A Three-Layer Model Based Estimation of Emotions in Multilingual Speech

Xingfeng LI

Japan Advanced Institute of Science and Technology

Doctoral Dissertation

A Three-Layer Model Based Estimation of Emotions in Multilingual Speech

Xingfeng LI

Supervisor: Professor Masato AKAGI

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

Information Science
June 2019

Abstract

The goal of this research is to design a computational model for recognizing emotional states using speech across multiple languages. Recognition of emotion in speech in general relied upon specific training data, and a different speaker and/or language may present significant challenges. Traditions on multilingual speech emotion recognition have been done to be extensively focused on reducing the differences between different speakers and/or languages by using different approaches, like normalization strategies, domain adaption methods, transfer learning algorithms, and so on. However, those previous studies usually required pre-knowledge on the target data that is challenging in practice; and most specific, they do not consider and clarify the effective features in common for different data sets. On the other hand, cross-lingual studies have revealed that human speech perception of emotion shares universal rules among people who speak different languages and provided a set of common emotions independent of languages, like anger, happy, sad, fear, disgust and surprise. By contrast, limited research focus has been gained to study the process of human emotion perception for multilingual speech emotion recognition. Thus, it motivated us to answer exactly what commonalities were in the speech perception of emotion across multiple languages.

In this regard, a three-layer model conducted with acoustic features, semantic primitives, and emotion dimensions was studied, taking inspiration from the fact that there exist multiple layers in a process of human emotion recognition. To achieve the research goal, 215 statistical acoustic features derived from prosodic and spectral domains, and seventeen semantic primitives were first explored. The shared appropriate parameter sets of acoustic features and semantic primitives were then determined by a feature selection algorithm. The implemented three-layer model suggested 22 acoustic features and four semantic primitives to be universal in the process of emotion perception in multilingual speech.

Based on studies implemented in this research it was demonstrated that the combination of acoustic features of prosodic and spectral domains is beneficial for multilingual speech emotion recognition; the proposed feature selection method can handle well in determining associations in a three-layer model, providing appropriate features across languages. The computational model for recognizing emotional states using speech that was implemented in this research was speaker and language independent where it does not require any pre-knowledge on the target speaker and/or language. This three-layer model yielded a comparable performance by multilingual and monolingual speech emotion recognition across three different languages; in addition, the cross-lingual tasks interestingly provided comparable results to those obtained by human evaluations in cross-lingual studies.

The outcome of this research is potential to contributing the recognition of emotion from speech irrespective of speakers and languages in many areas, such as affective speech-to-speech translation, call centre application, and healthcare.

Keywords: Multilingual perception of emotion, three-layer model, emotion space, modulation spectral feature, prosodic feature

Acknowledgments

First and foremost, I would like to express my sincerely appreciation to my supervisor, Professor Masato Akagi, for his tremendous guidance and generous support. It is a great honour to join the acoustic information science laboratory and furnishe my Master and PhD study with Professor Akagi. Beyond his great guidance, motivation, and kind care, my research work would never be goes well. This is a life-long wonderful memory for me to study at Japan Advanced Institute of Science and Technology, and to learn from the greater professors here.

I would like further to say gratefully thanks to my vice supervisor, Professor Masashi Unoki, for his great suggestions and comments in my research. I really learn quite much from Professor Unoki, especially the good manners in doing research. He is a really kind professor and does me a great favor on my research study and presentations.

Furthermore, my great appreciation goes to Professor Jianwu Dang, for his kind care of my study over master and doctor programs. Without his support, I may never have a chance to meet so nice people, and enjoy a research study here.

I would also like to express my sincere thanks to my dear friends Y. Li, Y. Xue, L. Li, H. Cai, J. Huang, K. Ren, X. Zhu and Z. Li and many other friends we all enjoy the life in JAIST.

I would like to acknowledge Researcher Rieko Kubo, for her great suggestions and comments in my research. She is quite nice, and warm-hearted.

Most importantly, I greatly appreciated my parents Fengzhen Wang and Baozhu Li. I dedicate this thesis to my parents and my friends, for their unfailing kindness, support and patience.

Table of Contents

List of Figures

List of Tables

Acronym and Abbreviation

PIDs	Personal Intelligence Devices
SER	Speech Emotion Recognition
MFCC	Mel Frequency Cepstral Coefficients
SFFS	Sequential Floating Forward Selection
LOSO	Leave-One-Speaker-Out
LOCO	Leave-One-Corpus-Out
NN	No Normalization
SN	Speaker Normalization
GMM	Guassian Mixture Model
SVM	Support Vector Machine
KNN	K-Nearest-Neighbour
HMM	Hidden Markov Model
DNN	Deep Neural Network
LSTM	Long Short-Term Memory
MSF	Modulation Spectral Features
PCC	Pearson Correlation Coefficient
GUI	Graphical User Interface
ANFIS	Adaptive Neuro Fuzzy Inference System
LMT	Logistic Model Trees
ERB_N	Equivalent Rectangular Bandwidth
MSF	Modulation Spectral Feature

Chapter 1

Introduction

1.1 Background

Speech is a great expression tool for delivering spoken messages of communication of thoughts in our daily lives. It carries two-channel information: one is the linguistic channel which relates to the spoken words; and the other is the non-linguistic channel which relates to other information like language, age, gender, and emotion, going a way that beyond just the spoken content. Presence of non-linguistic information like emotional state performs a natural speech in a conversation. Besides the words delivered by a spoken utterance, the way in which those words were expressed, carrying a vital message in connection with the intention of humans: for example, it is often the case that human express their loves generally in a sincere and hearty voice but depressions in a heavy and hopeless voice.

Over the last decades, the research of speech has been appearing to be an important research topic in the area of human-computer interaction, serving as a vital ingredient of 'artificial intelligence' of machines to understand human speech in the natural communication. There is a large body of personal intelligent devices (PIDs), such as Apples's Siri[1], Google's Google Now[2] and Microsoft's Cortana[3] have been done to be able to recognize and perform human speech effectively. Despite the fact that PIDs have the capability to interpret what we are saying advantageously, whereas they have only scratched the surface of what possible to understand human speech. This is due to the

[1]Apple's Siri. https://www.apple.com/ios/siri/

[2]Google's Google Now. https://www.google.com/landing/now/

[3]Microsoft's Cortana. https://www.windowsphone.com/en-us/features-8-1

fact that those PIDs were clueless about the emotional states in speech which are highly significant of communication concerning our feelings [8, 9].

