Doctoral Dissertation

Dimensional Speech Emotion Recognition by Fusing Acoustic and Linguistic Information

Bagus Tris Atmaja

Supervisor: Professor Masato Akagi

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology

Information Science
March 2021

Abstract

Humans perceive emotion in multimodal ways. Speech is one of sensory modalities in which emotion can be perceived. Within speech, human communicates emotion through acoustic and linguistic information. For automatic emotion recognition by a computer, there is a shift from unimodal acoustic analysis to multimodal information fusion. As in human speech emotion perception, computer should be able to perform speech emotion recognition (SER) from bimodal acoustic-linguistic information fusion.

This research aims to investigate the necessity to fuse acoustic with linguistic information for dimensional SER. To achieve this goal, three sub-goals were addressed: SER using acoustic features only, fusing acoustic and linguistic information at the feature level, and fusing acoustic and linguistic information at the decision level.

The first strategy aims at maximizing the potency of recognizing dimensional SER from acoustic information only by investigating region of analysis and effect of silent pause region. This study generalizes the effectiveness of means and standard deviations from acoustic features and prediction of the importance of silent pause region for dimensional SER. In addition, the aggregation of acoustic feature models valence and arousal prediction better than majority voting method. Although several approaches have been carried out, acoustic-based SER still has limitations such as the low performance of valence's prediction score.

The second and third strategies aim at improving valence prediction, investigating the necessity of bimodal information fusion, and evaluating the fusion framework for fusing acoustic and linguistic information. Two fusion methods for acoustic-linguistic information fusion are studied namely early-fusion approach and late-fusion approach. In the second strategy evaluating the feature level (FL) or early-fusion approach, two fusion methods are evaluated – feature concatenation and network concatenation. The FL methods show significant performance improvement over unimodal dimensional SER. In the third strategy using decision level (DL) or late-fusion approach, acoustic and linguistic information are trained independently, and the results are fused by SVM to make the final predictions. Although this proposal is more complex than the previous FL fusion, the results show improvements over the previous DL approach. These studies reveal the necessity to fuse acoustic with linguistic features for dimensional SER.

This study links the current problems in dimensional SER with its potential solutions. The fusion of acoustic and linguistic information fills the gap in dimensional SER. The FL approach improves the performance of unimodal SER significantly. The DL approach improves the FL approach's performance by fusing decisions from bimodal FL approaches. The results devote insights for future strategy in implementing SER, whether to use acoustic-only features (less complex, less accurate), an early-fusion method (more complex, more accurate), or a late-fusion method (most complex, most accurate).

Keywords: dimensional emotion, speech emotion recognition, information fusion, affective computing, acoustic, linguistic

Acknowledgments

The author wishes to express his sincere gratitude to his principal supervisor, Professor Masato Akagi of Japan Advanced Institute of Science and Technology, for his constant encouragement and kind guidance during this dissertation work. Prof. Akagi not only guides the author on the research direction but also pushes the timeline of the author's research to stay on track. Prof. Akagi is a role model of supervisor for granting "license" for prospectus researcher, a Ph.D. student.

The author also wishes to express his thanks to Professor Unoki, the co-supervisor of this dissertation. Without his critical questions, some part of this doctoral study will not exist. The author owes the idea of multitask learning and silent pause feature calculations to Prof. Unoki.

The author is grateful to Professor Kiyoaki Shirai for his helpful suggestions and discussions during minor research. With the acoustic information science background, the author has no sufficient knowledge of linguistic information processing until conducting minor research in his lab. His patience and kindness accelerate the author's understanding of the use of linguistic information for dimensional speech emotion recognition.

The kind, friendly, and warm environment at Acoustic Information Science-Lab (Akagi and Unoki Lab) was the ideal place for conducting research and study. Surrounded by forests and natural landscape, there is no better place for Ph.D. study except in JAIST. The author would like to thank all AIS members for their kindness support during his Ph.D. study.

Funding is one of the crucial factors when conducting research. The author would like to thank the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) for granting a scholarship for his studies.

Finally, the author would like to thank his family and friends. This dissertation is dedicated to them.

Contents

List of Figures

List of Tables

List of Abbreviations

ASR	Automatic Speech Recognition
CCC	Concordance Correlation Coefficient
CNN	Convolutional Neural Networks
DNN	Deep (learning) Neural Networks
FFT	Fast Fourier Transform
FC	Fully Connected (Networks)
GeMAPS	Geneva Minimalistic Acoustic Parameter Set
HSF	High-level Statistical Function
LLD	Low-level Descriptor
LSTM	Long Short-Term Memory
LOSO/SI	Leave One Session Out/Speaker Independent
MFCC	Mel Frequency Cepstral Coefficients
MLP	Multilayer Perceptron
MTL	Multitask Learning
pAA	pyAudioAnalysis
pAA_D	pyAudioAnalysis with their deltas
RMS	Root Mean Square
RMSE	Root Mean Square Energy
SER	Speech Emotion Recognition
SD	Speaker Dependent
Std/std	Standard deviation
STL	Single-task Learning
SVM/SVR	Support Vector Machine/Support Vector Regression
sil	Silence
STFT	Short Time Fourier Transform
ZCR	Zero Crossing Rate

Chapter 1

Introduction

This chapter introduces the necessary background to conduct research on dimensional speech emotion recognition. The aims, problems, concept, contributions, and structure of the dissertation are also presented briefly to ease following the successor chapters.

1.1 Background

Emotion can be regarded as the major difference between machines and humans. At the beginning of human-machine interaction (HMI) development, no machine (robot) would understand human emotion. Nowadays, there have been attempted to recognize human emotion automatically by machines. If the machine could correctly recognize emotion, HMI would be benefited greatly. For example, the vehicle could detect the driver's mood (long time emotion) to ensure safety. In other applications, the satisfaction and performance of user and operator through call center applications can be measured using speech emotion recognition technologies.

Speech emotion recognition (SER) is an emerging technology that is resulted from science in the multi-discipline area, including psychology, physiology, acoustics, and affective computing. The psychology of emotion focuses on how a human reacts to certain stimuli and how these stimuli affect both mentally and physically of humans. The physiology of emotion is related to arousal of the nervous system with various states and strengths of arousal relating to particular emotions. The acoustic of emotion studies on what acoustic features relate to human emotion. The latter affective computing, according to Picard [8], is termed as "computing that relates to, arises from, or influences emotion."

Aside from those disciplines, language also has an impact on the expression of emotion. Humans communicate emotion through speech and language [9]. Furthermore, language is argued to shape perceived emotion intrinsically [10]. Thus, utilizing linguistic information in automatic SER may be useful to make machines accurately recognize human emotions.

Information science is a study of using information processing to find a solution to an important social problem. In information science, a hierarchy of data-information-knowledge (DIK) is known to model the flow of data. This model can be used to model emotion recognition. The input to data is signal, which is an acoustic signal in SER. The data is the speech dataset. The information are features (acoustic and linguistic). The knowledge is the degree of dimensional emotion. This concept, which will be explained in more detail in Chapter 3, correlates information science with SER or models SER from an information science perspective.

Acoustic information science is the study of information science for acoustic phenomena – phenomena relate to sound. Speech is part of acoustic information science; thus, the study of speech involves concepts used in acoustic information science (e.g., acoustic signal processing). SER, although multidisciplinary research, involves two main fields, the acoustics of speech and the psychology of emotion. For automatic SER by computers, the understanding of speech acoustics with computer algorithms will provide a necessary foundation for building a better SER system.

This experimental study explores the necessity of fusing acoustic and linguistic information for dimensional emotion recognition. The study views SER from an acoustic information science point of view. Speech contains both linguistic (verbal) and acoustic (non-verbal) information. While the conventional SER method only uses acoustic information, fusing both acoustic and linguistic information exists in human communication and is feasible for human-machine interaction.

1.2 Research aims and problems

Speech is the primary modality for communication (known as speech communication) including communicating human emotion. Even if other modalities may influence on how human communicate emotions (e.g., facial expressions, gesture, posture, body's motion/-movement, and other physiological signals), in special cases, like in telephone calls or voice assistant applications, only speech can be used to determine a speaker's emotion.

In certain cases, using acoustic information only (e.g., prosody or intonation) to perceive human emotions may not enough. For instance, the happy and angry voices may have similarities in high intonation, sad and fear may have similarities in low intonation. In this case, knowing the semantic of spoken words will increase the possibility to recognize the perceived emotion from speech. If the words have positive meanings and are uttered with high intonation, then the chance that the speaker was happy is higher than angry. This bimodal information fusion of acoustic and linguistic could be implemented in computer algorithms to improve the performance of SER.

Thus, combining acoustic and linguistic information is relevant for improving SER performance by machines. There is no need to add additional modalities since linguistic information can be derived from speech. Modern automatic speech recognition (ASR) can produce text in almost real-time processing. The transcribed spoken text can be used to extract linguistic information. Both acoustic and linguistic information can be fused in such frameworks to evaluate the effectiveness of information fusion over unimodal information.

The main goal of this research is to investigate the necessity of fusing acoustic information with linguistic information for dimensional SER. To achieve this main goal, the following three sub-goals were addressed:

1. maximizing the potency of SER from acoustic information only by investigating the region of analysis and silence region for feature extraction,

2. studying the fuse of acoustic and linguistic information at the feature level (early fusion) and its effect, particularly for valence prediction improvement, and

3. studying the fuse of acoustic and linguistic information at the decision level (late fusion) and comparing the results with the previous approach.

There are five problems to be solved by these goals. The first problem is the region of analysis for acoustic feature extraction. The second problem is the effect of silent pause region in dimensional SER. The third problem is the low performance of valence prediction. The fourth problem is the necessity to fuse acoustic information and linguistic information for dimensional SER. The fifth problem is the fusion framework for combining acoustic and linguistic information. Figure 1.1 shows the connections between research aims and research problems. The details of these aims (which is transformed into research strategies) and problems (issues) are discussed further in Chapter 3.

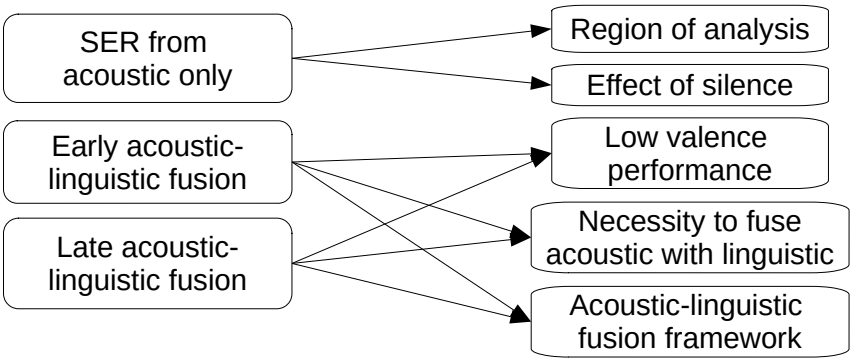

Figure 1.1: Connection between research aims (left) and research problems (right)

1.3 Research concept

Speech delivers a message that goes beyond words. In this understanding, word meaning is not enough to convey a message; acoustic information is needed. Acoustic information only is also not enough to deliver a message. It is not only how it is said (acoustic), but also what is said (linguistic). This concept is the foundation of this research, shown in Figure 1.2.

This research concept differs from the previous studies (e.g., [11, 12]). In these studies, the belief for speech is about how it is said rather than what is said. This study combines both "how" and "what" information. Acoustic features contain information on how it is being said. Linguistic features contain information on what is being said. Fusing both pieces of information, which are extracted from a speech, will improve the clarity of the message, including the expressed emotion. The perception of emotion will also improve by fusing this bimodal information. Figure 1.2 shows the automatic recognition of dimensional SER by fusing both information. This process is also inline with the previous DIK concept in information science.

The process of recognizing emotion from speech consists of two main steps. First is extracting information from speech data; second is extracting degree of dimensional emotions as knowledge from acoustic and linguistic information. Features extraction extracts two pieces of information from speech — acoustic and linguistic. Since dimensional SER is a regression task, a regression process will map extracted features to the ground truth labels. This process is commonly performed within machine learning or deep learning. The acoustic and linguistic information are fused in this step, which can be implemented in various ways.

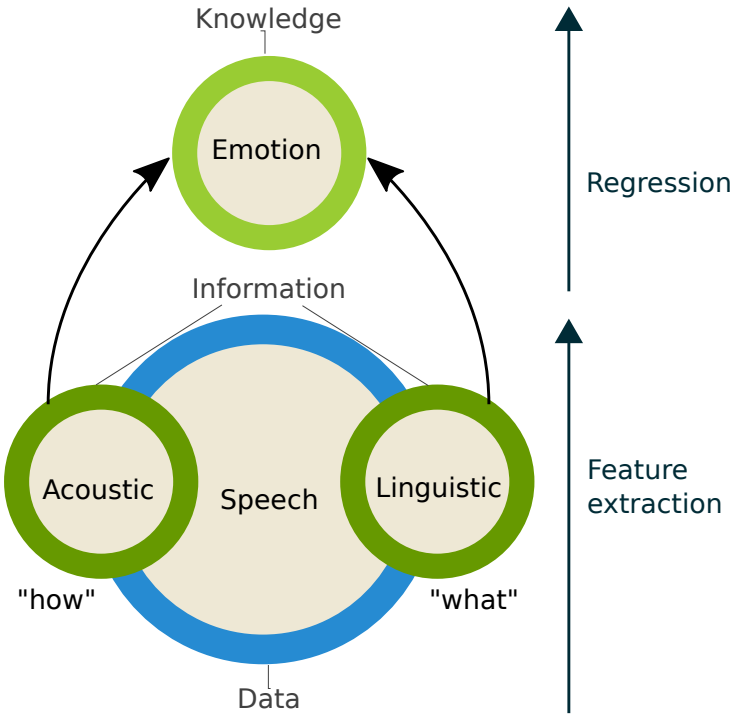

Figure 1.2: Research concept of dimensional speech emotion recognition by fusing acoustic and linguistic information

1.4 Contributions

The contributions of this dissertation can be traced to the published papers. These contributions can be divided into three areas, as follows.

1. Acoustic feature extraction
   In [13] the author evaluated categorical speech emotion recognition from silence-removed speech region. The result suggests that extracting acoustic features from the silence-removed region is better than from the whole speech region. In [14], the author utilized silence as an additional feature to statistical functions. The results achieved a better score than baseline raw speech. The author confirms and generalizes the effectiveness of mean and standard deviations of low-level acoustic features for SER [14, 15]. In [16], the author showed that acoustic feature aggregation leads to better performances than output aggregation. These contributions are explained in Chapter 4. In [17], the author found that the acoustic features performed better in SER will also perform better in song emotion recognition.

2. Information fusion
   In [18, 15, 16], the author proposed emotion recognition by fusing acoustic and linguistic information at feature level. The results significantly improved unimodal dimensional emotion recognition from either acoustic or linguistic information. Furthermore, the author discussed the improvement of valence prediction in [19]. While this contribution is discussed in Chapter 5, the improved version of the proposed method, the late fusion method, is explained in Chapter 6. The evaluated fusion methods are expanded not only for acoustic and linguistic fusion but also for acoustic and visual information fusion [20, 21].

3. Classification methods
   Modern classification methods utilized deep learning models. However, the conventional method, such as support vector machine (SVM) and multi-layer perceptron (MLP), are still used in many fields. The author showed that traditional MLP with deeper layers and proper configurations performed better than modern deep learning architecture [22]. For the SER task with deep learning, the author confirms the need for bigger data size to be fed to deep learning models [23]. The choice of the loss function is a matter in machine/deep learning. The author proposed correlation-based function to improve the performance of dimensional SER [20, 15, 24]. The author also evaluated multitask learning (MTL) for predicting valence, arousal, and dominance degrees based on this loss function [15, 25]. Furthermore, the author found that recurrent-based neural networks (RNN) are effective for the SER task [26]. More improvements were obtained when this RNN model is combined with the attention model [13].

1.5 Dissertation structure

This dissertation is organized in seven chapters. The rest of these chapters is organized as follows.

- **Chapter 2** presents a literature study on speech emotion recognition from bimodal acoustic and linguistic information fusion. An introduction that motivates the previous research by fusing acoustic and linguistic information is presented. This chapter reviews the models, features, classifiers, and fusion methods for the SER task.

- **Chapter 3** describes the research methodology — justification for using particular research methods. This chapter consists of the motivation of researching SER, research issues, research philosophy, research strategies, datasets, and a description of an evaluation metric.

- **Chapter 4** describes SER by using acoustic features. This chapter investigates the region of analysis, effect of silent pause features, and aggregation methods for acoustic-based SER.

- **Chapter 5** describes the fusion of acoustic and linguistic information at the feature level. This chapter evaluates the concatenation of features and networks for bimodal emotion recognition from acoustic and linguistic information. A SER evaluation from automatic transcription is also provided in addition to manual transcription.

- **Chapter 6** describes the fusion of acoustic and linguistic information at the decision level. This chapter evaluates the late-fusion approach by combining both information on two steps processing, including some related issues: speaker-dependent vs. speaker-independent scenarios and effect of lexical-controlled lexicons.

- **Chapter 7** compares the results within this study and with other studies.

- **Chapter 8** presents the overall conclusions of the dissertation. Some possible future research directions are proposed from the current research findings.

This dissertation's organization is summarized in Figure 1.3.

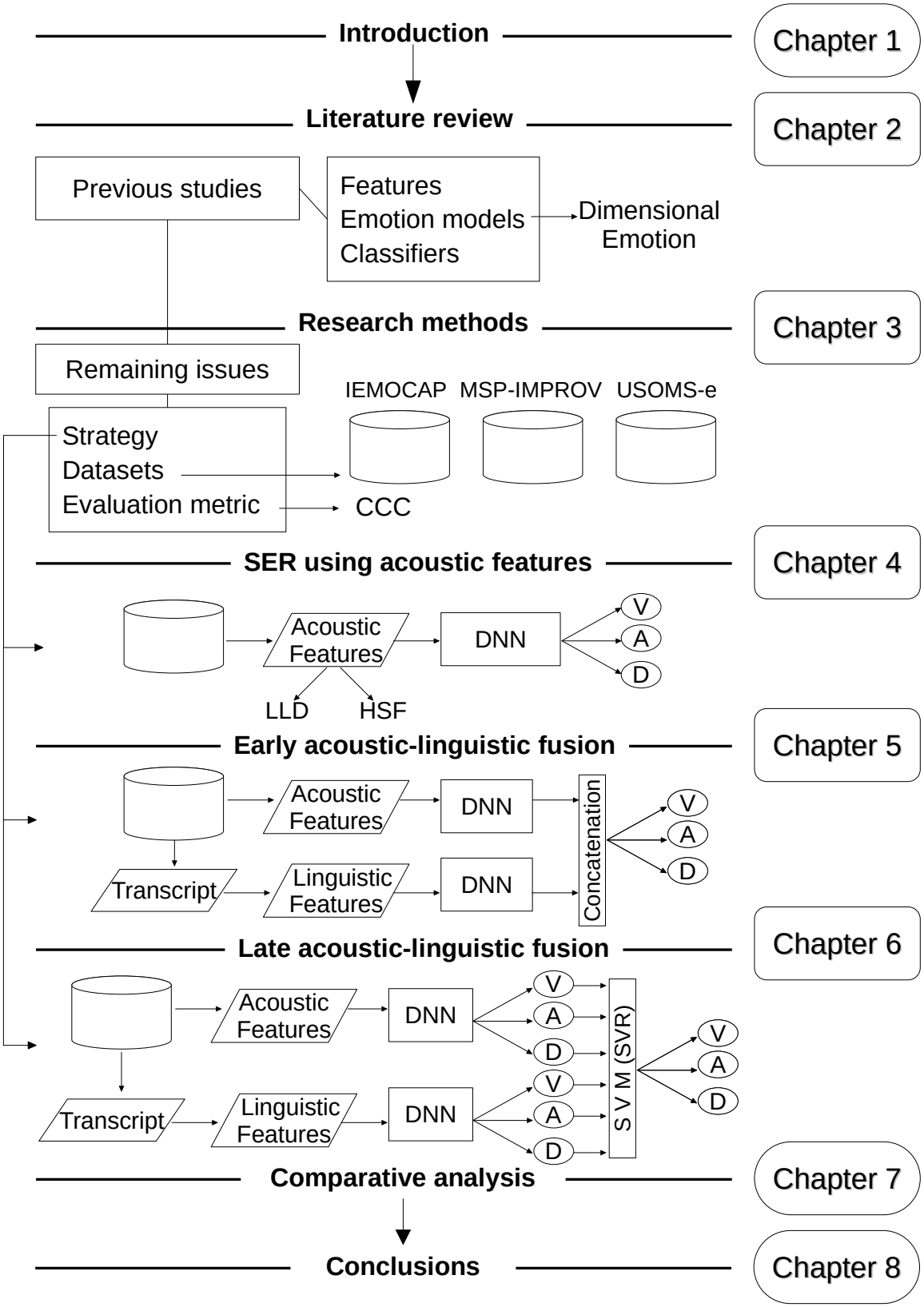

Figure 1.3: Organization of the dissertation

Chapter 2

Literature Review

This chapter reviews research on speech emotion recognition from two modalities, acoustic and linguistic information. Four main components of speech emotion recognition from multiodal fusion are described; including emotion models, features, classifications and fusion methods. Each section describes these components with their advancement and current limitation.

2.1 Introduction

Speech emotion recognition (SER) is a part of affective computing that relates to, arises from, or influences emotions [8] within speech. SER is an attempt to make the computer to be able to recognize expressed emotion in a given utterance. The earliest reported research on SER, perhaps, was the work of Dellaert et al. [27]. The study explores prosodic features with several statistical pattern recognition techniques to classify the emotional content of utterances. Under a limited number of data (1000 utterances), the system achieved a comparable performance close to humans.

Utilizing speech to identify humans' emotions roots from the correlation between voice and emotion. There is strong evidence that humans can recognize other's emotions from their voices. In addition, given that speech is less private than image and video, using speech to recognize emotion benefits future implementation. Another milestone conducted by [28] prove a pilot study to implement SER for call center application. This laboratory-scale study, at that time, showed a potential application for SER technologies. Nowadays, this SER technology is limited available in the commercial market, while its research is still ongoing.

The potential application of speech-based emotion recognition triggers the need for such datasets. During the 2000s, several datasets for emotion recognition have been published, including the availability of speech data in the datasets. Among many datasets, the following are commonly used in SER research: EmoDB [29], IEMOCAP [30], MSP-IMPROV [31], and RAVDESS [32]. The availability of these datasets accelerates research on the SER area.

The progressive research on SER led to practical implementation in the commercial industry. Nowadays, SER has been implemented in various applications, both web/cloud-based applications or standalone applications. Although it is useful to analyze the subject's affective state, these emerging affective recognizer technologies have been criticized by others. Researchers in psychology argued that due to individuals' high variability,

the emotional categories do not have an essence [33]. The correlation between particular facial expression and the corresponding basic emotion was not strongly supported[34].

Among many other issues, multimodal information fusion is a challenging task in pattern recognition. Recent studies (e.g., [35, 15, 20, 18]) confirm that multimodal classifiers outperform unimodal classifiers. In SER itself, one of the main issues in searching for a more predictive feature is whether it suffices to explore acoustic features only, or it is necessary to combine acoustic features with other modalities [36]. For speech, both acoustic and linguistic features can be extracted. Thus, two pieces of information can be fused to evaluate the effectiveness of information fusion from a single speech modality.

The use of linguistic information for SER is also reasonable from an affective computing point of view. In task-processing related tasks, linguistic information is extracted from the text for sentiment analysis. This textual information was also used to detect emotion in text [37, 38, 39]. In these works, textual information shows encouraging results on both categorical and dimensional emotion recognition from text. Fusing acoustic and linguistic information may improve the performance of SER more significantly than other strategies.

Indeed, bimodal emotion recognition by fusing acoustic and linguistic information shows significant performance improvement. References [40, 41, 42, 43, 44] show the usefulness of fusing acoustic and linguistic in different strategies to improve SER performance. Different acoustic and linguistic features were fused using different classifiers. Different fusing strategies were evaluated to investigate the effectiveness of the fusion method.

This chapter aims to review current studies of bimodal emotion recognition by utilizing acoustic and linguistic information. The scope of this study includes the emotion models, features, and classifiers used in bimodal SER. In the end of this chapter is a summary of works done in the past, including the remaining challenge for bimodal SER.

2.2 Emotion models

Before beginning to research emotion recognition, it is important to choose which emotion model to adopt. According to [45], there are at least three views to model humans' emotions: categorical emotion, dimensional emotion, and componential appraisal emotion. However, all SER research employed either first, second, or combination of both views as target emotion. No research was found on using speech data to obtain appraisal emotion. Thus, the following description describes major models in active affective computing research.

2.2.1 Categorical emotion

Categorical emotion, also known as basic emotions, is the discrete emotion that is independent of each other in its manifestations. Although the original idea is to organize affective state into their emotion families (rather than discrete emotion); however, most researchers agree that there are six basic emotions: anger, fear, enjoyment, sadness, disgust, and surprise. The first five emotions are backed by robust and consistent evidence, while the evidence for the surprise is not as firm [46]. Nevertheless, these six basic emotions have been standard in categorical emotions.

Before Ekman coined the terms of basic emotions, Plutchik [47] have defined basic eights bipolar emotions: joy (reproduction), sorrow/sadness (deprivation), accep-

tance/trust (incorporation), disgust (rejection), surprise (orientation), anticipation (exploration), anger (destruction), fear (protection). These eight emotions can be illustrated as a wheel of emotion, as shown in Figure 2.1. Each emotion can mix with other emotion to make up another emotion, as mixing colors.

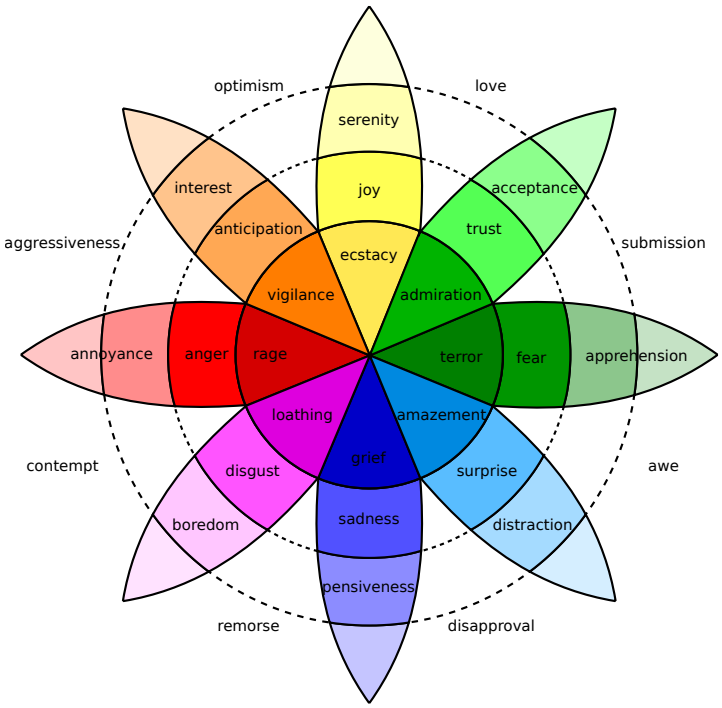

Figure 2.1: Plutchik wheel of emotions

Instead of six, recent research suggests that four latent expressive patterns were commonly observed in facial expressions [48]. However, instead of mentioning the name of basic emotions, the research utilized the term basic "action unit pattern" (AU Pattern), from one to four. Although backed by scientific evidence, this finding did not have any practical implementation yet.

Ekman revised the characteristics which distinguish basic emotion from 9 criteria [46] to 11 criteria [49]. The new criteria resulted in 15 emotions: amusement, anger, contempt, contentment, disgust, embarrassment, excitement, fear, guilt, pride in achievement, relief, sadness/distress, satisfaction, sensory pleasure, and shame. Du et al. shows 21 categories of facial expressions by a facial action coding system analysis. Furthermore, Cowen and Keltner [50] found 27 emotional experiences from facial expression by across self-report methods. The growth of the number of categorical emotions, based on facial expression measurement, confirms the high variability of humans expressed emotions. Darwin argued that the biological category, including the emotion category, does not have an essence; it is hard to map one-to-one facial expressions to emotional states.

2.2.2 Dimensional emotion

Instead of dividing emotion into several categories, a dimensional emotion views emotion as continuous values/degrees of attributes in valence-arousal space (VA) or valence-

arousal-dominance (VAD) space. Valence is the degree of positive or negative emotion, arousal refers to the level of activation from sleepiness (low) to awakeness (high), and dominance is the degree of control over the emotion [51]. In this theory, an emotion or affective state is not independent of one to another. Rather, they are related one to another in a systemic manner (in VA or VAD space). Russel argued that the previous categorical emotion could be mapped within VA spaces. An illustration of VA space with several emotion categories is shown in Figure 2.2.

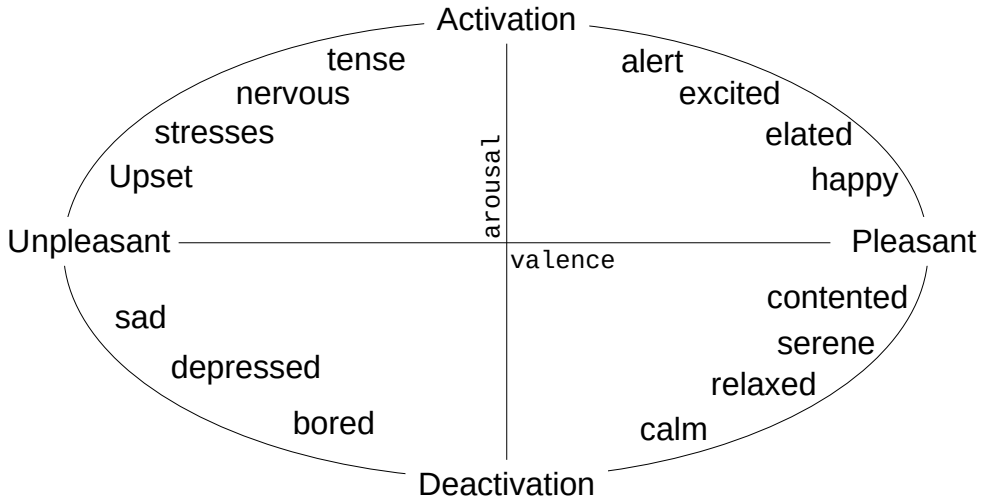

Figure 2.2: Graphical representation of circumplex model (VA space)[1]; vertical axis: arousal; horizontal axis: valence

The search for higher dimensions for dimensional emotion is a worthwhile study. Russel argued that all emotion categories could be mapped in 2D valence-arousal space [52]. However, Fontaine et al. have found that the world of dimensional emotion is not two or three dimensions, but four dimensions [53]. The fourth dimension is the unpredictability. In order of importance, the order of dimensional emotions is valence, dominance, arousal, and predictability. Fortunately, the similar fourth dimension is proposed in an emotion recognition challenge [54]. In this challenge, four-dimensional emotions were arousal, expectancy, power/dominance, and valence. Expectancy, which represents the predictiveness of the subject's feeling, is very similar to predictability/unpredictability in the previous report.

The third emotion model, hybrid model or appraisal model, can be viewed as an extension of the dimensional model. In this model, emotion categories are spanned between bipolar dimensions. For instance, "impatience" is located in the upper part of the arousal axis (see [55]). This study of appraisal-based emotion theory leads to the development of the Geneva emotion wheel (GEW) rating study. This hybrid model has two similarities with the previous 4D dimensional emotion model. First, the hybrid model also uses four attributes (i.e., dimensions00): valence, dominance/power, arousal, and conducive/obstructive (instead of predictiveness). Second, in version 2.0 of GEW, two axes used to draw emotion terms are valence and dominance/power, which are the two most important emotional attributes according to [53]. Nevertheless, the use of the hybrid model in SER is not familiar in the SER research community, perhaps due to these labels' availability in the dataset.

2.3 Features

The input features to the SER system is the most important issue for developing bimodal information SER. If the input is not informative for predicting emotion or does not correlate to the predicted emotion, the prediction results will suffer from low performance. In principle: garbage in, garbage out. The following divisions are useful features for SER from acoustics and linguistics.

2.3.1 Acoustic features

The correlation of acoustic features with emotion has been studied for many years [55, 56]. The main division of acoustic features for SER is the classical and modern approaches, i.e., hand-crafted features vs. deep learning-based features. Hand-crafted features employed acoustic features extracted per frame. These features often called local features or low-level descriptor (LLD). On the other hand, statistical features computed from LLDs are a new way to capture the dynamics among frames. This latter feature extraction method is called global features, suprasegmental features, high-level features, or high-level statistical function (HSF). Figure 2.3 shows the division of acoustic features for SER.

