КЛАССИФИКАЦИЯ ЭМОЦИОНАЛЬНОГО СОСТОЯНИЯ ДИКТОРА ПО ГОЛОСУ: ПРОБЛЕМЫ И РЕШЕНИЯ

А. Г. Давыдов (davydov-a@speetech.by)

В. В. Киселёв (kiselev-v@speetech.by)

Д. С. Кочетков (kochetkov-d@speetech.by)

ООО «Речевые технологии», Минск, Беларусь

Описан алгоритм, позволяющий автоматически определять эмоциональное состояние диктора по голосу. Обучение и тестирование проводилось на модельном корпусе эмоциональной речи, собранном в техническом университете Берлина. Точность распознавания состояний «*anger*» и «*neutral*» составила порядка 96 %.

Ключевые слова: голос, диктор, эмоции, эмоциональное состояние, эмоциональная речь.

VOICE EMOTION CLASSIFICATION: PROBLEMS AND SOLUTIONS

A. G. Davydov (davydov-a@speetech.by)

V. V. Kiselev (kiselev-v@speetech.by)

D. S. Kochetkov (kochetkov-d@speetech.by)

Speech Technologies LLC, Minsk, Belarus

An algorithm for automatic emotion recognition from the speaker's voice has been developed. A number of tests were performed using the widely known corpus of Emotional Speech — Berlin Database (Emo-DB). The classifi cation effi ciency for different acoustic features was estimated and a very small set of the most reliable characteristics was extracted in order to obtain a robust and quick emotion state classifi cation. Using the SVM classifi er with quadratic kernel and this feature set provides the recognition accuracy of approximately 96 %. between "anger" and "neutral" emotional states. GMM classifi er was less effective and demonstrates a classifi cation error of up to 6 %. A brief comparison of this feature set and SVM kernel effectiveness was performed using the Munich openEAR toolkit. A recommended set of 384 features and linear-kernel SVM was used to solve the same problem.

The classifi cation effi ciency of such algorithm reached 98%. This value is only ~2% higher than the respective value for the designed feature set and classifi er. Under the several conditions, such as in the case of obtaining a decision support factor in the systems of real-time speech analytics the simplifi ed classifi cation scheme would be more preferable than a complex one.

Key words: voice, emotions, emotional state, emotional speech.

1. Введение

Первые попытки автоматического определения эмоциональных состояний по голосу, были предприняты еще в середине 80-х годов. С тех пор эволюция компьютеров с одной стороны, и требования рынка с другой, неуклонно стимулируют дальнейшее развитие систем распознавания эмоций, а так же иных систем голосового анализа, детектирующих уровень стресса, депрессии, усталости, алкогольного опьянения и т. п. Несмотря на это, проблема взаимосвязи эмоциональных состояний диктора с параметрами его голоса до сих пор полностью не решена.

Трудности, встающие перед исследователем при решении этой задачи весьма многообразны, однако можно особо выделить две из них [1]. Прежде всего, чёткого определения эмоции не существует. Это приводит к различным формам классификации эмоциональных состояний и различной расстановке акцентов у разных исследователей [2]. Помимо этого, отсутствует однозначный ответ на вопрос о соотнесении акустических особенностей речи диктора и его эмоционального состояния. Порой различные авторы приводят полностью противоположные результаты. В настоящий момент многие исследователи приняли на вооружение классификацию эмоций, основанную на непрерывной их модели, согласно которой каждое эмоциональное состояние может быть описано точкой в эмоциональном пространстве. Чаще всего при этом координатная сетка задается двумя шкалами, определяющими уровень активации психики и валентность (устойчивость предпочтений человека относительно конкретного результата). Другой распространенной теоретической моделью является так называемая дискретная, «палитровая» теория, согласно которой любое эмоциональное состояние можно описать как совокупность действия ряда архетипических эмоций. Как правило, к ним относят гнев, раздражение, страх, радость, печаль и удивление. Выбор конкретной модели описания при решении задачи распознавания эмоционального состояния определяется в основном соображениями удобства. Вследствие отсутствия универсальной теоретической модели, и возникающей отсюда необходимости оперировать статистическими закономерностями, при практическом построении системы распознавания эмоций целесообразно проводить классификацию только наиболее существенных для решения конкретных задач эмоциональных состояний, что снижает ошибку классификации и повышает точность работы алгоритма.