Within this context, identifying an emotional state from human voices based on speech emotion recognition (SER) has been increasingly turned into a principal focus within affective computing research [10] for interpreting the semantics of a spoken utterance. SER enables PIDs with sufficient intelligence to listen to human speech exactly relative to how the speech was expressed besides what was spoken. SER is promising for many potential applications, one of which applies to a call-center service. The system can furnish a user-friendliness reply to a customer upon identifying emotional states from his or her voice [11]. Likewise, fatigue can be detected from a driver's voice by a car-board system and the driver can be alerted to ensure a secure driving [12]. Furthermore, it is potential for giving feedback in plays and human-robot interactions [13,14]. Other human-computer interactions such as web-movies [15], health care systems [16] and "Affective Mirror" [10] can also be enriched by the recognition of emotional states from a speaker's voice.

Most recent, the Interspeech Computational Paralinguistics ChallengE[4] (ComParE) has introduced groups of open challenges in the field of computational paralinguistics dealing with states and traits of speakers as embedded in their speech signal's properties since 2009 and has signified promising accomplishments. In spite of the substantial progress made in this area, SER still suffers plenty of not yet covered, but notably related challenges. This research work presented one of which on multilingual SER, specifically with a focus on designing a computational model for recognizing emotional states using speech across multiple languages.

1.2 Research Motivation

Speech emotion perception is universal between people who own different mother tongues, even in the event that they do not understand the verbal content being delivered [17, 18]. The evidence is mounting that there primary emotional states of joy, anger, sadness, fear, surprise, and disgust exist independent of languages and cultures [10, 19, 20]. In this light, studies on SER has lately been gaining more interests in recognizing speech emotional states in multilingual scenarios so as to provide a natural and human-equivalent emotion

[4]ComParE Series. http://www.compare.openaudio.eu/

recognizer for real-life situations [21–25].

Most traditions that have been implemented for SER in the last decade focused on classifying emotional states in different monolingual scenarios [26–28], achieving high performance in every single language. However, it has to be noted that they were performed on an assumption that the training and testing data were often collected under the same condition, leading to the optimal feature sets and classifiers were highly specific to each language. And even more restricted to the same language drawn from different scenarios, for as an example of speakers, age, gender, and noise level. Changing a source corpus to be recognized requires re-selecting the optimal acoustic features and re-training an SER system, making multilingual SER tasks very challenging.

In order to overcome the challenge as mentioned above, a number of researchers have taken into consideration the case in multilingual SER by reducing the discrepancies between source and target languages. In [23], Schuller et al. introduced to normalize speaker information, corpus information as well as the combined information to get an improvement over cross-lingual SER. Deng et al. in 2014 [29], performed an autoencoder based domain adaptation technique to handle in cross-corpus emotion recognition. Most recent, researchers also conducted other proposals for multilingual SER, like transfer learning algorithms [30]. Nonetheless, all these previous considerations required a priori knowledge on the target data of speakers and languages, that is a problem and usually challenging to get in practice. Despite the success of the reduction of differences among multiple languages, most notably, it is still a failure to answer exactly what commonalities were in the speech perception of emotion independent of languages.

Han and Akagi (2015) have recently conducted a cross-culture study [31], in the valence (positive versus negative emotional states) and arousal (calm versus aroused emotional states) space, which has attempted to study the commonalities and divergences in the speech perception of emotion among languages. Thirty subjects from three countries in Japan, China, and Vietnam were recruited to take part in rating three emotional corpora of Fujitsu dataset, Berlin EmoDB, and CASIA corpus relative to valence and arousal in countries of ten subjects each. The centroids and covariances for each emotional category varied for three listener groups were displayed in Figure 1.1. Most interestingly, they found that the directions and distances from a neutral voice to that of another emotional

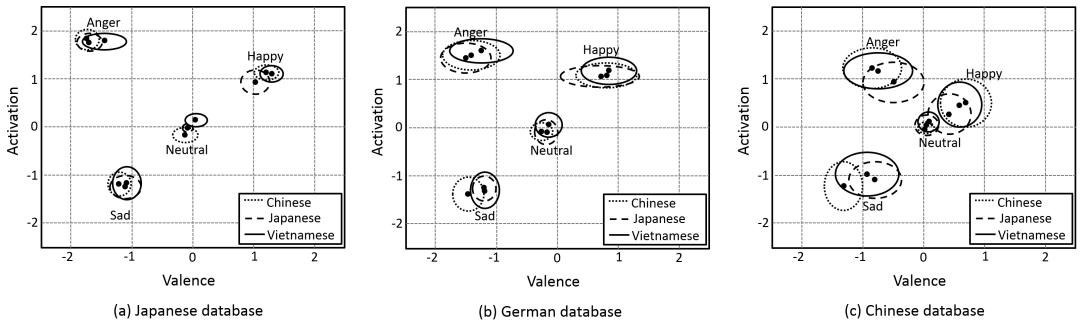


Figure 1.1: Emotional classes in the valence and activation space. For each of the four emotional classes, three emotion subspaces were shown as determined from the average ratings of 10 subjects in each group.

state are highly similar between human judge groups irrespective to languages; besides, the neutral positions are different.

As an alternative to others mostly studied various techniques to reduce the differences among languages, this exciting finding of commonalities in the human speech perception of emotion motivated us to think of adapting the direction and distance as standard features to deal with the multilingual SER tasks. The insight here is that humans share universal rules in the speech perception of emotion independent of languages and cultures. And indeed, it could be a vital and potential method to examine and define the process of human emotion perception for approaching multilingual SER tasks.

1.3 Research Concept

In line with these findings in previous literature relative to the multilingual SER, the process of human speech emotion perception seems to be an intriguing topic to be explored; on the grounds that this process is universal between humans even they own different mother-tongues, and is potential to advance the multilingual SER tasks.

The purpose of this research, therefore, was to design a multilingual emotional computation model by forming the process in human speech perception of emotion. Once achieved, this model has the capability to recognize emotional states in multilingual speech and even in a different language that is not trained so as to enable the recognizer with human-equivalent performance, outperforming others on SER.