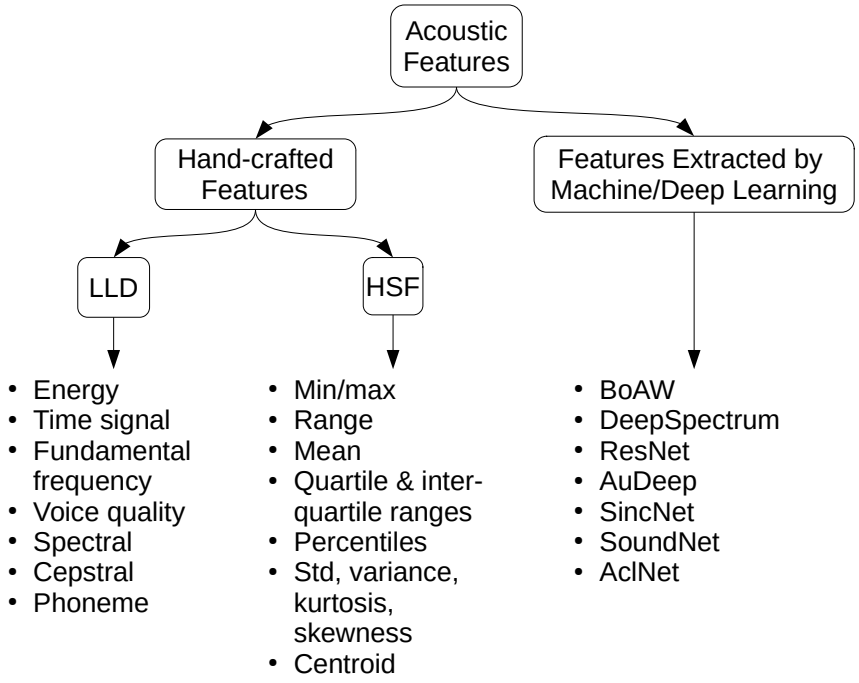

Figure 2.3: Division of acoustic features for SER

Eyben et al. [42] divided LLD and HSF into five groups: signal energy, fundamental frequency (perception: pitch), voice quality, cepstral, time signal, and spectral. Prosodic features (f_o, duration, intensity, voice quality) have been known to have a strong correlation with emotion [57, 58, 59] from a psychology point of view. In acoustics, prosody is implemented into several acoustic features, including LLD and HSF. Vayrynen [12] made a distinction between prosodic and acoustic (non-prosodic) features. His study reported that a combination of prosodic and acoustic features achieved comparable performance to human reference on basic emotion recognition.

Both references [60] and [40] employed f_o and energy-based acoustic features for SER. The former applied both LLD and HSF of f_o and energy features, while the latter only applied HSF of f_o and energy features. The latter reference found that f_o-based features correlated to SER performance more than energy-based features.

As a 'default' feature on most automatic speech recognition (ASR), MFCC has been explored for SER. Metze et al. [61] has found that MFCC is the most informative acoustic features compared to other evaluate acoustic features. Tripathi et al. [62] found that MFCC performed better than spectrogram features on unimodal acoustic SER.

The shift from MFCC to mel-filterbank (MFB) features in ASR motivates SER researchers to adopt a similar direction. Aldeneh et al. [63] extracted 40 MFB features for dimensional SER tasks on the IEMOCAP dataset. Zhang et al. [64] employed a similar MFB with 40-dimensional with z-normalization on categorical IEMOCAP and MSP-IMPROV datasets. Both research showed fair performances (50% – 65% accuracy) from MFB features for the SER task.

Phoneme, the smallest unit of speech, has been investigated to be useful for SER task. Zhang et al. [64] furthermore combined MFB with phoneme for the same SER task. A combination of phoneme with MFB outperforms MFB-only of phoneme-only input features. Yenigalla et al. [65] combined phoneme embedding with a spectrogram. The phoneme embedding is generated from the word2vec model [3] and IEMOCAP speech data. The combination of phoneme with spectrogram achieves the highest accuracy among individual features.

Since most classifiers in modern SER systems used deep learning methods, it is reasonable to extract an acoustic representation of speech in an end-to-end manner via deep learning methods. In INTERSPEECH 2020 ComParE challenge, two deep-learning-based features were given in the baseline system, DeepSpectrum and AuDeep. The provided DeepSpectrum features with ResNet50 network achieve the highest unweighted average recall (AUR) on the elderly emotion sub-challenge test set.

2.3.2 Linguistic features

Linguistic features are the realization of linguistic information. It is also called text features, textual features, lexical features, language features, or semantic features. Although linguistic and lexical terms have different meanings, i.e., language vs. word meaning, these terms in computer science (or information processing) also have a different meaning from the term used in book/article writing, particularly the term "text features." In book/article writing, the text features include writing components such as a glossary, bold typeface, title, headings, captions, and labels. In information science, text or linguistic features are features extracted from written or spoken text. Thus, the term linguistic features is a preferable term than text features to avoid confusion among readers.

Linguistics features used in emotion recognition represent numerical values related to the emotional states in a word. The simplest way to build linguistic features for emotion detection is emotional keyword spotter [66]. In this framework, every word is assumed to have a correlation with emotion categories. For instance, the word "disappointed" can be represented as [(2, 0.2), (3, 0.6)] where 2 represents "angry" emotion and 3 represent "sadness" emotion. Both 0.2 and 0.3 represent degrees of emotion's intensity. This emotional keyword spotter can be expanded into an emotional phrase spotter [40].

The first systematic linguistic representation of a document, perhaps, is TF-IDF (term-

frequency inverse document frequency). TF is defined as the frequency of a word in a particular document/utterance. IDF is defined as a logarithm of the total number of documents' ratio to the total number containing that word. TF-IDF is the multiplication of TF with IDF.

Bag-of-Words (BoW) is a numerical feature vector to represent "words in a bag." First, a fixed integer is assigned to each word occurring in any documents, i.e., building a dictionary from a corpus by assigning a word to integer indices. Second, count the number of occurrences of each word and store it as the value of feature j where j is the index of word w in the dictionary [67]. These BoW features can be expanded for acoustic and visual modalities (BoAW and BoVW).

Several lexicon dictionaries have been developed to inform the 'emotion score' of emotional words. These dictionaries include DAL [68], ANEW [69], VADER [70], and NRC [71]. Using these dictionaries allows direct measurement of emotional words in the given utterances. For instance, the word "arose" has values of 2.11, 2.00, and 1.40 for pleasantness, activation, and imagery. These values are on a 3-point scale; different dictionaries have different scales.

The search for vector representation from a word led to the research of word embedding or word vector. In this approach, a deep neural network is used to train a large corpus (i.e., a Wikipedia corpus) to generate word vectors based on an algorithm. This approach has resulted in a new paradigm in the vector representation of linguistic information of a word. Several models exist, including word2vec, GloVe, FastText, and BERT. These models are detailed in Chapter 5.

Figure 2.4 shows the division of linguistic features used in SER task. Similar to acoustic features, there is a tendency to move to deep learning-based features from hand-crafted features. The choice of linguistic feature is usually based on the task as in other text processing areas.

2.4 Classifiers

This section reviews the four most used classifiers in speech emotion recognition. One is a machine learning classifier, i.e., support vector machine (SVM). Others are three deep learning classifiers, i.e., multilayer perceptron (MLP), convolutional neural networks (CNN), and long short-term memory (LSTM) neural networks. The brief descriptions of these classifiers are given below.

2.4.1 SVM

SVM is a useful machine learning classifier for, generally, small datasets. For categorical emotion recognition, SVM applies acoustic or linguistic features for the given labels. This SVM applied to the classification task is called support vector classification (SVC). For dimensional emotion recognition, SVM applies regression analysis to map them to the given labels. This SVM for regression task is called as support vector regression (SVR).

SVM can accept unimodal or multimodal inputs. In bimodal emotion recognition from acoustic and linguistic information, SVM can be utilized in two-stage scheme for evaluation of the emotion recognition system from DNNs outputs. In bimodal information fusion, each prediction from the acoustic and text networks is fed into the SVM. From two values (e.g., valence predictions from the acoustic and text networks), the SVM learns to

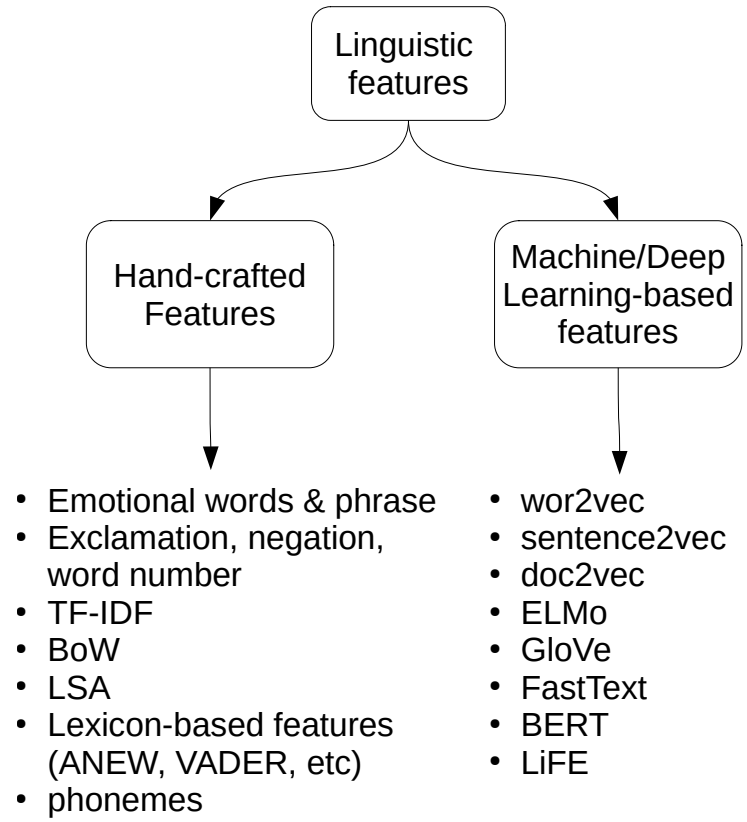

Figure 2.4: Divisions of linguistic features used in SER

generate a final predicted degree (e.g., for valence). The concept of using the SVM as the final classifier can be summarized as follows.

Suppose that two valence prediction outputs from the acoustic and text networks, $x_i = [x_{ser}[i], x_{ter}[i]]$, are generated by the DNNs, and that y_i is the corresponding valence label. The problem in dimensional SER fusing acoustic and text results is to minimize the following:

$$\begin{aligned} \min_{w,b,\zeta,\zeta^*} \quad & \frac{1}{2}w^T w + C\sum_{i=1}^{n}\zeta_i + C\sum_{i=1}^{n}\zeta_i^* \\ \text{subject to} \quad & w^T\phi(x_i) + b - y_i \leq \epsilon + \zeta_i, \\ & y_i - w^T\phi(x_i) - b \leq \epsilon + \zeta_i^*, \\ & \zeta_i, \zeta_i^* \geq 0, i = 1, \ldots, n, \end{aligned} \tag{2.1}$$

where w is a weighting vector, C is a penalty parameter, ζ and ζ^* is the distance between misclassified points and the corresponding marginal boundary (above or below). Here, ϕ is the kernel function. On the use for late fusion approach, the study choose a radial basis function (RBF) kernel because of its flexibility to model a nonlinear process with a dimensional emotion model close to this kernel. The function ϕ for the RBF kernel is formulated as:

$$K(x_i, x_j) = e^{\gamma(x_i - x_j)^2}, \tag{2.2}$$

where γ defines how much influence a single training has on the model. All parameters

in this SVM are obtained empirically via linear search in a specific range. Although the explanation above uses valence, the same also applies for arousal and dominance.

2.4.2 MLP

MLP is a classical feedforward neural network by projecting input data into linearly separable space using non-linear transformation. A hidden layer is an intermediate layer between inputs and outputs, containing many perceptrons (also called units or nodes). An MLP commonly refers to more than one hidden layer. The MLP used here is similar to the definition of connectionist learning proposed by Hinton [72]. A deeper layer MLP usually consists of many layers to enable deep learning hierarchically. This neural network architecture is also known as dense networks or fully-connected (FC) networks.

MLP is powerful for combining acoustic and linguistic network for network concatenation. Mathematically, the combined Acoustic+Text network, could be formulated as in equation 2.3. Here, $f(y)$ denotes the output of the corresponding layer; W_1, W_2 denote the weights from previous layers (a: acoustic; t: text), i.e., dense layer after LSTM for each network, and the current hidden layer, respectively; x_a and x_t are the acoustic features and word embeddings, respectively; b is a bias; and g is an activation function. Thus, the output of the that dense layer was

$$f(y) = W_2 g([W_{1a}^{\top} xa + b_{1a}; W_{1t}^{\top} xt + b_{1t}]) + b_2. \tag{2.3}$$

Schuller et al. [40] utilized MLP for combining acoustic and linguistic information. Their evaluation using MLP showed lower error than the fusion method by means of logical "OR". Callejar and Cozar [73] compared baseline majority-class method to MLP for evaluating the effect of context on categorical SER task. The result shows that MLP outperforms the baseline method in six out of eight scenarios. Zhang et al. [64] used MLP in all experiments involving acoustic features, phoneme, and combination of both; MLP showed its effectiveness on both single-stage and multi-stage SER tasks.

2.4.3 CNN

CNN is a class of neural networks that contain a convolutional layer. Convolution is a mathematical operation between two functions by measuring the overlap of both when one function ("input") is flipped and shifted by another function ("kernel"). The resulting output, which is the goal of a convolution layer, is a feature map. This convolution operation is similar to cross-correlation; cross-correlation does not flip the second function. Convolution is also can be seen as cross-correlation with a scalar bias. In deep learning literature, the convolution terminology views cross-correlation as convolution since many deep learning frameworks did not take bias into account by default.

The convolutional network is often applied to image-like data. Time-series data, including acoustic feature vector, can be fed into convolutional networks using 1-dimensional (1D) CNN. To take the most benefit of CNN, spectrogram and mel-filterbank (MFB) features are frequently used to input the SER system. For text processing, the main idea for CNN is to compute vectors for n-grams (e.g., 2-, 3-, and 4-grams) and group them afterward. CNN is commonly used for both speech and language processing.

Apart from convolutional layers, CNN typically still needs a fully-connected layer (FC or MLP). The feature map as the output of the convolution layer is fed into MLP to obtain

desired outputs. Although recently it is found unnecessary [74], a CNN commonly uses pooling layers after convolutional layers for mitigating and reducing spatial representation [2].

Yenigalla et al. [65] experimented with CNN for categorical SER by inputting phoneme, spectrogram, and combination of both. The combination of both phoneme embeddings and spectrogram achieved the highest performance. The architecture of each phoneme and spectrogram networks was convolution layer and max-pooling and FC layer. Both networks are concatenated with the FC layer to obtain the outputs.

Instead of phoneme and spectrogram, Huang et al. [75] proposed to use bag-of-audio-words for the input of the CNN-based SER system. The architecture was similar to [65], i.e., convolution, pooling, and FC layer. The result shows that the use of BoAW outperforms raw acoustic features.

Cho et al. [76] combined acoustic and linguistic information for categorical SER; the acoustic inputs used an LSTM network while the linguistic inputs used a multi-resolution CNN. A multi-resolution CNN is utilized to emotion words by employing word embedding, convolution layer, and global mean pooling. The combination of acoustic network with LSTM, linguistic network with CNN, and emotion vector (e-vector) achieved the highest performance compared to unimodal results.

While most bimodal SER research used CNN for linguistic and LSTM for acoustic information processing, Sebastian and Pierucci [77], used LSTM for text and CNN for speech. The CNN architecture contains two convolution layers and two FC layers. In this case, the performance of CNN-based text emotion recognition is the lowest among other models.

Cai et al. [78] replaced FC layers in most CNN architectures with bidirectional LSTM with attention. The improved architecture was called CNN-Bi-LSTM-Attention (CBLA). On both unimodal and multimodal, CBLA outperforms MLP models by considering both global and temporal information in the data.

2.4.4 LSTM

Long Short-Term Memory (LSTM) neural networks is an extension of a recurrent neural network. The idea of using LSTM networks comes from an approach that human has the persistence to keep memory long in a short-term period. Humans do not start their thinking from scratch every second. When reading a paper, a reader understands each word based on our understanding of previous words. Humans do not throw everything away and start thinking from scratch again. The thoughts have persistence.

Three gates are introduced in LSTMs: the input gate ($\mathbf{I}_t$), the forgetting gate ($\mathbf{F}_t$), and the output gate ($\mathbf{G}_t$). In addition to that we introduce memory cells that take the same shape as the hidden state. A memory cell is just a fancy version of a hidden state, custom engineered to record additional information. The three gates in LSTM are defined as,

$$\mathbf{I}_t = \sigma(\mathbf{X}_t\mathbf{W}_{xi} + \mathbf{H}_{t-1}\mathbf{W}_{hi} + \mathbf{b}_i), \tag{2.4}$$

$$\mathbf{F}_t = \sigma(\mathbf{X}_t\mathbf{W}_{xf} + \mathbf{H}_{t-1}\mathbf{W}_{hf} + \mathbf{b}_f), \tag{2.5}$$

$$\mathbf{O}_t = \sigma(\mathbf{X}_t\mathbf{W}_{xo} + \mathbf{H}_{t-1}\mathbf{W}_{ho} + \mathbf{b}_o), \tag{2.6}$$

Then the candidate memory cell, memory cell, and hidden state are calculated on the following equation.

$$\widetilde{C}_t = \tanh(\mathbf{X}_t\mathbf{W}_{xc} + \mathbf{H}_{t-1}\mathbf{W}_{hc} + \mathbf{b}_c) \tag{2.7}$$

$$\mathbf{C}_t = \mathbf{F}_t \odot \mathbf{C}_{t-1} + \mathbf{I}_t \odot \tilde{\mathbf{C}}_t. \tag{2.8}$$

$$\mathbf{H}_t = \mathbf{O}_t \odot \tanh(\mathbf{C}_t) \tag{2.9}$$

Graphical illustration of LSTM explained by equations above is shown in Figure 2.5.

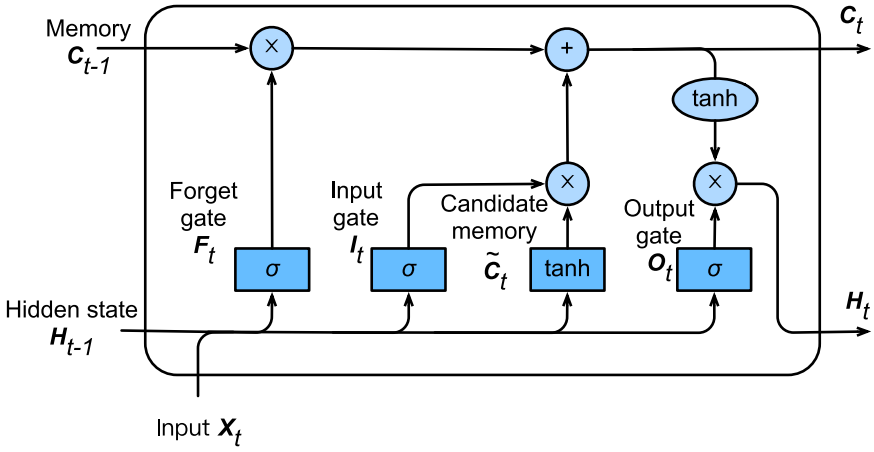

Figure 2.5: Graphical illustration of LSTM [2]

LSTM has dominated the classifier used in both ASR and SER. Since the data (i.e., input features) are sequence, a recurrent neural network is a straightforward way to process these data. In addition, LSTM is able to model long-range context in emotional features to map it with the emotional labels. Tian et al. [79] used the LSTM classifier to build hierarchical neural networks. In [76], LSTM was used to train acoustic features before the network was concatenated with a CNN-based linguistic network.

Instead of unidirectional LSTM, bidirectional LSTM (BLSTM) has been utilized to learn information both from the past and the future inside the network (unidirectional LSTM only learns from the past). In [78], BLSTM is used for the textual network rather than the acoustic network. This bidirectional LSTM is often combined with an attention model to boost the performance of the SER task [13]. However, using BLSTM doubles the model's complexity, making the model may not suitable for real-time applications.

2.5 Fusion Methods

Multimodal fusion in technology is the combination of information that comes from different sources [80]. This terminology is similar but different to the human multimodal perception. In human multimodal perception, the information comes from different sensory organs; in technology, this requirement is not necessary. Multimodal fusion can be viewed as multisensor data fusion. In this terminology, the 'sensor' is the soft sensor. Acoustic and linguistic feature extractors can be regarded as soft sensors in bimodal acoustic-linguistic information fusion.

Fusing acoustic and linguistic features have been attempted at the early stage of speech emotion recognition research. The first work on fusing acoustic with linguistic information has been performed by Lee et al. [60] by combining acoustic and language features at the decision level using logical "OR". If at least one decision corresponds to a specific emotion, then the result is this specific emotion. This earliest work only involved negative and non-negative emotion categories.

Fusing acoustic and linguistic information for SER can be accomplished in several ways, Figure 2.6 shows the classification. Early fusion combines acoustic and linguistic information at the feature level; late fusion combines results from acoustic and linguistic information at the decision level. Early fusion, furthermore, can be split into three main categories: feature concatenation, networks/model concatenation, and hierarchical model. Hierarchical model, as proposed in [81, 44], can be regarded as early fusion since the method fuses features at a different level of layers, not at the decision level.

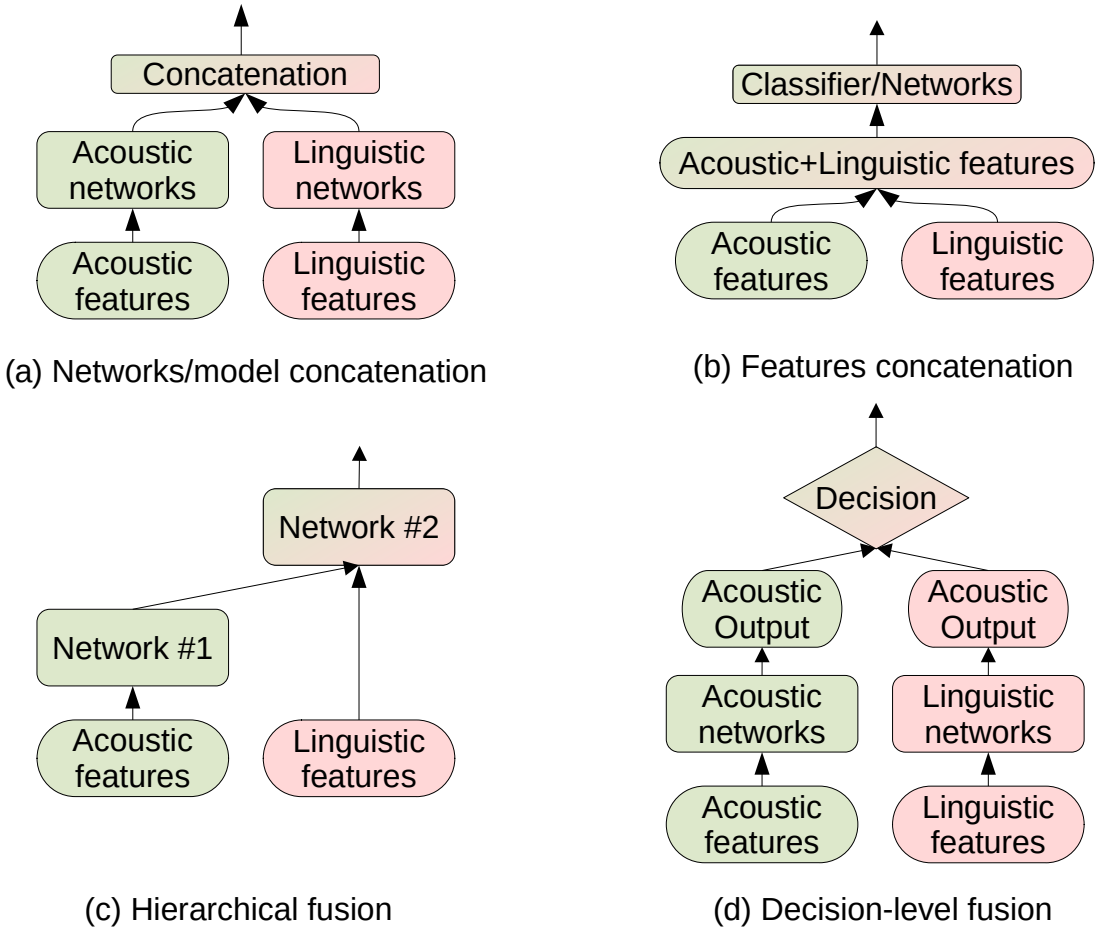

Figure 2.6: Different scheme of fusing acoustic with linguistic information; (a), (b), (c): early fusion approach; (d): late fusion approach

Eyben et al. [42] proposed an online method to detect not only valence and arousal but also the time when those emotion attributes are detected. They used a recurrent neural network (RNN) based on long short-term memory (LSTM) to recognize a frame-wise valence-arousal continuum with time. By adding a keyword spotter, they were able to improve the performance by using regression analysis. The results were measured in

Pearson correlation coefficient (PCC). They also found that keywords like "again," "angry," "assertive," and "very" were related to activation, while typical keywords correlated to valence were "good," "great," "lovely," and "totally." Similar to that idea, Karedogan and Larsen [82] used affective words from Affective Norms for English Words (ANEW) to determine a valence-arousal value and combine it with a result from acoustic features. The latter paper also obtained similar improvement over using a single modality.

Ye and Fan [43] used bimodal features from acoustic and text information to recognize emotion within speech. The acoustic features were trained in two parallel classifiers: an SVM and a backpropagation network. The text features were trained in two serial classifiers, which were both Naive Bayes classifiers. The second classifier acted as a filter for unreliable parts from the first classifier. Decision-level fusion (late fusion) was then implemented by combining the acoustic and text features with tree-weighting factors for the SVM, backpropagation network, and text classifiers. The resulting fusion method obtained 93% accuracy, as compared to 83% from the acoustic features only and 89% from the text features only. The task was categorical emotion detection from a Chinese database. Similar to that approach for a categorical task, Jin et al. [83] used the IEMOCAP dataset to test combinations of acoustic and text features for SER. The novelty of their method was the use of an emotion vector for lexical features, which improved the accuracy in four-class emotion recognition from 53.5% (acoustic) and 57.4% (text) to 69.2% (acoustic + linguistic).

Aldeneh et al. [63] used acoustic and lexical features to detect the degree of valence from speech. They used 40 mel-filterbanks (MFBs) as acoustic features and word vectors as linguistic features. Continuous valence values were then converted to three categorical classes: negative, neutral, and positive. Using that approach, they improved the weighted accuracy from 64.5% (text) and 58.9% (acoustic) to 69.2% (acoustic + linguistic).

Yoon et al. [84] used audio and text networks to predict emotion classes from the IEMOCAP dataset. Both networks used RNNs with inputs of mel-frequency cepstral coefficients (MFCCs) for audio and word vectors for text. The proposed multimodal dual recurrent encoder (MDRE) improved on the single-modality RNNs from 54.6% (audio) and 63.5% (text) to 71.8% (audio + text). Atmaja et al. [18] obtained a better result by using 34 acoustic features after silence removal and combining them with word embeddings. With LSTM used for the text and dense networks for speech, the latter paper obtained an accuracy of 75.49% on the same dataset and task (categorical emotion recognition).

Instead of using lexical features, Zhang et al. [64] used phonemes and combined them with acoustic features to recognize valence in speech. They used 39 unique phonemes from the IEMOCAP and MSP-IMPROV datasets and a 40-dimensional log-scale MFB energy for the acoustic features. Using a scaled version of valence, converted from a 5-point scale to three categorical classes, they showed that their multistage fusion model outperformed all other models on both IEMOCAP and MSP-IMPROV.

2.6 Summary

This chapter reviews research on bimodal speech emotion recognition from acoustic and linguistic information. The study focuses on four building blocks of SER from bimodal information: emotion models, features, classifiers, and fusion methods. There are three

emotion models developed in psychological research; however, most SER research focused on the categorical model. There is a move to extract acoustic features in the feature extraction step by using deep-learning methods, while deep-learning-based linguistic features already dominated text processing research, including SER from linguistic information. Finally, four common classifiers for SER are briefly described. Although more advanced DNN architectures have been developed, bimodal SER still relies heavily on SVM, MLP, CNN, and LSTM architectures. The fusion of different information can be performed in several methods; these methods can be divided into early and late fusions. While this literature study presents the current state of speech emotion recognition research in these four blocks, the raised issues in SER research will be highlighted in the next Research Methodology chapter along with other related sections.

Chapter 3

Research Methodology

The purpose of this chapter is to introduce the research methodology, the justification for using particular research methods, for this experimental study on dimensional speech emotion recognition by fusing acoustic and linguistic information. This approach allows the examination of the contribution of fusing bimodal information for more accurate dimensional speech emotion recognition. The research issues and philosophies used to tackle these issues are discussed in this chapter. The implementation of research philosophy through research strategies are highlighted. The experimental methods, including datasets and an evaluation metric, are also the primary components of the research methodology presented to close this chapter.

3.1 Research motivation

3.1.1 Why is SER difficult?

Speech emotion recognition (SER) is a difficult task. It differs from other traditional tasks like image recognition (cat vs. dog), digit recognition, or speech recognition. For instance, the features and labels for recognizing cat or dog in image recognition are both clear. The difference in eyes, ears, skin color and shape between cat and dog can be used as informative features. The labels also clear, either cat or dog, with a very high level of confidence. The digit recognition problem has similar properties; from zero to nine can be distinguished by their shapes; this input feature is an important factor for the obtained high accuracy classification. Automatic speech recognition (ASR) also has similar properties to both cases. SER is different from both.

In SER, both features and labels are not clear. Researchers have attempted to find useful features related to affective states (e.g., [85, 56]). In annotating labels for categorical or dimensional emotion, the datasets makers rely on subjective evaluation. This means that the labels are not exact values (e.g., compared to image labels). However, high agreements among evaluators show the reliability of the datasets. In this SER problem, it is almost impossible to obtain perfect accuracy, which is possible to obtain in the previous image recognition problems.

3.1.2 Why dimensional SER?

While most SER research focus on categorical emotion recognition, only a few focuses on dimensional SER. In contrast, the present evidence about the "fingerprint" in categorical emotion, particularly based on facial expression, is weak [34]. As Darwin argued that biological categories, including emotion categories, does not have an essence due to the high variability of individuals [86], so does emotion categories.

Dimensional emotion may represent humans' affective state better than categorical emotion. Humans do not perceive emotion categorically but in continuous space. In this case, Russel argued that emotion categories could be derived from valence-arousal space [87]. Given this understanding, dimensional SER is more challenging (since it predicts degree) and more useful (since categorical emotion also can be derived) than categorical emotion recognition.

3.1.3 Why fusing acoustic and linguistic information?

The third motivation is about the use of linguistic information. The simplest answer to this question is that linguistic information is also can be extracted from speech, and language is related to emotion. In other words, two information can be extracted from speech without adding other modalities to recognize expressed emotion. Hence, fusing acoustic and linguistic information is reasonable and feasible for future implementation.

There are other possible motivations for fusing acoustic and linguistic information for dimensional SER. One is from human multimodal processing, which uses linguistic information as a cue for emotion perception [10]. Another reason is that linguistic information is widely used in sentiment analysis tasks, which are closely related to emotion recognition. Sentiment analysis can be viewed as emotion recognition in text, which focuses on sentiment or valence prediction.

In the speech chain, acoustic and linguistic are connected by a physiological mechanism [88]. This chain shows a direct correlation between linguistic and acoustic information. Fusing both information may improve the prediction of the conveyed message. Since the message is elicited from the same sources (e.g., thought, information, including expressed emotion), using both information is a straightforward way to track the transported information, in this case, the expressed emotion from a speaker.

3.2 Research issues

There are five issues discussed in this dissertation. The issues appeared from the previous SER studies [36, 11]. These issues raise in SER from acoustic information, dimensional SER, and multimodal SER. Although these issues are important, the previous chapter on literature study has found no thorough study investigated these five issues. The importance of these issues and the contributions of this study to these issues are summarized below.

The first issue is the region of analysis used for feature extraction in dimensional SER, local areas in frames, or whole utterance processing. The importance of addressing this issue is to determine which region of analysis is informative to extract dimensional emotion from speech. The traditional way to extract acoustic features in acoustical signal processing is frame-based processing. In this way, an utterance is split into several frames

in a fixed duration, e.g., 25 ms. A window function applies to these frames, and the intended acoustic features are extracted on these frames. This local feature extraction is known as a low-level descriptor (LLD). In contrast, the newer research on speech emotion recognition proposed to extract statistical functions based on these LLDs. These global features are known as high-level statistical function (HSF). Form both features, it is unclear which one performs better; one claimed that LLDs are enough since it is highly correlated with emotion (e.g., tone and prosody [59]), others claimed that global features are better in classification accuracy and classification time (perhaps due to small feature size) [36]. This study reveals the significant contribution of specific method for solving the effectiveness of region analysis for acoustic feature extraction.