Важное влияние на форму проявления эмоционального состояния оказывают культурные, языковые особенности и окружение диктора. Попытки многоязычной классификации эмоций демонстрируют значительное снижение эффективности их распознавания [3]. Помимо этого, эмоциональный корпус может

содержать специально подготовленные с участием актеров записи, либо спонтанную эмоциональную речь, полученную в реальных условиях [4]. Переход от модельных эмоциональных баз к распознаванию эмоций в спонтанной речи так же неминуемо ведет к заметному снижению эффективности работы алгоритмов. Тем не менее, существует определенная общность в выражении эмоций у различных людей, которой уделяется особое внимание в рамках эволюционной биологии [2], и которая делает возможной создание систем голосового анализа эмоционального состояния. При этом модельные эмоциональные базы, записанные при помощи профессиональных актеров, служат неплохим плацдармом для первоначальной оценки работоспособности разрабатываемых алгоритмов, позволяя на время избежать сложностей работы со спонтанной речью, хотя их репрезентативность существенно ниже, чем в случае реальных записей. Тем не менее, использование известных модельных корпусов эмоциональной речи, с которыми ранее работали другие группы исследователей, позволяет выявить относительную эффективность функционирования разрабатываемых алгоритмов.

Диктор-независимые системы голосового детектирования эмоционального состояния, работающие со спонтанной речью, могут комбинировать акустические и лингвистические информативные признаки, использовать алгоритмы сегментации речевого потока на отдельные эмоциональные фрагменты, иерархическую классификацию и т. п. Однако в любом случае основой алгоритма голосового анализа является модуль выделения информативных признаков речевого сигнала и классификатор, относящий звуковой фрагмент, согласно этим признакам, к тому либо иному эмоциональному классу. Соответственно, выделение новых, по возможности родственных человеческому восприятию, информативных признаков, а так же поиск новых высокоэффективных техник классификации на текущий момент времени являются важнейшими задачами голосового распознавания эмоционального состояния.

Исследования в области психологии и психолингвистики предоставили сведения о множестве акустических, просодических и лингвистических характеристик речи, способных служить информативными признаками при распознавании эмоционального состояния, и проявляющихся на уровне голосовых сегментов, слогов и целых слов. Чаще всего в целях дальнейшего анализа из аудио сигнала выделяют [5]:

- различные параметры частоты основного тона и формант;
- кратковременную оценку мощности;
- темп речи (количество слов произносимых в единицу времени);
- контур основного тона.

На основе выделяемого набора информативных признаков строится классификатор, который обучается на предварительно подготовленном наборе звуковых фрагментов. Классификация эмоциональных состояний производится в соответствии либо с задачами построения анализатора (оценки удовлетворенности, уровня стресса, усталости и т. п.), либо с выбранной моделью описания (набор базовых эмоций, непрерывная классификация и т. п.). Как правило, с ростом числа возможных вариантов классификации, точность распознавания эмоциональных состояний

значительно снижается. Соответственно, количество классов, используемых для обучения выбирается небольшим. Наиболее популярными техниками классификации являются следующие [5]: поиск ближайших соседей, метод опорных векторов, скрытые марковские модели, модель смеси нормальных распределений, модели на основе нечеткой логики, байесовские классификаторы максимума вероятности.

Далее в работе рассматривается набор информативных признаков и алгоритм, позволяющий с высокой степенью надежности и малой временной задержкой определять эмоциональное состояние диктора по голосу. Для оценки эффективности работы алгоритма аналогичная задача классификации решалась с использованием разработанного в мюнхенском университете инструментария распознавания эмоций openEAR [6].

2. Методы

Для автоматического распознавания были выбраны два эмоциональных состояния — нейтральное и агрессивное, отмеченных как «*neutral*» и «*anger*». Такой выбор обусловлен интересами дальнейшего применения разрабатываемой технологии. Обучение и тестирование алгоритма проводилось на записях, взятых из берлинской базы данных эмоциональной речи (Emo-DB) [7]. Данный корпус был собран в техническом университете Берлина, и неоднократно использовался исследователями при разработке систем распознавания эмоционального состояния. Он содержит записи эмоциональной речи на немецком языке, полученные с привлечением профессиональных актеров. База включает 535 записей речи 10 дикторов (5 мужчин, 5 женщин), воспроизводящих набор дискретных эмоциональных состояний, называемых иногда «архетипическими» (гнев, раздражение, страх, радость, печаль, удивление и нейтральное состояние). Авторское исследование Берлинской базы показало [7], что эмоции в ней распознаются слушателями в 80 % случаев, и в 60 % признаются естественными.