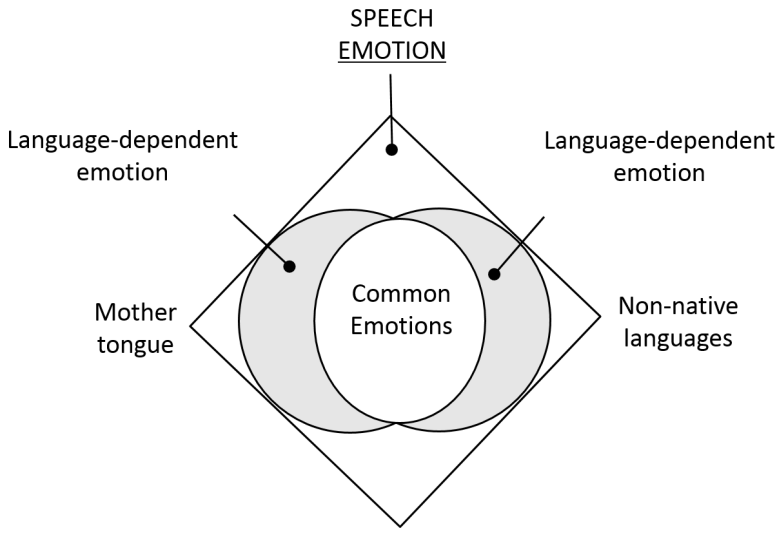


Figure 1.2: Human speech expression and perception of emotion between the mother tongue and non-native languages.

The design of a computational model for multilingual SER is three-fold, namely the determination on (1) emotion theory for characterizing universal speech emotion among multiple languages; (2) learning scheme for recognizing emotional states in the process of speech perception of emotion; (3) relevant features derived from speech that can generalize well across languages.

1.3.1 Emotion Theory

The first issue to be studied is how to determine and characterize universal emotion in multilingual scenarios. In order to detect the discrete label of an emotional state along with its changing degree, this study suggests a combined emotion theory on the basis of categorical and dimensional approaches.

Figure 1.2 demonstrates a model of how human listeners express and perceive the speech emotion culturally and universally: (1) each language/culture has its own distinct speech emotions that may go beyond some other particular emotions in another language/culture; (2) and there may also universal emotions exist in all languages and thus provides common types of emotion facilitating cross-culture recognition of speech emotions. The significant challenge that remains to be explored is which and how different emotional states can be defined universally independent of languages.

Table 1.1: *A Selection of List of "Basic" Emotions [6]*

Reference	Fundamental emotion	Basis for inclusion
Arnold (1960)	Anger, aversion, courage, dejection, desire,despair, fear, hate, hope, love, sadness	Relation to actioin tendencies
Ekman, Friesen, & Ellsworth (1982)	Anger, disgust, fear, joy, sadness, surprise	Universal facial expression
Frijda (personal communication, September 8, 1986)	Desire, happiness, interest, surprise, wonder, sorrow	Forms of action readiness
Gray (1982)	Rage and terror, anxiety, joy	Hardwired
Izard (1971)	Anger, contempt, disgust, distress, fear, guilt, interest,joy, shame, surprise	Hardwired
James (1884)	Fear, grief, love, rage	Bodily involvement
McDougall (1926)	Anger, disgust, elation, fear, subjection,tender-emotion, wonder	Relation to instincts
Mowrer (1960)	Pain, pleasure	Unclearned emotional states
Oatley & Johnson-Laird (1987)	Anger, disgust, anxiety, happiness, sadness	Do not resuire propositional content
Panksepp (1982)	Expectancy, fear, rage, panic	Hardwired
Plutchik (1980)	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise	Relation to adaptive biological processes
Tomkins (1984)	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise	Density of neural
Waston (1930)	Fear, love, rage	Hardwired
Weiner & Graham (1984)	Happiness, sadness	Attribution independent

Note. Not all the theorists represented in this table are equally strong advocates of the idea of basic emotions. For some it is crucial notion (e.g., Izard, 1977; Panksepp,1982; Plutchik, 1980; Tomkins,1984), whereas for others it is of peripheral interest only, and their discussions of basic emotiona re hedged (e.g., Mowrer, 1960; Weiber & Graham, 1984).

One of the theories that have generally driven the current research is the categorical emotion theory. Essential of this theory is the presence of primary emotions, asserting that any emotional state can be derived from a small number of basic emotions [10, 19, 32]. Table 1.1 lists some relevant categorical-based theories to emotion as summarized by Ortony and Turner [6]. Even the set of basic emotions seems to be different from one research to another; most notably, researchers have agreed in the 'big six theory,' suggesting the existence of basic emotions by frequently examining emotional states of happiness, anger, sadness, fear and disgust [19, 33]. These 'big six emotions' were found to be universal independent of languages/cultures. [34–38].

The most important benefit of the categorical emotion theory is to render and provide a simple and straightforward method to describe speech emotion. It could be incorporated easily within a human-machine interaction. The categorical theory seems to be too important an alternative in multilingual SER tasks. Whereas, the emotional state is transited dynamically and gradually in human speech; besides the discrete label, the categorical theory fails to capture the changing degrees within an emotional state that can fluently be communicated in our daily conversation.

Within this context, researchers have recently been paying growing focus to dimensional emotion theory, taking into consideration to capture the changing intensities of emotional states delivered in speech.The dimensional theory provides an attempt corresponding to human perception of emotions to define emotional states as points in a multi-dimensional space. This space was determined by examining the associations between emotions from speech. Human listeners were recruited to use many rating scales of affectionate terms to match the emotional states in the voice stimulus. The correlations between the rating scale next fix the important and primary dimensions using a factor analysis approach. Two dimensions generally spanned the dimensional space suggested by psychological studies, i.e., valence/pleasantness (positive versus negative emotional states) and arousal/activation (relaxed versus aroused emotional states), or three dimensions jointed the dominance/potency (sense of power versus freedom to act).

Figure 1.3 shows a two-dimensional emotion model as introduced by Russell [1], using dimensions of valence and arousal to mark 28 affective markers. One of the

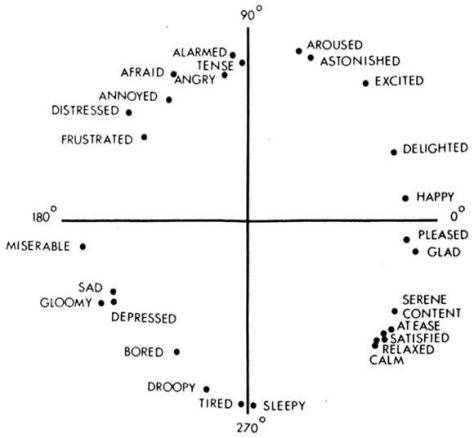


Figure 1.3: Russell's circumplex model [1].