The second issue is the effect of the silence region in dimensional speech emotion recognition. This issue is important to know whether such post-processing technique contributes to dimensional SER. In conventional ways, acoustic features are extracted from speech utterance, including the silence region. Some removed silence before extracting acoustic features (e.g., [13, 89, 56, 90]) and some used silence as a feature (e.g., [14, 91]) or as an emotion category [92]. Although silent pause is an important cue for human speech communication [93, 94], it is unclear whether silence contributes to human-computer interaction (HCI). This study contributes to this effect of silent pause region by predicting the role of the region to dimensional SER.

The third issue is the low score of valence prediction in dimensional SER. Among the three emotion dimensions, valence always obtained lower scores than arousal and dominance. The previous study confirms this evidence [11]. Considering valence is the most important emotion dimension [53], the need to improve valence prediction score is worthwhile of study. Although there are several attempts to improve valence prediction [64, 63, 95], the obtained score is still not comparable to scores achieved by arousal and dominance (e.g., in [95]). This study proposed and discussed a method to double the performance of valence prediction in dimensional SER.

The fourth issue is whether it suffices to use acoustic features for modeling emotions or if it is necessary to fuse them with linguistic features. Since linguistic information can be obtained from speech (via ASR), it is reasonable to fuse linguistic information with acoustic information. In other studies, it found that linguistic information helps to improve valence prediction [82]. Fusing acoustic and linguistic information may improve not only valence prediction but also other emotion dimensions. This study reveals the necessity of fusing both acoustic and linguistic information for dimensional SER.

The fifth issue is the scheme or framework to fuse linguistic and acoustic information. If the linguistic information contributes to dimensional emotion prediction, what the most appropriate approach to fuse acoustic and linguistic information is. In human multimodal emotion perception, how multimodal information are fused is not clear yet. Both acoustic and linguistic information are believed to process separately in different regions of the cortex (right and left hemisphere). In HCI, the simplest method to fuse multimodal information is by concatenating input features from all modalities. This study showed the effectiveness of late fusion approach over early fusion approach for combining acoustic and linguistic information for dimension SER.

3.3 Research philosophy

Research philosophy can be defined as "a belief about the way in which data about a phenomenon should be gathered, analyzed and used [96]." This research used data-information-knowledge hierarchy (DIK) concept, which is known as the canon of information science [97]. Figure 3.1 shows this DIK concept and its representation in the speech emotion recognition area. Although some researchers defined these concepts in different ways [98, 99], the following concepts are the proper and valid explanation used in this research.

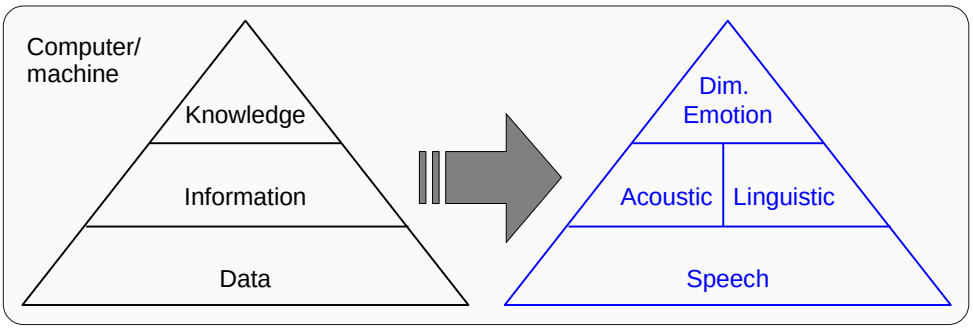

Figure 3.1: The DIK hierarchy and its representation in speech emotion recognition; information (I) is extracted from data (D); knowledge (K) is extracted from information.

Data: speech

Data is worth nothing. According to Cho [100], signals structure the data. Acoustics signal composes speech; speech is the data. In human communication, speech by a speaker is data for the listener. In HCI, the speech dataset is a collection of recorded utterances from speakers elicited intended expressions. An emotional speech dataset is a type of speech dataset that provides utterances with their emotional labels in categorical or/and dimensional emotions.

Information: acoustic and linguistic

Information is know-what. It is the relevant, usable, significant, meaningful, or processed data [101]. Information is extracted from data. Humans arguably perceive the speaker's emotions from their acoustics and linguistic information [102]. How both information fused is not clear until now; however, researchers suggest that both perceived information are processed separately [103, 9] (verbal information in the left cortex and non-verbal information in the right cortex). In HCI, acoustic information can be extracted directly from speech, while linguistic information needs a mediator, i.e., speech-to-text system, to extract words from speech. Both information are represented as features. Feature extraction is the process of extracting acoustic and linguistic information from the speech dataset and its transcription.

Knowledge: dimensional emotions degree

Knowledge is know-how. It is extracted from the information. Knowledge transforms information into instructions (mapping). In human communication, the knowledge to

perceive emotion is innate rather than learned [104]. In dimensional emotion, this knowledge is the degree of valence (V), arousal (A), and dominance (D) [e.g., V:4, A:4, D:2]. The process of mapping information (features) to dimensional emotion degree is regression task which is performed by such regressors.

3.4 Research strategy

A research strategy is the steps or ways in which research's goals could be achieved. The goal of this study is to answer the research issues presented previously. This research studies three strategies to answers these research issues. The following is the description and rationale of the strategies.

3.4.1 Dimensional SER by acoustic information

This study evaluated dimensional SER based on acoustic features only to tackle the first and second issues. This study aims at maximizing the potency of acoustic-based SER. First, the region of analysis is investigated; a thorough comparison among three conditions are performed: (1) extracting acoustic features from silence-removed regions, (2) extracting acoustic features from the whole region, including silence, and (3) extracting features from the whole region and utilizing a silence feature as an additional feature. Second, this study evaluated which aggregation method performs better: input features aggregation or outputs aggregation. The common approach in aggregation is output aggregation by majority voting; however, input aggregation may perform better, particularly dimensional SER. For instance, humans perceive emotion from acoustic information (e.g., tone) to recognize the emotion based on that information (rather than aggregating emotion/outputs). As found in other SER research, this study contributes the necessity to go beyond acoustic-only SER.

3.4.2 Fusing acoustic and linguistic information at feature level

In certain cases, it may be difficult for a listener to perceive the speaker's emotion from acoustic information only. For instance, both joy and angry may have a similar intonation (e.g., high tone); hence it is difficult to differentiate both emotions. By knowing the semantic of utterances, it may be easier to judge the expressed emotion for both human communication and human-computer interaction. This case raises an opportunity to investigate whether linguistic information contributes to dimensional emotion recognition. Although the study of this phenomenon has been performed previously (e.g., in [82]), several limitations exist. The emotion model, the used linguistic information, and the classification framework have evolved since the publication.

Apart from the need for multimodal/bimodal information fusion, linguistic information has been actively developed for sentiment analysis, analyzing text to obtain the affective state of the writer (positive or negative). This 'sentiment' term reflects directly to valence; hence, one possible solution to improve low valence score in dimensional SER is by utilizing linguistic information. Research showed that utilizing linguistic information improves both categorical [84] and dimensional [82] emotion recognition. Fusing acoustic and linguistic information tackle both the third and fourth issues: the necessity of using linguistic information and the low score valence prediction.

A simple approach to fusing acoustic and linguistic information is fusion at the feature level. In this approach, either features or networks can be concatenated to predict dimensional emotions. In the first method, all features are inputted to the same classifier, while in the latter, both information may have different classifiers (networks). In the latter method, additional networks are needed to fuse both networks, typically a type of dense networks (also called as fully connected [FC] networks or multilayer perceptron [MLP]).

Although there are several studies that focus on the linguistic and acoustic features fusion for SER at the feature level, this study differs in several aspects. First, this study evaluated both feature concatenation and network concatenation. Second, this study proposed correlation-based multitask learning (MTL) to predict valence, arousal, and simultaneously from both acoustic and linguistic information. Third, this study contributes to a comparison of manual and automatic transcription for acoustic-linguistic dimensional SER

3.4.3 Fusing acoustic and linguistic information at decision level

To extend the fourth issue, it is not only necessary to study the fusing of acoustic and linguistic information at the feature level but also at the decision level. This strategy is motivated by human multimodal processing. The neural mechanism on how the brain processes multimodal information suggests that each information is processed in a separate brain region. Hence, a late fusion approach, i.e., decision-level information fusion, may work better than feature-level fusion. Apart from investigating which fusion method is better to combine bimodal information (fifth issue), this strategy can also be used to investigate the third and fourth issues.

This last strategy contributes to investigate which framework performs better for fusing acoustic and linguistic information. Although there is an argument that any fusion approach will perform similarly [105], the opposite also has been argued [106]. Consistency found in this study (that late fusion is better than early fusion) may help the future research on dimensional SER and trigger more ways to fuse both acoustic and linguistic information for SER.

3.5 Datasets

The strategies to answer research issues need several instruments to experiment with. One key component in this dimensional SER research is the dataset. Three emotional speech datasets have been chosen for different experiments. These three datasets are explained below.

1. IEMOCAP
IEMOCAP, which stands for interactive emotional dyadic motion capture database, contains dyadic conversations with markers on the face, head, and hands. The recordings thus provide detailed information about the actors' facial expressions and hand movements during both scripted and spontaneous spoken communication scenarios [30]. This research only uses acoustic and linguistic features because the goal is bimodal speech emotion recognition. The IEMOCAP dataset is freely available upon request, including its labels for categorical and dimensional emotion. This study uses dimensional emotion labels (valence, arousal, dominance), which are average scores for two evaluators because

they enable deeper emotional states analysis. The dimensional emotion scores, for valence, arousal, and dominance, are meant to range from 1 to 5 as a result of Self-Assessment Manikin (SAM) evaluation. It has been found that some labels with scores lower than 1 or higher than 5. Either removing those data (seven samples) or converting them into neutral was chosen in different experiments. All labels are then converted from the 5-point scale to a floating-point values range [-1, 1] when fed to a DNN system.

The total length of the IEMOCAP dataset is about 12 hours, or 10039 turns/utterances, from ten actors in five dyadic sessions (two actors each). The speech modality used to extract acoustic features is a set of files in the dataset with a single channel per sentence. The sampling rate of the speech data was 16 kHz. The manual transcription in the dataset without additional preprocessing is used for text data except for comparing it with ASR outputs (chapter 5).

2. MSP-IMPROV

MSP-IMPROV [31], developed by the Multimodal Signal Processing (MSP) Lab at the University of Texas, Dallas, is a multimodal emotional database obtained by applying lexical and emotion control in the recording process while also promoting naturalness. The dataset provides audio and visual recordings, while text transcriptions are obtained via automatic speech recognition (ASR) provided by the authors. As with IEMOCAP, the speech and speech+text data with dimensional emotion labels were used in different experiments. The annotation method for the recordings was the same as for IEMOCAP, i.e., SAM evaluation, with rating by at least five evaluators. Some data with missing evaluations were treated as neutral speech (i.e., a score of 3 for valence, arousal, and dominance). Also, as with IEMOCAP, all labels are converted to floating-point values in the range [-1, 1] from the original 5-point scale.

The MSP-IMPROV dataset contains 8438 turns/utterances in more than 9 hours. Similar to IEMOCAP, there are two speakers for each session. The number of sessions is six. Originally the dataset is divided into four scenarios: "Target-improvised" and "Target-read," "Other-improvised," and "Natural-interaction." This allotment was designed to evaluate the effect of target sentences. The whole dataset is used for acoustic-only emotion recognition. The parts of MSP-IMPROV, excluding "Target-read," are used for acoustic-linguistic information fusion. A further explanation about this dataset will be added in the explanation of the experiment involving this dataset (Chapter 6).

3. USOMS-e

Ulm State of Mind in Speech-elderly (USOMS-e) dataset is the corpus used in the elderly emotion sub-challenge in the INTERSPEECH 2020 computational paralinguistic challenge. The whole dataset subset is used with 87 subjects aged 60 – 95 years; 55 of the subjects were male, and the rest 32 were female. The dimensional emotion labels were given in valence and arousal divided into three categories: low, medium, and high.

Table 3.1 shows the number of instances/stories and chunks in all partitions. The labels are given per each story. The label on the dataset is given on both alphabetic and numeric symbols, i.e., low ('L' or '0'), medium ('M' or '1'), and high ('H' or '2'). This research used alphabetic labels as given in the baseline paper. Note that the number of chunks is different for each story; for instance, there are 34 chunks in the first story and 46 chunks in the second story.

Table 3.1: Number of instances and chunks in each partition USOMS-e dataset

Partition	# Stories (text)	# Chunks (audio)
Train	87	2496
Dev	87	2466
Test	87	2816
Total	261	7778

3.6 Evaluation metric

Apart from the datasets, a metric to measure the performance of proposed/evaluated methods is needed to evaluate the research. Instead of using several metrics, this research focus on the use of concordance correlation coefficient (CCC) as a single metric to evaluate the performance of dimensional SER. This metric is proposed to be the standard metric for dimensional SER previously [107]. The formula to calculate CCC is given as,

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{3.1}$$

where ρ is the Pearson correlation coefficient (PCC/CC) between predicted emotion degree x and true emotion degree y, σ^2 is a variance and μ is a mean. This metric is more challenging than the correlation coefficient since it penalizes the score, even the correlation is well but shifted. The penalized values are in proportion to the deviation.

3.7 Summary

This chapter presents the research methodology for studying dimensional SER by fusing acoustic and linguistic information. The motivations to choose this research theme are discussed, and the raised issues are presented. These five issues are region of analysis for acoustic feature extraction, effect of silent pause features, low valence prediction, the necessity for fusing acoustic with linguistic information, and framework for fusing acoustic with linguistic information. These issues have never been studied thoroughly in the previous studies. The importance of each issue and the contribution of this study to each issue are briefly described. Three strategies are highlighted to address these issues, including the datasets to evaluate the strategies and a metric to measure the performance. The proposed strategies investigate the necessity of go to beyond acoustic information and the necessity to fuse acoustic with linguistic information for dimensional SER. The next three chapters discuss each strategy proposed in this chapter, followed by a chapter on Comparative Analysis and a Conclusions chapter to end the discussion.

Chapter 4

Speech Emotion Recognition Using Acoustic Features

The purpose of this chapter is three-fold: (1) to investigate the effective region of analysis for acoustic feature extraction, whether frame-based region (local features) or utterance-based region (global features); (2) to evaluate which action is the best with regard to the silence region in a dimensional speech emotion recognition (SER); and (3) to evaluate which aggregation method performs better for dimensional SER: acoustic input aggregation or output aggregation (e.g., majority voting method).

4.1 Which region of analysis to extract acoustic features in SER

4.1.1 SER using low-level acoustic features

SER in conventional ways are performed by extracting acoustic features on frame-based processing and then applied these features to a classifier. Let $y(n)$, with $n = 1, 2, 3, \ldots, L$, denotes acoustic signal with length L. In frame-based processing, this $y(n)$ signal is divided into frames by fixed length. A typical length for a single frame is 16-25 milliseconds (ms) with 10 ms to 15 ms hop length (stride). For 25 ms frame length and 10 ms hop length, which is equal to 60% overlap (15 ms), a window is applied to this frame to make the short-time signal behave as a quasistationary signal – near the stationary signal. In their original length, an acoustic signal varies with the time: non-stationary property. Windowing processes the acoustic signal in a short-term interval to remove this property. Figure 4.3 shows the windowing process; short-term windowed signals look stationary more than the original signal.

Windowing multiplies the spectrum of an input signal with window signal $w(n)$. A typical window function for an acoustic signal is Hann and Hamming windows (named after Julius von Hann and Richard W. Hamming). The others are rectangular, Bartlet, Kaiser, and Blackman. The choice of the window function is based on two aspects: the width of the main lobe and the additional lobes. Hann and Hamming windows only differ in weighting factors with similar concept: cosine-sum windows

$$w[n] = A + B\cos\left(\frac{2\pi n}{M}\right), \qquad n = -M/2, \ldots, M/2, \tag{4.1}$$

where A is 0.5 for Hann and 0.54 for Hamming. B is 0.5 for Hann and 0.46 for Hamming. Both window functions are widely used in speech processing due to a good trade-off between time and frequency resolution (effect of side lobes). Figure 4.1 shows a Hann window and its spectrum, while Figure 4.2 shows an example of a Hamming window applied to a sinusoid signal and its result.

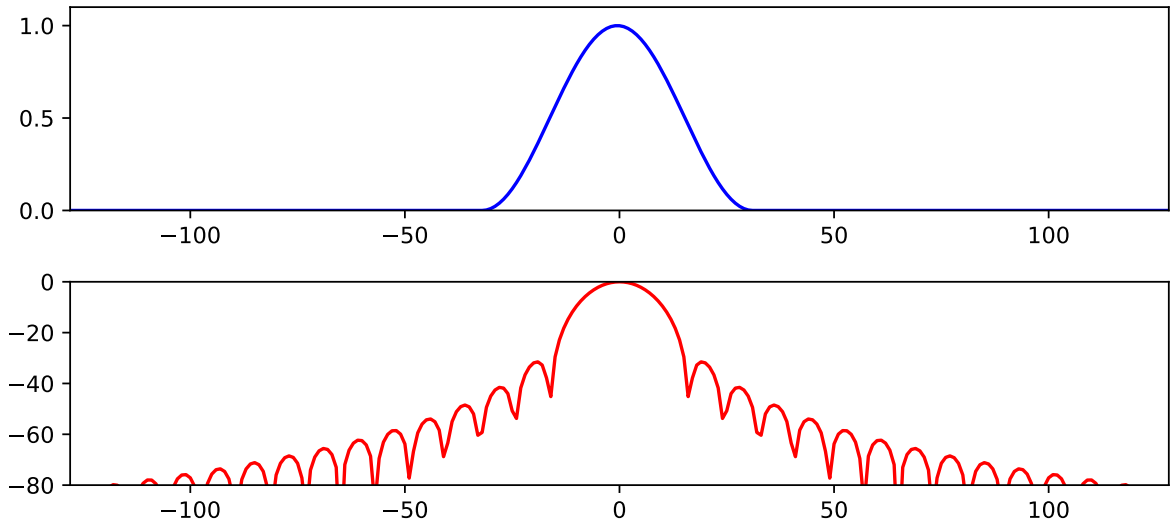

Figure 4.1: Hann window and its spectrum

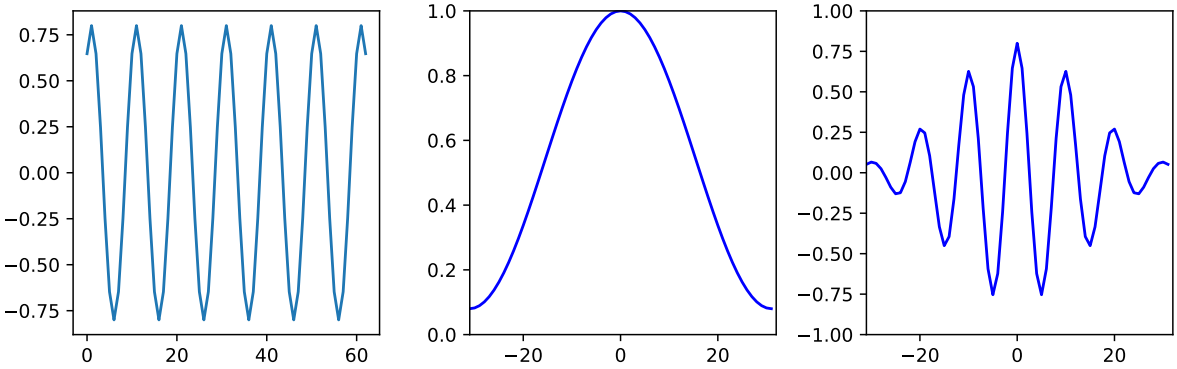

Figure 4.2: An example of Hamming window (middle) applied to sinusoid signal (left); the resulted windowed signal (right) is multiplication of both.

The length of a window is usually equal to the length of the frame: one window per frame. If the length of a window is smaller than a frame, each frame will be windowed with window length and padded with zeros to match the frame's length. In speech emotion recognition, a short window is used to capture short dynamics context while a longer window is used to capture mid and longer dynamics. A common approach used a short window to extract acoustic features in short-term time while statistical functions model long-term dynamics. Figure 4.3 shows the frame-based processing of an acoustic signal (speech), which windows short-term signals using the Hamming window.

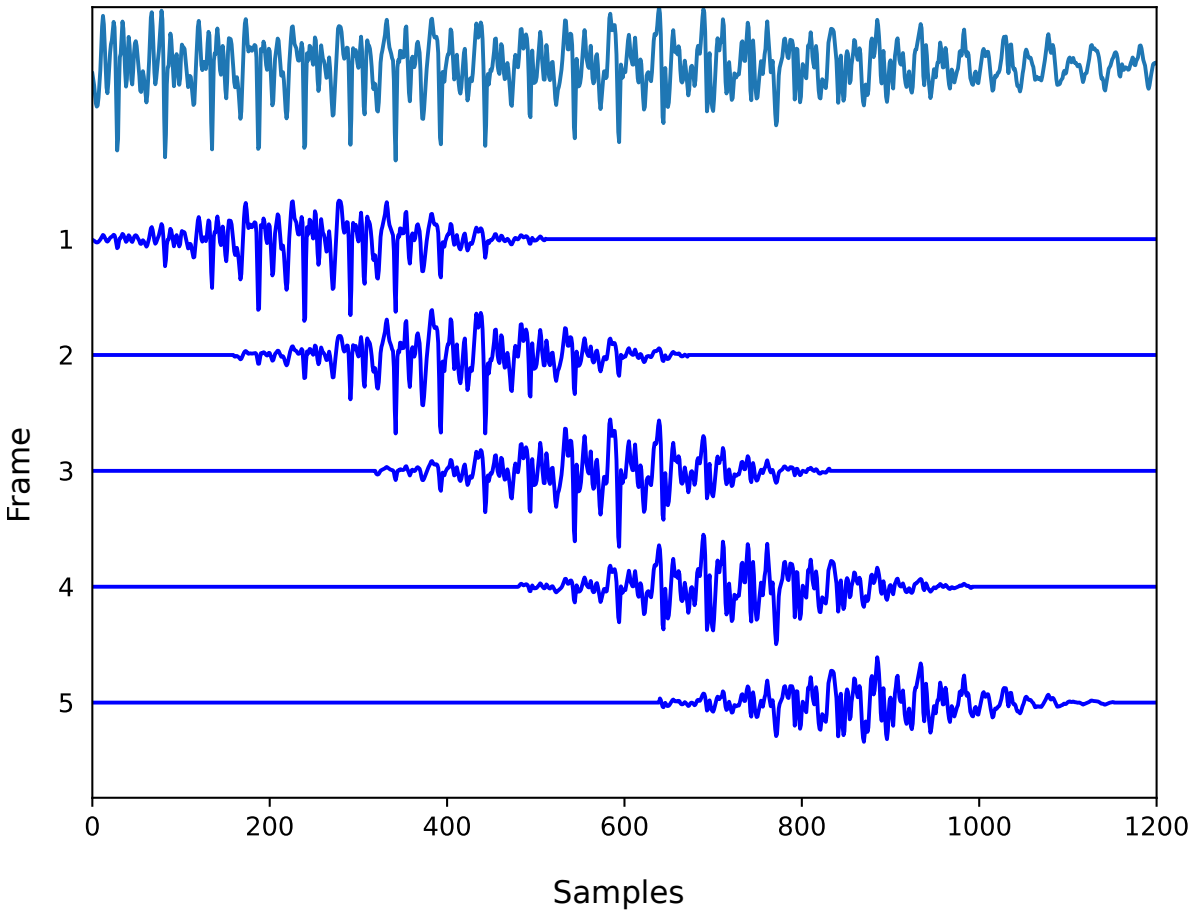

Figure 4.3: Frame-based processing for extracting low-level descriptors of an acoustic signal; the signal is an excerpt of IEMOCAP utterance with 400 samples frame length and 160 samples hop length; sampling frequency is 16 kHz.

The acoustic features extracted on each frame are known as local features or low-level descriptors (LLD) [108]. The most common LLD for speech processing is mel-frequency cepstral coefficients (MFCC). MFCC captures different aspects of the spectral shape of a speech. The following steps compute MFCC in sequences. First, FFT/DFT transformed a time-domain signal into a frequency domain signal (spectra). Second, the mel frequency warping function converts spectra in linear scale into the mel scale. Although several functions have been proposed, a common approach keeps linear scale for acoustic frequencies below 1 kHz and converts to logarithmic scale for acoustic frequencies above 1 kHz. This conversion imitates the human perceptual system. Third, convert a power spectrogram (amplitude squared) to decibel (dB) units (log). Finally, DCT computes MFCC as amplitude cepstra.

One of the important parameters in MFCC is the number of coefficients. A number of 13 to 40 coefficients are common for speech processing. For each frame, 13 MFCCs are extracted. If there are 40 frames in an utterance, the dimension of MFCC features will be (40, 13). The number of frames corresponds to the number of samples divided by hop length (in samples). If an utterance comprises 1 second (s) with a 16 kHz sampling rate, the number of samples is 16000. Using 25 ms (400 samples) window/frame length and 10 ms (160 samples) hop length, the number of frames is 16000/160, i.e., 100 frames. Figure 4.4 top shows an MFCC spectrogram of an IEMOCAP utterance with 13 coefficients.

Recently, researchers found that mel-spectrogram, also called as (mel) filterbank or mel-frequency spectral coefficients (MFSC), yields better performance for deep learning-based automatic speech recognition (ASR) (e.g., [109]). Given that a deep learning system is less susceptible to highly correlated input, the DCT step in the previous MFCC calculation is not necessary since it is a linear transformation. DCT discards some information in speech signals, which are highly non-linear [110]. Furthermore, a log version (in decibel unit) of mel-spectrogram, i.e., log mel-spectrogram, is preferable since deep learning learns better in this unit. The conversion from mel-spectrogram to log mel-spectrogram is given by

$$S_{dB} = 10 \log\left(\frac{S}{ref}\right), \tag{4.2}$$

where S is the input power spectrogram and ref is reference power. A value of 1.0 is a common ref value for 32-bit floating-point wav data ('float32').

Figure 4.4 shows visualization of MFCC, mel-spectrogram and log mel-spectrogram. From this figure, it is clear that the log mel-spectrogram is more informative than the mel-spectrogram and MFCC. This visualization may support the previous argument that log mel-spectrogram may works better in DNN-based speech emotion recognition.

Apart from the use of one type of acoustic features for speech processing, some researchers have proposed a set of acoustic features for speech emotion recognition. Eyben et al. [4] proposed Geneva minimalistic parameter set (GeMAPS) as standard acoustic features for affective voice research. The proposed acoustic features are based on (1) physiological changes in voice production, (2) proven significance in previous studies, and (3) theoretical significance. The proposed acoustic feature set comprises 23 LLDs, as shown in Table 4.1. This acoustic feature set is extracted on a frame-processing basis with 25 ms frame length and 10 ms hop length.

Giannakopulos [5] proposed pyAudioanalysis as an open-source Python library for audio signal analysis. The library supports a wide range of audio analysis procedures such

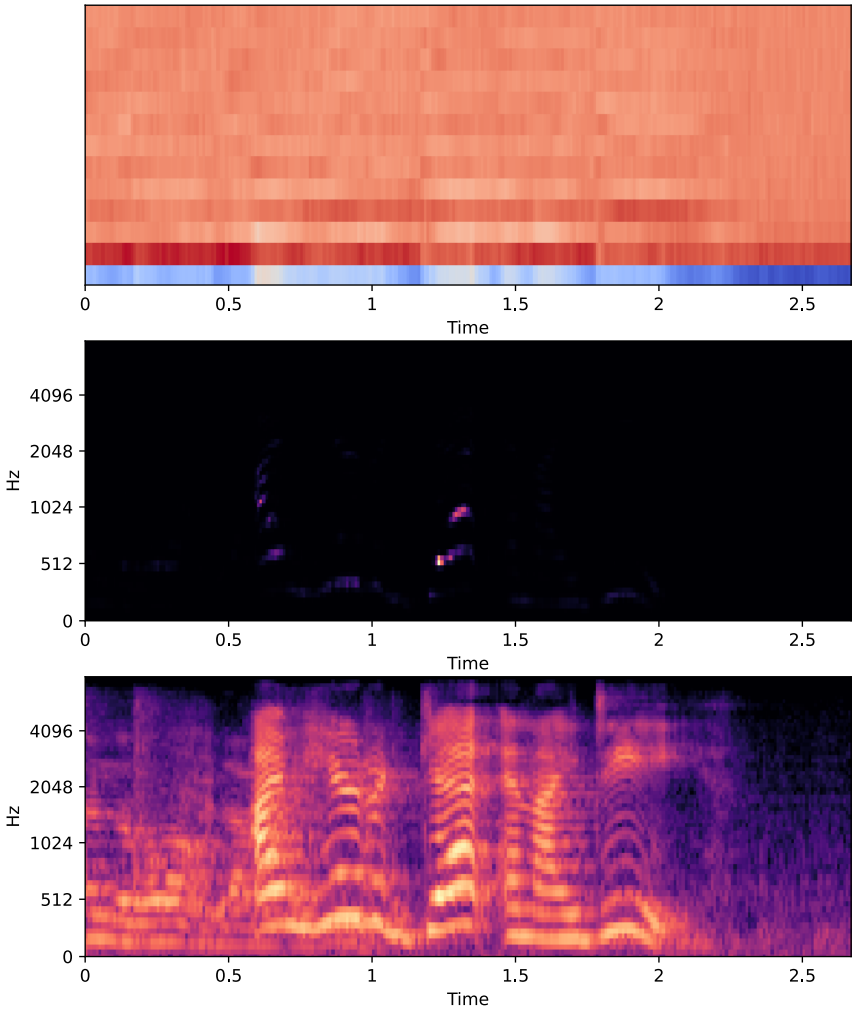

Figure 4.4: Visualization of MFCC features with 13 coefficients (top), mel-spectrogram (middle), and log mel-spectrogram with 64 mels (bottom)

as feature extraction, classification, supervised and unsupervised segmentation, and visualization. Different from GeMAPS feature set, pyAudioanalysis targets a wide range of voice applications like audio event detection, speech emotion recognition, music segmentation, and health application. The short-term feature set, which is extracted on a frame-processing basis, consists of 34 LLDs. These LLDs are shown in Table 4.1.

Table 4.1: Acoustic feature sets: GeMAPS [4] and pyAudioAnalysis [5]. The numbers in parentheses indicate the total numbers of features (LLDs).

GeMAPs (23)	pyAudioAnalysis (34)
intensity, alpha ratio, Hammarberg index, spectral slope 0-500 Hz, spectral slope 500-1500 Hz, spectral flux, 4 MFCCs, f_o, jitter, shimmer, harmonics-to-noise ratio (HNR), harmonic difference H1-H2, harmonic difference H1-A3, F1, F1 bandwidth, F1 amplitude, F2, F2 amplitude, F3, and F3 amplitude.	zero crossing rate, energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off, 13 MFCCs, 12 chroma vectors, chroma deviation.

As additional features sets, a temporal difference on pyAudioAnalysis were computed in the first order, referred to *deltas*. The addition of the first-order regression coefficients show better performances than original LLDs (MFCC and MFSC) in ASR. In dimensional SER, these temporal differences may show the dynamics between frames. Together with the previous four feature sets, these LDDs are compared to evaluate the effectiveness of frame-based LLDs in dimensional SER.

Table 4.2 shows performance of dimensional SER from IEMOCAP dataset in CCC scores. There is no remarkable difference in the use of different common acoustic features used in acoustic signal processing. This presented results also challenged the specially designed acoustic features namely GeMAPS [30] which is proposed to be the standard feature set for voice research and affective computing [4]. Although it achieves the highest score in LLD comparison, the general-purpose pyAudioAnalysis (pAA) features set attained a comparable performance to GeMAPS, and in later analysis it will be shown that this feature set achieve a higher performance than GeMAPS on utterance-based feature extraction.

4.1.2 SER using high-level acoustic features

In the previous subsection, it is shown that frame-based acoustic features work with limited performance. In this subsection, the effectiveness of two statistical functions is shown. Two high-level acoustic features, i.e., mean values and standard deviations from LLDs, are evaluated from the previous five acoustic feature sets.