Из записей эмоциональной речи Берлинской базы данных выделялся ряд информативных признаков. Для каждого из них производилась оценка эффективности классификации. По полученным данным выбирался ряд наиболее эффективных показателей, по значениям которых производилось обучение классификатора, либо процедура классификации.

В блоке выделения признаков каждая запись предварительно умножалась на случайный коэффициент усиления от -20 до +20 дБ, чтобы исключить привязку к абсолютному уровню сигнала. В качестве возможных информативных признаков был выделен ряд прямых акустических характеристик сигнала, а так же набор метапризнаков, определяемых косвенно на основе данных прямых измерений. Прямыми характеристиками звукового сигнала являлись: оценка мощности, частота основного тона (ЧОТ), асимметрия от медианы ЧОТ, линейные спектральные частоты, кепстральные коэффициенты, вычисленные по коэффициентам предсказания, статистики высшего порядка, энергетический оператор Тигера [8]. Затем по известным данным прямых наблюдений рассчитывались 1я и 2я производные, энергетический оператор Тигера и ряд других параметров.

Оценка эффективности классификации наблюдения, необходимая для выявления наиболее существенных информативных признаков, осуществлялась следующим образом. В начале для каждого класса вычислялась медианная функция распределения на основе всех функций распределения (CDF) данного класса (Рисунок 1). Затем все функции распределения классифицировались по минимуму функции расстояния до медианных функций распределения. В качестве функции расстояния использовалась сумма модулей разностей функции распределения и медианной функции распределения. После этого вычислялась матрица неточностей. Эффективность классификации определялась как среднее значение диагональных элементов этой матрицы.

Классификация производилась с использованием двух моделей: опорных векторов и смеси нормальных распределений. В методе опорных векторов использовалось квадратичное ядро.

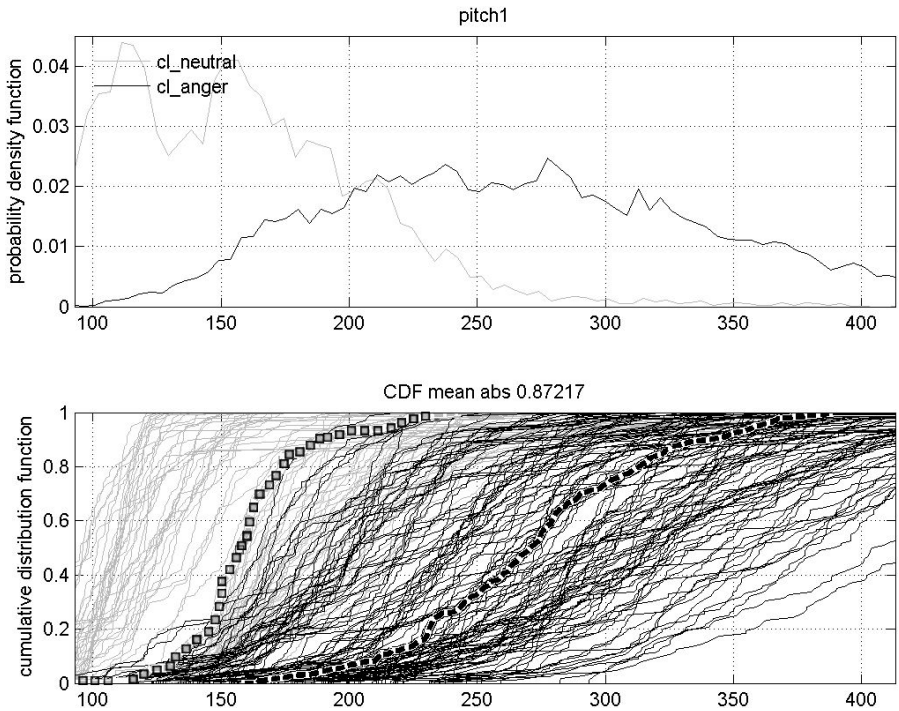

Рис. 1. Функции плотности вероятности (вверху) и распределения (внизу) частоты основного тона для состояния «neutral» (серые кривые) и «anger» (черные). Жирными прерывистыми линиями на нижнем рисунке отмечены медианные функции распределения для этих состояний. Видна область перекрытия функций распределения для двух эмоциональных классов. Оценка эффективности классификации для данного информативного признака ~0,87