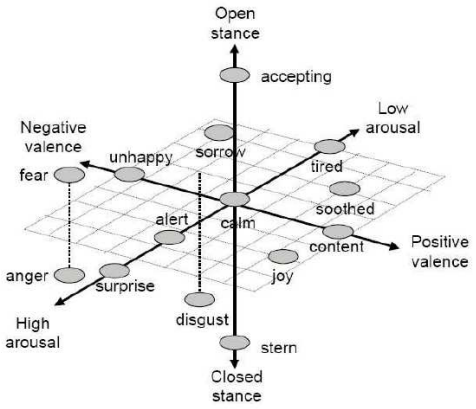


Figure 1.4: The 3-D emotion space.

important advantages of this circumplex model is to provide a direct and persuasive method of defining different emotions concerning their affect assessments (valence) and physiological responses (arousal). This space allows for comparing different emotions on two standard and essential dimensions. Other researchers studied speech emotion in a three-dimensional space whose dimensions are valence, arousal, and dominance was displayed in Figure 1.4.The third dimension is joined to render a more full image of emotion. It succeeds a deficiency as argued that the 2-D model overwhelms significant psychological differentiation and consequently obscures essential aspects of the emotion process. For instance, anger and fear are rated adjacent to each other in the 2-D emotion space; however, they are much varied relative to their implication for organism [39–41]. From both the theoretical and realistic perspectives, more and more studies like to represent emotions using the 2-D or 3-D emotion space.

Following the findings of the cross-culture study in the emotion dimensional space [31], this study suggests a two-fold scheme to estimate emotional states in multilingual speech. It brought together the dimensional and categorical emotion theories: first, estimate emotion dimensions of valence and arousal accurately among languages; second, classify the estimated values in the dimensional space into primary emotional classes. The most important advantage in combining two emotion theories provided sufficient identification for emotional states in the multilingual speech. Besides recognizing the discrete label per emotion, it also advanced the estimation of changing degrees within an emotional state.

1.3.2 Emotion Computational Model

The second issue is an effort to design a computational model that works well for multiple languages. In this regard, this study proposed a three-layer model to explore the universal inference rules of human speech perception in multilingual scenarios.

Many computational models, widely used in machine learning and pattern recognition, have been previously constructed in traditions for acoustic SER tasks, such as Gaussian mixture model [33, 42], support vector machine/regression [9, 43], k-nearest-neighbour [44], hidden Markov model [45] and so on. However, it was found that each model has specific optimal patterns like the types of the kernel, varying significantly with respect to the different situations of language, gender, speaker, and noise [22]. This limitation in traditions, in turn, was hard to generate estimation for multilingual SER tasks.

All those traditions have been done to recognize emotional states directly estimated from acoustic features, treating human perception of emotion as a two-layer process. Whereas, those models notably led to a poorer estimation of valence dimension when compared with that of arousal. This poor estimation might be mainly attributed to the fact that such a framework may not match the complex cognitive processes utilized to judge emotional states as humans do. Researchers also introduced other models for SER in a multi-layer process at the present time, such as the extreme learning machine [26], deep neural network [46], and long short-term memory [47].Despite the success of the multi-layer process in mimicking human perception of emotion, this success, in general, required a massive set of training data that is another problem in data scarcity for multiple languages. Most specifically, many scientists recently argued that deep learning still cannot sufficiently clarify the well fit in mimicking the mechanism of human speech perception of emotion by black-box testing.

Alternatively, Scherer (1978) suggested that the inference of personality from human voice can be interpreted in a modified version of Brunswik's lens model. The brief overview of this lens model can be found in Figure 1.5, which started with the process of encoding, or expression of the speaker's states. Afterward, the author believed that the emotional states of the speaker were externalized by the physiological changes producing distinct distal cues, i.e., acoustic features. The variation of acoustic features to speaker's states were next perceived by a listener and internally named 'proximal percepts.' The

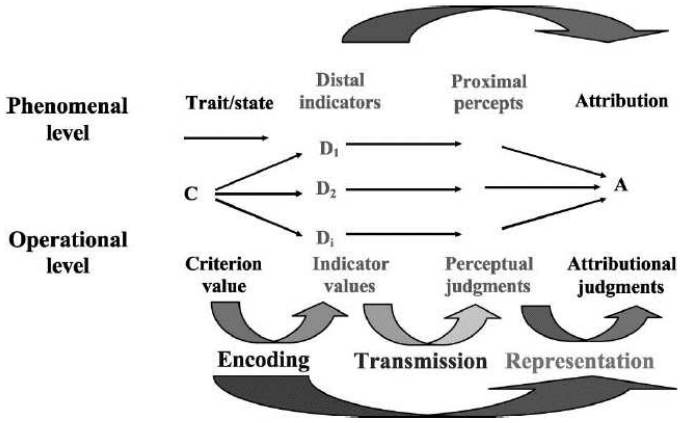


Figure 1.5: A Brunswikian lens model of the vocal communication of emotion [2].

attribution of states can be eventually defined based on the proximal percepts of distal cues. This version of Brunswik's model seems to be too engaging a focus towards consistency of research results in the field of vocal emotional perception and communication. It advances the clarification of what possible to define and operationalize distal cues and proximal percepts in the channel of speech emotion communication.

Following the concept of the Brunswik's lens model, Huang and Akagi (2008) introduced a three-layer model lately for the perception of expressive speech as shown in Figure 1.6. This three-layer model has further reinforced that human perception of speech emotion includes multiple layers, providing some insights into the process of human speech emotion perception. It was found that human subjects judge expressive speech by a small set of perceptions that are expressed by seventeen semantic primitives instead of directly from acoustic features. For example, the low arousal and negative valence speech (such as sadness) in general easily makes an impression on listeners with dark and heavy feelings, but high arousal and positive valence speech (pleasant or happiness) is frequently perceived as bright and well-modulated. Most interestingly, they confirmed that people with different language-tongue have some universal characteristics for the perception of categorical expressive speech, as well as some language-dependent aspects.