The first high-level acoustic features used for this dimensional SER task are mean values. The idea of using these mean values is to capture the shared information across all frames. This mean values can be formulated as:

Table 4.2: Results of frame-based LLDs for dimensional SER in IEMOCAP dataset

Feature	Dim	Val	Aro	Dom	Mean
MFCC	(3414, 40)	0.148	0.488	0.419	0.352
Log mel	(3414, 128)	0.103	0.543	0.438	0.362
GeMAPS	(3409, 23)	0.164	0.527	0.454	0.382
pAA	(3412, 34)	0.130	0.513	0.419	0.354
pAA_D	(3412, 68)	0.145	0.526	0.439	0.370

$$\mu_F = \frac{1}{\mathcal{K}} \sum_{i=1}^{n} F_i \tag{4.3}$$

where $\mathcal{K}$ is the number of frames, and F is the corresponding feature. For instance, in pyAudioAnalysis, the first feature is a zero-crossing rate (ZCR). The ZCR feature's mean value is the arithmetical mean of all ZCR values in all frames within an utterance.

The second high-level acoustic features are standard deviation (std). This statistical function shows the dispersion of feature values from its mean. While mean is intended to capture the commonalities among features values in all frames within an utterance, std is intended to capture the dynamics of feature values in an utterance. Accordingly, std is formulated as follows,

$$\sigma_F^2 = \frac{1}{\mathcal{K}} \sum_{i=1}^{n} (F_i - \mu_F)^2. \tag{4.4}$$

Both mean and std (Mean+Std) are known as valuable functions in SER. References [77, 111, 112] have used Mean+Std for categorical SER. However, most references did not use only Mean+Std, but other statistical functions like median, quartiles, minimum, maximum, and other features. It is interesting to experiment with Mean+Std only for dimensional SER given the fact that both high-level features are two most informative descriptors, among other statistical functions. Besides reducing the size or dimension of features significantly, using Mean+Std consequently speed up the computation of SER with regard to their small input features.

Another advantage of using Mean+Std features is no need for zero paddings. Although zero paddings are useful for FFT calculation (spectral smoothness), it is unclear the effect of zero paddings on LLDs for acoustic feature extraction. Zero padded values may impact information represented by features when such processing is performed, e.g., standardization or normalization. Hence, acoustic features represented by Mean+Std features are more informative than LLDs since it only contains information from speech.

An illustration of Mean+Std extraction from LLDs is shown in Figure 4.5. For instance, an MFCC feature set from an utterance consists of 3414 frames with 40 MFCC coefficients. For each mean and std features, 40 values are calculated. Both mean and std are concatenated to form Mean+Std features after transposing both statistical features.

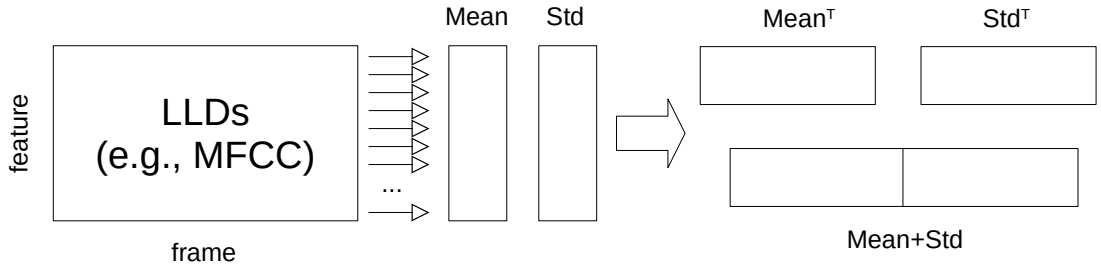

Figure 4.5: Illustration of Mean+Std extraction from LLDs (e.g., MFCCs)

Table 4.3: Results of utterance-based HSF for dimensional SER in IEMOCAP dataset

Feature	Dim	Val	Aro	Dom	Mean
MFCC	80	0.155	0.580	0.456	0.397
Log Mel	256	0.151	0.549	0.455	0.385
GeMAPS	46	0.191	0.523	0.452	0.389
pAA	68	0.145	0.563	0.445	0.384
pAA_D	128	0.173	0.612	0.455	0.413

4.1.3 Optimizing dimensional SER using different classifiers

In the previous subsections, the study focuses on the search for relevant acoustic features (regions) for feature extraction. As described in the previous chapter, two important components of SER are features and classifiers. In this subsection, a study to optimize dimensional SER using different classifiers is presented.

Long short-term memory (LSTM) networks are used to obtain the previous results on both LLDs and Mean+Std features. The LSTM network consists of 3 layers with 256 units each. This configuration is based on [15]. Since the input is a sequence of acoustic features, using a recurrent-based LSTM network is a straightforward approach. The use of LSTM as a classifier for SER has been found useful for both dimensional [113] and categorical task [13]. While the previous results used three layers with the same units, this optimization section varies one to five layers with a different number of units.

Apart from LSTM networks, convolutional neural networks (CNN) and multilayer perceptrons (MLP) were accommodated. Both CNN and MLP use varying layers and their corresponding units, as shown in Table 4.4. CNN has been found to be useful for image-like input. Log mel-spectrogram is an example of this input. MLP, one of the oldest neural network architecture, remains useful due to its simplicity to model complex internal representation of their environment [72]. While the implementation of both LSTM and CNN were performed by using Keras toolkit [114], the implementation of MLP was performed using the scikit-learn toolkit [67].

Table 4.5 and 4.6 show results of optimizing dimensional SER using different classifiers from IEMOCAP and MSP-IMPROV datasets. As for the input, all networks take the previous pyAudioAnalysis with deltas (pAA_D). Changing the number of layers and their units changes their performances, as well as changing the classifiers. Using CNN, an improvement from the previous LSTM result was obtained with a single layer with 16 nodes. On using MLP, significant improvements were obtained; the highest average CCC

Table 4.4: Number of layer and corresponding units on each layer

# layers	# units
1	(16)
2	(32, 16)
3	(64, 32, 16)
4	(128, 64, 32, 16)
5	(256, 128, 64, 32, 16)
6	(512, 256, 128, 64, 32, 16)

score was obtained using five layers for IEMOCAP and four layers for MSP-IMPROV. The average CCC score of this architecture is 0.472 for IEMOCAP and 0.433 for MSP-IMPROV. These high results should be evaluated in the same framework (toolkit) in the future; this study evaluated LSTM and CNN using Keras with TensorFlow backend while MLP is performed using scikit-learn toolkit.

Table 4.5: Average CCC score on IEMOCAP dataset using different classifiers (features: pAA_D)

Classifier	1lay	2lay	3lay	4lay	5lay	6lay
LSTM	0.389	0.403	0.385	0.401	0.395	0.399
CNN	0.415	0.399	0.380	0.376	0.390	0.379
MLP	0.450	0.469	0.448	0.462	0.472	0.452

Table 4.6: Average CCC score on MSP-IMPROV dataset using different classifiers (features: pAA_D)

Classifier	1lay	2lay	3lay	4lay	5lay	6lay
LSTM	0.350	0.378	0.372	0.343	0.317	0.354
CNN	0.356	0.335	0.326	0.349	0.382	0.296
MLP	0.413	0.420	0.421	0.433	0.359	0.369

To this end, several steps to observe which region of analysis to extract acoustic features were performed. At first, the common LLDs, like MFCC features, were evaluated. Later, the Mean+Std of these LLDs were used as input features to the same classifier. The results clearly show that extracting statistical functions over frame-based LLDs is better. Mean+Std with small size (80 vs. (3414×40)) consistently performs better than LLDs. An optimization using different classifiers shows improvements from the previous LSTM networks with the same high-level acoustic features.

4.2 Effect of silent pause features in dimensional SER

In the previous section, analysis of region for acoustic feature extraction was investigated. This section investigates the second issue in dimensional SER using acoustic features: which action is better to treat silent pause region in dimensional SER. There are three actions that can be taken regarding silent pause in SER:

- removing silence: extract acoustic features from speech-segment only,
- keeping silence: extract acoustic features from the whole utterance, including speech and silence regions,
- utilizing silence: utilize silent pause regions as acoustic features.

The goal of this section is to examine which action from these three serves the best for dimensional SER. The second action was already evaluated in the previous results; therefore, only first and third actions will be explained and evaluated in this section. The baseline uses pAA features (with 68 HSFs) to observe the difference among the silence-removed region, silence-kept region, and silent pause features as an additional feature.

4.2.1 Dimensional SER on silence-removed region

Removing silence is a common practice in speech processing. ASR avoids recording silent voice and only uses voiced speech to save power and computational load. In automatic SER, the contribution of silence in speech is not clear until now. Aguilar et al. [90] evaluated both removing and keeping silence for unimodal and multimodal categorical emotion recognition. The result is different. Keeping silences lead to better performance in unimodal emotion recognition, while multimodal shows that removing silence is better than keeping silence. Atmaja and Akagi [13] shows that removing silence leads to higher accuracies score than using whole speech in categorical speech emotion recognition.

A naive way to calculate silence region within speech is by using root mean square (RMS) energy. Given a threshold τ, if the RMS energy of a speech frame below this τ threshold, then that frame is categorized as a silence. The RMS energy is given by the following equation:

$$x_{rms} = \sqrt{\frac{1}{n}(x_1^2 + x_2^2 + \ldots + x_n^2)} \tag{4.5}$$

As shown in Figure 4.6 (b), a threshold of 0.065 is a reasonable choice for given speech utterance (an excerpt from IEMOCAP dataset). However, this RMS energy curve occasionally fall below the threshold for a moment and these values are not counted as silence. A better way to detect silence is by probability mapping, a conversion from raw RMS energy to a likelihood/probability. The probability mapping is formulated as

$$P[NS = 1|x_{rms}] = \frac{\exp(x_{rms} - \tau)}{1 + \exp(x_{rms} - \tau)} \tag{4.6}$$

The result shown in Figure 4.6 (c). The final part in the bottom of the figure shows the result in binary value, an NS of 1 for non-silence and 0 for silence.

In practice, using a threshold τ as a percentage from maximum values is more intuitive while it gives a similar result. Additionally, (minimum) duration of silence is another

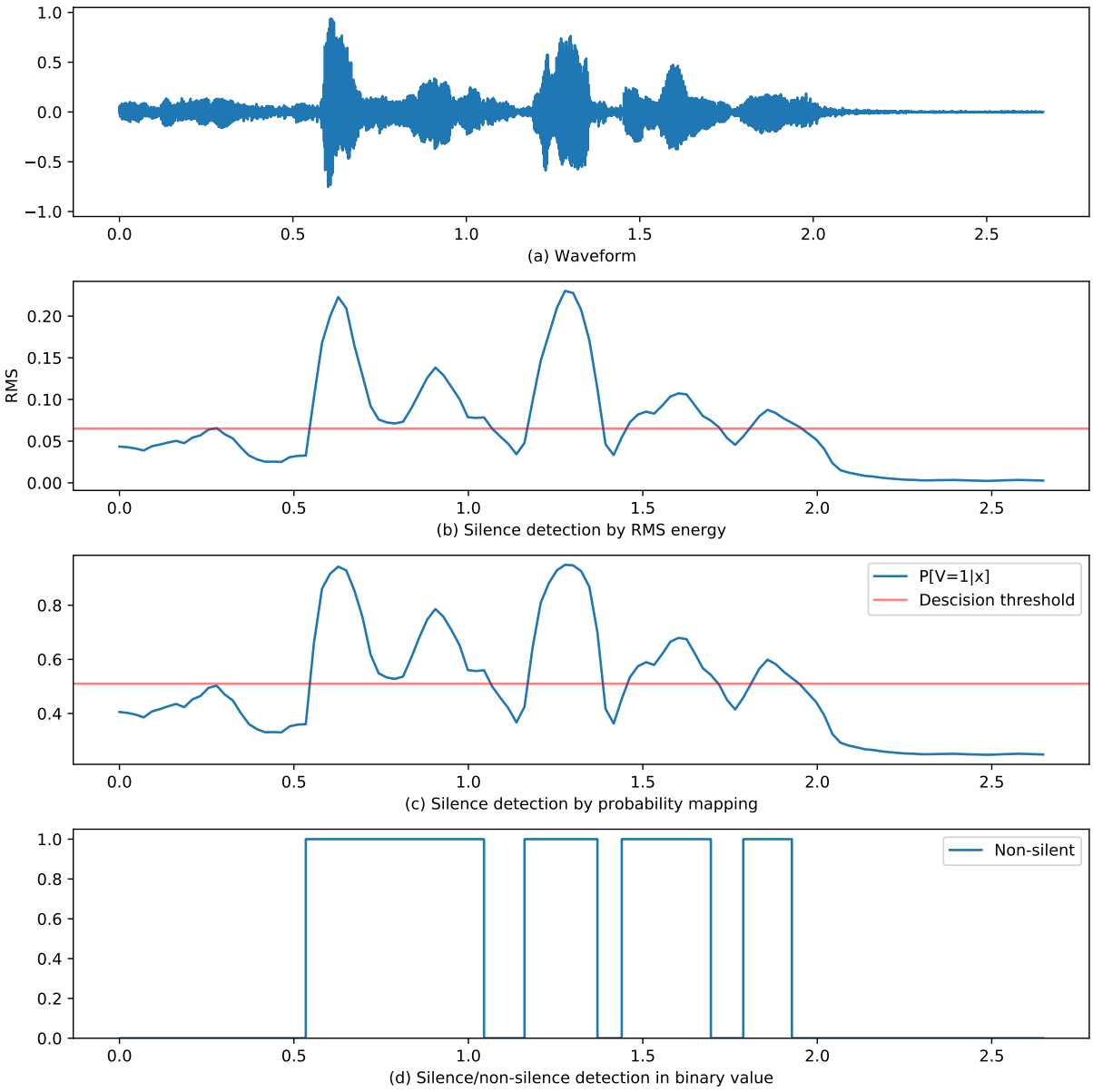

Figure 4.6: Calculation of silent region in speech

Table 4.7: Result of using different duration and threshold factor for removing silence on IEMOCAP dataset; bold-typed scores indicate a higher value than baseline.

Duration (ms)	Threshold (%)	V	A	D	Mean
10	0.1	0.283	0.640	0.454	**0.459**
10	1	0.234	0.560	0.418	0.404
10	5	0.205	0.568	0.393	0.389
60	0.1	0.279	0.625	0.453	0.452
60	1	0.255	0.574	0.425	0.418
60	5	0.209	0.567	0.398	0.391
100	0.1	0.281	0.629	0.456	0.455
100	1	0.276	0.571	0.429	0.425
100	5	0.205	0.557	0.393	0.385

Table 4.8: Result of using different duration and threshold factor for removing silence on MSP-IMPROV dataset; bold-typed scores indicate a higher value than baseline.

Duration (ms)	Threshold (%)	V	A	D	Mean
10	0.1	0.228	0.575	0.437	**0.413**
10	1	0.246	0.581	0.431	**0.420**
10	5	0.148	0.569	0.414	0.377
60	0.1	0.241	0.588	0.442	**0.424**
60	1	0.259	0.586	0.441	**0.429**
60	5	0.184	0.569	0.415	0.389
100	0.1	0.239	0.587	0.438	**0.421**
100	1	0.252	0.580	0.430	**0.421**
100	5	0.201	0.574	0.422	0.399

important parameter. The minimum number of samples to be removed represents the duration of the pause in speech communication, which has been studied thoroughly [115]. These two parameters, silence τ and duration d can be used to experiment with for removing the silence region and extract the acoustic features from these regions.

Three different thresholds and durations were performed to remove silence from the speech dataset. For the threshold, the values of 0.01%, 0.1%, and 5% were examined, while for durations were 10 ms, 60 ms, and 100 ms. The results are shown in Table 4.7. As the baseline method, MLP with Mean+Std of pyAudioAnalysis features (without deltas, 68 dimensions) were used. The baseline method gains 0.458 of the average CCC score. Among nine combinations of duration and threshold for removing silences where acoustic features were extracted, only one result shows a higher performance than the baseline. This result is inline with previous research ([13, 91, 90, 92]), in which special adjustments are needed to take the benefit of treating silence pause region in speech.

4.2.2 Dimensional SER with silent pause features

Tian et al. [91] argued that silence is an effective cue for recognizing emotion. Using this idea, Fayek et al. used silence as an additional category for detecting emotion categories from the speech signal. Silent pause length also plays an important role in ascribing emotions based on psychoacoustics experiment [94]. These assumptions along with their results are motivation to use silence as a feature for dimensional SER.

There are several ways to count silent pause features in speech. A straightforward way is by detecting the number of silent regions and compared them to the whole utterance. The result is a portion of silence region over speech. Although this method may represent silent pause more precisely, there is more effort needed to align the timing of spoken words and silence region manually to obtain a more accurate result. Alternatively, silent pause detection can be done on frame basis calculation with fixed-length samples (of speech signals). A frame, then, can be categorized as silence or non-silence by a specific rule.

A silent pause feature, in this research, is defined as the proportion of silent frames among all frames in an utterance. In human communication, the proportion of silence in speaking depends on the speaker's emotion. For example, a happy speaker may have fewer silences (or pauses) than a sad speaker. The proportion of silence in an utterance can be calculated as

$$sf = \frac{N_s}{N_t}, \tag{4.7}$$

where N_s is the number of frames categorized as silence (silent frames), and N_t is the total number of frames. A frame is categorized as silent if it does not exceed a threshold value (th) defined by multiplying a factor (α) by a root mean square (RMS) energy, X_{rms}. Mathematically, this is formulated as

$$th = \alpha \times \tilde{x}_{rms}, \tag{4.8}$$

where $\tilde{x}_{rms}$ is the median value of RMS energy. This calculation differs from the previous study [14] which used mean values. Median value is more similar to the previous silence removal calculation (by percentage from maximum amplitude) than a mean value.

This silence feature is similar to the disfluency feature proposed in [116]. In that paper, the author divided the total duration of disfluency by the total utterance length for n words. Figure 4.7 illustrates the calculation of the silence feature. If x_{rms} from a frame is below th, then it is categorized as a silence, and the calculation of equation 4.7 is applied. Two important parameters for this silent pause features then can be investigated: (1) threshold factor (α), and (2) silent pause duration.

This study evaluates three α values, i.e., 0.1, 0.2, and 0.3 based on the previous finding [14]. Silent pause duration of 10 ms, 60 ms, 100 ms, 200 ms, 500 ms, and 1 s are also investigated based on the study of the division of pause [115].

Figure 4.8 shows the use of different threshold factors in determining silent pause features. The lower threshold factor, the smaller number of silence frames correspond to silent pause features. Thus, the choice of silence threshold factor is also critical when calculating silent pause features apart from the silent pause duration. Notice that leading and trailing silences have been trimmed; hence, the calculated silent pause features are only within the trimmed region.

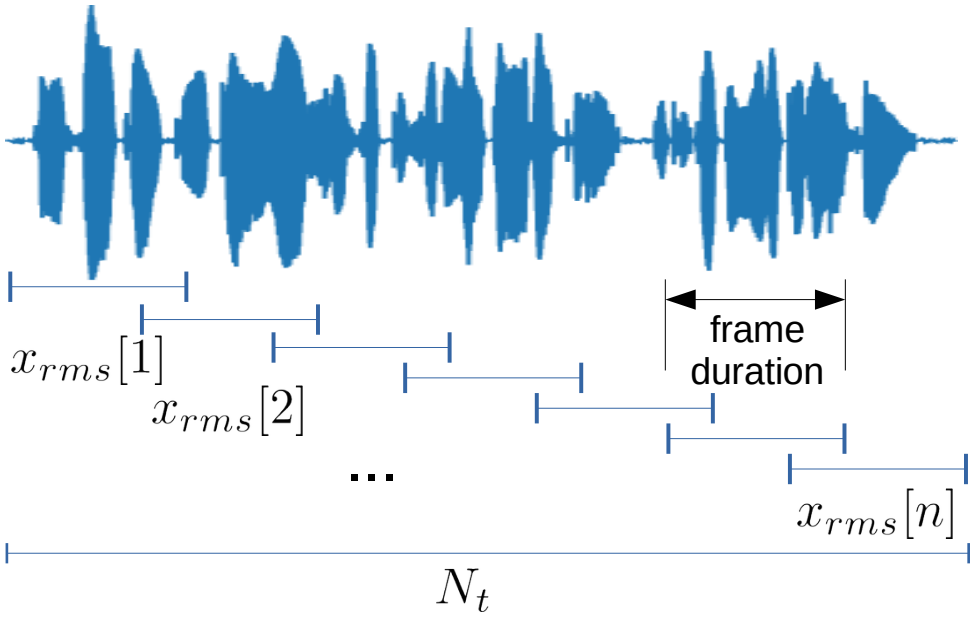

Figure 4.7: Silent pause features calculation

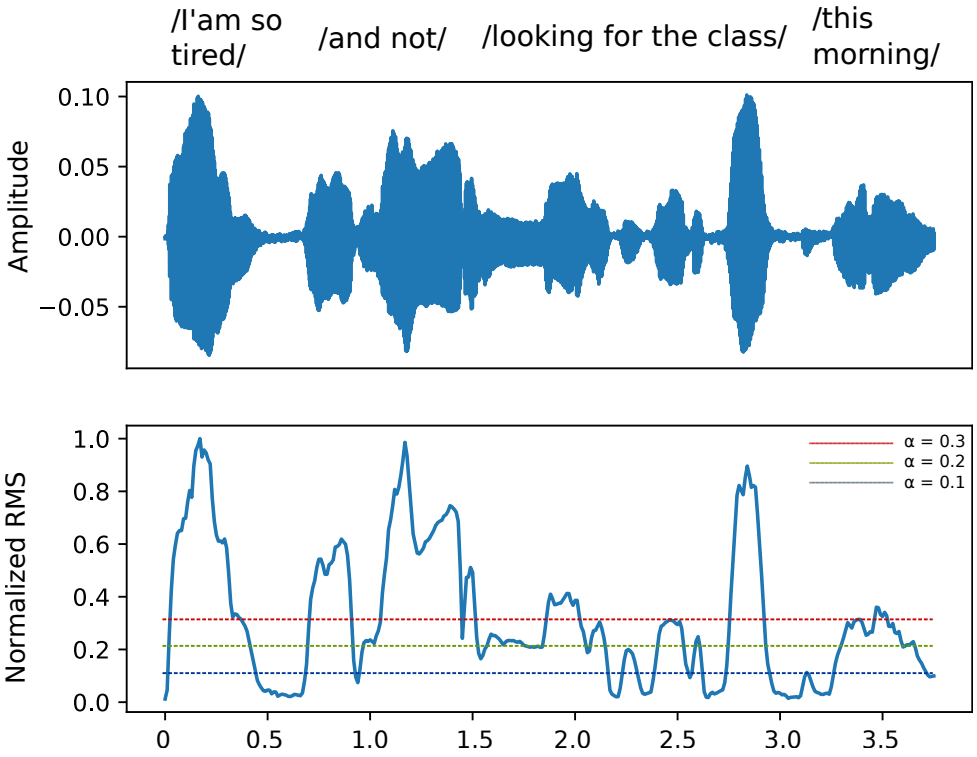

Figure 4.8: Different silent threshold factors on normalized RMS with trimmed leading and trailing silences

Table 4.9: Result of using silence as an additional feature on pAA feature set on IEMOCAP dataset; bold-typed scores indicate a higher mean value than baseline.

Duration (ms)	Threshold	V	A	D	Mean
10	0.2	0.273	0.607	0.424	0.435
10	0.3	0.288	0.626	0.448	0.454
60	0.1	0.277	0.606	0.424	0.436
60	0.2	0.273	0.606	0.422	0.434
60	0.3	0.283	0.624	0.447	0.451
100	0.1	0.278	0.604	0.421	0.434
100	0.2	0.284	0.624	0.446	0.451
100	0.3	0.298	0.641	0.460	**0.466**

Table 4.10: Result of using silence as an additional feature on pAA feature set on MSP-IMPROV dataset; bold-typed scores indicate a higher mean value than baseline.

Duration (ms)	Threshold	V	A	D	Mean
10	0.1	0.227	0.601	0.443	**0.424**
10	0.2	0.211	0.586	0.428	0.408
10	0.3	0.209	0.584	0.427	0.407
60	0.1	0.219	0.601	0.436	**0.419**
60	0.2	0.209	0.586	0.426	0.407
60	0.3	0.207	0.585	0.430	0.407
100	0.1	0.212	0.585	0.425	0.407
100	0.2	0.208	0.586	0.430	0.408
100	0.3	0.207	0.585	0.430	0.407

Table 4.9 and 4.10 show the result of using silence pauses as a feature on IEMOCAP and MSP-IMPROV datasets. Both results confirm the improvement of CCC scores from the baseline. While result on Table 4.9 were obtained using 5 layers MLP, result on Table 4.10 were obtained using 3 layers MLP.

Finally, Table 4.11 shows a summary of three strategies to observe the effect of silence on dimensional speech emotion recognition. In the IEMOCAP dataset, utilizing silence leads to higher performance than keeping silence. In the MSP-IMPROV dataset, removing silence leads to higher performance than keeping silence. Both tables show the advantage of either removing silence or utilizing silence; both lead to better performances than the baseline score.

4.3 Acoustic feature aggregation

The final issue to discuss in this chapter is to choose which aggregation method works best for dimensional SER from acoustic features. It is common in processing audio to split an utterance (or story) into chunks. The goal is for fast processing as well as for reducing

Table 4.11: Comparison of three conditions for investigating the effect of silence in dimensional SER

Strategy	V	A	D	Mean
IEMOCAP				
Removing silence	0.283	0.640	0.454	0.459
Keeping silence	0.268	0.641	0.458	0.456
Utilizing silence	0.298	0.641	0.460	0.466
MSP-IMPROV				
Removing silence	0.259	0.586	0.441	0.429
Keeping silence	0.217	0.586	0.425	0.409
Utilizing silence	0.227	0.601	0.443	0.424

the size of recorded/analyzed audio data. While the label is only given per utterance, acoustic features extraction is performed on chunk-based processing, either using LLDs or HSFs. Thus, two options exist: whether aggregating input features to have single label per utterance or aggregating outputs with many labels for a single utterance. For the latter method, the label to represent a single story from many chunk labels can be performed by a such method, e.g., majority voting.

Seven types of LLDs from LibROSA features extractor [117] extracted for acoustic input features: MFCCs (40 coefficients), chroma (12), mel-spectrogram (128), spectral contrast (7), tonal centroid (6), deltas of MFCCs (40), and deltas-deltas of MFCCs (40). This feature set is adopted from [17]. In total, there are 273 features on each frame. Following the previous success in using global features for determining region of analysis, Mean+Std from these 273 LLDs were extracted, resulting in 546-dimensional functional features.

Input feature aggregation is a method to choose which features to represent a set of data (story) given many recordings (chunks). Statistical functions were widely used to aggregate many measurements. The choice of mean and maximum values for acoustic feature aggregation is based on the assumption that acoustic features representing emotion either from mean values (e.g., mean intonation) or maximum values (e.g., high pitch in specific speech region when expressing fear or happy). In maximum aggregation, the highest column vector value of acoustic features (Mean+Std from LLDs) for each chunk on the same stories. By using these methods, each story has the same n-dimensional feature vector depends on extracted acoustic features. A similar approach was conducted for mean values feature aggregation. Figure 4.9 shows acoustic input aggregation from chunks to story.

Output aggregation is often performed by majority voting. The use of majority voting in SER has been implemented in various techniques [118, 89]. The majority voting method was often used to choose the final label over different classifiers (known as *ensemble* method). However, the majority voting defined in this study closer its original term; the most frequent classes is chosen among them to represent the data. In the INTERSPEECH 2020 elderly sub-challenge (ESC), the dataset provided audio files as chunks, parts of an utterance/story. Acoustic features (frame-based processing and statistical functions) were extracted per this chunk and forwarded to a classifier. Thus, in a single story, there are

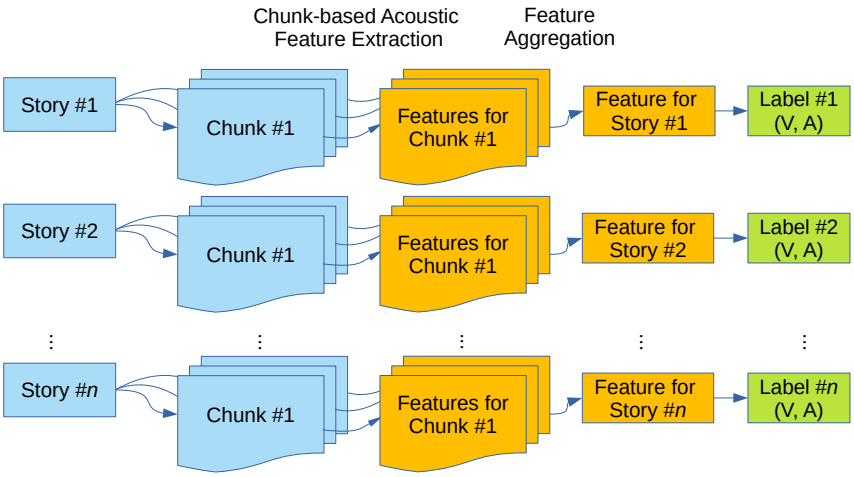

Figure 4.9: Flow diagram of acoustic input feature aggregation

many labels with regard to the number of chunks. The majority voting chooses the most frequent labels to represent a story (Figure 4.10).

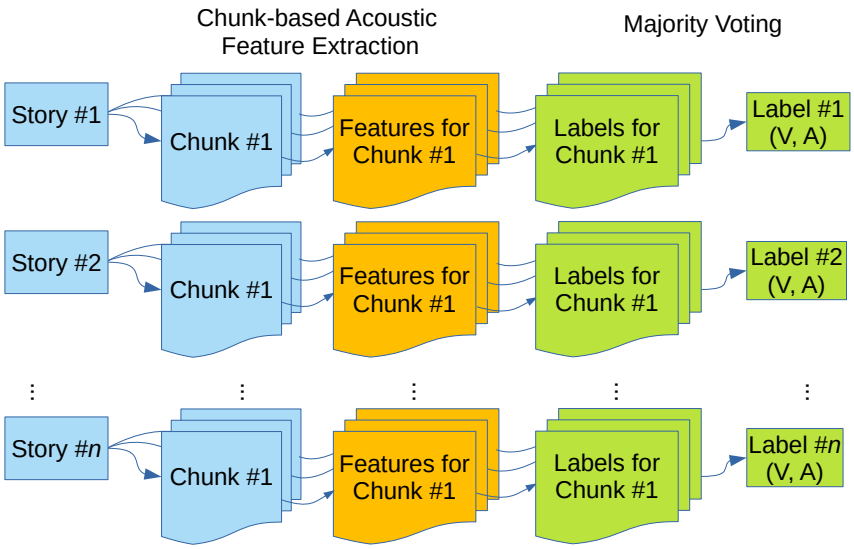

Figure 4.10: Flow diagram of acoustic output aggregation (majority voting)

In comparing mean vs. maximum aggregation methods, it is found that mean aggregation leads to higher UAR scores than maximum aggregation in development partition. All results from mean input aggregation attain higher scores than baseline majority voting. Table 4.9 also shows that mean input aggregation works better than mean output aggregation. This finding suggests that using all chunks (by averaging) is better than choosing one value from a chunk (by maximum value). This evidence also supports the previous global features approach as a solution to choose the region of analysis of acoustic features.

The use of aggregation methods reduces feature dimension (for input to classifier) and computational complexity. Using all chunks to process the data, e.g., without feature aggregation, increase computational load and complexity. The number of samples became larger according to the number of chunks. However, the UAR score is low. Using feature

Table 4.12: UAR results on development set: unimodal acoustic feature aggregation vs. baseline [6] (INTERSPEECH 2020 ComParE Elderly Emotion Sub-Challenge dataset)

Features	Majority Voting [6]		Mean Input Agg.		Max Input Agg.	
	V	A	V	A	V	A
LibROSA Mean+Std	-	-	**45.1**	38.3	42.7	39.7
ComParE	33.3	39.1	43.4	42.7	**45.3**	37.0
BoAW-125	38.9	42.0	44.6	**45.7**	44.6	40.1
BoAW-250	33.3	40.5	43.0	40.8	39.6	37.6
BoAW-500	38.9	41.0	42.6	41.0	42.9	37.9
BoAW-1000	38.7	30.5	43.5	41.5	40.2	39.8
BoAW-2000	**40.6**	39.7	41.9	44.8	43.4	40.1
ResNet50	31.6	35.0	36.5	36.7	37.1	39.0
AuDeep-30	35.4	36.2	38.4	42.1	42.8	35.6
AuDeep-45	36.7	34.9	39.5	40.5	39.3	33.3
AuDeep-60	35.1	**41.6**	43.4	42.1	40.7	41.4
AuDeep-75	32.7	40.4	41.9	44.4	40.9	**43.3**
AuDeep-fused	29.2	36.3	43.6	39.5	42.2	39.3

aggregation not only reduces complexity and feature dimension but also increases the performance score.

