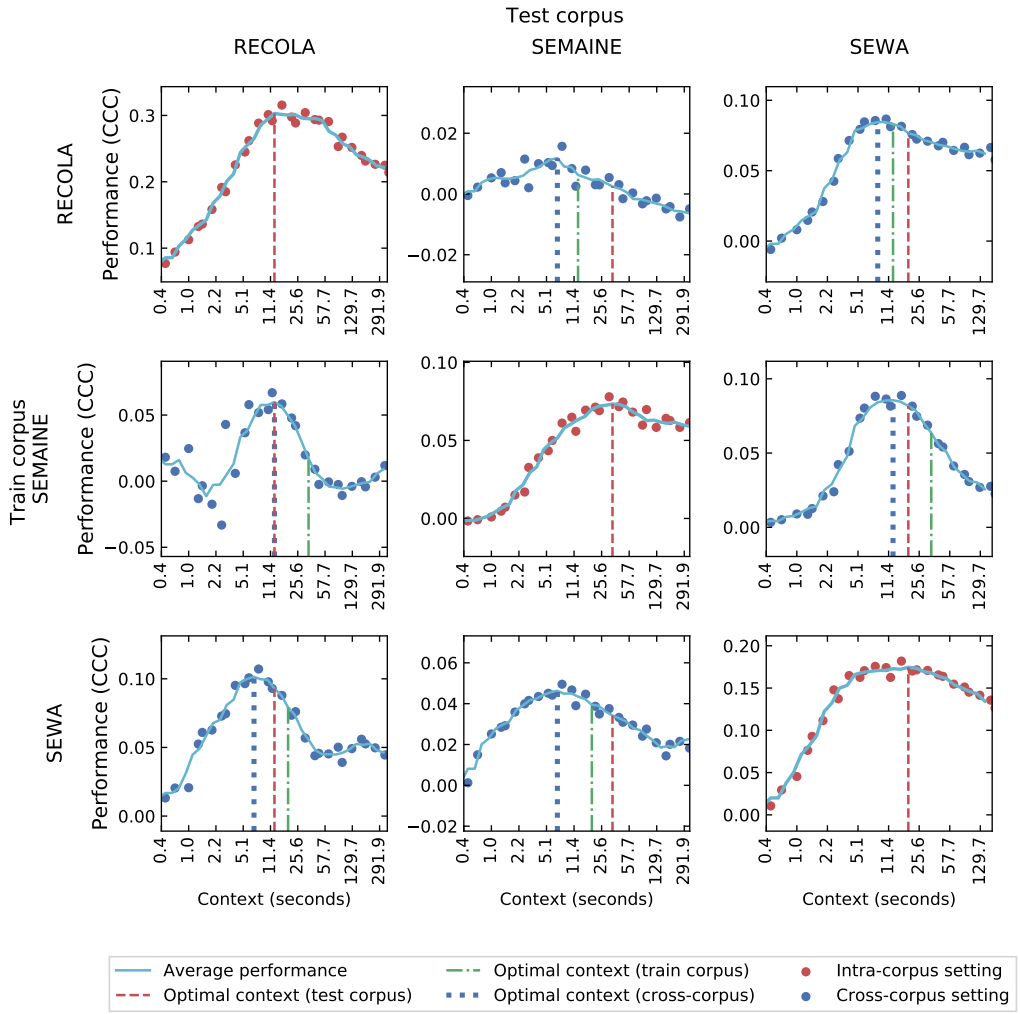


Figure B.1: Results for data sparsing approach with varying feature window to cross-corpus context modeling on audio modality (eGeMAPS), valence dimension with feature based time-dependent models. Red dashed line represents optimal context amount for test corpus, green dot-dashed line – for train corpus, blue dotted line – for cross corpus setting. The performance is measured in terms of CCC and is dependent on amount of context represented by number of time steps used in RNN, data frequency and sparsing coefficient.

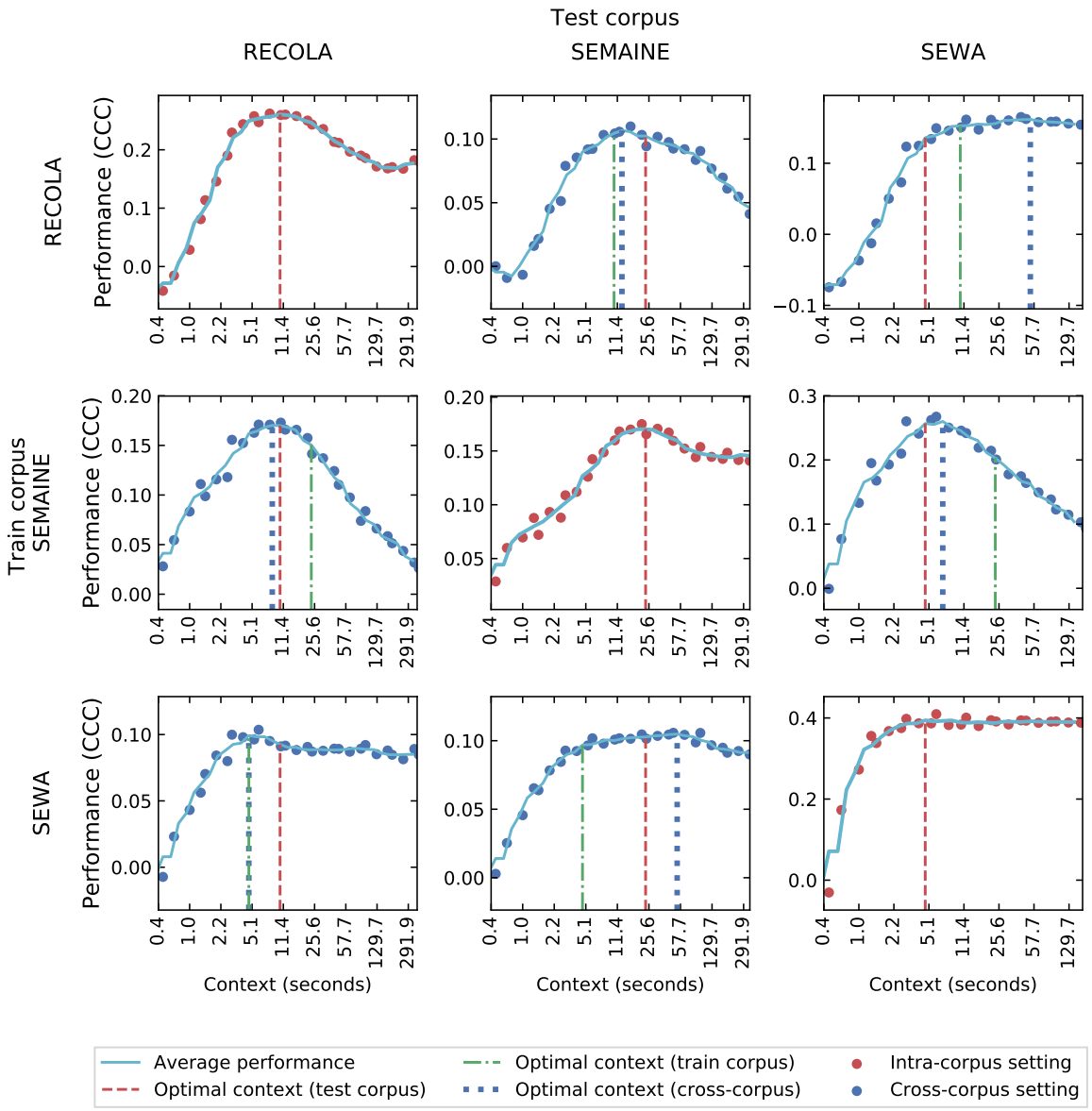


Figure B.2: Results for data sparsing approach with varying feature window to cross-corpus context modeling on video modality (AUs), arousal dimension with feature based time-dependent models. Red dashed line represents optimal context amount for test corpus, green dot-dashed line – for train corpus, blue dotted line – for cross corpus setting. The performance is measured in terms of CCC and is dependent on amount of context represented by number of time steps used in RNN, data frequency and sparsing coefficient.