Besides, Elbarougy and Akagi (2014) improved a three-layer model in a dimensional approach, by modifying acoustic features in the bottom layer, semantic primitives in the

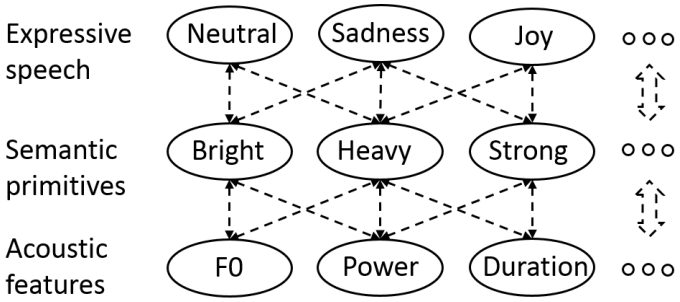


Figure 1.6: Conceptual diagram of the three-layer perceptual model [3].

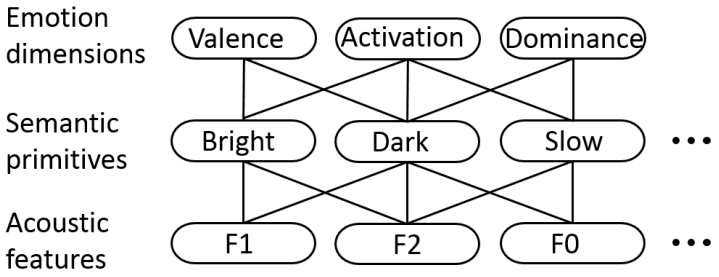


Figure 1.7: The improved three-layer model for human perception [4].

middle layer and emotion dimensions in the top layer (Figure 1.7). Essential to this model was taking into account the human emotion perception processing for improving the accuracies of estimation of emotion dimensions, particularly in the estimation of the valence dimension from which most of the essential efforts suffered.

In line with these findings, most importantly, it demonstrated that the nature of the underlying speech emotion perception mechanism advanced the performance on SER. This study was an effort on the basis of human speech perception of emotion renewing a three-layer model to provide a special insight into universal inference rules from acoustic features to specific emotions among multiple languages.

1.3.3 Emotion Speech Feature

The third issue of determining the relevant features to implement a three-layer emotion computation model is two-fold. The first challenge is a determination on measuring which types of features in a three-layer model relative to acoustic features and semantic primitives. The second challenge is to define the optimal ones from a pre-defined raw set.

Specifically, regarding examining relevant acoustic features to different emotions in multilingual speech, an increasing amount of studies have been previously conducted [33, 48, 49]. The nonlinear Teager-energy-operator-based features, for instance, have also been investigated as a new attempt in the acoustic domain [50, 51]. Unfortunately, the conclusions derived from those works were inconsistent, leading to the best vocal features to be language-dependent [33]. In this regard, other efforts have tried to reduce the differences between the source and target languages by normalization, adaptation, and transferring algorithms, like supervised domain adaptation, maximum a posteriori, etc. [29, 42, 52]. In spite of the promising performance reported, they fail to know what is exactly universal among different languages.

At the same time, many researchers commonly held that prosodic features such as energy, fundamental frequency, and duration often provide much of the emotional information in speech [33, 53–56]. Despite the substantial performance, however, it appeared that there was ambiguousness between the acoustical properties of some emotional states. For instance, anger and joy shared similar characteristics over the fundamental frequency, but they convey different emotional states. As argued by current literature [4, 7, 20, 33], prosodic features formed the profitable feature type for predicting arousal dimension, however, associations of these features to valence dimension have generally been observed to be weak; and yielding a comparatively poor performance in the estimation of valence.

At this point in time, increasing research efforts, alternatively, have been put into studying powerful speech features from the spectral domain. This is because spectral features are strongly correlated with the shape of the vocal tract and the rate of change in articulator movement [20, 57, 58], varying from one emotional state to another [59]. It has also been reported the valence dimension was reflected in the acoustic correlates of spectral cues [60–62].

In line with these findings, this research examined a set of combined acoustic features from both prosodic and spectral domains to refine the estimation accuracies of emotion dimensions in multilingual tasks. Those features involved parameters of duration, voice quality, format, fundamental frequency, power, and spectral modulation features as following:

Duration: Speech rate is one of the fundamental cues to carry a particular emotional trait/state. For instance, a sadness speech can be usually noticed in a low speech rate; In contrast, an exciting voice is often uttered in a higher speech rate. Duration is of importance to produce the prosodic rhythm through words, phrases, and pauses as well.

Voice quality: Many researchers believed that the state of an expressive speech is highly correlated with voice quality [53,63]. Experimental efforts have shown that human listeners can identify emotional state using only voice quality [2,64]. Consequently, voice quality seems to be too important aspect study the relation to speech emotions.

Formants: Formants can represent the natural resonance frequencies of the vocal tract. The commonly used formants are first and second formants and their bandwidths. Recent effort in [65] have found that higher values of F1 in vowel /*a*/ yielding higher values in both valence and arousal, confirming the significance of formants in speech communication of emotion.

Fundamental frequency: The fundamental frequency provides crucial information to speech emotion and is one the most commonly applied features in the prosodic domain [3, 4, 33]. The fundamental frequency is one of the essential features related to the pitch signal, denoting the vibration rate of the vocal folds.

Power: Concerning the physical meaning of power, it principally observed as power in an utterance by a human listener, which is defined through the volume of the air flow of breath sent out by the lungs. Power has been demonstrated to be relevant to the arousal level of emotions [33, 45].

Modulation spectral: Modulation spectral features are on the basis of frequency analysis of the amplitude modulation of multiple acoustic frequency bins, thus catching both spectral and temporal characteristics of the speech signal. These features improved the weakness in capturing temporal behavior information by traditional spectral features such as MFCC that coveys the short-term spectral properties only and trying to gain insight into perceptual-inspired spectral features [7, 62, 66]. It is an intriguing attempt to study spectral features in this domain to distinguish types of speech emotion.