Although this evaluation of the aggregation method for speech emotion recognition is not intended to mimic human auditory perception, there may be a similarity in human auditory perception on aggregating different cues. Humans may use the aggregation of prosodic information from short term voices for longer time emotion perception. The initial goal of this feature aggregation is to concatenate acoustic features with linguistic features, which will be explained in the next chapter.

4.4 Summary

This chapter presents an evaluation of speech emotion recognition from acoustic information. Three problems are investigated, including the region of analysis for feature extraction, the silent pause region's effect, and the aggregation methods. Table 4.13 presents the results of these investigations. On the first problem, it was found consistently that high-level statistical functions obtained better performance than low-level descriptors. This result shows that small-feature size is not a problem for DNN-based classifiers (instead, the number of data still a problem). Mean and standard deviation from acoustic features showed meaningful representation for acoustic-based emotion recognition. The second evaluation of the silent pause region's effect showed that either removing silence or utilizing silence as a feature leads to a better performance than using acoustic features from the whole speech region. Between the two, it is difficult to choose which one is better based on the current results. The results predict the important role of silent in emotion. The third evaluation showed that the input aggregation method showed better performances than output aggregation by majority voting. Not only improving the performance, this aggregation technique made the ability for concatenating acoustic features with other fea-

tures. The use of acoustic features only still shows some lacks in dimensional SER; one of them is the low performance of valence prediction. This drawback of acoustic-based SER led to the investigation of fusing acoustic information with other modalities. The next chapter presents fusion of acoustic with linguistic information at feature level.

Table 4.13: Summary of study on dimensional SER using acoustic features

Issue	Proposed method		
Region of analysis	frames		utterance (fixed length)
Silence region	removing silence	keeping silence	utilizing silence
Aggregation method	input aggregation		output aggregation

Chapter 5

Fusing Acoustic and Linguistic Information at Feature Level

This chapter evaluates the fusion of acoustic and linguistic information at the feature level. Two approaches were evaluated, network concatenation and feature concatenation. A comparison of manual and automatic transcriptions from the IEMOCAP dataset provides insight into the current automatic speech recognition system's contribution to speech emotion recognition (SER).

5.1 Extracting linguistic information

5.1.1 Word embedding

A classifier needs a set of input features to model input-output relation. One of the common features used in text processing is word embeddings. A word embedding is a vector representation of a word. A numerical value in the form of a vector is used to make the computer process text data as it only processes numerical values. This value is the points (numeric data) in the space of a dimension, in which the size of the dimension is equal to the vocabulary size. The word representations embed these points in a feature space of lower dimension [119]. A one-hot vector represents every word; a value of 1 corresponds to this word and 0 for others. This element with a value of 1 will be converted into a point in the range of vocabulary size.

To obtain a vector of each word in an utterance, first, this utterance in the dataset must be tokenized. Tokenization is a process to divide an utterance by the number of constituent words. For example, the text "That's out of control." from IEMOCAP dataset will be tokenized as ["That's," "out," "of," "control"]. Suppose the number of vocabulary is 2182 (number of words in IEMOCAP dataset with six emotion categories), then the obtained word vector is something similar to

```
text_vector = [42, 44, 11, 471].
```

An embedding layer will convert those positive fixed integers into dense vectors of fixed size. For instance, 1-dimensional word vector in the utterance will be converted into 2-dimensional dense vector,

$$[42, 44, 11, 471] \rightarrow [[0.12, 0.3], [0.12, 0.29], [-0.54, 0.2], [0.71, 0.23]].$$

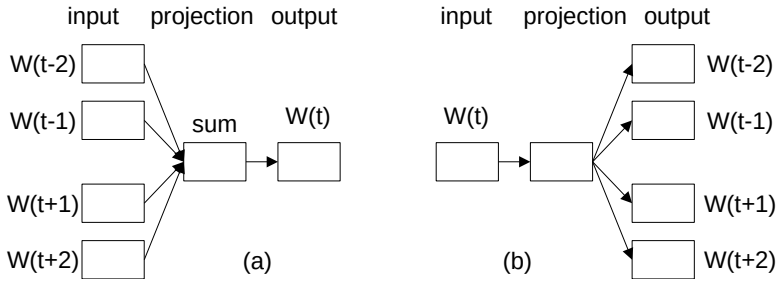

Figure 5.1: Two architectures of word2vec: (a) CBOW and (b) Skip-gram [3]

The higher dimensions are used to obtain a better representation of a word vector. A number of 50-, 100-, and 300-dimensional vectors are commonly employed to build pre-trained word vectors from a large corpus.

A set of zeros can be padded in front of or behind the obtained vector to obtain the fixed-length vector for all utterances. The size of this zeros sequence can be obtained from the longest sequence, i.e., an utterance within the dataset which has the longest words, subtracted by the length of a vector in the current utterance.

5.1.2 Pre-trained word embeddings

A study to vectorize certain words has been performed by several researchers [3, 7, 120]. The vector of those words can be used to weight the word vector obtained previously. The following word embedding techniques were used in this research.

word2vec

Classical word embedding paradigm used unsupervised (hand-crafted) learning algorithm such as LSA, N-gram, and similar methods. Due to advancements in neural network theory supported by computer hardware's speedup, word vector search shifted to deep learning-based algorithms. Mikolov et al. [3] developed word representation using so-called word2vec (word to vector) using a neural network language model trained in two steps. First, continuous word vectors are learned by using a simple model, and then the n-gram neural net language model (NNLM) is trained on top of these distributed representations of words [3]. Two new model architectures are proposed to obtain word vector: the Continuous-Bag-of-Word (CBOW) architecture to predict the current word based on the context. The Skip-gram predicts surrounding words given the current word. Figure 5.1 shows those two different architectures and how they process the input to the output.

From those two approaches, skip-gram was founded as an efficient method for learning high-quality distributed vector representations that capture precise syntactic and semantic word relationships [3]. The objective of the Skip-gram model is to maximize the average log probability,

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c<j<c,c\neq 0}\log p(w_{t+j}|w_t) \tag{5.1}$$

where c is the size of the training context (which can be a function of the center word w_t). Larger c results in more training examples and can lead to higher accuracy, at the expense

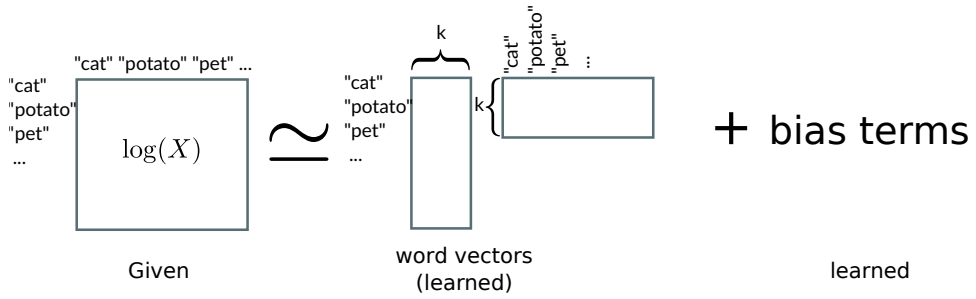

Figure 5.2: Illustration of GloVe representation

of the training time. The basic Skip-gram formulation of $p(w_{t+j}|w_t)$ can be defined using the softmax function, and computational efficiency can be approached by a hierarchical softmax [3].

GloVe
Pennington et al. [7] combined global matrix factorization and local context window methods for learning the space representation of a word. In GloVe (Global Vectors) model, the statistics of word occurrences in a corpus are the primary source of information available to all unsupervised methods for learning the word representations. Although man methods now exist, the question remains as to how meaning is generated from these statistics and how the resulting word vectors might represent that meaning. Glove captured the global statistics from the corpus, for example, a Wikipedia document or a common crawl document.

In GloVe model, the cost function is given by

$$\sum_{i,j=1}^{V} f(X_{i,j})(u_{i,j}^T v_j + b_i + c_j - \log X_{i,j})^2, \tag{5.2}$$

where

- V is the size of the vocabulary,
- X denotes the word co-occurrence matrix (so $X_{i,j}$ is the number of times that word j occurs in the context of word i),
- the weighting f is given by $f(x) = (x/x_{\max})^\alpha$ if $x < x_{\max}$ and 1 otherwise,
- $x_{\max} = 100$ and $\alpha = 0.75$ (determined empirically),
- u_i, v_j are the two layers of word vectors,
- b_i, c_j are bias terms.

In a simple way, GloVe is a weighted matrix factorization with the bias terms, as shown in Figure 5.2.

FastText
Mikolov et al. [120] improved word2vec CBOW model by using some strategies including subsample frequent words technique. This modification of word2vec is trained on large

text corpus such as news collection, Wikipedia and web crawl. They named the pre-trained model with that modification as FastText. The following probability p_{disc} of discarding a word is used by FastText to subsample the frequent words:

$$P_{disc}(w) = 1 - \sqrt{t/f_w}, \tag{5.3}$$

where f_w is the frequency of the word w, and t is a parameter > 0.

FastText also counts the classical N-gram word representation by enriching word vector with a bag of character n-gram vectors learned from a large corpus. In this computation, each word is decomposed into its character n-grams N, and each n-gram n is represented by a vector x_n. The new word vector is then simply the sum of both representations,

$$v_w + \frac{1}{|N|} \sum_{n \in N} x_n, \tag{5.4}$$

where v_w is the old word vector. The set of n-grams N is limited to 3 to 6 characters in practical implementation.

BERT

The previous aforementioned word embeddings – word2vec, GloVe, FastText – are generated word representation in a context-free model. It means, the same word appears in a different phrase has the same word representation, e.g., word "book" in "mathematics book" and "book a hotel." Instead of using a context-free model, BERT (bidirectional transformers language understanding) was built upon pre-training contextual representation [121].

BERT is different in many ways from its predecessors. Apart from contextual representation, the main contribution of BERT is to employ bidirectional pre-training for language representation. Unlike its predecessors, which model languages in a unidirectional way, i.e., from left to right as a writing/reading system, BERT used two unsupervised tasks for pre-training models. The first task is the masked language model; the second task is the next sentence prediction (NSP). The BERT model's dimension for each word depends on the number of hidden layers used in the architecture. This number is either 768-dimensions for the base model, or 1024-dimension for the large model.

Apart from the pre-trained model, BERT provides a fine-tuning model. Fine-tuning allows BERT to model several tasks, single or text pairs, by swapping out the corresponding inputs and outputs. Fine-tuning can be seen as adjusting the pre-trained model according to the context, i.e., the dataset. Hence, fine-tuning can only be done after obtaining the pre-trained model and is relatively expensive. Fine-tuning is suitable for a specific task rather than general linguistic tasks.

5.1.3 CCC loss function

In the Chapter 3, a metric to measure the performance of dimensional emotion recognition was introduced, namely concordance correlation coefficient. Since, the goal is CCC, using CCC loss instead of conventional regression loss function such as mean square error (MSE) and mean absolute error (MAE) is more beneficial than both error-based loss functions. The CCC loss function (CCCL) to maximize the agreement between the true value and

the predicted emotion can be defined as

$$CCCL = 1 - CCC. \tag{5.5}$$

In single-task learning, the loss function would be one for either valence ($CCCL_V$), arousal ($CCCL_A$), or dominance ($CCCL_D$). In multitask learning (MTL), when the total CCC loss is used as a single metric for predicting valence, arousal, and dominance simultaneously, $CCCL_{tot}$ is the following combination of those three CCC loss functions:

$$CCCL_{tot} = CCCL_V + CCCL_A + CCCL_D. \tag{5.6}$$

This MTL equation is referred as "MTL without parameters," because there is no weighting among valence, arousal, and dominance. In this case, the relation among the three emotional dimensions is determined by joint learning in the training process. As it has been stated that these three emotional dimensions are related in a systemic manner [51], two parameters are introduced to weight the valence and arousal, with the weight for dominance determined by subtracting those two parameters from 1. This MTL with two parameters is defined as

$$\begin{aligned} CCCL_{tot} = \alpha\ CCCL_V + \beta\ CCCL_A \\ + (1 - \alpha - \beta)\ CCCL_D, \end{aligned} \tag{5.7}$$

where α and β are the weighting factors for the valence and arousal loss functions, respectively. This proposed MTL is similar to that defined in [122]. While those authors used the mean squared error (MSE) as the loss function, this study have proposed using this CCC-based loss function. In addition, a parameter γ is added for dominance to obtain independent scales among valence, arousal, and dominance. The resulting MTL with three parameters is defined as

$$CCCL_{tot} = \alpha\ CCCL_V + \beta\ CCCL_A + \gamma\ CCCL_D. \tag{5.8}$$

For comparison with the previous MTL without parameters; α, β, and γ were set to 1 in that equation 5.6, which can be seen as a special case in this MTL with three parameters.

These MTL approaches compare the predicted output from the three one-unit dense layers with the ground truth labels. The training process mechanism relies on the above loss function. Hence, the performance of the produced model is based on this mechanism, too. The loss function's choice is a critical aspect of machine learning, and this study thus proposed this MTL based on the CCC loss to learn valence, arousal, and dominance concurrently.

5.2 Early fusion by networks concatenation

A simple way to fuse two different information is by concatenating two DNNs of different network modalities. Different features shape is not a problem in this fusion; the concatenation branch only requires the same dimension of DNN outputs from each network. For instance, a 3D vector can solely be concatenated with another 3D vector.

A unimodal feature is a feature set from either acoustic or linguistic (e.g., pAA LLD). At first, the system trained each feature set on both LSTM and CNN classifiers independently. Tables 5.1 and 5.2 list the unimodal dimensional emotion results from those acoustic and linguistic networks, respectively. Each table lists the scores for valence, arousal,

and dominance in terms of the CCC, along with averaged CCC scores to determine which method performed better. The results were grouped by modality and architecture. They all used the same metric scale and were obtained under the same conditions. Hence, these results can be compared directly with each other.

In the acoustic-based modality, the obtained results are consistent among the feature sets on both architectures. From bottom to top, the performance order was pAA LLD, GeMAPS LLD, GeMAPS HSF, and pAA HSF. Thus, although GeMAPS performed better for LLDs, the HSF for pAA performed best on both the LSTM and CNN architectures. This result supports the previous finding that the mean and standard deviation outperform the low-level descriptors (LLDs) defined in GeMAPS. Furthermore, this finding can be generalized to the means and standard deviations from the other feature sets. In this case, the HSF for pAA performed better than the HSF for the affective-designed GeMAPS.

Comparing the LSTM and CNN architectures, it is found that the LSTM performed better than CNN did. In terms of three emotional dimensions and an average of three, the score obtained by the highest-performing LSTM was higher than that obtained by the highest-performing CNN. The best architecture in the acoustic networks was chosen to combine with the linguistic networks' best architectures.

As for the linguistic networks, word embeddings with pre-trained GloVe embeddings performed better than either word embeddings without weighting or word embeddings weighted by the FastText model did. The linguistic networks also showed that the LSTM with GloVe embedding is better than the CNN with the same input feature. However, in this dimensional emotion recognition, the linguistic network's highest performance was lower than an acoustic network's highest performance. As with the acoustic networks, two networks were chosen, GloVe with LSTM and GloVe with CNN, to combine in the bimodal network fusion.

Table 5.1: CCC score results on the acoustic networks

Feature set	V	A	D	Mean
LSTM				
pAA LLD	0.0987	0.5175	0.3536	0.3233
pAA HSF	0.1729	**0.5804**	**0.4476**	**0.4003**
GeMAPS LLD	0.1629	0.5070	0.4433	0.3711
GeMAPS HSF	**0.1818**	0.5306	0.4332	0.3819
CNN				
pAA LLD	0.0687	0.3665	0.3382	0.2578
pAA HSF	**0.1310**	**0.5553**	**0.4431**	**0.3764**
GeMAPS LLD	0.0581	0.4751	0.4203	0.3178
GeMAPS HSF	0.0975	0.4658	0.4170	0.3268

Table 5.2: CCC score results on the linguistic networks

Feature set	V	A	D	Mean
		LSTM		
WE	0.3784	0.3412	0.3638	0.3611
word2vec	0.3937	0.3811	0.3824	0.3857
GloVe	**0.4096**	**0.3886**	**0.3790**	**0.3924**
FastText	0.4017	0.3718	0.3771	0.3835
BERT	0.3858	0.3675	0.3722	0.3752
		CNN		
WE	0.3740	0.3285	0.3144	0.3390
word2vec	0.3692	0.3589	0.3613	0.3631
GloVe	**0.3843**	0.3646	**0.3911**	**0.3800**
FastText	0.3786	**0.3648**	0.3147	0.3527
BERT	0.3598	0.3479	0.3530	0.3535

5.2.1 Results on bimodal feature fusion

Performance of bimodal networks

According to their unimodal network performance, eight pairs of bimodal acoustic-linguistic networks were evaluated. Table 5.3 lists their performance results in the same way as for the unimodal results. Among the eight pairs, the LSTM acoustic networks and the LSTM linguistic networks achieved the best performance. This result in bimodal feature fusion is linear with respect to the obtained results for the unimodal networks, in which the LSTM performed best on both the acoustic and linguistic networks.

In terms of both emotional dimensions and average CCC scores, the LSTM+LSTM pair outperformed the other bimodal pairs. Moreover, the deviation of the LSTM+LSTM pair was also the lowest. It can be stated that, apart from attaining the highest performance, the LSTM+LSTM pair also gave the most stable results. This result suggests that the LSTM not only attained comparable results to the CNN with a similar number of trainable parameters but also attained better performances, which differs from what was reported in [123].

One reasonable explanation for why the LSTM performed better is the use of the full sequence instead of the final sequence in the last LSTM layer. In most applications, the last layer in an LSTM stack only returns the final sequence to be combined with the outputs of other layers (e.g., a dense layer). In this implementation, however, all sequences outputs were returned from the last LSTM layer and flattened before combining them with another dense layer's output (from the linguistic network). This strategy may keep more relevant information than what is returned by the final sequence of the last LSTM layer. On the other hand, this phenomenon is only observed on the acoustic network. In the linguistic network case, the last LSTM layer returning the final sequence performed better than the LSTM returning all sequences. In that case, the last LSTM layer was directly coupled with that of a dense layer.

If the highest unimodal score is chosen as a baseline, i.e., the HSF of pAA, then the highest bimodal score's relative improvement was 23.97%. A significance test among

the bimodal pair results was performed. A significant difference was observed between an LSTM+LSTM pair and other pairs, such as a CNN+LSTM pair, on a two-tail paired test. The small p-value ($\simeq 10^{-5}$) indicated the strong difference obtained by the LSTM+LSTM and CNN+LSTM pairs. While the CNN+LSTM pair obtained the third-highest score, the second-best performance was by a Dense+CNN pair with CCC = 0.485. The significance test result between the LSTM+LSTM pair and this pair was $p = 0.0006$. The more similar the performance of two acoustic-linguistic networks pairs was, the higher the p-value between them was. The assertion that the LSTM+LSTM pair had a big difference from the other pairs was set with $p < 0.05$.

Table 5.3: Results of bimodal feature fusion (without parameters) by concatenating the acoustic and linguistic networks; each modality used either an LSTM, CNN, or dense network; batch size = 8

Acoustic+Linguistic	V	A	D	Mean
LSTM + LSTM	**0.418 ± 0.01**	**0.571 ± 0.017**	**0.5 ± 0.017**	**0.496 ± 0.01**
LSTM + CNN	0.256 ± 0.052	0.531 ± 0.031	0.450 ± 0.036	0.412 ± 0.030
CNN + LSTM	0.401 ± 0.020	0.545 ± 0.016	0.478 ± 0.015	0.476 ± 0.012
CNN + CNN	0.399 ± 0.015	0.541 ± 0.020	0.475 ± 0.014	0.472 ± 0.012
LSTM + Dense	0.274 ± 0.050	0.553 ± 0.019	0.484 ± 0.015	0.437 ± 0.018
CNN + Dense	0.266 ± 0.038	0.497 ± 0.059	0.457 ± 0.047	0.407 ± 0.040
Dense + LSTM	0.368 ± 0.105	0.564 ± 0.015	0.478 ± 0.025	0.470 ± 0.043
Dense + CNN	0.398 ± 0.015	0.570 ± 0.013	0.487 ± 0.015	0.485 ± 0.013

Evaluation of MTL with weighting factors

As an extension of the main proposal to jointly learn the valence, arousal, and dominance from acoustic features and word embeddings by using MTL, this study also evaluated some weighting factors for the MTL formulation (equations 4, 5, and 6). In contrast, the above results were obtained using MTL with no parameters (equation 5.6). Thus, the following results show the effect of the weighting parameters on the MTL method.

MTL with two parameters is an approach to capture the interrelation among valence, arousal, and dominance. In equation 5.7, the gains of valence and arousal are provided independently, while the gain of dominance depends on the other gains. This simple weighting strategy may represent the relation among the emotional dimensions if the obtained results are better than the results without this weighting strategy.

Figure 5.3 shows a surface plot of the impact of varying α and β from 0.0 to 1.0 with the corresponding average CCC score. Performance improvement could be obtained by using proper weighting factors in two-parameter MTL. It is found that $\alpha = 0.7$ and $\beta = 0.2$ were the best weighting factors, and the linguistic network also used them. In the unimodal network, the best factors for MTL with two parameters were $\alpha = 0.7$ and $\beta = 0.2$ for the linguistic networks, and $\alpha = 0.1$ and $\beta = 0.5$ for the acoustic network. These factors were used to obtain the unimodal results above. In this case, it is difficult to judge whether these same obtained factors for the bimodal network were contributed

by the unimodal network or caused by other factors. Investigation on the cause of this finding is a challenging issue for both theoretical and empirical studies.

Next, MTL with three parameters provided all factors for three-dimensional emotions, with every emotional dimension's factor independent of each other. MTL with no parameters is also a subset of MTL with three parameters, with $\alpha = 1.0$, $\beta = 1.0$, and $\gamma = 1.0$. The experiments optimized the weighting factors with three parameters by using linear search independently on each emotion dimension. Figure 5.4 shows the impact of the weighting factors on MTL with three parameters. In this scaling strategy, the best weighting factors were $\alpha = 0.9$, $\beta = 0.9$, and $\gamma = 0.2$. The obtained result of $CCC = 0.497$ with these factors was lower than that obtained by MTL with two parameters, i.e., $CCC = 0.508$. While the previous Table 5.3 presented results with batch size = 8, results in Table 5.4 are obtained with batch size = 256, to speed up computation process. The results listed in Table 5.4 show that MTL with two parameters obtained the best performance among the MTL methods. This result suggests that MTL with two parameters may better represent the interrelation among the emotional dimensions.

Figure 5.3: Surface plot of different α and β factors for MTL with two parameters; The best mean CCC score of 0.51 was obtained using $\alpha = 0.7$ and $\beta = 0.2$; Both factors were searched simultaneously/dependently

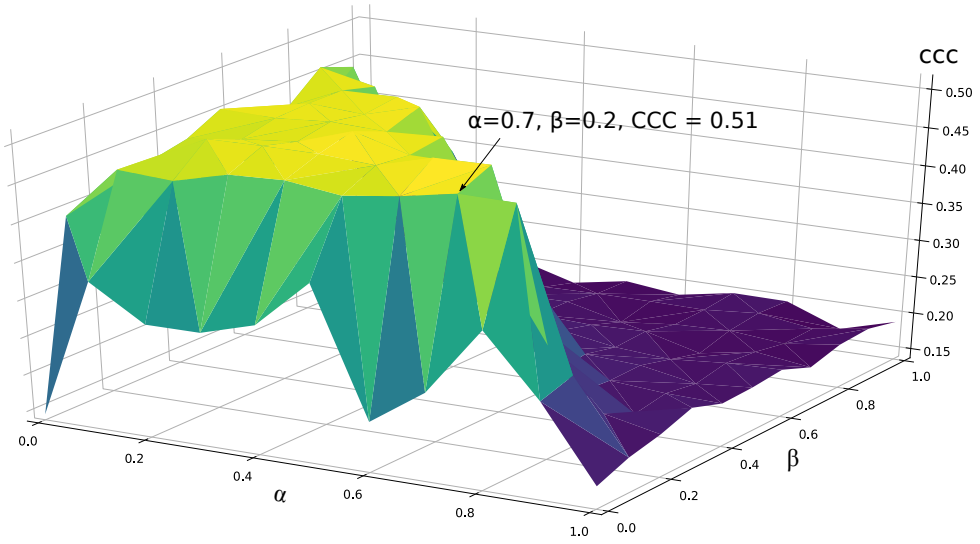

Table 5.4: Results of MTL with and without parameters for bimodal feature fusion (LSTM+LSTM); batch size = 256

MTL method	V	A	D	Mean
No parameter	0.409 ± 0.015	0.585 ± 0.011	**0.486 ± 0.016**	0.493 ± 0.01
2 parameters	**0.446 ± 0.002**	**0.594 ± 0.003**	0.485 ± 0.003	**0.508 ± 0.002**
3 parameters	0.419 ± 0.012	0.589 ± 0.012	0.483 ± 0.011	0.497 ± 0.008

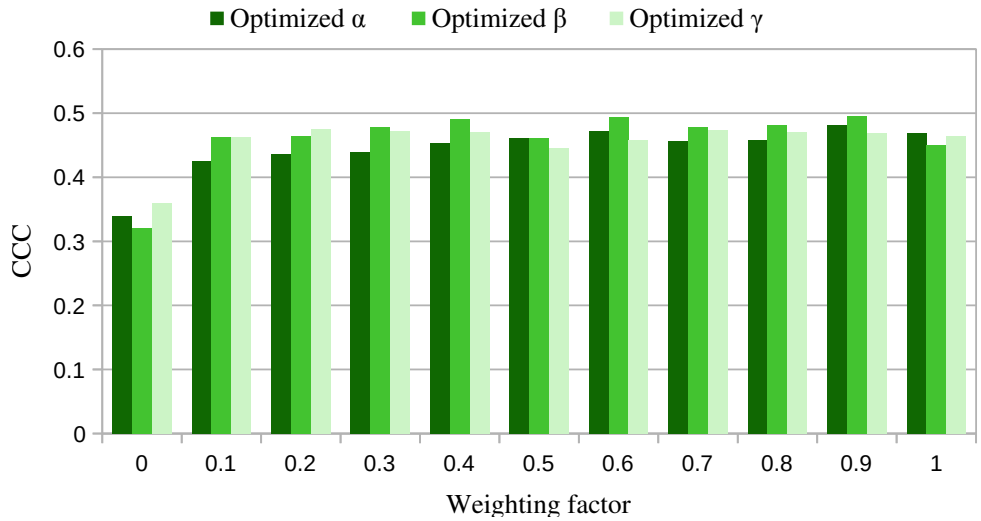

Figure 5.4: CCC scores for MTL with three parameters, obtained to find the optimal weighting factors; linear search was performed independently on each parameter; The best weighting factors for the three parameters were $\alpha = 0.9$, $\beta = 0.9$ and $\gamma = 0.2$.

Evaluation of dropout for different modalities

An investigation of the impact of the dropout rate for the acoustic and text networks in bimodal feature fusion was performed to extend the discussion. In this evaluation, the dropout rates were varied from each modality before concatenating them. The goal of the evaluation, at first, was to investigate the dropout rates for the different modalities.

Figure 5.5 shows the impact of different dropout rates and the obtained CCC scores. From the figure, using dropout rates of $p = 0.1$ and $p = 0.4$ for the acoustic and linguistic networks, respectively, achieved the best score of $CCC = 0.510$. These dropout rates were used to obtain the above results on the bimodal network.

From the obtained dropout rates, it is believed that this factor depends on the size of the feature/input rather than on modality differences. The acoustic network used the smaller HSF for pAA, a 68-dimensional vector, compared to the word embedding's size of 100 sequences $\times$ 300-dimensional word vectors. Because the goal of using dropout is to avoid overfitting, it is reasonable that, on small data, the dropout rate is low, while on larger data, the dropout rate increases. Hence, in this research, it can be believed that dropout rates depend on the input size rather than its modality.

5.2.2 Discussion in terms of categorical emotions

This study on dimensional speech emotion recognition using bimodal features is an extension of a similar categorical method. It is found both similarities and differences as compared to the previous categorical research. Here, the discussion is limited to the best bimodal pairs and the impact of feature sets from different modalities.

In dimensional speech emotion recognition, it was found the more consistent results. This study observed low variation among the experiments, while the previous categorical research only used the highest accuracy from many experiments. Both the categorical and dimensional approaches gained performance improvement over unimodal emotion recognition by combining acoustic features and word embeddings. It was found that LSTM

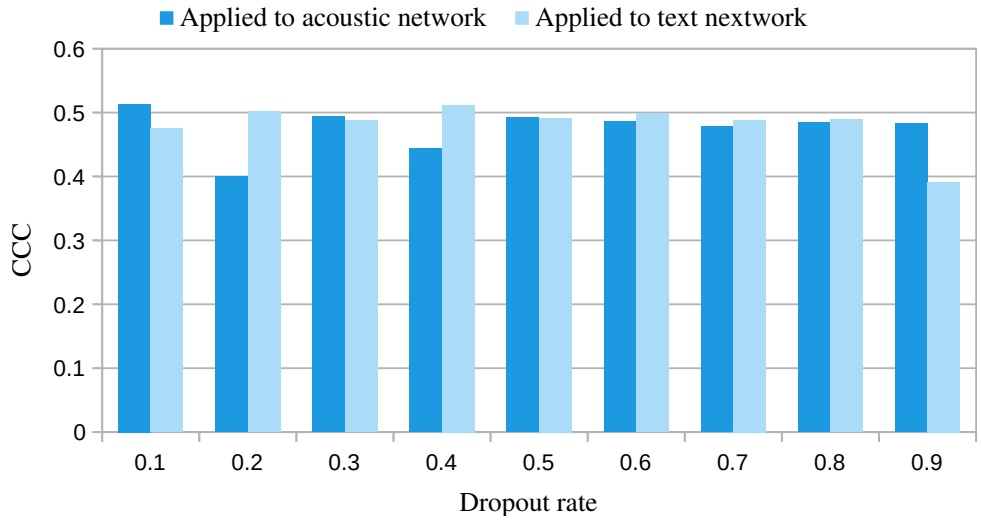

Figure 5.5: Analysis of dropout rates applied to the acoustic and linguistic networks before concatenating them; the dropout rates were applied independently on either network while keeping a fixed rate for the other network.

networks on both modalities performed best in dimensional emotion recognition. This result was also supported by a small standard deviation and significant differences with respect to other results. In the categorical research, a Dense+LSTM pair attained the highest result, followed by a Dense+CNN pair. It is observed high performance in some of the 20 experiments with the Dense+LSTM pair. Its average performance ranked fifth, however, among the eight acoustic-linguistic networks pairs. The Dense+CNN pair, which was the second-best in the categorical emotion research, also ranked second in this dimensional emotion approach. This result from dimensional emotion recognition was supported by the fact that the LSTM also attained the highest performance on unimodal emotion recognition. Similar unimodal results were also observed in the categorical approach, in which the LSTM architecture performed the best among all the architectures.

A second important finding is the different results between categorical and dimensional emotion recognition from the feature/modality perspective. Feature set/modality, which attained the highest performance in the categorical approach, is different from the dimensional approach. In the categorical approach with the IEMOCAP dataset, word embeddings gave the highest performance in the unimodal model, as reported in [18, 124, 125, 126]. In contrast, in the dimensional approach, acoustic features' average performance gave better performance over linguistic features. This phenomenon can be explained by the fact that linguistic features (word embeddings) contribute to valence more than acoustic features do (see Tables 5.1 and 5.2). While the authors in [63] found this result, the authors in [127, 42, 82] extended it to find that, for arousal, acoustic features contribute more than linguistic features do. The results here further extend the evidence that linguistic features contribute more in valence prediction, while acoustic features give more accuracy in arousal and dominance prediction. Given this evidence, it is more likely that acoustic features will obtain higher performance than linguistic features in the unimodal case since they provide better performance for two of the three emotional dimensions. As suggested by Russell [52], however, a categorical emotion can be characterized by its valence and arousal only. This relation shows why linguistic features achieve

better performance than acoustic features do on categorical emotion.

As a final remark for this section, some important findings can be emphasized in this feature-level fusion of acoustic-linguistic information for dimensional emotion recognition. Dimensional emotion recognition is scientifically more challenging than categorical emotion recognition. This work achieved more consistent results than what it did in categorical emotion recognition. The combination of LSTM networks for both the acoustic and linguistic networks achieved the highest performance on bimodal feature fusion, as the same architecture did on unimodal emotion recognition. The proposal on using MTL for simultaneously predicting valence, arousal, and dominance worked as expected, and it is found that MTL with two parameters represented the interrelation among the emotional dimensions better than other MTL methods did.