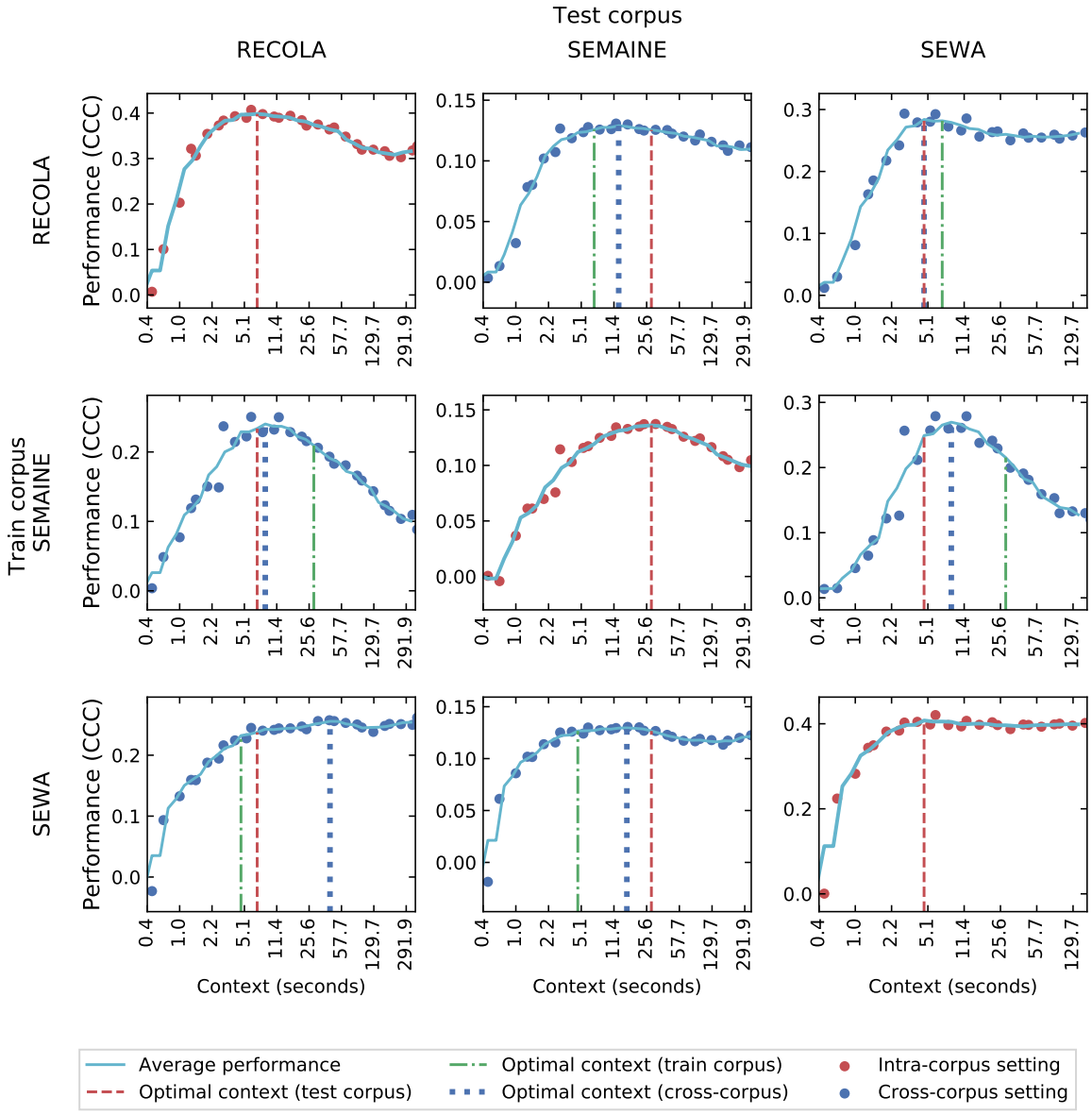


Figure B.3: Results for data sparsing approach with varying feature window to cross-corpus context modeling on video modality (AUs), valence dimension with feature based time-dependent models. Red dashed line represents optimal context amount for test corpus, green dot-dashed line – for train corpus, blue dotted line – for cross corpus setting. The performance is measured in terms of CCC and is dependent on amount of context represented by number of time steps used in RNN, data frequency and sparsing coefficient.

C Additional results for sparsing analysis in speaker context modeling

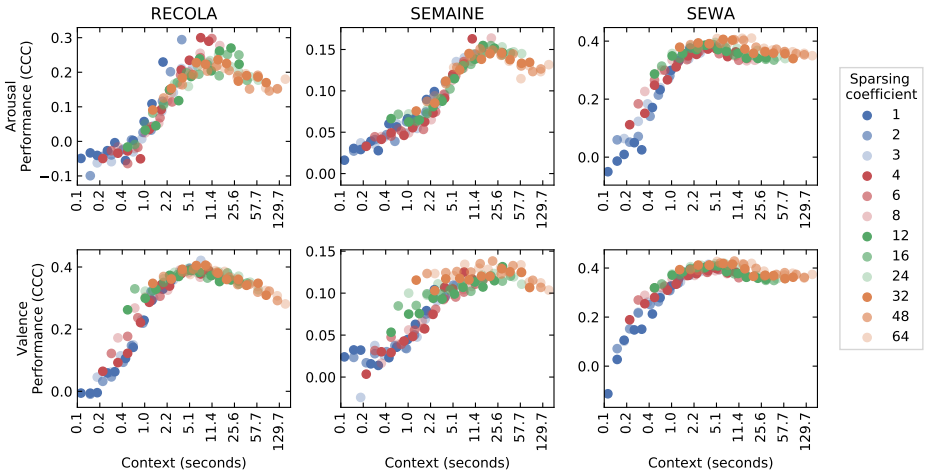


Figure C.1: Full graph for results of context modeling for video modality (AUs) presented at Fig. 4.20, separated by sparsing coefficient used (depicted with color) for the frequency of 25Hz