Semantic primitives: As an aside, the raw set of semantic primitives for describing emotional speech derives from the previous study [3], as suggested by Huang and Akagi (2008), involving 17 adjectives, i.e., bright, dark, high, low, strong, weak, calm, unstable,

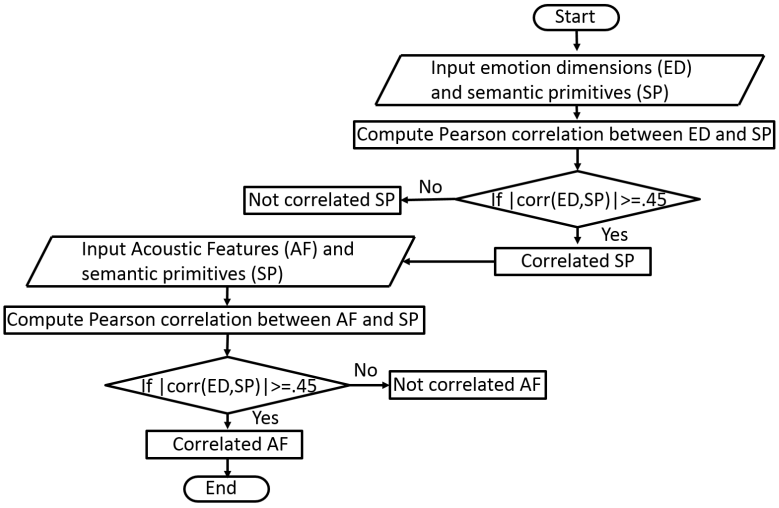


Figure 1.8: Feature selection to implementation of a three-layer model [4].

monotonous, well-modulated, heavy, clear, noisy, quiet, sharp, fast and slow. These semantic primitives were reported to be suited for different languages [4].

On the second uncertainty on determining optimal patterns in a three-layer model, previous effort suggested in [4] adopted a regulation by incorporating the Pearson correlation coefficients (PCC) to determine the relative features in the process of SER by a linear measure as shown in Figure 1.8 . Regarding the implementation of a human-perceptual-based three-layer model consisting of acoustic features, semantic primitives, and emotion dimensions, they first calculated the PCC between emotion dimensions and semantic primitives individually, and the degree of relevance was quantified within the range [-1,1], where -1 is total negative linear correlation, 1 is total positive linear correlation, and 0 is no linear correlation. These correlation coefficients were then ranked on the basis of their absolute scale; those greater than 0.45 were finally selected as relevant semantic primitives for emotion dimensions. The relevant acoustic features were defined following the same scheme by calculating the PCC between each semantic primitive and acoustic features.

The number of correlated acoustic features selected by this procedure seems to be greater compared with most of the related efforts which identifying correlated acoustic features for emotion dimensions directly by correlating the relation between acoustic features and emotion dimensions [67–70]. Despite the substantial advances in the

three-layer model, the limitation of this regulation, however, is considerable due to following reason, i.e., the PCC can naively capture the linear relationship between features and target, but cannot capture correlations that are not linear. Human emotion perception, for instance, is vague, complex, and has multi-processes; it does not suffice always to use linear correlation to capture the association between semantic primitives and emotion dimensions, and acoustic feature and semantic primitives.

Human speech perception of emotion is vague, involving both functional and nonfunctional associations. Beyond the filtered based approach captured linear associations simply, this research adopted a wrapper-based feature selection algorithm, taking into account it can evaluate the selected subset and combined effects of features irrespective of association of linear and nonlinear relations.

To solve this problem, concretely, the assumption in this study is that the emotional traits/states of a speaker are externalized by distinct patterns from a low-level to high-level domain varied in acoustic features, semantic primitives and emotion dimensions.Those emotional traits/states are basically externalized in the form of physiological changes that producing patterns of acoustic features in a low-level; and accompanied by psychological changes that will effect perceptual judgments in a manner to provide patterns of proximal percepts from semantic primitives to emotion dimensions. This work provides an attempt to use the SFFS to define the best features from a raw feature set of acoustic features and semantic primitives.

1.4 Research Purpose

The final goal of this research is to design a computation model for SER, possessing the ability to recognize emotional state across multiple languages, and even for a new target language that was not trained. To this end, three sub-goals were addressed with respect to study the universal process of human speech emotion recognition in a three-layer model. (i) estimating emotional state relative to the discrete labels as well as the changing degrees within spoken utterances; (ii) studying appropriate features that can generalize well across languages to an accurate estimation of emotion dimensions; (iii) presenting a framework to determine relevant features to implement the proposed computation SER model.

1.5 Contribution

On the one hand, this research studied on multilingual SER can incredibly enhance the full development of speech emotion systems in a real-world-context. At present, there is a number of emotional speech corpora; however, they differ from one to another relative to the spoken language, kind of delivered emotional state and labeling method [23]. More than 5000 spoken languages exist around the world, and 389 languages could account for 94% of the world's population[5]. Whereas, even for 389 languages, decidedly fewer resources are available for language and speech processing research. This fact suggests that language and speech research must overcome the obstacles of data scarcity for many languages. This research consequently contributes to investigating similarities between languages which in turn can solve the issue of data sparsity in training emotion models by combining more speech instances from different databases. The equilibrium, fixation, and resemblance among speech and language corpora suggest that it is desirable to study a model in multilingual scenarios and then expect it to be beneficial commonly in different languages in practice.

On the other hand, there are many countries around the world have peoples who speak multiple languages, such as Singapore, South Africa, India, Australia and so on [71–73]. In most multilingual nations and regions, everybody speaks at least one language than his or her mother tongue. Nonetheless, many people can converse fluently in four or five tongues, sometimes using multiple languages in the same conversation, or even in the same sentence. The multilingual SER challenge, as suggested in this research can advance social communication in multilingualism scenarios; yet, that most of the relevant studies on SER are still suffering.

Besides those contributions as mentioned above to the open challenges in the field of speech emotion recognition, the main contribution of this work is to implement an emotion computation model that has the capability to recognize emotional speech states across multiple languages and speakers; and even for a different language and speaker beyond training. More specific contributions can be included as following regarding how to design, implement, and validate a multilingual SER model.

- This study proposed to incorporate the universal knowledge of human emotion

[5]Ethnologue: language of the world. https://www.ethnologue.com/statistics

perception in a 2-D emotion space to estimate emotional speech attributions relative to not only categorical labels but gradual degrees within an emotional state.

◦ This study presented an acoustic model consisting of acoustic features, semantic primitives, and emotion dimensions in a three-layer scheme to approach the mechanism of the process of human multilingual SER rather than a traditional two-layer scheme in a study of recognition of speech emotion from acoustic features directly.

◦ This study explored and clarified the existence of universal acoustic features and semantic primitives from human speech to certain emotional states independent of different languages.