5.3 Dimensional SER with ASR outputs

In the previous section, the linguistic information provided for the fusion with acoustic information came from manual transcription. It is difficult to obtain the linguistic information (i.e., correct transcription of spoken words) from the speech in a real implementation. Hence, this study evaluates the fusion of acoustic and linguistic information from ASR outputs to provide insight into the early-fusion method's achievement with current ASR technology.

Figure 5.6 shows an architecture of dimensional SER with ASR outputs. While acoustic information can be processed directly from speech, the linguistic information must wait until text transcription are generated by ASR. This bottleneck between acoustic and linguistic processing is a worth of study for future research direction.

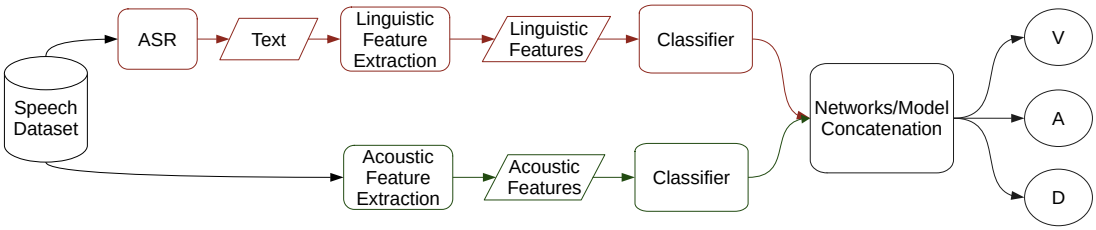

Figure 5.6: SER architecture by fusing acoustic and linguistic features from ASR outputs

An open-source project, namely DeepSpeech, was used to produce ASR outputs: text transcription [128]. The system was built upon [129], which used well-optimized end-to-end recurrent neural networks (RNN) to recognize spoken words. The system achieved a 45% word error rate (WER) on the IEMOCAP dataset. This loss in recognizing original (manual) transcription impacts on the lower performance of linguistic-based dimensional SER since some words cannot be obtained correctly.

An evaluation of five word embeddings of ASR outputs from IEMOCAP datasets has been performed. Table 5.5 shows performance of linguistic information on dimensional SER while Table 5.6 shows performances of these embeddings when fused with acoustic information. Compared to manual transcription (Table 5.2), it is clear that these results of ASR outputs are worse than manual transcription. There is no significant difference in the use of pre-trained models compared to the original word embeddings from these ASR

outputs. In this case, the BERT model achieves the highest performance on linguistic-only dimensional SER.

The addition of linguistic information from ASR outputs only improves the baseline acoustic information when it utilized pre-trained models. The fusion of acoustic information (pAA_D features) with original word embeddings (WE) attain a lower score than the baseline (average CCC score of 0.4 vs. 0.41). On the use of ASR outputs, pAA_D + FastText is the best pair for Acoustic+Linguistic information from ASR outputs. This highest score from ASR outputs is 0.05 (10% of relative loss), lower than the highest in manual transcription (0.453 vs. 0.508).

Table 5.5: Evaluation results on emotion recognition using linguistic information from ASR outputs

Feature set	V	A	D	Mean
WE	0.212	0.303	0.351	0.288
word2vec	0.218	0.293	0.350	0.287
GloVe	0.226	0.279	0.349	0.285
FastText	0.218	0.284	0.350	0.284
BERT	0.220	0.300	0.360	0.293

Table 5.6: Evaluation results on emotion recognition using acoustic and linguistic information from ASR outputs

Feature set	V	A	D	Mean
pAA_D + WE	0.221	0.550	0.428	0.400
pAA_D + word2vec	0.286	0.582	0.470	0.446
pAA_D + GloVe	0.275	0.582	0.472	0.443
pAA_D + FastText	0.277	0.602	0.479	0.453
pAA_D + BERT	0.263	0.599	0.469	0.444

5.3.1 Effect of word embeddings dimension

Since it is observed that BERT attains the highest performance on linguistic-only dimensional SER from ASR outputs, a higher dimension of word embedding may lead to better performance for dimensional SER from linguistic information. A BERT model has 768-dimension while others have 300-dimension. An investigation for the effect of word embeddings dimension has been performed by varying the original word embedding to 768- and 1024-dimension. Table 5.7 shows the difference of word embeddings dimension on linguistic and Acoustic+Linguistic dimensional SER performances. While the use of a higher dimension shows no significant differences in linguistic-only dimensional SER, the use of 768- and 1024-dimension improve the Acoustic+Linguistic pairs (pAA_D with original WE) to surpass the baseline acoustic-only (pAA_D) performance. The larger inputs (WE) may help the network learn better to achieve these results.

Table 5.7: Evaluation of different word embedding dimensions

Feature set	Dimension	V	A	D	Mean
WE	300	0.212	0.303	0.351	0.288
	768	0.199	0.307	0.352	0.286
	1024	0.203	0.293	0.347	0.281
pAA_D + We	300	0.221	0.550	0.428	0.400
	768	0.255	0.596	0.464	0.438
	1024	0.239	0.564	0.450	0.418

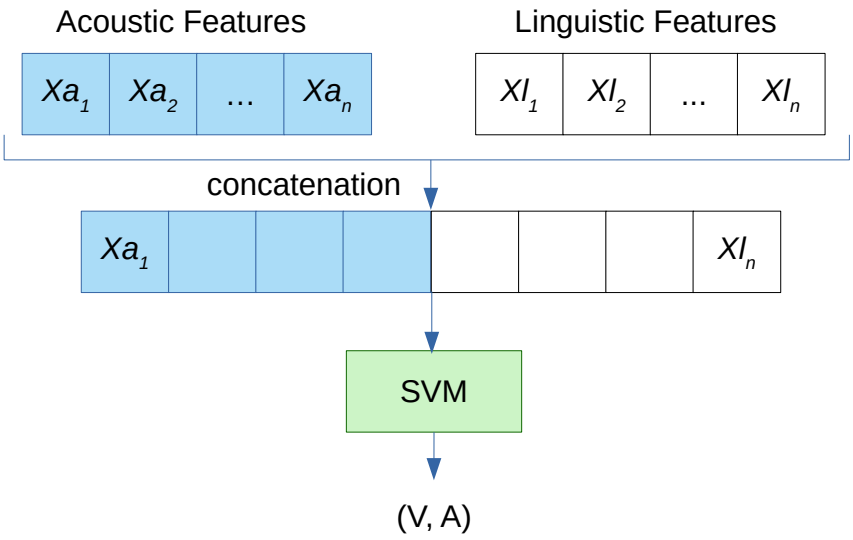

Figure 5.7: Acoustic-linguistic feature concatenation with SVM

5.4 Early fusion by network concatenation

5.4.1 Bimodal acoustic-linguistic feature fusion

To evaluate a different approach for early-fusion dimensional SER by fusing acoustic-linguistic information, this study employs another dataset, the Ulm State of Mind in Speech-elderly (USOMS-e) corpus. The dataset is part of Interspeech 2020 ComParE challenge [6]. The task is to predict categories of valence and arousal, which is converted from a 0-10 scale to low (0-6), medium (7-8), and high (9-10) classes.

In the previous chapter, evaluation of acoustic-only valence and arousal predictions were evaluated. It is shown that feature-based aggregation is better than outputs-based aggregation (majority voting). Since the feature aggregation's goal is to have the same dimension ($n \times 1$) for both acoustic and linguistic features, it is easy to concatenate both features to improve valence and arousal prediction. Figure 5.7 shows an approach on acoustic-linguistic features fusion. Two feature sets are stacked horizontally to build a new feature vector for the SVM classifier's input.

Table 5.8: Result of bimodal valence and arousal prediction on development and test partition: official baselines vs. proposed method

Features		Dev		Test	
Acoustic	Linguistic	V	A	V	A
ResNet50 [6]	-	31.6	35.0	40.3	**50.4**
-	BLAtt [6]	49.2	40.6	49.0	44.0
LibROSA	Gmax	**58.2**	34.6	40.5	34.8
ResNet50	Gmax	**58.2**	51.0	40.9	**50.4**
ResNet50	BLAtt	47.6	**52.5**	**56.3**	46.4
BoAW-250	BLAtt	**58.2**	44.4	49.0	47.4

5.4.2 Feature concatenation results

Given a set of acoustic-linguistic features pair (x_a and x_l) with valence and arousal category labels ('L', 'M', 'H'), the task of SVM is to classify whether a given feature set belongs to a category of valence and arousal. This classification task is performed using support vector classification (SVC) in scikit-learn toolkit [67] with a linear SVC kernel, 10^6 of maximum iteration, and optimized complexities (C) values in the range $[10^{-6}, 10^1]$ with 10^1 step size. For data balancing, imbalanced-learn toolkit was used [130]; however, no significant difference was found between balanced and imbalanced data. The other parameters are left as default. The SVC classification is performed separately to predict valence and arousal categories for the same feature set.

Table 5.8 shows the results on using acoustic-linguistic feature concatenation for valence and arousal category prediction on development and test partitions. This study improved the UAR score on development partition from 49.2 to 58.2 for valence and from 40.6 to 52.5 for arousal. On test partition, the UAR scores were improved from 49.8 to 56.3 for valence and from 49.0 to 50.4 for arousal. Although the gain was small, it is shown that bimodal acoustic-linguistic feature concatenation improved the UAR scores of valence and arousal in most combinations of acoustic-linguistic feature pairs. Table 5.8 shows that evidence on both development and test partitions.

5.5 Summary

This chapter reports an investigation of using acoustic features and word embeddings for dimensional speech emotion recognition with multitask learning (MTL). First, it can be concluded that using acoustic features and word embeddings can improve the prediction of valence, arousal, and dominance. Word embeddings help improve valence prediction, while acoustic features contribute more to arousal and dominance prediction. All the emotional dimensions gained prediction improvements on bimodal acoustic and linguistic networks; the greatest improvement was obtained using LSTM+LSTM architectures pair. Second, the proposed MTL with two parameters could improve all emotional dimensions' prediction compared to MTL with no parameters. The weighting factors given to valence and dominance may represent the interrelation among the emotional dimensions. This formulation only partially represents that interrelation because the obtained

improvement was still small. The formulation can be improved for future research by implementing other strategies, particularly those based on psychological theories and experiments. Third, a mismatch between categorical and dimensional emotion recognition can be explained as follows. Linguistic-based emotion recognition obtained better results than acoustic features did in categorical emotion, but the result was the opposite for dimensional emotion. This contrast can be explained by the fact that categorical emotion only relies on the valence-arousal space. The higher valence prediction obtained by word embeddings may result in better categorical emotion prediction than the prediction by acoustic features. Fourth, in comparing manual transcription with ASR outputs, a 10% loss in CCC score was obtained using a word error rate of 45%. Fifth, the feature concatenation of acoustic and linguistic features on the USOMS-e dataset obtained higher performances than a single modality emotion recognition. This feature concatenation was performed using acoustic features aggregation explained in the acoustic features side from the previous chapter.

In summary, a combination of speech features and word embeddings can solve the drawback of dimensional speech emotion recognition. Word embeddings improve the low score of the valence dimension in acoustic-based speech emotion recognition. The combination of both features not just improved valence but arousal and dominance dimensions too. Multitask learning also works as expected; it can simultaneously predict three emotion dimensions' degrees instead of predicting one by one dimension using single-task learning. This strategy may similar to human bimodal emotion perception from voice and linguistic information. Based on the obtained performances, however, there is room for improvement, e.g., a fine-tuned BERT model may improve the current results. In the next chapter, another framework fusion is explored, i.e., a late-fusion based approach.

Chapter 6

Fusing Acoustic and Linguistic Information at Decision Level

This chapter evaluates the fusion of acoustic and linguistic information at the decision level to improve speech emotion recognition (SER) performance. The evaluated method consists of two stages. First, deep neural networks (DNN) process the unimodal training data to predict the output (emotion degrees) of development/validation data. Second, the output of DNNs from acoustic and linguistic networks are fed into SVM to obtain the final prediction of emotion degrees.

6.1 Datasets partition

Since this study also evaluates some condition of the dataset (semi lexical-controlled vs. lexical-uncontrolled speaker, dependent vs. speaker independent), the datasets are split into four partitions and two scenarios. Two datasets were used in this part of the study. The first is the IEMOCAP dataset, which was used in the previous chapters. The dimensional labels are valence (V), arousal (A), and dominance (D) in a 5-point integer scale. However, it was found that some labels have values lower than 1 (e.g., 0.5) and higher than 5. These outlier data were removed; the remaining data were converted from a 5-point scale to [-1, 1] scale.

In addition to IEMOCAP dataset, MSP-IMPROV dataset [31] was used. The MSP-IMPROV dataset was designed within a dialogue framework to elicit target sentences with the same semantic content but was produced with different emotional expressions. In one recording, the target sentences were produced ad-lib; for another recording, the target sentences were read. These two recordings are referred to as "Target-improvised" and "Target-read", respectively. Since the goal is to examine the effect of both linguistic and acoustic information on emotional ratings at the late-fusion stage, these recordings were not appropriate for this study. However, two sets of recordings, which did not have the same semantic content, were used, called "Other-improvised" and "Natural-interaction." The former included conversations of the actors during improvisation sessions; the latter included the exchanges during the breaks. This natural-interaction is recorded while the actors were not acting. Zhang et al. [64] used a similar protocol, and this study followed their lead in referring to this subset of the MSP-IMPROV dataset as MSP-I+N (MSP improvised and natural interaction) or MSPIN. In this work, the same text transcriptions used by Zhang et al. was used (the authors of the dataset provide transcriptions); for

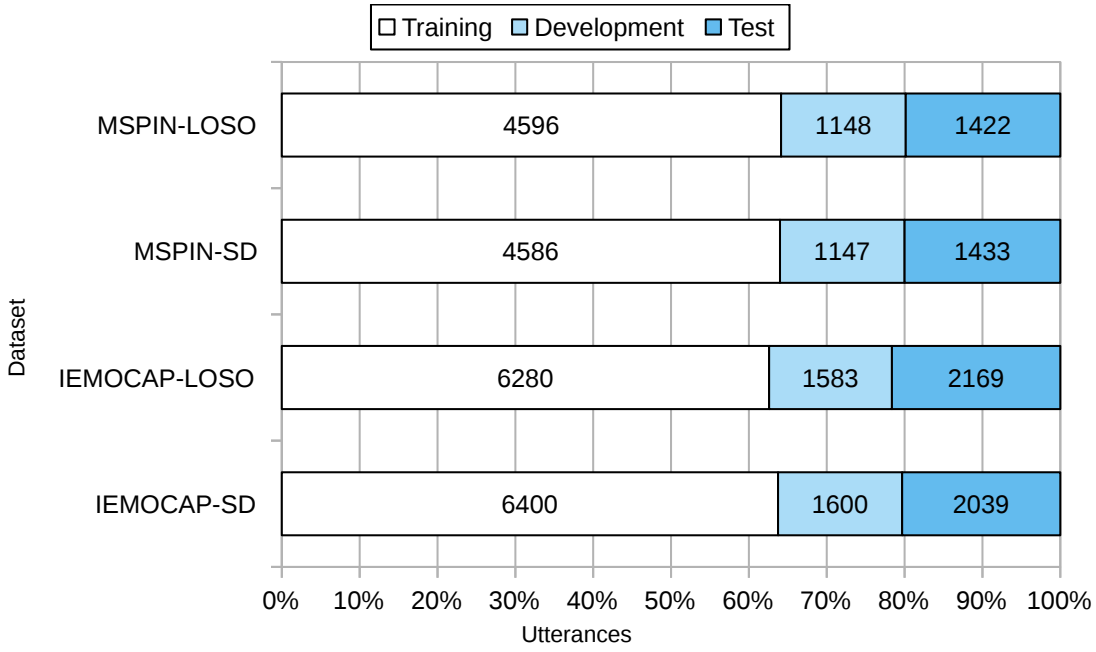

Figure 6.1: Proportions of data splitting for each partition of each dataset. In one-stage LSTM processing, the outputs of the model are both development and test data. In the second stage, i.e., the SVM processing, the input data is the prediction from the development set of the previous stage, and the output is the prediction of test data.

the additional utterances not included in the Zang study, transcriptions were obtained using Mozilla's DeepSpeech [128]. This study thus uses 7166 utterances from a total of 8438. The speech data in the dataset was sampled in mono at 44.1 kHz, with one file per utterance/sentence.

This study split each dataset into two partitions to observe any differences between a speaker-dependent (SD) partition and a speaker-independent partition made by leaving one session out (LOSO) for each dataset. For example, for the IEMOCAP dataset, the last session (i.e., session 5), recorded from two different actors (out of 10), is only used for testing. Similarly, all utterances from session 6 (two speakers out of 12) are used for the MSP-I+N test set. The rule for data splitting is to divide between the training + development and test sets in a ratio close to 80:20. This rule applies to both the SD and LOSO partitions. Then, of the training + development data, 80% is used for training, and the remaining 20% is used for development, as shown in Figure 6.1. Both methods are evaluated with the same unseen test sets to compare the performance and measure the improvement. Note that the dataset was not validated using a cross-validation technique (but instead divided into training and test data) for evaluation since the number of samples for both datasets is adequate (10039 and 7166 samples). This strategy is also utilized to keep the same test set for LSTM (one-stage processing) and SVM (two-stage processing) which is difficult if the samples are shuffled/cross-validated.

Table 6.1: Acoustic feature sets derived from the GeMAPS features by [4] and the statistical functions used for dimensional SER in this research

LLDs	HSF1	HSF2
intensity, alpha ratio, Hammarberg index, spectral slope 0-500 Hz, spectral slope 500-1500 Hz, spectral flux, 4 MFCCs, F0, jitter, shimmer, harmonics-to-noise ratio (HNR), harmonic difference H1-H2, harmonic difference H1-A3, F1, F1 bandwidth, F1 amplitude, F2, F2 amplitude, F3, and F3 amplitude.	mean (of LLDs), standard deviation (of LLDs)	mean (of LLDs), standard deviation (of LLDs), silence

6.2 Two-stage dimensional SER

6.2.1 LSTM network for unimodal prediction

Acoustic emotion recognition

Most SER research uses only acoustic features. The approach to acoustic SER in the study reported in the previous chapter is similar to that research. The acoustic networks input mean + std features at the utterance level and then use a silence feature with these statistical functions to investigate any improvement. This evaluation is a continuation of [14] with an extension on different feature sets and datasets. Apart from acoustic features used in the previous chapter, this study evaluates three acoustic different feature sets: LLDs of GeMAPS, HSF1, and HSF2.

The LLD features are the 23 acoustic features listed in Table 6.1. For each frame (25 ms), these 23 acoustic features are extracted. With a hop size of 10 ms, the maximum number of sequences is 3409 for the IEMOCAP dataset and 3812 for the MSP-I+N dataset. Hence, the input size is 3409×23 for IEMOCAP and 3812×23 for MSP-I+N. The extraction process uses the openSMILE toolkit [131].

Figure 6.2 shows an overview of the acoustic network. LSTM is chosen because the number of training samples is adequate (> 5000 samples). Furthermore, it shows promising results in the previous research [113]. Before entering the LSTM layers, the LLD features at the input layer are fed into a batch normalization layer to speed up the computation process. The three subsequent LSTM layers are stacked with 256 nodes in each layer, following one of the configurations in [132]. Instead of returning the last LSTM layer's final output, the networks were designed to return the full sequence and flatten it before inputting it to three dense layers representing valence, arousal, and dominance. The outputs of these last dense layers are then the predictions for those emotional attributes, i.e., the degrees of valence, arousal, and dominance in the range [-1, 1].

The tuning of hyper-parameters follows the previous research [18, 15]. A batch size of 8 was used with a maximum of 50 epochs. An early stop criterion with ten patiences would stop the training process if no improvement was made in 10 epochs (before the

Table 6.2: The hyper-parameter used in experiments

Hyper-parameter	Acoustic network	linguistic networks
network type	LSTM	LSTM
number of layers	3	3
number of units	256	256
fourth layer	Flatten	Dense
hidden activation	linear	linear
output activation	linear (LLD) / tanh (HSF)	linear
dropout rate	0.3 (LLD) / 0 (HSF)	0.3
learning rate	0.001	0.001
batch size	8	8
maximum epochs	50	50
optimizer	RMSprop	RMSprop

maximum epoch) and used the last highest-score model to predict the development data. An RMSprop optimizer was used with its default learning rate, i.e., 0.001. Table 6.2 shows the setups on acoustic and linguistic networks. These setups were obtained based on experiments with regard to the size of networks. For instance, the smaller acoustic networks with HSF features employed tanh output activation function did not use the dropout rate while the larger acoustic networks (with LLD) and linguistic networks employed linear activation function and dropout rate.

For the HSF1 and HSF2 inputs on acoustic networks, the same setup applies. These two feature sets are very small as compared to the LLDs: HSF1 has a size of 1×46, while HSF2 has a size of 1×47. This big difference in input size (1:1800) leads to faster computation on HSF1 and HSF2 than on the LLDs. Note that, although Figure 6.2 shows HSF2 as the input feature, the same architecture also applies for the LLDs and HSF1.

The idea of using LSTM is to hold the last output in memory and use that output as a successive step. For instance, LLD with $(3409, 23)$ feature size will process the first time step 1 to the last time step 3409. For HSF1 and HSF2, which contains a single timestamp, the data is processed only once ([1, 46] and [1, 47] for HSF1 and HSF2). Here, the only difference, from multiple time steps, is that the network performs three passes (forget gate, input gate, and output gate) instead of a single pass (see [42]). This information will include all information from the networks' memory.

Linguistic emotion recognition

The linguistic networks, shown in Figure 6.3 for the MSP-I+N dataset, uses the same input size for the three different linguistic features. The WE, WE with pre-trained word2vec, and WE with pre-trained GloVe embedding were used on the basis of the previous results with 300 dimensions for each word. The longest sequence in the IEMOCAP dataset is 100 sequences (words), while for MSP-I+N, the longest is 300 sequences. Hence, the input feature sizes for the LSTM layers are 100×300 for IEMOCAP and 300×300 for MSP-I+N with its corresponding number of samples. The same three LSTM layers are stacked as in the acoustic network, but the last LSTM layer only returns the last output.

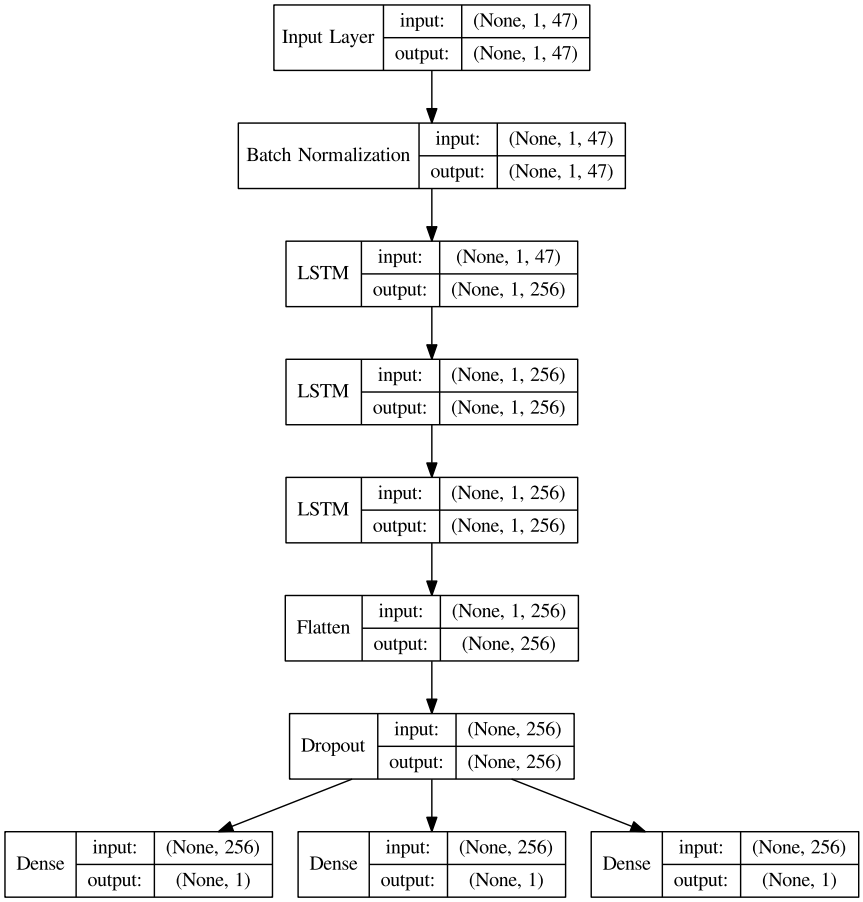

Figure 6.2: Structure of acoustic network to process acoustic features

A dense layer with a size of 128 nodes is added after the LSTM layers and before the last three dense layers. Between the dense layers is a dropout layer with the same probability of 0.3 to avoid overfitting.

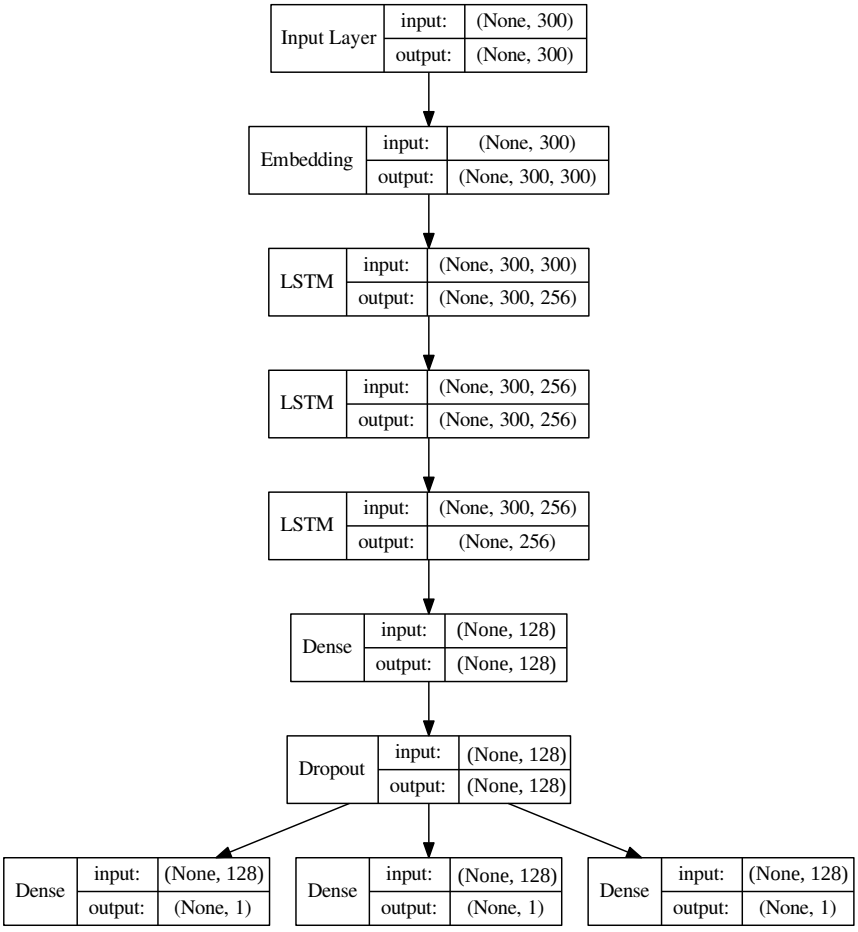

Figure 6.3: Structure of linguistic networks to process word embeddings/vectors

6.2.2 SVM for results fusion

The choice of an SVM (in this case, support vector regression, SVR) as the final classifier to fuse the outputs of the acoustic and linguistic networks is due to its effectiveness in handling smaller data (compared to a DNN) and its computation speed. The data points produced by LSTM processing as the input of SVM is small; i.e., 1600, 1538, 1147, and 1148 for IEMOCAP-SD, IEMOCAP-LOSO, MSPIN-SD and MSPIN-LOS0, respectively. The SVM then applies regression analysis to map them to the given labels. Figure 6.4 shows the architecture of this two-stage emotion recognition system using DNNs and an SVM. Each prediction from the acoustic and linguistic networks is fed into the SVM. From two values (e.g., valence predictions from the acoustic and text networks), the SVM learns to generate a final predicted degree (e.g., for valence). The concept of using the SVM as the final classifier is summarized as Chapter 2.

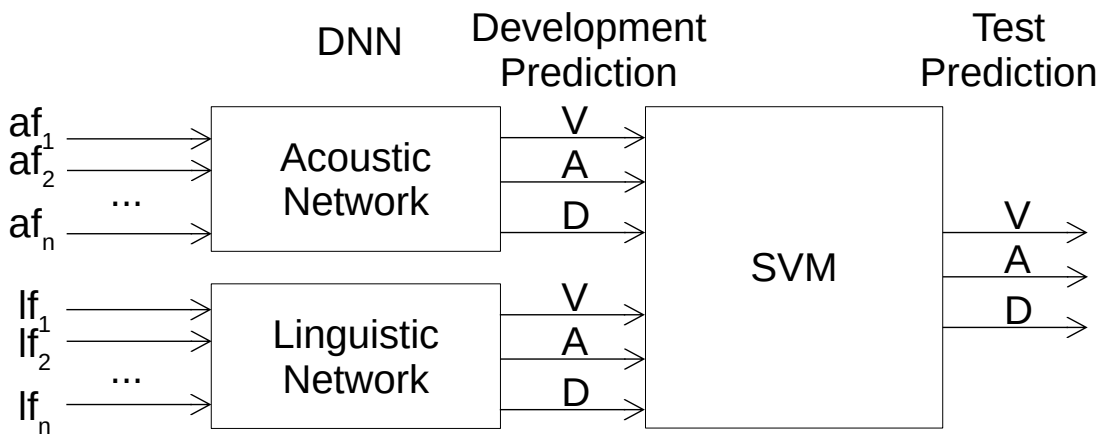

Figure 6.4: Proposed two-stage dimensional emotion recognition method using DNNs and an SVM. The inputs are acoustic features (af) and linguistic features (lf); the outputs are valence (V), arousal (A), and dominance (D).

6.3 Results and discussion

6.3.1 Results from single modality

Before presenting the bimodal feature-fusion results, it is important to show the results of unimodal emotion recognition. The goals here are (1) to observe the (relative) improvement of bimodal feature fusion over using a single modality, and (2) to observe the effects of different features on different emotion attributes.

Tables 6.3 and 6.4 list the single-modality results of dimensional emotion recognition from the acoustic and linguistic networks, respectively. In general, acoustic-based SER gave better results than text-based SER in terms of the average CCC score. For particular emotion attributes, the linguistic network gave a higher CCC score for valence prediction than those obtained by the acoustic network, except on the MSPIN datasets. These results confirm the previous finding by [82] that valence is better estimated by semantic features, while acoustic features better predict arousal. It is also found that acoustic features better predicted the dominance dimension than linguistic features. This finding can be inferred from both tables, in which the CCC scores for the dominance dimension are frequently higher from the acoustic network than from the linguistic networks.

The exception of a higher valence score on the MSPIN-SD dataset by the acoustic networks can be seen as the effect of either the DNN architecture or the dataset's characteristics. In [127], the obtained score was higher for valence than for arousal or liking (the third dimension, instead of dominance) with their strategy on acoustic features. In contrast, [132] obtained a lower score for valence than for arousal and dominance by using their proposed DANN method on the same MSP-IMPROV dataset (whole data, all four scenarios). Given this comparison, it can be concluded that the higher valence score obtained here was an effect of the DNN architecture, because of the multitask learning. The result on a single modality (acoustic network) outperformed the DANN result on MSP-IMPROV, where their highest CCC scores were (0.303, 0.176, 0.476) as compared to the obtained scores of (0.404, 0.605, 0.517) for valence, arousal, and dominance, respectively.