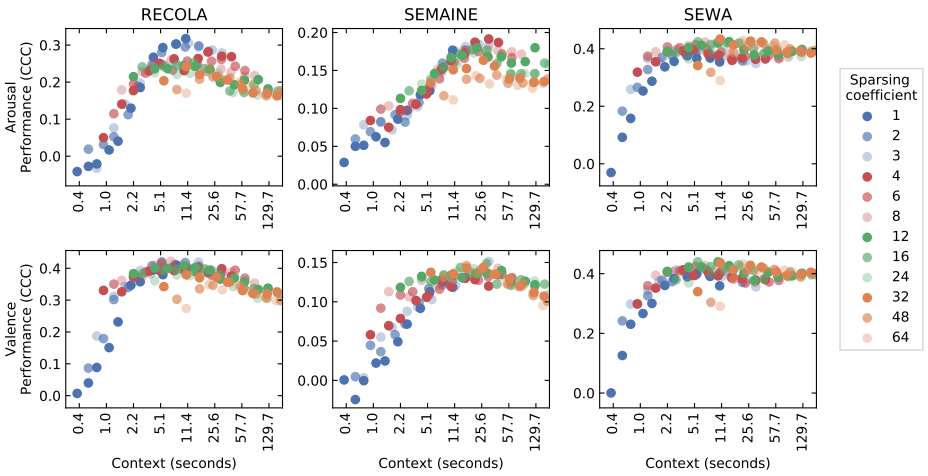


Figure C.2: Full graph for results of context modeling for video modality (AUs) presented at Fig. 4.20, separated by sparsing coefficient used (depicted with color) for the frequency of 6Hz

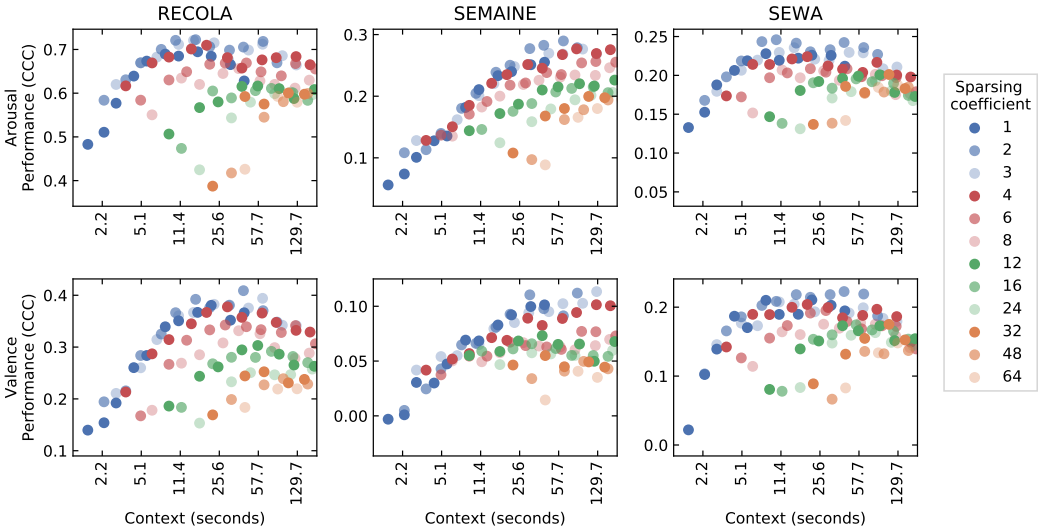


Figure C.3: Full graph for results of context modeling for audio modality (eGeMAPS) presented at Fig. 4.20, separated by sparsing coefficient used (depicted with color) for the frequency of 1.5Hz

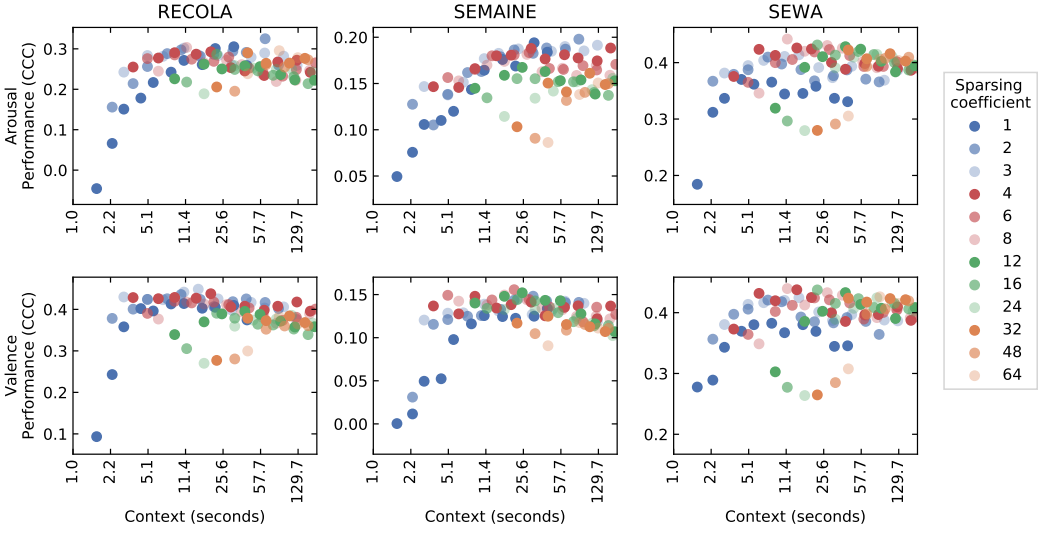


Figure C.4: Full graph for results of context modeling for video modality (AUs) presented at Fig. 4.20, separated by sparsing coefficient used (depicted with color) for the frequency of 1.5Hz

References

Akhtiamov, O., Sidorov, M., Karpov, A. A., and Minker, W. (2017). Speech and text analysis for multimodal addressee detection in human-human-computer interaction. In *INTERSPEECH*, (pp. 2521–2525).

Altshuller, G., Altov, G., and Altov, H. (1996). *And suddenly the inventor appeared: TRIZ, the theory of inventive problem solving*. Technical Innovation Center, Inc.

Alvarado, N. (1997). Arousal and valence in the direct scaling of emotional response to film clips. *Motivation and Emotion*, *21*(4), 323–348.

Amornpashara, N., Arakawa, Y., Tamai, M., and Yasumoto, K. (2015). Landscape photo classification mechanism for context-aware photography support system. In *2015 IEEE International Conference on Consumer Electronics (ICCE)*, (pp. 663–666). IEEE.

Arakawa, Y. (2019). Sensing and changing human behavior for workplace wellness. *Journal of Information Processing*, *27*, 614–623.

Aspandi, D., Mallol-Ragolta, A., Schuller, B., and Binefa, X. (2020). Adversarial-based neural network for affect estimations in the wild. *arXiv preprint arXiv:2002.00883*.

Atmaja, B. T. and Akagi, M. (2020). Multitask learning and multistage fusion for dimensional audiovisual emotion recognition. *arXiv preprint arXiv:2002.11312*.

Aytar, Y., Vondrick, C., and Torralba, A. (2016). Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*, (pp. 892–900).

Baron-Cohen, S. (2007). Mind reading: the interactive guide to emotions–version 1.3. *Jessica Kingsley, London*.

Beilock, S. (2015). *How the body knows its mind: The surprising power of the physical environment to influence how you think and feel*. Simon and Schuster.

Bengio, Y., Frasconi, P., and Simard, P. (1993). The problem of learning long-term dependencies in recurrent networks. In *IEEE international conference on neural networks*, (pp. 1183–1188). IEEE.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, *5*(2), 157–166.