1.6 Organization of dissertation

The rest of this dissertation consists of five chapters and is organized as follows.

Chapter 2 introduced three emotional speech corpora that were used in this research. Cross-lingual emotional evaluations were carried out in this chapter to ensure those corpora were well suited to the study of SER in multilingual scenarios.

Chapter 3 detailed the extraction of acoustic features and human evaluations relative to semantic primitives and emotional dimensions. We firstly extracted 215 acoustic features from two different domains involving prosodic-related features and spectral-related features. Followed by carrying out two listening experiments,i.e., the first one is to evaluate 17 semantic primitives for each spoken utterance per corpus; the second one is to evaluate the emotional dimensions of valence and arousal. The inter-rater agreement was obtained by correlating each listener's evaluations and the averaged evaluations overall human subjects each corpus.

Chapter 4 presented the implementation of the three-layer multilingual SER model from feature selection to emotional dimensions estimation over emotional categories classification. Verification of the proposed system was given from the four aspects: i) emotion theory to multilingual speech emotion; ii) multilingual computation emotion model; iii)relevant features of the acoustic domain and perceptual domain relative to semantic primitives; iv) methodology to implement the three-layer model for mimicking human speech perception of emotion.

Chapter 5 discussed the evaluations of the proposed multilingual SER from the following three aspects: i) effectiveness to speaker variability; ii) effectiveness to language variability; iii) superior to the traditions on SER.

Chapter 6 concluded this research and emphasized its contributions in the research field of speech emotion recognition. Besides, future works on a full development to speech emotion recognition were summarized.

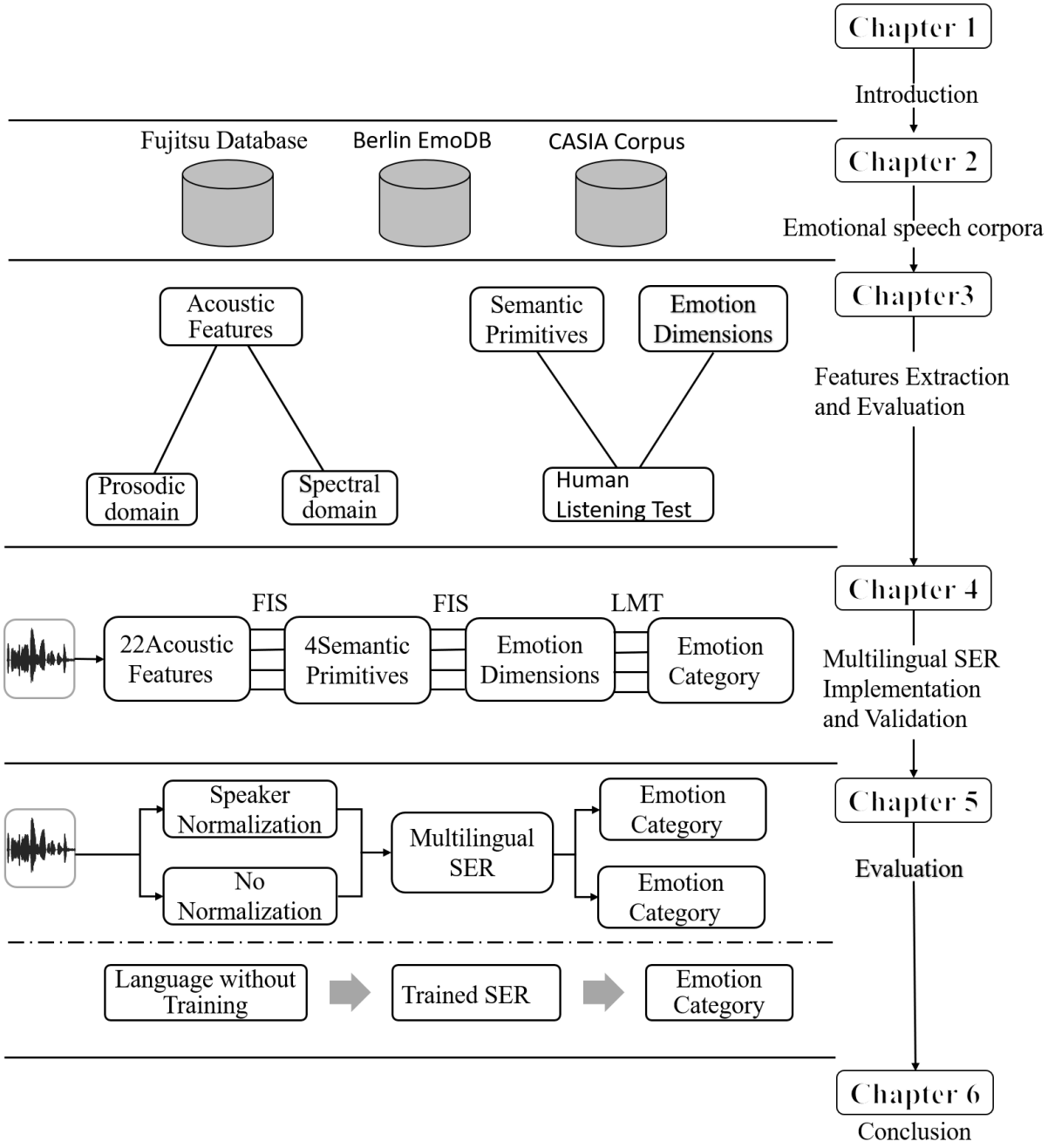


Figure 1.9: Organization of this dissertation.

Chapter 2

Emotional Speech Corpus

2.1 Introduction

The purpose of this research was to analyze the universal inference rules in the speech perception of emotion across different languages and design an SER model for identifying emotional traits/states from multilingual speech. To this end, this chapter introduced three emotional speech corpora in groups of two corpora were from oriental languages in Japanese and Chinese; and the other was from western language in German. The two traditions of emotional corpora that have been most regularly used in the area of SER are commonly collected from the spontaneous and acted speech. As suggested by Scherer (1992) [74], stereotypes of emotional states in speech often generalized inference rules that enable the cognitive studies of making inferences via very limiting data; and many social psychiatrists have additionally pointed out that stereotypes usually have a core of certainty. However, there are conflicting studies on the proper of emotional corpora on acted speech debated that stereotypes hold a very negative implication and are not adequate, suggesting that the corresponding inference rules are misleading. In this context, we learned three corpora from the acted domain and the other from the semi-spontaneous domain to assess the effectiveness of the constructed model in approaching the inference process of emotion in multilingual scenarios.