A linear search algorithm was performed on the scale [0.0, 1.0] with 0.1 steps to find the optimal parameter values for α and β. Using this. It was found that four sets

Table 6.3: CCC score results of dimensional emotion recognition using an acoustic network. The best results on the test set are in bold. LLDs: low-level descriptors from GeMAPS [4]; HSF1: mean + std of LLDs; HSF2: mean + std + silence

Feature set	V	A	D	Mean
IEMOCAP-SD				
LLD	0.153	0.522	0.534	**0.403**
HSF1	0.186	0.535	0.466	0.396
HSF2	0.192	0.539	0.469	0.400
MSPIN-SD				
LLD	0.299	0.545	0.441	0.428
HSF1	0.400	0.603	0.506	0.503
HSF2	0.404	0.605	0.517	**0.508**
IEMOCAP-LOSO				
LLD	0.168	0.486	0.442	0.365
HSF1	0.206	0.526	0.442	0.391
HSF2	0.204	0.543	0.442	**0.396**
MSPIN-LOSO				
LLD	0.176	0.454	0.369	0.333
HSF1	0.201	0.506	0.357	**0.355**
HSF2	0.206	0.503	0.346	0.352

Table 6.4: CCC score results of dimensional emotion recognition using text networks; each score is an averaged score of 20 runs with its standard deviation. WE: word embeddings; word2vec: WE weighted by pre-trained word vectors [3]; GloVe: WE weighted by pre-trained global vectors [7]

Feature set	V	A	D	Mean
IEMOCAP-SD				
WE	0.389 ± 0.008	0.373 ± 0.010	0.398 ± 0.017	0.387 ± 0.010
word2vec	0.393 ± 0.012	0.371 ± 0.018	0.366 ± 0.024	0.377 ± 0.016
GloVe	0.410 ± 0.007	0.381 ± 0.013	0.393 ± 0.016	**0.395 ± 0.010**
MSPIN-SD				
WE	0.120 ± 0.047	0.148 ± 0.023	0.084 ± 0.024	0.105 ± 0.026
word2vec	0.138 ± 0.031	0.108 ± 0.024	0.101 ± 0.024	0.116 ± 0.017
GloVe	0.147 ± 0.043	0.141 ± 0.019	0.098 ± 0.017	**0.128 ± 0.015**
IEMOCAP-LOSO				
WE	0.376 ± 0.008	0.359 ± 0.018	0.370 ± 0.020	0.368 ± 0.013
word2vec	0.375 ± 0.058	0.357 ± 0.058	0.365 ± 0.065	0.366 ± 0.059
GloVe	0.405 ± 0.009	0.382 ± 0.020	0.378 ± 0.021	**0.389 ± 0.014**
MSPIN-LOSO				
WE	0.076 ± 0.013	0.196 ± 0.011	0.136 ± 0.015	0.136 ± 0.009
word2vec	0.162 ± 0.008	0.202 ± 0.005	0.147 ± 0.003	**0.170 ± 0.000**
GloVe	0.192 ± 0.004	0.189 ± 0.007	0.129 ± 0.004	**0.170 ± 0.003**

of optimal parameters for the acoustic and text networks. Note that, while only the improvised and natural scenarios (MSP-I+N) were used to find the optimal text-network parameters for the MSP-IMPROV dataset, the whole dataset was used to find the optimal acoustic-network parameters. Table 6.5 lists the optimal parameter values for α and β.

Table 6.5: Optimal parameters for multitask learning

Dataset	Modality	α	β
IEMOCAP	acoustic	0.1	0.5
	linguistic	0.7	0.2
MSP-IMPROV	acoustic	0.3	0.6
	linguistic	0.1	0.6

To summarize the single-modality results, average CCC scores from three emotion dimensions can be used to justify which features perform better, among others. The results show that HSF2 was the most useful of the acoustic feature sets (in two of four datasets), while the word embedding (WE) with pre-trained GloVe embedding was the most useful of the linguistic feature sets. The performance of dimensional emotion recognition in the speaker-independent (LOSO) case was lower than in the speaker-dependent (SD) case, as predicted. Note that both acoustic and linguistic emotion networks used a fixed seed number to achieve the same result for each run; however, the linguistic networks resulted in different scores. Hence, standard deviations were given to measure fluctuation in 20 runs.

6.3.2 Results from SVM-based fusion

The main proposal of this research is the late-fusion approach combining the results from acoustic and linguistic networks for dimensional emotion recognition. This subsection presents the results for the late-fusion approach, including the obtained performances, comparison with the single-modality results, which pairs of acoustic-linguistic results performed better, and the overall findings.

For each dataset (IEMOCAP-SD, MSPIN-SD, IEMOCAP-LOSO, MSPIN-LOSO), nine combinations of acoustic-linguistic result pairs could be fed to the SVM system. Tables 6.6, 6.7, 6.8, and 6.9 list the respective CCC results for these datasets. Generally, the proposed two-stage dimensional emotion recognition improved the CCC score from single-modality emotion recognition. The pair of results from HSF2 (acoustic) and word2vec (text) gave the highest CCC score on speaker-dependent scenarios.

On the speaker-independent IEMOCAP dataset (IEMOCAP-LOSO), the result from the pair of HSF2 and GloVe gave the highest CCC score. This result linearly correlated with the single-modality results for that dataset, in which HSF2 obtained the highest CCC score among the acoustic features, and GloVe was the best among the linguistic features. On the four datasets, the results from HSF2 obtained the highest CCC score for two out of four datasets while GloVe obtained the highest CCC score for all four datasets. Hence, it can be concluded that the highest result from a single modality, when paired with the highest result from another modality, will achieve the highest performance among possible pairs.

Table 6.6: CCC score results of the late-fusion SVM on the IEMOCAP-SD test set

Inputs	V	A	D	Mean
LLD + WE	0.520	0.602	0.519	0.547
LLD + word2vec	0.552	0.613	0.524	0.563
LLD + GloVe	0.546	0.606	0.520	0.557
HSF1 + WE	0.578	0.575	0.490	0.548
HSF1 + word2vec	0.599	0.590	0.491	0.560
HSF1 + GloVe	0.595	0.582	0.495	0.557
HSF2 + WE	0.598	0.591	0.502	0.564
HSF2 + word2vec	0.595	0.601	0.499	**0.565**
HSF2 + GloVe	0.598	0.591	0.502	0.564

Table 6.7: CCC score results of the late-fusion SVM on the MSPIN-SD dataset

Inputs	V	A	D	Mean
LLD + WE	0.344	0.591	0.447	0.461
LLD + word2vec	0.326	0.586	0.439	0.450
LLD + GloVe	0.344	0.585	0.439	0.456
HSF1 + WE	0.461	0.637	0.517	0.538
HSF1 + word2vec	0.464	0.634	0.518	0.539
HSF1 + GloVe	0.466	0.630	0.510	0.535
HSF2 + WE	0.475	0.640	0.522	0.546
HSF2 + word2vec	0.486	0.641	0.524	**0.550**
HSF2 + GloVe	0.485	0.638	0.523	0.549

Table 6.8: CCC score results of late-fusion SVM on the IEMOCAP-LOSO test set

Inputs	V	A	D	Mean
LLD + WE	0.537	0.583	0.431	0.517
LLD + word2vec	0.528	0.580	0.421	0.510
LLD + GloVe	0.539	0.587	0.430	0.518
HSF1 + WE	0.565	0.565	0.453	0.528
HSF1 + word2vec	0.536	0.559	0.434	0.510
HSF1 + GloVe	0.559	0.570	0.452	0.527
HSF2 + WE	0.524	0.566	0.452	0.514
HSF2 + word2vec	0.531	0.571	0.445	0.516
HSF2 + GloVe	0.553	0.579	0.465	**0.532**

Table 6.9: CCC score results of late-fusion SVM on the MSPIN-LOSO test set

Inputs	V	A	D	Mean
LLD + WE	0.204	0.485	0.387	0.358
LLD + word2vec	0.267	0.487	0.386	0.380
LLD + GloVe	0.269	0.482	0.375	0.376
HSF1 + WE	0.224	0.565	0.410	0.400
HSF1 + word2vec	0.286	0.558	0.411	0.418
HSF1 + GloVe	0.282	0.555	0.409	0.415
HSF2 + WE	0.232	0.566	0.421	0.406
HSF2 + word2vec	0.287	0.562	0.411	0.420
HSF2 + GloVe	0.291	0.570	0.405	**0.422**

An average CCC score from three emotion dimensions can be used as a single metric to evaluate the improvement obtained by SVM-based late fusion. The right-most column in Table 6.6, 6.7, 6.8, and 6.9 shows the average CCC scores obtained from the nine pairs of acoustic and linguistic results on the four different datasets. Comparing these bimodal results to unimodal results (Chapter 4 and Chapter 5) shows the difference. All results from SVM improved unimodal results. In speaker-independent (LOSO) results (which are more appropriate for real-life analysis), the scores resulted from pairs of HSF with any word vector obtain remarkable improvements, particularly in the MSPIN-LOSO dataset. For any other pair involving LLDs, the obtained score was also lower as compared to other pairs. Considering all low scores involved LLD results, improving dimensional emotion recognition by using LLDs is more complicated than using HSF1 and HSF2 due to the larger feature size and the longer training time. The large network size created by an LLD input as a result of its much bigger feature dimension (e.g., 3409 $\times$ 23 on IEMOCAP) did not help either the single-modality or late-fusion performance. In contrast, the small sizes of the functional features (HSF1 and HSF2) enabled better performance on a single modality, which led to better performance for the late-fusion score. To obtain functional features, however, a set of LLD features must be obtained first. This problem is a challenging future research direction, especially for implementing dimensional emotion recognition with real-time processing.

In addition to the fact that a speaker-independent dataset is usually more difficult than a speaker-dependent dataset, the low score on MSPIN-LOSO was due to its low scores on a single modality. In other words, lower pair performance from a single modality will result in low performance in late fusion. In particular, these low results derive from low CCC scores from the text modality. The average CCC score for the linguistic modality on the MSPIN-LOSO dataset was less than 0.16, compared to an average score higher than 0.34 for the acoustic modality. All nine pairs in late-fusion approaches improved on the single-modality results because of the two-stage DNN and SVM regression analysis. Thus, out of 36 trials (9 pairs $\times$ 4 datasets), the proposed two-stage dimensional emotion recognition outperforms any single modality result (used in a pair).

The low score on MSPIN for the linguistic modality can be tracked to the origin of the dataset. There may have been a number of sentences semantically identical to the target sentences in the dataset used in this study. Although this study already chooses sentences

from the improvised dialogues and the natural interactions only, some of the sentences were identical to that of the target sentences in the "Target-Improvised" data set. This evidence was confirmed retroactively by manually checking the provided transcription and the automatic transcription. Given the nature of the elicitation task in a dialogue framework, this is not surprising. A similar low result for the linguistic modality on this MSPIN dataset was also shown in [64]. In general, compared to the IEMOCAP dataset, the MSPIN dataset suffers from low accuracy in recognizing the valence category by using acoustic and lexical properties. Interestingly, however, those authors also did not show improvement on the IEMOCAP scripted dataset, another text-based session in which linguistic features do not contribute significantly.

A relative improvement can be calculated to measure the performance of the proposed two-stage late fusion and by single modalities. For example, the pair of LLD + WE used the results from the LLDs in the acoustic network and the WE in the linguistic networks. This study compared the result for LLD + WE with that of the LLDs, as it had a higher score than the WE did. Figure 6.5 thus shows the relative improvement for all nine pairs. All of 36 trials showed improvements ranging from 5.11% to 40.32%. Table 6.10 lists the statistics for the obtained relative improvement. The obtained results show higher relative accuracy improvement as compared to those obtained by [64] for valence prediction, which ranged from 6% to 9%. Nevertheless, their multistage fusion method also showed benefits over the multimodal and single-modality approaches. These findings confirm the benefits of using bimodal/multimodal fusion instead of single-modality processing for valence, arousal, and dominance predictions.

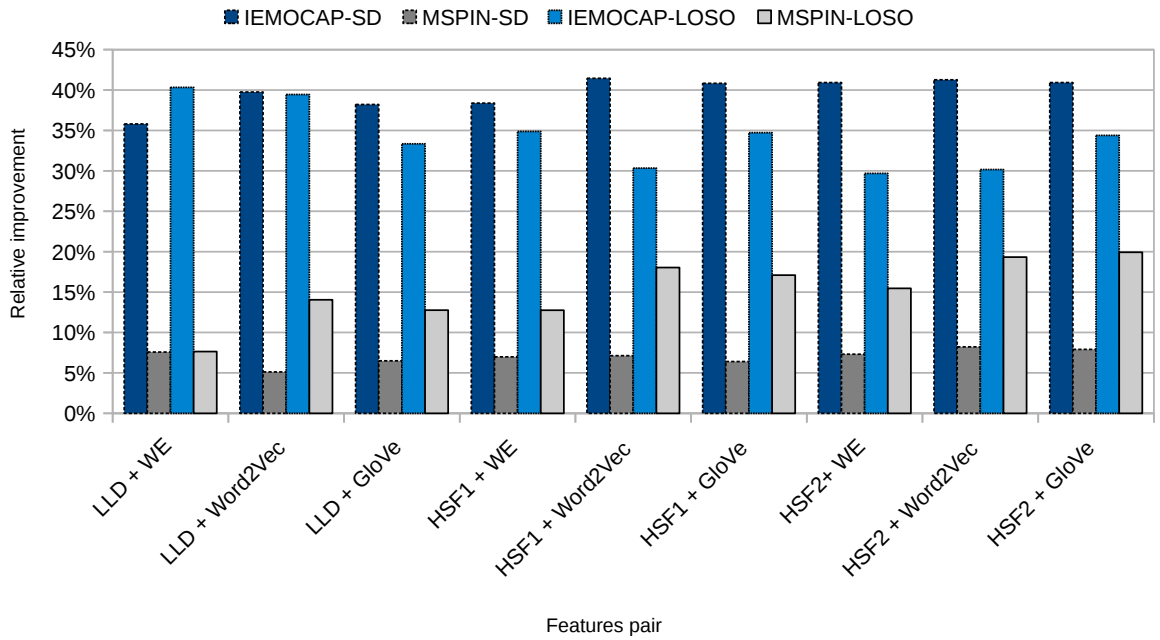

Figure 6.5: Relative improvement in average CCC scores from late fusion using an SVM as compared to the highest average CCC scores from a single modality

Table 6.10: Statistics of relative improvement by late fusion using an SVM as compared to the highest scores for a single modality across datasets; the scores were extracted from the data shown in Figure 6.5.

Statistic	IEMOCAP-SD	MSPIN-SD	IEMOCAP-LOSO	MSPIN-LOSO
Average	39.73%	7.01%	34.15%	15.23%
Max	41.45%	8.22%	40.32%	19.93%
Min	35.80%	5.11%	29.69%	7.64%
Std	1.90%	0.93%	3.84%	3.90%

Table 6.11: Significant difference between speaker-dependent and speaker-independent scenario on the same linguistic feature set; statistical tests were performed using two-tail paired $t-$test with $p-$value $= 0.05$.

Feature	IEMOCAP	MSPIN
WE	Yes	Yes
word2vec	No	Yes
GloVe	Yes	Yes

6.3.3 Speaker-dependent vs. speaker-independent linguistic emotion recognition

While speech-based emotion recognition is performed with a fixed random seed to generate the same result for each run, linguistic-based emotion recognition resulted in different scores for each run. The different results on text emotion recognition probably is caused by the initiation of weightings on embedding layers. In this case, statistical tests can be performed on text emotion results to observe the difference between speaker-dependent and speaker-independent scenario. In contrast, statistical tests cannot be performed between acoustic results and bimodal acoustic-linguistic results due to differences in the data (deterministic vs. non-deterministic).

Table 6.11 shows if there is a significant difference between speaker-dependent and speaker-independent results on the same feature set. The $p-$value was set at 0.05 with a two-tail paired t-test between mean scores of speaker-dependent and speaker-independent results. This paired t-test was based on the assumption that there are no outliers (after pre-processing), and two different inputs are fed into the same system. Only one result from text emotion recognition shows no significant difference in the IEMOCAP dataset. In contrast, all results from the MSPIN dataset shows a significant difference between speaker-dependent and speaker-independent results. This result reveals a tendency for a difference in evaluating speaker-dependent and speaker-independent data. The results from speaker-dependent data did not represent speaker-independent data. In other words, results from speaker-dependent data cannot be used to justify speaker-independent or whole data.

6.3.4 Effect of removing target sentence from MSP-IMPROV dataset

Since this research aims to evaluate the contribution of both acoustic and linguistic information in affective expressions, it is necessary to have sentences in the dataset that are free from any stimuli control. However, the original MSP-IMPROV dataset contains 20 "target" sentences, the same sentence that is elicited for different emotions. These parts of MSP-IMPROV dataset are irrelevant to this study; hence, it can be removed from the dataset (*Target - improvised* and *Target - read* parts). However, it was found that the results show low CCC scores indicating the influences of the target sentence. These results may be explained, as mentioned in section 6.3.2, that some of the utterances in the data analyzed in this study also inadvertently included sentences semantically the same as those in the improvised target sentences. Hence, it is necessary to compare the contribution of linguistic information in lexical-controlled and lexical uncontrolled datasets.

6.3.5 Final remarks

It has been tried to benchmark the obtained results in this study and others on the same datasets, scenarios, and metrics. Comparing to an early-fusion described in the previous chapter, which reports an early fusion method on the IEMOCAP dataset, this study improves the average CCC score from 0.508 to 0.532. This higher result suggests that late fusion is better than early fusion to model how humans fuse multimodal information, which is in line with neuropsychological research. This late-fusion approach can be embedded with current speech technology, i.e., ASR, in which the text output can be processed to weigh emotion prediction from acoustic features.

Abdelwahab et al. [132] used MSP-Podcast [133] as a target corpora, which is not available for the public yet, and IEMOCAP with MSP-IMPROV a source corpus to implement their DANN for cross-corpus speech emotion recognition. Although the goal is different, it was observed similar patterns between theirs and the acoustic-only speech emotion recognition int this study. First, it was observed that the order of highest to lowest CCC scores is arousal, dominance, and valence. This pattern is also consistent when IEMOCAP is mixed with MSP-IMPROV as reported by [122] (in Table 2). Second, it was observed that the CCC scores obtained in IEMOCAP are higher than those obtained in MSP-IMPROV; this lower score in MSP-IMPROV was due to the smaller size of the dataset.

Along with the SVM architecture, this study also explored the parameters C and γ, because both parameters are important for an RBF-kernel-based SVM architecture [67]. Linear search was used in the ranges of $[10^{-2}, 1, 10^2, 2 \times 10^2, 3 \times 10^2]$ for C and $[10^{-2}, 10^{-1}, 1, 10, 10^2]$ for γ with a fixed value of ϵ, i.e., 0.01. The best parameter values were $C = 200$ and $\gamma = 0.1$. The repository includes the detailed implementation of the SVM architecture.

Per the stated objective in this chapter, this study applied two-stage processing by using DNNs and an SVM for dimensional emotion recognition from acoustic and linguistic features on four different datasets. It is found that the combination of mean + std + silence from the acoustic features and word embeddings weighted by pre-trained GloVe embeddings achieved the highest result among the nine pairs of acoustic-linguistic results from DNNs trained with multitask learning. When the performance in obtaining one input

to the SVM is very low, the resulting relative improvement due to the SVM is also low. For instance, the lowest improvement on MSPIN-LOSO was from LLD + WE features, in which WE obtained a low score (CCC_{avg} or $\overline{CCC} = 0.136$) on linguistic networks. This phenomenon suggests a challenging future research direction for dealing with very little information, particularly linguistic information, in the fusion strategy. One strategy applied in this research was to use a pre-trained GloVe embedding on linguistic features with HSF2 on acoustic features, which improved the $\overline{CCC}$ score from 0.358 (relative improvement = 7.64%) to 0.422 (relative improvement = 19.93%). Other strategies should also be proposed, such as how to handle the data differently when the same sentence elicits different emotions (i.e., whole MSP-IMPROV dataset). In contrast, the current word-embedding feature treats the same sentence as the same representation, even when it conveys different emotions. Although BERT was used in the previous chapter, the configuration only utilized a pre-trained model instead of fine-tuned model.

6.4 Summary

This chapter presents a late-fusion approach to acoustic-linguistic emotion recognition. Several findings can be emphasized in this chapter. First, it was found a linear correlation between the single-modality and late-fusion methods in dimensional emotion recognition. The best results from each modality, when they were paired, gave the best fusion result. Similarly, the worst results obtained from each network, when they were paired, gave the worst fusion results for bimodal emotion recognition. This finding differs from that reported in [18], which used an early-fusion approach for categorical emotion recognition. In their work, the best pair differs from then best methods in single modalities.

Second, linguistic features strongly influenced dimensional SER's score on the valence dimension, while acoustic features strongly influenced arousal and dominance scores. Accordingly, the proposed two-stage processing can take advantage of linguistic features, which are commonly used in predicting sentiment (valence) for the dimensional emotion recognition task. The proposed fusion method improves all three emotion dimensions without attenuating the performance of any dimension. The proposed method elevates valence scores, arousal, and dominance, subsequently from the highest to the lowest gain.

Third, the combination of input pairs does not matter in the proposed fusion method, as indicated by the low deviation in relative improvement across the nine possible input pairs. What does matter is the performance of the input in the DNN stage. If the performance of a feature set in the DNN stage is low ($\overline{CCC} \leq 0.2$), it will also result in low performance when paired with another low-performance input in the SVM stage.

Future research can be directed to generalize the evaluated method presented in this chapter to other datasets. While the SVM stage in this study only performed once, it can be extended to be performed many times to observe such improvements. These broad research directions are open challenges for researchers in human-computer interaction.

Chapter 7

Comparative Analysis

This chapter aims at summarizing the results obtained in this research and comparing with others.

7.1 Comparison within this study

To this end, several methods have been proposed, and several results have been obtained. Summarizing the results could be used to evaluate the research and its trend. Comparing one method to another, either from the feature side or classifier/model side, could be used to judge the effectiveness of the method. The use of a single metric CCC makes it easy to track the performance of different methods.

Improvements in CCC scores were obtained using different methods described from Chapter 4 to Chapter 6. Table 7.1 shows averaged CCC scores among valence, arousal, and dominance using different acoustic features, linguistic features, and classifiers on the IEMOCAP dataset. Clearly, it shows the gradation of the performance scores (average CCC [CCC_{ave}]) from unimodal acoustic SER to bimodal acoustic-linguistic SER.

The first eight rows in Table 7.1 shows dimensional SER from acoustic features only. HSF consistently obtained higher CCC scores than LLD. An optimization of this feature set can be achieved using MLP architecture with a deeper layer (a maximum score was obtained using five layers). A comparison of ignoring (keeping) silence, removing silence, and utilizing silence as an additional feature showed that the latter two methods are better than the first. Since linguistic-only dimensional SER is not the focus of this study, the discussion of linguistic-only dimensional SER is not discussed thoroughly.

Instead, the use of linguistic information, in addition to acoustic information, improved the performance scores. Using pAA feature set with LSTM as the baseline, the early fusion method improved the average CCC score from 0.400 to 0.508. Furthermore, the late fusion method improved the early fusion method from 0.508 to 0.532. As suggested in the next chapter, there is a bottleneck between acoustic and linguistic processing in bimodal SER fusion; the process needs to wait for ASR outputs in real application for information fusion. However, bimodal fusion consistently gains higher scores than acoustic-only SER. Although there is room for improvement, this study limits the discussion to this point since the goal is studying the fuse of acoustic and linguistic information for dimensional emotion recognition. Further investigations are recommended in the next chapter's Future research direction section.

Table 7.1: Reported results on the IEMOCAP dataset test set (Session 5); the number inside bracket represents the number of layers; sil: silence

Acoustic	Linguistic	Classifier	CCC_{avg}
GeMAPS LLD	-	LSTM (3)	0.382
pAA LLD	-	LSTM (3)	0.354
pAA_D LLD	-	LSTM (3)	0.370
GeMAPS HSF	-	LSTM (3)	0.389
pAA HSF	-	LSTM (3)	0.384
pAA_D HSF	-	LSTM (3)	0.413
pAA_D HSF	-	MLP (6)	0.452
pAA_D HSF sil-removed	-	MLP (6)	0.459
pAA_D HSF + sil	-	MLP (6)	0.466
-	WE	LSTM (3)	0.361
-	word2vec	LSTM (3)	0.386
-	GloVe	LSTM (3)	0.392
-	FastText	LSTM (3)	0.384
-	BERT	LSTM (3)	0.375
pAA HSF	GloVe	LSTM (3) – LSTM (3) [early fusion]	0.508
pAA HSF + sil (HSF2)	Glove	LSTM (3) – SVM [late fusion]	0.532

7.2 Comparison with other studies

One of the motivations to use the dimensional model over the categorical model is that only a little research has been conducted using this emotion model. This small number of research leads to the difficulties of comparing this research to similar studies. In INTERSPEECH 2020, 11 papers proposed bimodal acoustic-linguistic fusion for SER. Among these papers, only four papers evaluated dimensional SER. However, no paper reported the results in CCC score.

Table 7.2 compares this research with others. Although the exact condition cannot be performed for ideal comparison, the performance of valence (V), arousal (A), and dominance (D) in CCC scores can be used to judge the rough performance among several methods. The first four rows are the results obtained in this study. The rest are from other research with different datasets and methods.

The closest comparison can be made between this research and the ones proposed by Zhao et al. [134, 64]. In both papers, the authors proposed to fuse acoustic features with gender and age information. Using the other non-linguistic information, they improved the CCC scores of dimensional SER except for valence prediction. The fusion of acoustic, age, and gender information is performed in a hierarchical manner. Since the scenario of the IEMOCAP is not explained; it is assumed the results obtained by Zhao et al. (rows fifth and sixth) are in speaker-independent (SI) scenarios. In [135], the authors copied the parts of the dataset for augmentation or balancing, since they also evaluated categorical emotion. The addition of these data improved the CCC scores; however, this technique should be avoided since the model learns the same data twice.

To overcome the problem of mismatch among datasets, Abdelwahab and Busso [132] proposed domain adversarial neural network (DANN) for acoustic emotion recognition.

Table 7.2: Comparison of this study with others; SD: speaker-dependent; SI: speaker-independent; Ac: acoustic, Li: linguistic, Vi: visual

No.	Dataset	Authors	Modalities	V	A	D
1	IEMOCAP SD	Atmaja	Ac+Li	0.596	0.601	0.499
2	IEMOCAP SI	Atmaja	Ac+Li	0.553	0.579	0.465
3	MSPIN SD	Atmaja	Ac+Li	0.486	0.641	0.524
4	MSPIN SI	Atmaja	Ac+Li	0.291	0.570	0.405
5	IEMOCAP	Zhao et al. [134]	Ac	0.715	0.392	0.539
6	IEMOCAP	Zhao et al. [135]	Ac	0.590	0.689	0.591
7	IEMOCAP (train) & MSP-Podcast (test)	Abdelwahab & Busso [132]	Ac	0.140	0.305	0.181
8	MSP-Podcast (train) & IEMOCAP (test)	Partasarathy & Busso [136]	Ac	0.235	0.623	0.441
9	MSP-Podcast SI	Sridhar et al. [95]	Ac	0.291	0.711	0.690
10	SEMAINE	Yang & Hirschber [137]	Ac	0.506	0.680	-
11	RECOLA	Bakshi et al. [138]	Ac	0.314	0.660	-
12	SEWA (DE)	Schmitt et al. [113]	Ac	0.489	0.499	-
13	SEWA (DE+HU)	Atmaja & Akagi [20]	Ac+Vis	0.656	0.680	-
14	SEWA (DE+HU)	Chen et al. [127]	Ac+Vi+Li	0.755	0.672	-

They obtained low CCC scores by using different datasets for training and test, as shown in Table (row No.7). These results were achieved using three layers of DANN. Parthasarathy and Busso [136] also took into account the problem of generalization across datasets by proposing a semi-supervised method with the reconstruction of intermediate feature representation that does not require labels. One of the results, using opposite datasets for testing and test as used by Abdelwahab and Busso, shows significant improvement on CCC scores. However, both research [132, 136] showed low valence prediction performance, which is tackled in this study.

Another way to improve valence prediction is by utilizing different regularization for different emotion attributes, as proposed by Sridhar et al. [95]. However, as shown in row No. 9, the improved valence prediction for valence is not comparable to arousal and dominance. In this study, we achieve comparable performances among valence, arousal, and dominance.

Using other datasets, SEMAINE and RECOLA, comparable CCC scores were observed between the results in these datasets and this study. Yang and Hirschber [137] combined waveform and spectrogram for predicting valence and arousal from speech. Although the results are shown for SEMAINE (row No. 10), similar scores were observed for RECOLA. Using similar methods, Bakhshi et al. [138] combined time and frequency information using different networks for predicting valence and arousal. In this case, the score of valence is about half from that of arousal.

SEWA is another dataset designed for emotion and sentiment research. Using this dataset Schmitt et al. [113] reveal the importance of mean and standard deviation from

GeMAPS feature set for dimensional SER. Significant improvements were observed in the German (DE, Deutsche) sub-corpus. Atmaja and Akagi [20] added visual features in addition to acoustic features to improve CCC scores. Finally, the last row in Table 7.2 shows that the fusion of acoustic, linguistic, and visual information attained the highest average CCC score for dimensional SER.

The proposed methods in this study show advantages among those other methods. First, the bimodal acoustic-linguistic fusion doubles the amount of information from the unimodal acoustic analysis. More data improves the effectiveness of the SER system since the system can learn from more resources. The results proved this hypothesis. Second, there is no need to add other modalities. Since linguistic information could also be obtained from speech, the proposed method only relies on speech data. Unlike audiovisual emotion recognition and addition of age and gender information, additional modalities are needed for the fusion method. Third, the fusion approach is performed automatically based on the data (bottom-up approach). Although this approach has several disadvantages, the implementation is less complicated than model-driven approach and the results show modest improvement from other methods.

Several drawbacks of the evaluated methods have been found during this study. A bottleneck between acoustic and linguistic processing is the major shortcoming of this study. This drawback triggers a future study to predict linguistic information from acoustic information only without the need for the transcription. Another challenge is to reduce the complexity of the proposed two-stage processing. In practice, the SER system should be able to recognize emotion within a speech in almost real-time. This requirement is difficult to be accomplished within the current late fusion approach.

Aside from the comparison among different methods, Table 7.2 shows other trends in dimensional SER. First, the addition of other non-linguistic information significantly improved the performance. This significant improvement is evidence for relations among non-linguistic information. Second, there is a mismatch among SER datasets, which are currently being tackled by SER researchers. This problem is a challenging opportunity for testing the proposed SER method for future research. Finally, more modalities tend to improve SER performance. However, per the stated objective of this study, some cases cannot provide the measurement of other modalities. This study is intended to maximize the performance of emotion recognition by fusing acoustic and linguistic information. The result shows substantial improvement for dimensional emotion recognition from speech.

Chapter 8

Conclusions

This closing chapter is divided into two sections, General summary and Future research directions. Both sections are described below.

8.1 General summary

This dissertation demonstrates the necessity of fusing acoustic with linguistic information for dimensional speech emotion recognition (SER). The results of acoustic-only SER showed the necessity to goes beyond unimodal acoustic analysis only. Two evaluated fusion methods, early and late fusions, confirm the effectiveness of fusing acoustic with linguistic information for dimensional SER.

Aside from the main goal, three strategies were evaluated to investigate the potential solution of partial problems in dimensional SER. These strategies were dimensional SER using acoustic information only, fusing acoustic and linguistic at the feature level, and fusing acoustic and linguistic information at the decision level. These strategies yield answers to the following five issues:

1. region of analysis for feature extractions: high-level statistics (HSF) using mean and standard deviation consistently shows more meaningful representation for speech than low-level descriptors (LLD);

2. effect of silent pause region: silence regions are predicted to contribute in dimensional speech emotion recognition; both removing silence or using silence feature as an additional feature slightly improves the performance score of the baseline whole speech regions;

3. low valence prediction score on dimensional SER: fusing linguistic information with acoustic information could improve the performance of valence prediction;

4. the necessity of fusing acoustic and linguistic information: fusing both acoustic and linguistic information consistently and significantly improved performance score;

5. framework for fusing acoustic and linguistic information: the late-fusion (decision-level) approach obtained slightly better performance than an early-fusion (feature-level) approach.

Furthermore, this research not only contributes to solve these issues. Several new insights are gained including the following findings.

In the first strategy, the research is extended to evaluate the aggregation methods for chunks to a story, the many-to-one problem. It is found that input feature aggregation, either by mean or maximum values, consistently obtained better performance than output aggregation by majority voting. This result reveals the importance of statistical functions as feature representation. Similarly, the analysis of region for feature extraction showed that statistical function on fixed length or utterance could represent emotional contents in speech better than frame-based acoustic features.

The second important insight is the importance of correlation-based loss function. Since the goal is concordance correlation coefficient (CCC), this research is developed by inverting CCC as loss function and then accommodating three dimensional emotions by summing up them. This straightforward flow, as expected, improves the performance of dimensional SER.

Speaker independent scenario is different from speaker dependent scenario. The results in Chapter 6 reveal this finding. The significant different between speaker dependent and speaker independent should guide the future research on SER to choose speaker independent scenario for evaluating the model. Speaker independent also not enough. The repetition of linguistic information (word or phrase) may make the model shows the higher performance than its original performance. A SER model should be able to recognize emotion from speech regardless the speaker information.

The previous research shows the strong correlation of linguistic information with valence and acoustic information with arousal. This research adds the finding of strong correlation of acoustic with dominance. However, adding more information, i.e., linguistic information, not only improves the prediction of valence, but also arousal and dominance. The contribution from each modality to each dimensional emotion is worth study for future, particularly on psychological side. The cross relation between modality and dimensional emotion should be also studied from neuroscience side.