Berntson, G. G., Thomas Bigger Jr, J., Eckberg, D. L., Grossman, P., Kaufmann, P. G., Malik, M., Nagaraja, H. N., Porges, S. W., Saul, J. P., Stone, P. H., et al. (1997). Heart rate variability: origins, methods, and interpretive caveats. *Psychophysiology*, *34*(6), 623–648.

Bone, D., Lee, C.-C., and Narayanan, S. (2014). Robust unsupervised arousal rating: A rule-based framework withknowledge-inspired vocal features. *IEEE transactions on affective computing*, *5*(2), 201–213.

Borràs, J., Moreno, A., and Valls, A. (2014). Intelligent tourism recommender systems: A survey. *Expert Systems with Applications*, *41*(16), 7370 – 7389.

Brester, C., Semenkin, E., and Sidorov, M. (2016). Multi-objective heuristic feature selection for speech-based multilingual emotion recognition. *Journal of Artificial Intelligence and Soft Computing Research*, *6*(4), 243–253.

Buhalis, D. and Amaranggana, A. (2013). Smart tourism destinations. In *Information and communication technologies in tourism 2014* (pp. 553–564). Springer.

Buhalis, D. and Amaranggana, A. (2015). Smart tourism destinations enhancing tourism experience through personalisation of services. In *Information and communication technologies in tourism 2015* (pp. 377–389). Springer.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B. (2005). A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, *42*(4), 335.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, (pp. 67–74). IEEE.

Chanel, G., Kierkels, J. J., Soleymani, M., and Pun, T. (2009). Short-term emotion assessment in a recall paradigm. *International Journal of Human-Computer Studies*, *67*(8), 607–627.

Chang, C.-Y., Chang, C.-W., Zheng, J.-Y., and Chung, P.-C. (2013). Physiological emotion analysis using support vector regression. *Neurocomputing*, *122*, 79–87.

Chen, H., Deng, Y., Cheng, S., Wang, Y., Jiang, D., and Sahli, H. (2019). Efficient spatial temporal convolutional features for audiovisual continuous affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, (pp. 19–26).

Chen, M., Zhang, Y., Li, Y., Mao, S., and Leung, V. C. (2015). Emc: Emotion-aware mobile cloud computing in 5g. *IEEE Network*, *29*(2), 32–38.

Chen, S., Jin, Q., Zhao, J., and Wang, S. (2017). Multimodal multi-task learning for dimensional and continuous emotion recognition. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, (pp. 19–26).

Chen, S.-Y., Hsu, C.-C., Kuo, C.-C., Ku, L.-W., et al. (2018). Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, (pp. 785–794).

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 8789–8797).

Chou, H.-C., Lin, W.-C., Chang, L.-C., Li, C.-C., Ma, H.-P., and Lee, C.-C. (2017). Nnime: The nthu-ntua chinese interactive multimodal emotion corpus. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, (pp. 292–298). IEEE.

Chronaki, G., Hadwin, J. A., Garner, M., Maurage, P., and Sonuga-Barke, E. J. (2015). The development of emotion recognition from facial expressions and non-linguistic vocalizations during childhood. *British Journal of Developmental Psychology*, *33*(2), 218–236.

Chu, W.-S., De la Torre, F., and Cohn, J. F. (2017). Selective transfer machine for personalized facial expression analysis. *IEEE transactions on pattern analysis and machine intelligence*, *39*(3), 529–545.

Coulson, M. (2004). Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of nonverbal behavior*, *28*(2), 117–139.

Cowie, R., Douglas-Cowie, E., Savvidou*, S., McMahon, E., Sawey, M., and Schröder, M. (2000). 'feeltrace': An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*.

Degottex, G., Kane, J., Drugman, T., Raitio, T., and Scherer, S. (2014). Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, (pp. 960–964). IEEE.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788–798.

Denham, S. A. (1998). *Emotional development in young children.* Guilford Press.

Dhall, A. (2019). Emotiw 2019: Automatic emotion, engagement and cohesion prediction tasks. In *2019 International Conference on Multimodal Interaction*, (pp. 546–550).

Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J., and Gedeon, T. (2017). From individual to group-level emotion recognition: Emotiw 5.0. In *Proceedings of the 19th ACM international conference on multimodal interaction*, (pp. 524–528).

Dhall, A., Goecke, R., Joshi, J., Hoey, J., and Gedeon, T. (2016). Emotiw 2016: Video and group-level emotion recognition challenges. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, (pp. 427–432). ACM.

Dhall, A., Goecke, R., Joshi, J., Sikka, K., and Gedeon, T. (2014). Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In *Proceedings of the 16th international conference on multimodal interaction*, (pp. 461–466).

Dhall, A., Goecke, R., Joshi, J., Wagner, M., and Gedeon, T. (2013). Emotion recognition in the wild challenge (emotiw) challenge and workshop summary. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, (pp. 371–372).

Dhall, A., Kaur, A., Goecke, R., and Gedeon, T. (2018). Emotiw 2018: Audio-video, student engagement and group-level affect prediction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, (pp. 653–656).

Dhall, A., Ramana Murthy, O., Goecke, R., Joshi, J., and Gedeon, T. (2015). Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, (pp. 423–426).

Ekman, P. (1999). Basic emotions. *Handbook of cognition and emotion*, *98*(45-60), 16.

Ekman, P., Friesen, W., and Hager, J. (2002). Facs manual. *A human face*.

Ekman, P. and Friesen, W. V. (1978). *Manual for the facial action coding system*. Consulting Psychologists Press.

Elbarougy, R. and Akagi, M. (2014). Improving speech emotion dimensions estimation using a three-layer model of human perception. *Acoustical science and technology*, *35*(2), 86–98.

Eskenazi, M., Black, A. W., Raux, A., and Langner, B. (2008). Let's go lab: a platform for evaluation of spoken dialog systems with real world users. In *Ninth Annual Conference of the International Speech Communication Association*.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al. (2016). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, *7*(2), 190–202.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: the Munich versatile and fast open-source audio feature extractor. In *Proc. of the 18th ACM international conference on Multimedia*, (pp. 1459–1462). ACM.

Fedotov, D., Ivanko, D., Sidorov, M., and Minker, W. (2018). Contextual dependencies in time-continuous multidimensional affect recognition. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, (pp. 1220–1224).

Fedotov, D., Kaya, H., and Karpov, A. (2018b). Context modeling for cross-corpus dimensional acoustic emotion recognition: Challenges and mixup. In *International Conference on Speech and Computer*, (pp. 155–165). Springer.

Fedotov, D., Matsuda, Y., Takahashi, Y., Arakawa, Y., Yasumoto, K., and Minker, W. (2018). Towards estimating emotions and satisfaction level of tourist based on eye gaze and head movement. In *2018 IEEE International Conference on Smart Computing (SMARTCOMP)*, (pp. 399–404). IEEE.

Fischer, L., Brauns, D., and Belschak, F. (2002). Zur messung von emotionen in der angewandten forschung. *Lengerich: Pabst Science Publishers*.

Fontaine, J. R., Scherer, K. R., Roesch, E. B., and Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological science*, *18*(12), 1050–1057.