Для оценки эффективности работы алгоритма, эта же задача распознавания эмоционального состояния решалась с использованием инструментария openEAR (open-source emotion and affect recognition toolkit) [6]. Данный набор программ был разработан в мюнхенском университете для нужд исследователей, работающих в сфере голосового анализа эмоционального состояния. Пакет доступен на условиях GNU General Public License, и включает в себя средства чтения и записи аудио файлов, выделения из речевого сигнала паралингвистических информативных признаков, и инструментарий для построения классификатора на основе библиотеки libSVM (широко известная библиотека, реализующая метод опорных векторов). Программа способна выделить свыше 6500 характеристик звукового сигнала, перечень которых задается при помощи файлов конфигурации. Файлы эмоциональной базы анализировались с использованием набора из 384 информативных признаков, сформированного авторами пакета openEAR по итогам конференции Interspeech'09. Количество записей в обучающей выборке при этом составляло 80 штук — 40 записей для нейтрального состояния, и 40 для состояния гнева. Документация к пакету libSVM [9], рекомендует в таких случаях, когда количество информативных признаков сравнимо либо превышает количество образцов в обучающей выборке, применять для классификации линейное ядро.

3. Результаты

Тестирование алгоритма позволило выявить ряд информативных признаков, эффективность классификации эмоциональной речи по которым оказалась максимальной (Таблица 1). Наиболее значимыми для принятия решения о принадлежности записи к классу «neutral» либо «anger» информативными признаками оказались частота основного тона, 2ой коэффициент линейных спектральных частот, вторая производная оценки мощности и эксцесс ошибки линейного предсказания (Таблица 1). В литературе существуют разногласия касательно взаимосвязи эмоциональных состояний с просодическими и акустическими характеристиками речи [1]. Однако влияние состояния гнева, помимо прочего, на частоту основного тона и форму огибающей энергии сигнала хорошо известно и не подлежит сомнению [10]. Таким образом, выделенные информативные признаки находятся в хорошем согласии с литературными данными.

Таблица 1. Эффективность классификации выделенного набора информативных признаков

Информативный признак	Эффективность классификации
Частота основного тона	0.87
Второй коэффициент линейных спектральных частот	0.90
Вторая производная оценки мощности	0.74
Эксцесс ошибки линейного предсказания	0.78

Для выбранного набора параметров, оценки точности работы алгоритма классификации оказались следующими. Ошибка классификации в случае использования метода опорных векторов составила порядка 4%. При применении модели смеси нормальных распределений соответствующее значение оказалось равным примерно 6%. Таким образом, использование выделенных информативных признаков позволило распознавать в базе Emo-DB состояния «neutral» и «anger» с точностью порядка 94–96% в зависимости от используемого алгоритма классификации.

В свою очередь, модель, построенная и обученная с использованием инструментария openEAR, задействующая набор из 384 информативных признаков и классификатор, работающий по методу опорных векторов с линейным ядром, имела для этой же задачи точность классификации порядка 98%. Иными словами, точность классификации, достигаемая с использованием этой методики лишь на 2% превосходит точность, достигнутую при помощи ограниченного набора информативных признаков, отобранных по принципу наибольшей эффективности классификации. Данный факт свидетельствует о том, что при отнесении голоса диктора к одному из классов «anger» либо «neutral» целесообразно пользоваться небольшим набором информативных признаков, важную роль в котором играют характеристики основного тона и мощности сигнала.

4. Анализ и выводы

Автоматическое распознавание эмоционального состояния окажется полезным в любой сфере человеческой деятельности, где требуется его оперативная оценка — в маркетинге, медицине, психологии, обеспечении безопасности и т. п. Разработка такой технологии позволит качественно изменить форму коммуникации между человеком и машиной. Более того, разрабатываемые здесь подходы находят свое применение не только в сфере анализа эмоционального состояния, но и при распознавании других состояний, например — алкогольной интоксикации, усталости, подавленности и т. п. Тем не менее, общие теории взаимосвязи эмоций с характеристиками голоса на данный момент отсутствуют, что вынуждает исследователей каждый раз заниматься разработкой новых и тонкой подстройкой существующих алгоритмов под условия конкретной задачи.