Although several solutions have been proposed and several insights have been gained, it is known that current understanding on dimensional emotion is limited. Fusing acoustic and linguistic information may reflect humans multimodal perception works. However, extracting the "real" emotion from speech measurement is a long journey research. As stated in the philosophy of this research, it is impossible to reach perfect accuracy to recognize human emotion. The possibility is to maximize the recognition rate from the given information, acoustic and linguistic.

8.2 Future research directions

While this research contributes to several areas, the following issues are suggested for future research on automatic speech emotion recognition based on this study.

Accelerating high-level feature extraction for speech emotion recognition

In this research, it was found that HSF consistently obtained higher performance score than LLD. However, to obtain HSF, LLD first must be extracted. In practice, this is not an efficient method. A strategy to avoid this time lag should be proposed. For instance,

dividing frames into chunks and aggregating these HSF (as evaluated in Chapter 4). The computation time must be considered apart from the performance score.

Bimodal late-fusion approach by output aggregation

In chapter four, it was found that input aggregation is better than output aggregation by majority voting. The goal of the input aggregation for the acoustic features is to be able to concatenate with linguistic features. The acoustic and linguistic information can be processed separately through different classifiers. The output prediction by both modalities can be fused by output aggregation methods such as majority voting. Since a late fusion showed a better performance than an early fusion, the obtained score may improve the previously reported score in bimodal feature concatenation (Chapter 4).

Bimodal acoustic-linguistic fusion by two spaces resultant

In this research, the best results were obtained by late fusion with SVM. It means that the decision function is taken automatically by measuring the distance of the prediction from support vector line. In Chapter 5, the optimization has been performed to find the optimum parameters for α, β, and γ. Instead of concatenating models for finding optimum parameters, two spaces (acoustic and linguistic) can be modeled statistically/-mathematically. Another approach is by a late fusion. Each acoustic and linguistic model will predict vector of valence, arousal, and dominance. The fusion decision can be taken by adding both vectors from the same space (e.g., dominance space from acoustic and linguistic) by some weightings or modifications, if necessary.

Lexical controlled vs. lexical uncontrolled emotion recognition

While this study performed an evaluation on parts of MSP-IMPROV datasets, it was found that these parts of the dataset (lexical uncontrolled) has been influenced by other parts (lexical controlled). This proposal will evaluate the necessity of linguistic information for SER: does it always need linguistic information for SER? In some cases, linguistic information may not be needed (e.g., in the condition in which the intonation to express the emotion is clear in short utterance). The trade-off between the performance improvement and model complexities should be carried out to judge "when is linguistic information needed?," "in what condition?," and "what is the cues to use linguistic information?."

Bottleneck between acoustic and linguistic processing

The goal of the research is to enhance human life. While this study focuses on a proof-of-concept of fusing acoustic and linguistic information for emotion recognition, the real problem may appear on its implementation. One of the spotted problems is the bottleneck between acoustic and linguistic processing. Acoustic features can be extracted directly from speech, while linguistic information must wait for ASR output in practice. This time different processing will make the (emotion recognition) system occupies a longer time if no strategies are proposed to minimize the bottleneck.

Model generalization

A common view in emotion recognition has been challenged, particularly based on facial expression. The weak evidence and model-specific results have raised the need for a generalization for automatic emotion recognition. The models and their results reported here can be applied to other datasets. For instance, to check the consistency of efficient high-level features, removing silence for feature extraction, and comparing early fusion to late fusion approaches. While this research evaluated English, an extension to other languages should be made within the minimum effort since many linguistic models for these languages are available. The solution to the problem appeared in these multilingual approaches, for instance, "when is the linguistic information needed?", should be evaluated in future research.

References

[1] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Dev. Psychopathol.*, vol. 17, no. 3, pp. 715–734, 2005.

[2] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, *Dive into Deep Learning*, 2020.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," in *Int. Conf. Learn. Represent.*, jan 2013.

[4] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, apr 2016.

[5] T. Giannakopoulos, "pyAudioAnalysis: An open-source python library for audio signal analysis," *PLoS One*, vol. 10, no. 12, pp. 1–17, 2015. [Online]. Available: https://github.com/tyiannak/pyAudioAnalysis/

[6] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen, H. Baumeister, A. D. Macintyre, and S. Hantke, "The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks," in *INTERSPEECH*, 2020, pp. 2042–2046.

[7] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in *Conf. Empir. Methods Nat. Lang. Process.*, 2014, pp. 1532–1543.

[8] R. W. Picard, "Affective Computing," Tech. Rep., 1995. [Online]. Available: http://www.media.mit.edu/œpicard/

[9] S. A. Kotz and S. Paulmann, "Emotion, Language, and the Brain," *Lang. Linguist. Compass*, vol. 5, no. 3, pp. 108–125, mar 2011. [Online]. Available: http://doi.wiley.com/10.1111/j.1749-818X.2010.00267.x

[10] K. A. Lindquist, L. F. Barrett, E. Bliss-Moreau, and J. A. Russell, "Language and the perception of emotion," *Emotion*, vol. 6, no. 1, pp. 125–138, 2006.

[11] X. Li, "A Three-Layer Model Based Estimation of Emotions in Multilingual Speech," Ph.D. dissertation, Japan Advanced Institute of Science and Technology, 2019.

[12] E. Väyrynen, "Emotion recognition from speech using prosodic features," Ph.D. dissertation, University of Oulu, 2014.

[13] B. T. Atmaja and M. Akagi, "Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model," in *2019 IEEE Int. Conf. Signals Syst.* IEEE, jul 2019, pp. 40–44.

[14] ——, "The Effect of Silence Feature in Dimensional Speech Emotion Recognition," in *10th Int. Conf. Speech Prosody 2020*, no. May. ISCA: ISCA, may 2020, pp. 26–30.

[15] ——, "Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning," *APSIPA Trans. Signal Inf. Process.*, vol. 9, no. May, p. e17, may 2020.

[16] B. T. Atmaja, Y. Hamada, and M. Akagi, "Predicting Valence and Arousal by Aggregating Acoustic Features for Acoustic-Linguistic Information Fusion," in *TENCON*, 2020.

[17] B. T. Atmaja and M. Akagi, "On The Differences Between Song and Speech Emotion Recognition: Effect of Feature Sets, Feature Types, and Classifiers," *TENCON*, 2020.

[18] B. T. Atmaja, K. Shirai, and M. Akagi, "Speech Emotion Recognition Using Speech Feature and Word Embedding," in *2019 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Lanzhou, 2019, pp. 519—-523.

[19] B. T. Atmaja and M. Akagi, "Improving Valence Prediction in Dimensional Speech Emotion Recognition Using Linguistic Information," in *Orient. COCOSDA*, 2020, pp. 166–171.

[20] ——, "Multitask Learning and Multistage Fusion for Dimensional Audiovisual Emotion Recognition," in *ICASSP 2020 - 2020 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, may 2020, pp. 4482—-4486. [Online]. Available: https://ieeexplore.ieee.org/document/9052916/

[21] R. Elbarougy, B. T. Atmaja, and M. Akagi, "Continuous Audiovisual Emotion Recognition Using Feature Selection and LSTM," *J. Signal Process.*, vol. 24, no. 6, 2020.

[22] B. T. Atmaja and M. Akagi, "Deep Multilayer Perceptrons for Dimensional Speech Emotion Recognition," Auckland, pp. 325–331, 2020.

[23] B. T. Atmaja, K. Shirai, and M. Akagi, "Deep Learning-based Categorical and Dimensional Emotion Recognition for Written and Spoken Text," *IPTEK J. Proc. Ser.*, 2019.

[24] B. T. Atmaja and M. Akagi, "Evaluation of Error and Correlation-Based Loss Functions For Multitask Learning Dimensional Speech Emotion Recognition," *Annu. Conf. Acoust. Vib.*, mar 2020. [Online]. Available: http://arxiv.org/abs/2003.10724

[25] ——, "Dimensional Speech Emotion Recognition from Acoustic and Text Features Using Multitask Learning," in *ASJ Spring Meet.*, 2020, pp. 1003–1004.

[26] B. T. Atmaja, R. Elbarougy, and M. Akagi, "RNN-based Dimensional Speech Emotion Recognition," in *ASJ Autumn Meet.*, Shiga, 2019, pp. 743–744.

[27] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Int. Conf. Spok. Lang. Process. ICSLP, Proc.*, vol. 3, 1996, pp. 1970–1973.

[28] V. A. Petrushin, "Emotion In Speech: Recognition And Application To Call Centers," *Proc. Artif. neural networks Eng.*, vol. 710, pp. 22–30, 1999.

[29] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *INTERSPEECH*, 2005, pp. 1517–1520. [Online]. Available: http://www.expressive-speech.net/emodb/

[30] C. Busso, M. Bulut, C.-C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[31] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, jan 2017.

[32] S. Livingstone and F. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)," *PLoS One*, pp. 1–35, 2018.

[33] J. Vincent, "AI EMOTION RECOGNITION' CAN'T BE TRUSTED," 2019. [Online]. Available: https://www.theverge.com/2019/7/25/8929793/emotion-recognition-analysis-ai-machine-learning-facial-expression-review

[34] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements," *Psychol. Sci. Public Interes.*, vol. 20, no. 1, pp. 1–68, jul 2019.

[35] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. Fusion*, vol. 37, pp. 98–125, sep 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1566253517300738

[36] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.

[37] C. O. Alm, D. Roth, and R. Sproat, "Emotions from text: machine learning for text-based emotion prediction," in *Proc. Conf. Hum. Lang. Technol. Empir. Methods Nat. Lang. Process. - HLT '05*, no. October, 2005, pp. 579–586.

[38] K. Mulcrone, "Detecting Emotion in Text," in *UMM CSci Sr. Semin. Conf.*, 2012.

[39] R. A. Calvo and S. M. Kim, "Emotions in Text: Dimensional and Categorical Models," Tech. Rep. 3, 2013.

[40] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *2004 IEEE Int. Conf. Acoust. Speech, Signal Process.*, vol. 1. IEEE, 2004, pp. 577–580.

[41] B. Schuller, R. Müller, M. Lang, and G. Rigoll, "Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features within Ensembles," in *Ninth Eur. Conf. Speech Commun. Technol.*, 2005.

[42] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues," *J. Multimodal User Interfaces*, vol. 3, no. 1, pp. 7–19, mar 2010.

[43] W. Ye and X. Fan, "Bimodal Emotion Recognition from Speech and Text," *Int. J. Adv. Comput. Sci. Appl.*, vol. 5, no. 2, 2014.

[44] L. Tian, J. Moore, and C. Lai, "Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features," in *2016 IEEE Spok. Lang. Technol. Work.* IEEE, dec 2016, pp. 565–572.

[45] D. Grandjean, D. Sander, and K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Conscious. Cogn.*, vol. 17, no. 2, pp. 484–495, 2008.

[46] P. Ekman, "An Argument for Basic Emotions," *Cogn. Emot.*, vol. 6, no. 3-4, pp. 169–200, 1992.

[47] R. Plutchik and H. Kellerman, *Emotion, theory, research, and experience.* Academic press, 1980.

[48] R. E. Jack, W. Sun, I. Delis, O. G. B. Garrod, and P. G. Schyns, "Four not six: Revealing culturally common facial expressions of emotion," *J. Exp. Psychol. Gen.*, vol. 145, no. 6, pp. 708–730, 2016.

[49] P. Ekman, "Basic Emotions," in *Handb. Cogn. Emot.*, 2005.

[50] A. S. Cowen and D. Keltner, "Self-report captures 27 distinct categories of emotion bridged by continuous gradients," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, no. 38, pp. E7900–E7909, 2017.

[51] H. Gunes and M. Pantic, "Automatic, Dimensional and Continuous Emotion Recognition," *Int. J. Synth. Emot.*, vol. 1, no. 1, pp. 68–99, apr 2010.

[52] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980. [Online]. Available: http://content.apa.org/journals/psp/39/6/1161

[53] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, C. Phoebe, J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The World of Emotions Is Not Two-Dimensional," *Psychol. Sci.*, vol. 18, no. 12, pp. 1050–1057, 2017.

[54] B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic, "AVEC 2012 - The continuous audio/visual emotion challenge," in *ICMI'12 - Proc. ACM Int. Conf. Multimodal Interact.*, 2012, pp. 449–456.

[55] K. R. Scherer, "What are emotions? And how can they be measured?" *Soc. Sci. Inf.*, vol. 44, no. 4, pp. 695–729, 2005. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0539018405058216

[56] P. Mairano, E. Zovato, and V. Quinci, "Do sentiment analysis scores correlate with acoustic features of emotional speech?" in *AISV Conf.*, 2019.

[57] R. W. Frick, "Communicating Emotion: The Role of Prosodic Features," *Psychol. Bull.*, vol. 97, no. 3, pp. 412–429, 1985.

[58] S. Mozziconacci, "Prosody and Emotions," *Speech Prosody 2002*, pp. 1–9, 2002.

[59] E. Liebenthal, D. A. Silbersweig, E. Stern, J. B. Fritz, D. Zhang, E. Liebenthal, D. A. Silbersweig, and E. Stern, "The Language, Tone and Prosody of Emotions: Neural Substrates and Dynamics of Spoken-Word Emotion Perception," *Front. Neurosci. — www.frontiersin.org*, vol. 10, no. NOV, p. 506, 2016.

[60] C. M. Lee, S. S. Narayanan, L. Angeles, and R. Pieraccini, "Combining Acoustic and Language Information for Emotion Recognition," *Icslp 2002*, vol. 2002, pp. 6–9, 2002.

[61] F. Metze, T. Polzehl, and M. Wagner, "Fusion of Acoustic and Linguistic Speech Features for Emotion Detection," in *Proc. Int. Conf. Semant. Comput. (ICSC).*, Berkeley, CA, 2009.

[62] S. Tripathi, A. Kumar, A. Ramesh, C. Singh, and P. Yenigalla, "Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions," in *Int. Conf. Comput. Linguist. Intell. Text Process.*, 2019.

[63] Z. Aldeneh, S. Khorram, D. Dimitriadis, and E. M. Provost, "Pooling acoustic and lexical features for the prediction of valence," in *ICMI 2017 - Proc. 19th ACM Int. Conf. Multimodal Interact.* ACM, 2017, pp. 68–72.

[64] B. Zhang, S. Khorram, and E. M. Provost, "Exploiting Acoustic and Lexical Properties of Phonemes to Recognize Valence from Speech," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2019-May. Institute of Electrical and Electronics Engineers Inc., may 2019, pp. 5871–5875.

[65] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram & phoneme embedding," in *Interspeech 2018*, 2018, pp. 3688–3692.

[66] Z. J. Chuang and C.-h. Wu, "Multi-modal emotion recognition from speech and text," *J. Comput. Linguist. Chinese Lang. Process.*, vol. 9, no. 2, pp. 45–62, 2004. [Online]. Available: http://www.aclweb.org/anthology/O/O04/O04-3004.pdf

[67] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825—-2830, 2011.

[68] C. Whissell, "Using the Revised Dictionary of Affect in Language to Quantify the Emotional Undertones of Samples of Natural Language," *Psychol. Rep.*, vol. 105, no. 2, pp. 509–521, 2009. [Online]. Available: http://journals.sagepub.com/doi/10.2466/PR0.105.2.509-521

[69] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas," *Behav. Res. Methods*, vol. 45, no. 4, pp. 1191–1207, 2013. [Online]. Available: http://link.springer.com/10.3758/s13428-012-0314-x

[70] C. J. Hutto and E. E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in *Eighth Int. Conf. Weblogs Soc. Media (ICWSM-14).*, 2014. [Online]. Available: http://sentic.net/

[71] S. M. Mohammad, "Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words," *Proc. 56th Annu. Meet. ofthe Assoc. Comput. Linguist.*, vol. 0, pp. 1–11, 2018.

[72] G. E. Hinton, "Connectionist learning procedures," *Artif. Intell.*, vol. 40, no. 1-3, pp. 185–234, 1989.

[73] D. Griol, J. M. Molina, and Z. Callejas, "Combining speech-based and linguistic classifiers to recognize emotion in user spoken utterances," *Neurocomputing*, vol. 326-327, pp. 132–140, jan 2019.

[74] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for Simplicity: The All Convolutional Net," in *3rd Int. Conf. Learn. Represent. ICLR 2015 - Work. Track Proc.*, dec 2014, pp. 1–14. [Online]. Available: http://arxiv.org/abs/1412.6806

[75] K.-Y. Huang, C.-H. Wu, Q.-B. Hong, M.-H. Su, and Y.-R. Zeng, "Speech Emotion Recognition using Convolutional Neural Network with Audio Word-based Embedding," in *2018 11th Int. Symp. Chinese Spok. Lang. Process.* IEEE, nov 2018, pp. 265–269. [Online]. Available: https://ieeexplore.ieee.org/document/8706610/

[76] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2018-Septe, no. September, pp. 247–251, 2018.

[77] J. Sebastian and P. Pierucci, "Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts," *Interspeech*, pp. 51–55, 2019. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-3201

[78] L. Cai, Y. Hu, J. Dong, and S. Zhou, "Audio-Textual Emotion Recognition Based on Improved Neural Networks," *Math. Probl. Eng.*, vol. 2019, 2019.

[79] L. Tian, J. D. Moore, and C. Lai, "Recognizing emotions in dialogues with acoustic and lexical features," in *2015 Int. Conf. Affect. Comput. Intell. Interact.* IEEE, sep 2015, pp. 737–742. [Online]. Available: http://ieeexplore.ieee.org/document/7344651/

[80] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Inf. Fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[81] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowledge-Based Syst.*, vol. 161, pp. 124–133, dec 2018. [Online]. Available: https://doi.org/10.1016/j.knosys.2018.07.041https://linkinghub.elsevier.com/retrieve/pii/S0950705118303897

[82] S. G. Karadogan and J. Larsen, "Combining semantic and acoustic features for valence and arousal recognition in speech," in *2012 3rd Int. Work. Cogn. Inf. Process.* IEEE, may 2012, pp. 1–6.

[83] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *2015 IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2015-Augus. IEEE, apr 2015, pp. 4749–4753.

[84] S. Yoon, S. Byun, and K. Jung, "Multimodal Speech Emotion Recognition Using Audio and Text," *Spok. Lang. Technol. Work.*, pp. 112–118, oct 2018. [Online]. Available: http://arxiv.org/abs/1810.04635

[85] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, and N. Amir, "Whodunnit - Searching for the most important feature types signalling emotion-related user states in speech," *Comput. Speech Lang.*, vol. 25, no. 1, pp. 4–28, 2011.

[86] D. Charles, E. Paul, and P. Phillip, "The expression of the emotions in man and animals," *Electron. Text Center, Univ. Virginia Libr.*, 1872.

[87] J. A. Russel, "A Circumplex Model of Affect," pp. 1161–1178, 1980. [Online]. Available: https://www2.bc.edu/james-russell/publications/Russell1980.pdf

[88] P. B. Denes and E. Pinson, *The speech chain.* Macmillan, 1993.

[89] R. Elbarougy, "Speech Emotion Recognition based on Voiced Emotion Unit," *Int. J. Comput. Appl.*, vol. 178, no. 47, pp. 22–28, 2019.

[90] G. Aguilar, V. Rozgic, W. Wang, and C. Wang, "Multimodal and multi-view models for emotion recognition," *ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf.*, pp. 991–1002, 2020.

[91] L. Tian, C. Lai, and J. Moore, "Recognizing emotions in dialogues with disfluencies and non-verbal vocalisations," in *Proc. ofthe 4th Interdiscip. Work. Laugh. Other Non-verbal Vocalisations Speech,*, 2015, pp. 39–41.

[92] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," *Neural Networks*, vol. 92, pp. 60–68, 2017. [Online]. Available: http://dx.doi.org/10.1016/j.neunet.2017.02.013

[93] M. Ephratt, "The functions of silence," *J. Pragmat.*, vol. 40, no. 11, pp. 1909–1938, 2008.

[94] E. Tisljár-Szabó and C. Pléh, "Ascribing emotions depending on pause length in native and foreign language speech," *Speech Commun.*, vol. 56, no. 1, pp. 35–48, 2014.

[95] K. Sridhar, S. Parthasarathy, and C. Busso, "Role of Regularization in the Prediction of Valence from Speech," in *Interspeech 2018*. ISCA: ISCA, sep 2018, pp. 941–945.

[96] R. M. Davison, "An Action Research Perspective of Group Support Systems: How to Improve Meetings in Hong Kong," Ph.D. dissertation, City University of Hong Kong, 1998. [Online]. Available: http://www.is.cityu.edu.hk/staff/isrobert/phd/phd.htm

[97] M. Frické, "The knowledge pyramid: A critique of the DIKW hierarchy," *J. Inf. Sci.*, vol. 35, no. 2, pp. 131–142, 2009.

[98] A. Badia, "Data, Information, Knowledge: An Information Science Analysis," *J. Am. Soc. Inf. Sci. Technol.*, vol. 64, no. July, pp. 1852–1863, 2013. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/asi.22883/abstract

[99] C. Zins, "Conceptual Approaches for Defining Data, Information, and Knowledge," *J. Am. Soc. Inf. Sci. Technol.*, vol. 64, no. July, pp. 1852–1863, 2013. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1002/asi.22883/abstract

[100] C. W. Choo, *The Knowing Organization: How Organizations Use Information to Construct Meaning, Create Knowledge, and Make Decisions*, 2007.

[101] J. Rowley, "The wisdom hierarchy: Representations of the DIKW hierarchy," *J. Inf. Sci.*, vol. 33, no. 2, pp. 163–180, 2007.

[102] L. C. Nygaard and J. S. Queen, "Communicating emotion: Linking affective prosody and word meaning," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 34, no. 4, pp. 1017–1030, 2008.

[103] C. Berckmoes and G. Vingerhoets, "Neural foundations of emotional speech processing," pp. 182–185, 2004.

[104] D. Matsumoto and B. Willingham, "Spontaneous Facial Expressions of Emotion of Congenitally and Noncongenitally Blind Individuals," *J. Pers. Soc. Psychol.*, 2009.

[105] L. Pepino, P. Riera, L. Ferrer, and A. Gravano, "Fusion Approaches for Emotion Recognition from Speech Using Acoustic and Text-Based Features," in *ICASSP 2020 - 2020 IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 6484–6488.

[106] S. Planet and I. Iriondo, "Comparative Study on Feature Selection and Fusion Schemes for Emotion Recognition from Speech," *Int. J. Interact. Multimed. Artif. Intell.*, vol. 1, no. 6, p. 44, 2012.

[107] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "AV+EC 2015 - The first affect recognition challenge bridging across audio, video, and physiological data," in *AVEC 2015 - Proc. 5th Int. Work. Audio/Visual Emot. Challenge, co-Located with MM 2015*. Association for Computing Machinery, Inc, oct 2015, pp. 3–8.

[108] P. Herrera, X. Serra, and G. Peeters, "Audio Descriptors and Descriptor Schemes in the Context of MPEG-7," in *Int. Comput. Music Conf.*, 1999, pp. 581–584. [Online]. Available: http://mtg.upf.edu/files/publications/icmc99-perfe.pdf

[109] A. Mohamed, "Deep Neural Network acoustic models for ASR," Ph.D. dissertation, University of Toronto, 2014.

[110] H. M. Fayek, "Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between," 2016. [Online]. Available: https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html

[111] M. R. Morales and R. Levita, "Speech Vs . Text : A Comparative Analysis Of Features For Depression Detection Systems," in *2016 IEEE Spok. Lang. Technol. Work.*, 2016, pp. 136–143.

[112] G. S. Tomas, "Speech Emotion Recognition Using Convolutional Neural Networks," Ph.D. dissertation, 2019.

[113] M. Schmitt and B. Schuller, "Deep Recurrent Neural Networks for Emotion Recognition in Speech," in *DAGA*, 2018, pp. 1537–1540.

[114] F. Chollet and Others, "Keras," https://keras.io, 2015.

[115] E. Campione and J. Véronis, "A large-scale multilingual study of pause duration," in *Speech Prosody 2002. Proc. the1st Int. Conf. Speech Prosody*, 2002, pp. 199–202.

[116] J. D. Moore, L. Tian, and C. Lai, "Word-level emotion recognition using high-level features," in *Int. Conf. Intell. Text Process. Comput. Linguist.* Springer, 2014, pp. 17–31.

[117] B. McFee, V. Lostanlen, M. McVicar, A. Metsai, S. Balke, C. Thomé, C. Raffel, A. Malek, D. Lee, F. Zalkow, K. Lee, O. Nieto, J. Mason, D. Ellis, R. Yamamoto, S. Seyfarth, E. Battenberg, V. Morozov, R. Bittner, K. Choi, J. Moore, Z. Wei, S. Hidaka, Nullmightybofo, P. Friesch, F.-R. Stöter, D. Hereñú, T. Kim, M. Vollrath, and A. Weiss, "librosa/librosa: 0.7.2," jan 2020. [Online]. Available: https://zenodo.org/record/3606573

[118] M. C. Sezgin, B. Gunsel, and G. K. Kurt, "Perceptual audio features for emotion detection," *EURASIP J. Audio, Speech, Music Process.*, vol. 2012, no. 1, p. 16, 2012. [Online]. Available: https://asmp-eurasipjournals.springeropen.com/articles/10.1186/1687-4722-2012-16

[119] A. C. Ian Goodfellow, Yoshua Bengio, *Deep Learning Book*, 2015.

[120] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Lr. 2018 - 11th Int. Conf. Lang. Resour. Eval.*, dec 2019, pp. 52–55.

[121] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv Prepr. arXiv*, oct 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[122] S. Parthasarathy and C. Busso, "Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning," in *Interspeech*, 2017, pp. 1103–1107.

[123] M. Schmitt, N. Cummins, and B. W. Schuller, "Continuous Emotion Recognition in Speech - Do We Need Recurrence?" in *Interspeech 2019*. ISCA: ISCA, sep 2019, pp. 2808–2812.

[124] S. S. Tripathi, H. Beigi, S. S. Tripathi, and H. Beigi, "Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning," apr 2018. [Online]. Available: https://arxiv.org/pdf/1804.05788.pdfhttp://arxiv.org/abs/1804.05788

[125] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech Emotion Recognition Using Multi-hop Attention Mechanism," in *ICASSP 2019 - 2019 IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, may 2019, pp. 2822–2826.

[126] S. Sahu, V. Mitra, N. Seneviratne, and C. Espy-Wilson, "Multi-modal learning for speech emotion recognition: An analysis and comparison of ASR outputs with ground truth transcription," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, 2019, pp. 3302–3306.

[127] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Proc. 7th Annu. Work. Audio/Visual Emot. Chall.* ACM, 2017, pp. 19–26.

[128] Mozilla, "Project DeepSpeech: A TensorFlow implementation of Baidu's DeepSpeech architecture, https://github.com/mozilla/DeepSpeech," 2019. [Online]. Available: https://github.com/mozilla/DeepSpeech

[129] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep Speech: Scaling up end-to-end speech recognition," pp. 1–12, 2014. [Online]. Available: http://arxiv.org/abs/1412.5567

[130] G. Lemaitre, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," *J. Mach. Learn. Res.*, vol. 18, pp. 1–5, sep 2016.

[131] F. Eyben, M. Wöllmer, and B. Schuller, *Opensmile*. New York, New York, USA: ACM Press, 2016, no. November.

[132] M. AbdelWahab and C. Busso, "Domain Adversarial for Acoustic Emotion Recognition," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 12, pp. 2423–2435, dec 2018.

[133] R. Lotfian and C. Busso, "Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings," *IEEE Trans. Affect. Comput.*, vol. 10, no. 4, pp. 471–483, 2019.

[134] H. Zhao, N. Ye, and R. Wang, "Transferring Age and Gender Attributes for Dimensional Emotion Prediction from Big Speech Data Using Hierarchical Deep Learning," in *2018 IEEE 4th Int. Conf. Big Data Secur. Cloud (BigDataSecurity), IEEE Int. Conf. High Perform. Smart Comput. IEEE Int. Conf. Intell. Data Secur.* IEEE, may 2018, pp. 20–24. [Online]. Available: https://ieeexplore.ieee.org/document/8552276/

[135] ——, "Speech emotion recognition based on hierarchical attributes using feature nets," *Int. J. Parallel, Emergent Distrib. Syst.*, vol. 35, no. 3, pp. 354–364, 2019.

[136] S. Parthasarathy and C. Busso, "Semi-Supervised Speech Emotion Recognition With Ladder Networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2697–2709, aug 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9195799/

[137] Z. Yang and J. Hirschberg, "Predicting Arousal and Valence from Waveforms and Spectrograms Using Deep Neural Networks," in *Interspeech 2018*. ISCA: ISCA, sep 2018, pp. 3092–3096. [Online]. Available: http://www.isca-speech.org/archive/Interspeech{_}2018/abstracts/2397.html

[138] A. Bakhshi, A. S. W. Wong, and S. Chalup, "End-To-End Speech Emotion Recognition Based on Time and Frequency Information Using Deep Neural Networks," in *Proceeding Eur. Conf. Artif. Intell.*, 2020.

Publications

Journals

[1] B. T. Atmaja and M. Akagi, "Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning," APSIPA Trans. Signal Inf. Process., Vol. 9, No. May, p. e17, May 2020.

[2] R. Elbarougy, B.T. Atmaja and M. Akagi, "Continuous Audiovisual Emotion Recognition Using Feature Selection and LSTM," Journal of Signal Processing, Vol. 24, No. 6, November 2020.

[3] B.T. Atmaja, and M. Akagi. "Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using SVM," Speech Communication, vol 126, February 2021, pp 9-21. doi:10.1016/j.specom.2020.11.003.

International Conferences

[4] B.T. Atmaja, K. Shirai, and M. Akagi, "Deep Learning-based Categorical and Dimensional Emotion Recognition for Written and Spoken Text," *International Seminar on Science and Technology*, Surabaya, 2019.

[5] B. T. Atmaja, K. Shirai, and M. Akagi, "Speech Emotion Recognition Using Speech Feature and Word Embedding," in 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2019, pp. 519–523.

[6] B. T. Atmaja and M. Akagi, "Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model," in 2019 IEEE International Conference on Signals and Systems (ICSigSys), 2019, pp. 40–44.

[7] B. T. Atmaja and M. Akagi, "Multitask Learning and Multistage Fusion for Dimensional Audiovisual Emotion Recognition," in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 4482–4486.

[8] B. T. Atmaja and M. Akagi, "The Effect of Silence Feature in Dimensional Speech Emotion Recognition," in 10th International Conference on Speech Prosody 2020, 2020, May, pp. 26–30.

[9] B.T. Atmaja and M. Akagi, "Improving Valence Prediction in Dimensional Speech Emotion Recognition Using Linguistic Information, " in 2020 23nd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), pp. 166-171. IEEE, 2020. [**Awarded as best student paper**]

[10] B.T. Atmaja and M. Akagi, "On The Differences Between Song and Speech Emotion Recognition: Effect of Feature Sets, Feature Types, and Classifiers", TENCON 2020 - 2020 IEEE Region 10 Conference (TENCON), Osaka, Japan, 2020.

[11] B.T. Atmaja, Y. Hamada and M. Akagi, "Predicting Valence and Arousal by Aggregating Acoustic Features for Acoustic-Linguistic Information Fusion" TENCON 2020 - 2020 IEEE Region 10 Conference (TENCON), Osaka, Japan, 2020.

[12] B.T. Atmaja and M. Akagi, "Deep Multilayer Perceptrons for Dimensional Speech Emotion Recognition," in 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Auckland, New Zealand, 2020.

[13] B.T. Atmaja, M. Akagi. "Evaluation of Error and Correlation-based Loss Functions For Multitask Learning Dimensional Speech Emotion Recognition," International Conference on Acoustic and Vibration, Bali, Indonesia, 2020. [**Awarded as best student paper and presentation**]

Domestic Conferences

[14] R. Elbarougy, B.T. Atmaja, M. Akagi, "Continuous Tracking of Emotional State from Speech Based on Emotion Unit," *In Proceeding ASJ Autumn Meeting*, 2018.

[15] B.T. Atmaja, A.N.F. Fandy, D. Arifianto, M. Akagi, "Speech recognition on Indonesian language by using time delay neural network," *In Proceeding ASJ Spring Meeting*, 2019, pp. 1291–1294.

[16] B.T. Atmaja, R. Elbarougy, M. Akagi, "RNN-based dimensional speech emotion recognition," *In Proceeding ASJ Autumn Meeting*, 2019, pp. 743–744.

[17] B.T. Atmaja, M. Akagi, "Dimensional Speech Emotion Recognition from Acoustic and Text Features Using Multitask Learning," *In Proceeding ASJ Spring Meeting*, 2020, pp. 1003–1004.