Friesen, E. and Ekman, P. (1978). Facial action coding system: a technique for the measurement of facial movement. *Palo Alto*, *3*.

Gao, Z., Wang, X., Yang, Y., Mu, C., Cai, Q., Dang, W., and Zuo, S. (2019). Eeg-based spatio–temporal convolutional neural network for driver fatigue evaluation. *IEEE transactions on neural networks and learning systems*, *30*(9), 2755–2763.

Garbarino, M., Lai, M., Bender, D., Picard, R. W., and Tognetti, S. (2014). Empatica e3—a wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *2014 4th International Conference on Wireless Mobile Communication and Healthcare-Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH)*, (pp. 39–42). IEEE.

Gosling, S. D., Rentfrow, P. J., and Swann Jr, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in personality*, *37*(6), 504–528.

Gretzel, U., Mitsche, N., Hwang, Y.-H., and Fesenmaier, D. R. (2004). Tell me who you are and i will tell you where to go: Use of travel personalities in destination recommendation systems. *Information Technology & Tourism*, *7*(1), 3–12.

Gretzel, U., Reino, S., Kopera, S., and Koo, C. (2015). Smart tourism challenges. *Journal of Tourism*, *16*(1), 41–47.

Gretzel, U., Sigala, M., Xiang, Z., and Koo, C. (2015). Smart tourism: foundations and developments. *Electronic Markets*, *25*(3), 179–188.

Gretzel, U., Werthner, H., Koo, C., and Lamsfus, C. (2015). Conceptual foundations for understanding smart tourism ecosystems. *Computers in Human Behavior*, *50*, 558–563.

Grimm, M., Kroschel, K., Mower, E., and Narayanan, S. (2007). Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, *49*(10-11), 787–800.

Grimm, M., Kroschel, K., and Narayanan, S. (2007). Support vector regression for automatic recognition of spontaneous emotions in speech. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, (pp. IV–1085). IEEE.

Grimm, M., Kroschel, K., and Narayanan, S. (2008a). The vera am mittag german audio-visual emotional speech database. In *2008 IEEE international conference on multimedia and expo*, (pp. 865–868). IEEE.

Grimm, M., Kroschel, K., and Narayanan, S. (2008b). The vera am mittag german audio-visual emotional speech database. In *2008 IEEE international conference on multimedia and expo*, (pp. 865–868). IEEE.

Gunes, H. and Schuller, B. (2013). Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, *31*(2), 120–136.

Guo, X., Polanía, L. F., and Barner, K. E. (2017). Group-level emotion recognition using deep models on image scene, faces, and skeletons. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, (pp. 603–608).

Hall, J. and Watson, W. H. (1970). The effects of a normative intervention on group decision-making performance. *Human relations*, *23*(4), 299–317.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10–18.

Han, K., Yu, D., and Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Fifteenth annual conference of the international speech communication association*.

Haq, S. and Jackson, P. (2010). Machine audition: Principles, algorithms and systems, chapter multimodal emotion recognition. *IGI Global, Hershey PA*, 398–423.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 770–778).

He, L., Jiang, D., Yang, L., Pei, E., Wu, P., and Sahli, H. (2015). Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, (pp. 73–80).

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., et al. (2017). Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, (pp. 131–135). IEEE.

Hidaka, M., Matsuda, Y., Kawanaka, S., Nakamura, Y., Fujimoto, M., Arakawa, Y., and Yasumoto, K. (2017). A system for collecting and curating sightseeing information toward

satisfactory tour plan creation. *The Second International Workshop on Smart Sensing Systems (IWSSS '17)*.

Hjortsjö, C.-H. (1969). *Man's face and mimic language*. Studen litteratur.

Hochreiter, S. (1991). Untersuchungen zu dynamischen neuronalen netzen. *Diploma, Technische Universität München*, *91*(1).

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Holatka, A.-K., Suwa, H., and Yasumoto, K. (2019). Volleyball setting technique assessment using a single point sensor. In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, (pp. 567–572). IEEE.

Huang, C.-F. and Akagi, M. (2008). A three-layered model for expressive speech perception. *Speech Communication*, *50*(10), 810–828.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 4700–4708).

Huang, J., Li, Y., Tao, J., Lian, Z., Niu, M., and Yang, M. (2018). Multimodal continuous emotion recognition with data augmentation using recurrent neural networks. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, (pp. 57–64).

Huang, J., Li, Y., Tao, J., Lian, Z., Wen, Z., Yang, M., and Yi, J. (2017). Continuous multimodal emotion prediction based on long short term memory recurrent neural network. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, (pp. 11–18).

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., et al. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 7310–7311).

Huang, Y. and Bian, L. (2009). A bayesian network and analytic hierarchy process based personalized recommendations for tourist attractions over the internet. *Expert Systems with Applications*, *36*(1), 933 – 943.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Jani, D. (2014). Relating travel personality to big five factors of personality. *Turizam: međunarodni znanstveno-stručni časopis*, *62*(4), 347–359.

Jenke, R., Peer, A., and Buss, M. (2013). A comparison of evaluation measures for emotion recognition in dimensional space. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, (pp. 822–826). IEEE.

Jeong, M. and Shin, H. H. (2019). Tourists' experiences with smart tourism technology at smart destinations and their behavior intentions. *Journal of Travel Research*, 0047287519883034.

Jovicic, D. Z. (2019). From the traditional understanding of tourism destination to the smart tourism destination. *Current Issues in Tourism*, *22*(3), 276–282.

Kahou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç., Memisevic, R., Vincent, P., Courville, A., Bengio, Y., Ferrari, R. C., et al. (2013). Combining modality specific

deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, (pp. 543–550).

Kanaya, Y., Kawanaka, S., Suwa, H., Arakawa, Y., and Yasumoto, K. (2019). Automatic tour video summarization focusing on scene change for advance touristic experience. In *IEEE 2nd International Conference on Multimedia Information Processing and Retrieval*, MIPR'19. IEEE.

Kassner, M., Patera, W., and Bulling, A. (2014). Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, UbiComp '14 Adjunct, (pp. 1151–1160).

Kaya, H., Fedotov, D., Dresvyanskiy, D., Doyran, M., Mamontov, D., Markitantov, M., Akdag Salah, A. A., Kavcar, E., Karpov, A., and Salah, A. A. (2019). Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, (pp. 27–35).

Kaya, H., Fedotov, D., Yeşilkanat, A., Verkholyak, O., Zhang, Y., and Karpov, A. (2018). Lstm based cross-corpus and cross-task acoustic emotion recognition. *Proc. Interspeech 2018*, 521–525.

Keren, G., Deng, J., Pohjalainen, J., and Schuller, B. W. (2016). Convolutional neural networks with data augmentation for classifying speakers' native language. In *INTERSPEECH*, (pp. 2393–2397).

Kessler, V., Schels, M., Kächele, M., Palm, G., and Schwenker, F. (2015). On the effects of continuous annotation tools and the human factor on the annotation outcome. In *ISCT*, (pp. 174–180).

Kipp, M. (2001). Anvil-a generic annotation tool for multimodal dialogue. In *Seventh European Conference on Speech Communication and Technology*.

Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., Han, J., Pandit, V., Schuller, B., Star, K., et al. (2019). Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *arXiv preprint arXiv:1901.02839*.