Из литературы известно, какие из параметров голоса могут служить индикаторами состояния гнева [10]. Прежде всего, исследователи отмечают его влияние на характеристики частоты основного тона. Возрастает его медианное значение, скорость изменений, расширяется его диапазон. В состоянии гнева диктор издает звуки с более открытым речевым трактом, что приводит к возрастанию средней частоты первой форманты. Так же, по отношению к ней, возрастают амплитуды второй и третьей формант, повышается неоднородность формантных контуров. Кроме частотных параметров голоса важную роль играют характеристики огибающей его энергии. В состоянии гнева энергия

речевого сигнала увеличивается. К вышеназванным признакам можно добавить еще увеличение скорости речи, а так же ряд других показателей.

Анализ акустических характеристик записей берлинской базы эмоциональной речи позволил выделить ряд параметров с наибольшей эффективностью классификации состояний «neutral» и «anger». Включение выделенных информативных признаков в алгоритм, работающий на основе метода опорных векторов с квадратичной разделяющей функцией, позволило добиться точности классификации порядка 96%. Данный показатель лишь незначительно (~2%) уступает точности, полученной при помощи набора из 384 информативных признаков, выделенных и обработанных при помощи стандартного инструментария пакета openEAR.

Столь высокий процент распознавания можно объяснить, прежде всего, идеальностью использованного корпуса, и тем фактом, что классифицировались только лишь два эмоциональных состояния. Как известно из литературы [1, 5], увеличение числа распознаваемых эмоций, равно как и переход от модельных эмоциональных баз данных к реальным, необходимо ведет к возрастанию ошибки классификации. Так, средствами openEAR, можно получить точность распознавания одного из семи эмоциональных состояний на берлинской базе эмоциональной речи порядка 89% [6]. Кроме того, даже в модельных условиях эффективность работы алгоритмов распознавания эмоций может существенно варьироваться для различных языков [3]. Тем не менее, предполагается, что описанный набор информативных признаков и алгоритм классификации, работоспособность которых была проверена на корпусе Emo-DB, после соответствующей адаптации, будет, в качестве составного элемента, включен в систему, работающую с русскоязычными голосовыми базами данных в реальных условиях.

References

1. *Ayadi M. El, Kamel M. S., Karray F.* 2011. Survey on Speech Emotion Recognition:
2. *Features,* Classification Schemes, and Databases. Pattern Recognition, 44(3) : 572–587.
3. *Burkhardt F., Paeschke A., Rolfes M., Sendlmeier W., Weiss B.* 2005. A Database of German Emotional Speech. Proc. Interspeech.
4. *Chih-Chung C., Chih-Jen L.* 2001. LIBSVM: a Library for Support Vector Machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
5. *Cornelius R. R.* 1996. The Science of Emotion: Research and Tradition in the Psychology of Emotions.
6. *Eyben F., Wöllmer M. and Schuller B.* 2009. OpenEAR — Introducing the Munich OpenSource Emotion and Affect Recognition Toolkit. Proc. ACII : 576–581.
7. *Hozjan V., Kacic Z.* Context-independent Multilingual Emotion Recognition from Speech Signal. Int. J. Speech Technol, 6 : 311–320.
8. *Morrison D., Wang R., De Silva L. C.* 2007. Ensemble Methods for Spoken Emotion Recognition in Call-Centres. Speech Communication, 49 : 98–112.

9. *Ververidis D., Kotropoulos C.* 2003. A Review of Emotional Speech Databases. Proc. Panhellenic Conference on Informatics (PCI) : 560–574.
10. *Pantic M., Rothkrantz L. J. M.* 2003. Toward an Affect-Sensitive Multimodal Human–Computer Interaction. Proc. of the IEEE, 91(9) : 1370–1390.
11. *Zhuikov V. Ia., Kuznetsov N. N., Kharchenko A. N.* 2010. Signals Change Evaluation with Differential Energetic Operators [Otsenka Izmeneniia Signalov s pomoshch'iu Differentsial'nykh Energeticheskikh Operatorov]. Elektronika I Sviaz'. 3' Tematicheskii Vypusk "Elektronika I Nanotekhnologii" : 63–67.