Koutsombogera, M. and Vogel, C. (2018). Modeling collaborative multimodal behavior in group dialogues: the multisimo corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, (pp. 2945–2951).

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, (pp. 1097–1105).

Lasdon, L. S. (2002). *Optimization theory for large systems*. Courier Corporation.

Lawrence, I. and Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 255–268.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), 541–551.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lee, C.-C., Busso, C., Lee, S., and Narayanan, S. S. (2009). Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions. In *Tenth Annual Conference of the International Speech Communication Association*.

Lee, J.-H. and Hashimoto, H. (2002). Intelligent space—concept and contents. *Advanced Robotics*, *16*(3), 265–280.

Li, J.-L. and Lee, C.-C. (2018). Encoding individual acoustic features using dyad-augmented deep variational representations for dialog-level emotion recognition. *Proc. Interspeech 2018*, 3102–3106.

Li, S., Zheng, W., Zong, Y., Lu, C., Tang, C., Jiang, X., Liu, J., and Xia, W. (2019). Bimodality fusion for emotion recognition in the wild. In *2019 International Conference on Multimodal Interaction*, (pp. 589–594).

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017). Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Lim, K. H., Chan, J., Leckie, C., and Karunasekera, S. (2015). Personalized Tour Recommendation Based on User Interests and Points of Interest Visit Durations. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, (pp. 1778–1784). AAAI Press.

Liu, C., Tang, T., Lv, K., and Wang, M. (2018). Multi-feature based emotion recognition for video clips. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, (pp. 630–634).

Lopez de Avila, A. (2015). Smart destinations: Xxi century tourism. In *ENTER2015 conference on information and communication technologies in tourism, Lugano, Switzerland*, (pp. 4–6).

Lu, X., Wang, C., Yang, J.-M., Pang, Y., and Zhang, L. (2010). Photo2Trip: Generating Travel Routes from Geo-tagged Photos for Trip Planning. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, (pp. 143–152). ACM.

Makarova, V. and Petrushin, V. A. (2002). Ruslana: A database of russian emotional utterances. In *Seventh international conference on spoken language processing*.

Mangal, A. and Kumar, N. (2016). Using big data to enhance the bosch production line performance: A kaggle challenge. In *2016 IEEE International Conference on Big Data (Big Data)*, (pp. 2029–2035). IEEE.

Mariooryad, S. and Busso, C. (2013). Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, (pp. 85–90). IEEE.

Mariooryad, S. and Busso, C. (2015). Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. *IEEE Transactions on Affective Computing*, *6*(2), 97–108.

Mathieu, B., Essid, S., Fillon, T., Prado, J., and Richard, G. (2010). Yaafe, an easy to use and efficient audio feature extraction software. In *ISMIR*, (pp. 441–446).

Matsuda, Y., Fedotov, D., Takahashi, Y., Arakawa, Y., Yasumoto, K., and Minker, W. (2018a). Emotour: Estimating emotion and satisfaction of users based on behavioral cues and audiovisual data. *Sensors*, *18*(11), 3978.

Matsuda, Y., Fedotov, D., Takahashi, Y., Arakawa, Y., Yasumoto, K., and Minker, W.

(2018b). Emotour: Estimating emotion and satisfaction of users based on behavioral cues and audiovisual data. *Sensors*, *18*(11), 3978.

Matsuda, Y., Fedotov, D., Takahashi, Y., Arakawa, Y., Yasumoto, K., and Minker, W. (2018c). Emotour: Multimodal emotion recognition using physiological and audio-visual features. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, (pp. 946–951). ACM.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schroder, M. (2011). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, *3*(1), 5–17.

McKeown, G., Valstar, M. F., Cowie, R., and Pantic, M. (2010). The semaine corpus of emotionally coloured character interactions. In *2010 IEEE International Conference on Multimedia and Expo*, (pp. 1079–1084). IEEE.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, *14*(4), 261–292.

Mencattini, A., Martinelli, E., Ringeval, F., Schuller, B., and Di Natale, C. (2017). Continuous estimation of emotions in speech by dynamic cooperative speaker models. *IEEE transactions on affective computing*, *8*(3), 314–327.

Menghini, L., Gianfranchi, E., Cellini, N., Patron, E., Tagliabue, M., and Sarlo, M. (2019). Stressing the accuracy: Wrist-worn wearable sensor validation over different conditions. *Psychophysiology*, *56*(11), e13441.

Metallinou, A., Lee, C.-C., Busso, C., Carnicke, S., and Narayanan, S. (2010). The usc creativeit database: A multimodal database of theatrical improvisation. *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, 55.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mirsamadi, S., Barsoum, E., and Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 2227–2231). IEEE.

Mizumoto, T., Otoda, Y., Nakajima, C., Kohana, M., Uenishi, M., Yasumoto, K., and Arakawa, Y. (2020). Design and implementation of sensor-embedded chair for continuous sitting posture recognition. *IEICE TRANSACTIONS on Information and Systems*, *103*(5), 1067–1077.

Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, *10*(1), 18–31.

Mori, H. (2009). An analysis of switching pause duration as a paralinguistic feature in expressive dialogues. *Acoustical science and technology*, *30*(5), 376–378.

Mori, H., Satake, T., Nakamura, M., and Kasuya, H. (2008). Uu database: A spoken dialogue corpus for studies on paralinguistic information in expressive conversation. In *International conference on text, speech and dialogue*, (pp. 427–434). Springer.

Mori, H., Satake, T., Nakamura, M., and Kasuya, H. (2011). Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics. *Speech Communication*, *53*(1), 36–50.

Nakamura, Y., Arakawa, Y., Kanehira, T., Fujiwara, M., and Yasumoto, K. (2017). Senstick: Comprehensive sensing platform with an ultra tiny all-in-one sensor board for iot research. *Journal of Sensors*, *2017*.

Nakamura, Y., Arakawa, Y., Kanehira, T., and Yasumoto, K. (2016). Senstick 2: Ultra tiny all-in-one sensor with wireless charging. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, (pp. 337–340).

Nakamura, Y., Matsuda, Y., Arakawa, Y., and Yasumoto, K. (2019). Waistonbelt x: A belt-type wearable device with sensing and intervention toward health behavior change. *Sensors*, *19*(20), 4600.

Neuhofer, B., Buhalis, D., and Ladkin, A. (2015). Smart technologies for personalized experiences: a case study in the hospitality domain. *Electronic Markets*, *25*(3), 243–254.

Nicolaou, M. A., Gunes, H., and Pantic, M. (2011). Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, *2*(2), 92–105.

Nishigaki, K., Yasumoto, K., Shibata, N., Ito, M., and Higashino, T. (2005). Framework and rule-based language for facilitating context-aware computing using information appliances. In *25th IEEE International Conference on Distributed Computing Systems Workshops*, (pp. 345–351). IEEE.

Nishimura, T., Higuchi, T., Yamaguchi, H., and Higashino, T. (2014). Detecting smoothness of pedestrian flows by participatory sensing with mobile phones. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers*, ISWC '14, (pp. 15–18)., New York, NY, USA. ACM.

Okamoto, M. and Yanai, K. (2013). Summarization of egocentric moving videos for generating walking route guidance. In *Pacific-Rim Symposium on Image and Video Technology*, (pp. 431–442). Springer.

Oliveira, A. M., Teixeira, M. P., Fonseca, I. B., and Oliveira, M. (2006). Joint model-parameter validation of self-estimates of valence and arousal: Probing a differential-weighting model of affective intensity. *Proceedings of Fechner Day*, *22*, 245–250.

Otoda, Y., Mizumoto, T., Arakawa, Y., Nakajima, C., Kohana, M., Uenishi, M., and Yasumoto, K. (2018). Census: Continuous posture sensing chair for office workers. In *2018 IEEE International Conference on Consumer Electronics (ICCE)*, (pp. 1–2). IEEE.

Ouyang, A., Dang, T., Sethu, V., and Ambikairajah, E. (2019). Speech based emotion prediction: Can a linear model work? *Proc. Interspeech 2019*, 2813–2817.

Pandey, A. and Wang, D. (2019). Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 6875–6879). IEEE.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition.

Pelletier, C., Webb, G. I., and Petitjean, F. (2019). Temporal convolutional neural network for the classification of satellite image time series. *Remote Sensing*, *11*(5), 523.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, (pp. 1532–1543).

Perepelkina, O., Kazimirova, E., and Konstantinova, M. (2018). Ramas: Russian multimodal

corpus of dyadic interaction for affective computing. In *International Conference on Speech and Computer*, (pp. 501–510). Springer.

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, *10*(3), 61–74.

Plutchik, R. and Kellerman, H. (1980). *Emotion, theory, research, and experience*. Academic press.

Popescu, A. and Grefenstette, G. (2011). Mining Social Media to Create Personalized Recommendations for Tourist Visits. In *Proceedings of the 2nd International Conference on Computing for Geospatial Research & Applications*, COM.Geo '11, (pp. 37:1–37:6). ACM.

Poria, S., Cambria, E., Hazarika, D., Majumder, N., Zadeh, A., and Morency, L.-P. (2017). Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, (pp. 873–883).

Poria, S., Chaturvedi, I., Cambria, E., and Hussain, A. (2016). Convolutional mkl based multimodal emotion recognition and sentiment analysis. In *2016 IEEE 16th international conference on data mining (ICDM)*, (pp. 439–448). IEEE.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2018). Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Raux, A., Bohus, D., Langner, B., Black, A. W., and Eskenazi, M. (2006). Doing research on a deployed spoken dialogue system: One year of let's go! experience. In *Ninth International Conference on Spoken Language Processing*.

Ringeval, F., Eyben, F., Kroupi, E., Yuce, A., Thiran, J.-P., Ebrahimi, T., Lalanne, D., and Schuller, B. (2015). Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, *66*, 22–30.

Ringeval, F., Schuller, B., Valstar, M., Cowie, R., Kaya, H., Schmitt, M., Amiriparian, S., Cummins, N., Lalanne, D., Michaud, A., et al. (2018). Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, (pp. 3–13). ACM.

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., and Pantic, M. (2019). Avec'19: Audio/visual emotion challenge and workshop. In *Proceedings of the 27th ACM International Conference on Multimedia*, (pp. 2718–2719). ACM.

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., Mozgai, S., Cummins, N., Schmitt, M., and Pantic, M. (2017a). Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, (pp. 3–9). ACM.

Ringeval, F., Schuller, B., Valstar, M., Jaiswal, S., Marchi, E., Lalanne, D., Cowie, R., and Pantic, M. (2015). Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, (pp. 3–8). ACM.

Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, (pp. 1–8). IEEE.

Rozgić, V., Ananthakrishnan, S., Saleem, S., Kumar, R., and Prasad, R. (2012). Ensemble of svm trees for multimodal emotion recognition. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, (pp. 1–4). IEEE.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211–252.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

Sagha, H., Deng, J., Gavryukova, M., Han, J., and Schuller, B. (2016). Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (pp. 5800–5804). IEEE.

Salahuddin, L., Cho, J., Jeong, M. G., and Kim, D. (2007). Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings. In *2007 29th annual international conference of the ieee engineering in medicine and biology society*, (pp. 4656–4659). IEEE.

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Social science information*, *44*(4), 695–729.

Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., and Van Laerhoven, K. (2018). Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, (pp. 400–408).

Schmitt, M., Cummins, N., and Schuller, B. (2019). Continuous emotion recognition in speech–do we need recurrence? *Training*, *34*(93), 12.

Schuller, B., Steidl, S., and Batliner, A. (2009). The interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*.

Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., Amatuni, A., Casillas, M., Seidl, A., Soderstrom, M., et al. (2017). The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring. In *Computational Paralinguistics Challenge (ComParE), Interspeech 2017*, (pp. 3442–3446).

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., and Narayanan, S. S. (2010). The interspeech 2010 paralinguistic challenge. In *Eleventh Annual Conference of the International Speech Communication Association*.

Schuller, B., Steidl, S., Batliner, A., Epps, J., Eyben, F., Ringeval, F., Marchi, E., and Zhang, Y. (2014b). The interspeech 2014 computational paralinguistics challenge: Cognitive & physical load. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Schuller, B., Steidl, S., Batliner, A., Hantke, S., Hönig, F., Orozco-Arroyave, J. R., Nöth, E., Zhang, Y., and Weninger, F. (2015). The interspeech 2015 computational paralinguistics challenge: nativeness, parkinson's & eating condition. In *Sixteenth annual conference of the international speech communication association*.

Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., Son, R. v., Weninger, F., Eyben, F., Bocklet, T., et al. (2012). The interspeech 2012 speaker trait challenge. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Schuller, B., Steidl, S., Batliner, A., Schiel, F., and Krajewski, J. (2011). The interspeech 2011 speaker state challenge. In *Twelfth Annual Conference of the International Speech Communication Association*.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., et al. (2013). The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., and Pantic, M. (2011). Avec 2011–the first international audio/visual emotion challenge. In *International Conference on Affective Computing and Intelligent Interaction*, (pp. 415–424). Springer.

Schuller, B., Valster, M., Eyben, F., Cowie, R., and Pantic, M. (2012). Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, (pp. 449–456). ACM.

Schuller, B. W., Batliner, A., Bergler, C., Pokorny, F. B., Krajewski, J., Cychosz, M., Schmitt, M., et al. (2019). The interspeech 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity. In *Proceedings of Interspeech*.

Schuller, B. W., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J. K., Baird, A., Elkins, A. C., Zhang, Y., Coutinho, E., and Evanini, K. (2016). The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language. In *Interspeech*, volume 2016, (pp. 2001–2005).

Schuller, B. W., Steidl, S., Batliner, A., Marschik, P. B., Baumeister, H., Dong, F., Hantke, S., Pokorny, F. B., Rathner, E.-M., Bartl-Pokorny, K. D., et al. (2018). The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats. In *Interspeech*, (pp. 122–126).

Sidorov, M., Ultes, S., and Schmitt, A. (2014). Emotions are a personal thing: Towards speaker-adaptive emotion recognition. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (pp. 4803–4807). IEEE.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Soleymani, M., Koelstra, S., Patras, I., and Pun, T. (2011). Continuous emotion detection in response to music videos. In *Face and Gesture 2011*, (pp. 803–808). IEEE.

Steventon, A. and Wright, S. (2010). *Intelligent spaces: The application of pervasive ICT*. Springer Science & Business Media.

Strauss, P.-M. and Minker, W. (2010). *Proactive spoken dialogue interaction in multi-party environments*. Springer Science & Business Media.

Subrahmanya, N. and Shin, Y. C. (2009). Sparse multiple kernel learning for signal processing applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(5), 788–798.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V.,

and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 1–9).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 2818–2826).

Tan, L., Zhang, K., Wang, K., Zeng, X., Peng, X., and Qiao, Y. (2017). Group emotion recognition with individual facial emotion cnns and global image based cnns. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, (pp. 549–552).

Tani, Y., Fukuda, S., Matsuda, Y., Inoue, S., and Arakawa, Y. (2020). Workersense: Mobile sensing platform for collecting physiological, mental, and environmental state of office workers. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, (pp. 1–6). IEEE.

Tarnowski, P., Kołodziej, M., Majkowski, A., and Rak, R. J. (2017). Emotion recognition using facial expressions. *Procedia Computer Science*, *108*, 1175–1184.

Tian, L., Moore, J., and Lai, C. (2016). Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, (pp. 565–572). IEEE.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology-volume 1*, (pp. 173–180). Association for Computational Linguistics.

Toyama, S., Saito, D., and Minematsu, N. (2017). Use of global and acoustic features associated with contextual factors to adapt language models for spontaneous speech recognition. In *INTERSPEECH*, (pp. 543–547).

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., and Zafeiriou, S. (2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, (pp. 5200–5204). IEEE.

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., and Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, *11*(8), 1301–1309.

Ueda, K., Tamai, M., and Yasumoto, K. (2015). A method for recognizing living activities in homes using positioning sensor and power meters. In *2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, (pp. 354–359). IEEE.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., and Pantic, M. (2016a). Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th international workshop on audio/visual emotion challenge*, (pp. 3–10). ACM.

Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., and Pantic, M. (2014). Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, (pp. 3–10). ACM.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., and Pantic, M. (2013). Avec 2013: the continuous audio/visual emotion and depres-

sion recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, (pp. 3–10). ACM.

Verkholyak, O., Fedotov, D., Kaya, H., Zhang, Y., and Karpov, A. (2019). Hierarchical two-level modelling of emotional states in spoken dialog systems. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 6700–6704). IEEE.

Wagner, J., Lingenfelser, F., and André, E. (2015). Building a robust system for multimodal emotion recognition. *Emotion recognition: A pattern analysis approach*, 379–410.

Wang, S.-H., Li, H.-T., Chang, E.-J., and Wu, A.-Y. A. (2018). Entropy-assisted emotion recognition of valence and arousal using xgboost classifier. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, (pp. 249–260). Springer.

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, *45*(4), 1191–1207.

Weninger, F., Ringeval, F., Marchi, E., and Schuller, B. W. (2016). Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio. In *IJCAI*, volume 2016, (pp. 2196–2202).

Wöllmer, M., Metallinou, A., Eyben, F., Schuller, B., and Narayanan, S. (2010). Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *Proc. INTERSPEECH 2010, Makuhari, Japan*, (pp. 2362–2365).

Xing, B., Zhang, H., Zhang, K., Zhang, L., Wu, X., Shi, X., Yu, S., and Zhang, S. (2019). Exploiting eeg signals and audiovisual feature fusion for video emotion recognition. *IEEE Access*, *7*, 59844–59861.

Xu, Y., Hu, T., and Li, Y. (2016). A travel route recommendation algorithm with personal preference. In *2016 12th international conference on natural computation, fuzzy systems and knowledge discovery (icnc-fskd)*, (pp. 390–396). IEEE.

Yamamoto, S., Kouyama, N., Yasumoto, K., and Ito, M. (2011). Maximizing users comfort levels through user preference estimation in public smartspaces. In *2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, (pp. 572–577). IEEE.

Yan, W.-J., Wu, Q., Liang, J., Chen, Y.-H., and Fu, X. (2013). How fast are the leaked facial expressions: The duration of micro-expressions. *Journal of Nonverbal Behavior*, *37*(4), 217–230.

Ying, H., Silex, C., Schnitzer, A., Leonhardt, S., and Schiek, M. (2007). Automatic step detection in the accelerometer signal. In *4th International Workshop on Wearable and Implantable Body Sensor Networks (BSN 2007)*, (pp. 80–85).

Zadeh, A., Liang, P. P., Poria, S., Vij, P., Cambria, E., and Morency, L.-P. (2018). Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhang, Y., Ma, H., and Zimmermann, R. (2013). Dynamic multi-video summarization of sensor-rich videos in geo-space. In *International Conference on Multimedia Modeling*, (pp. 380–390). Springer.

Zhao, J., Li, R., Chen, S., and Jin, Q. (2018). Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, (pp. 65–72). ACM.

Zhao, M., Adib, F., and Katabi, D. (2016). Emotion recognition using wireless signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, (pp. 95–108).

Zhou, Y.-T. and Chellappa, R. (1988). Computation of optical flow using a neural network. In *ICNN*, (pp. 71–78).

Acronyms

AEG Auto-Encoder based Generator
AER Automatic Emotion Recognition
AHP Analytic Hierarchy Process
ANN Artificial Neural Network
ARX Autoregressive Exogenous model
AU Action Unit

BLSTM Bidirectional (Recurrent Neural Network with) Long Short-Term Memory
BPTT Backpropagation Through Time

CCA Canonical Correlation Analysis
CCC Concordance Correlation Coefficient
CFS Correlation-based Feature Selection
CNN Convolutional Neural Network

DLF Decision-level fusion
DNN Deep Neural Network

ECG Electrocardiogram
EDA Electro-dermal Activity
EEG Electroencephalography
ELM Extreme Learning Machine
EWE Evaluator Weighted Procedure

FACS Facial Action Coding System
FLF Feature-level fusion

GPS Global Positioning System

HCI Human-Computer Interaction
HMM Hidden Markov Model
HRI Human Robot Interaction

KDE Kernel Density Estimate

LLD Low-level Descriptor
LM Language Model
LPQ-TOP Local Phase Quantisation from Three Orthogonal Planes
LSTM Long Short-Term Memory

MAE Mean Absolute Error
MCC Mobile Cloud Computing

MFCC Mel Frequency Cepstral Coefficients
MI Mutual Information
MKL Multiple Kernel Learning
MLP Multilayer Perceptron
MMI Maximal Mutual Information

PCA Principal Component Analysis
PCC Pearson's Correlation Coefficient
PPG Photoplethysmography

QBTD Quadrant-Based Temporal Division

RAM Random Access Memory
RL Reaction Lag
RMSE Root Mean Square Error
RNN Recurrent Neural Network

SAL Sensitive Artificial Listener
SDS Spoken Dialogue System
SELU Scaled Exponential Linear Unit
SMO Sequential Minimal Optimization
SVM Support Vector Machine
SVR Support Vector Regression

TCCN Temporal Convolution Neural Network
TEQ Touristic Experience Quality

UAR Unweighted Average Recall
UTC Coordinated Universal Time
UV Ultraviolet

List of Figures

List of Tables

Pages 203-262 with publication reprints are removed due to copyright reasons.