2.2 Fujitsu database

The Fujitsu database was chosen as the Japanese emotional corpus. It was recorded by the Fujitsu Laboratory and acted by a professional actress. The female speaker was asked to express 20 different sentences as shown in Table 2.1 nine times with five emotional states: neutral, happiness, cold anger, sadness, and hot anger. Each sentence was repeated once with neutral and two times with the other four emotions. Since one sentence in cold anger lost, this database contained 179 utterances in total. Specifically, 140 utterances in four emotional categories of neutral, joy, hot anger, and sadness were finally utilized exclude cold anger so as to be consistent in the use of primary emotions independent of languages.

Table 2.1: A list of transcriptions in the Fujitsu database in Japanese and English

Utterance Id	Japanese	Translation in English
1	Atarashi meru ga todoite imasu	You have got a new message
2	Atama ni kuru koto nante arimasen	There is something frustrating
3	Machiawase wa Aoyamarashin desu	I heard that we would meet in Aoyama
4	Atarashi kuruma o kaimashita	I bought a new car
5	Iranai meru ga attara sutete kudasai	Please delete any unwanted e-mails
6	Sonna no furui meishin desuyo	That is an old superstition
7	Minna kara eru ga okura retan desu	Many people sent cheers
8	Tegami ga todoita hazu desu	You should have received a letter
9	Zutto mite imasu	I will think about you
10	Watashi no tokoro ni wa todoite imasu	I have received it
11	Arigatogozaimashita	Thank you
12	Moshiwake gozaimasen	I am sorry
13	Arigato wa iimasen	I would not say thank you
14	Ryoko suru ni wa futari ga ino desu	I would like to travel just the two of us
15	Ki ga toku nari-sodeshita	I felt like fainting
16	Kochira no techigai mo gozaimashita	There were our mistakes
17	Hanabi o miru no ni goza ga irimasu ka	Do we need a straw mat to watch fireworks
18	Mo shinai to itta janai desu ka	You said you would not do it again
19	Jikandorini konai wake o oshiete kudasai	Tell me the reason why you do not come on time please
20	Sabisueria de goryo shimashou	Meet me at the service area

2.2.1 Human perception of speech emotion in the Fujitsu database

To examine the Fujitsu database could be identified universally across different mother-tongue listeners, we carried out cross-lingual studies by human listening tests. These tests mainly provided insight into whether the Japanese database was suited for the task in multilingual SER. The categorical evaluations were performed by Japanese native speakers and Chinese native speakers (who neither speak nor understand Japanese), providing a comparable performance between two groups of listeners. It stressed the fact that the Fujitsu database could be utilized in the research of multilingual SER.

◦ **Stimulus material**

Table 2.2 detailed the chosen utterances in each of the four emotional states, involving 20 neutral, 40 happiness, 40 anger, and 40 sadness.

Table 2.2: Stimulus material chosen in the listening test

Emotional State	Number of stimulus
Neutral	20
Happiness	40
Anger	40
Sadness	40
Total	140

◦ **Listeners**

Two Japanese native speakers (one male and one female, mean age: 35.5 years old, std: 13.4 years old) and four male Chinese native speakers who neither speak nor understand Japanese (mean age:27.75 years old, std: 2.87 years old) were recruited to participate this listening test. All these listeners were from the Japan Advanced Institute of Science and Technology under the master, doctor, or researcher course, and no subjects had hearing impairments or mental disorders.

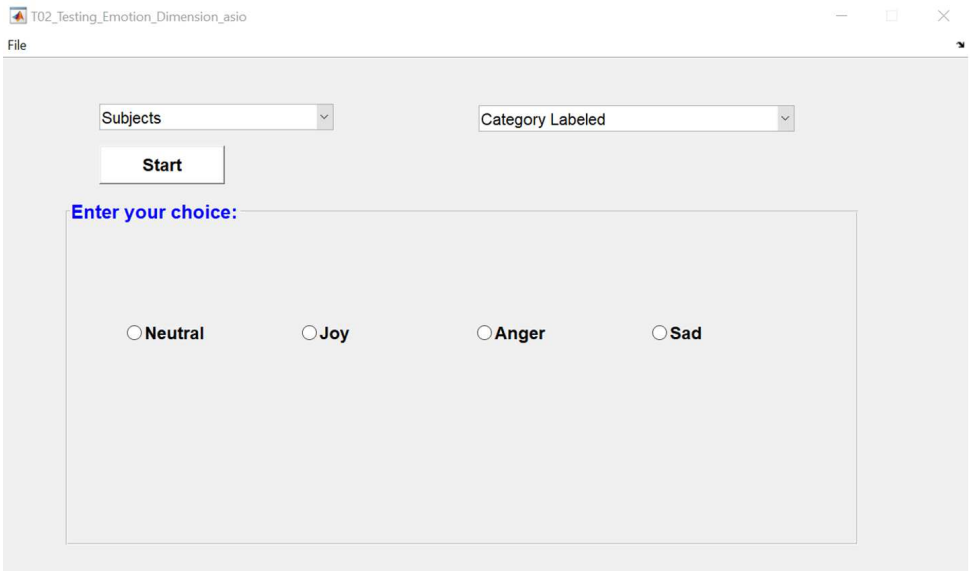


Figure 2.1: MATLAB GUI for labeling emotional categories in speech.

◦ Method

This experiment was carried out in a soundproof room. Listeners were recruited to label the 140 stimuli into an identified emotional state of each of the four categories. Those stimulus materials were randomly and repeatedly play. Figure 2.1 demonstrated a MATLAB Graphical User Interface that listeners used in this test.

◦ Metrics

The recall, precision, and F-measure were reported on each emotional state for assessing the categorical emotion perception. Formulate, let C_i be an emotional class to be identified, where $i \in \{neutral, happiness, anger, sadness\}$, and N_i was the total number of utterances for class C_i. Supposing a classifier predicted correctly NC_i^T utterances for class C_i, and predicted NC_i^F utterances to be in C_i where in fact those utterances belong to other emotional classes, then the recall, precision, and F-measure were defined as:

$$Recall = \frac{NC_i^T}{N_i} \tag{2.1}$$

$$Precision = \frac{NC_i^T}{NC_i^T + NC_i^F} \tag{2.2}$$

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2.3}$$

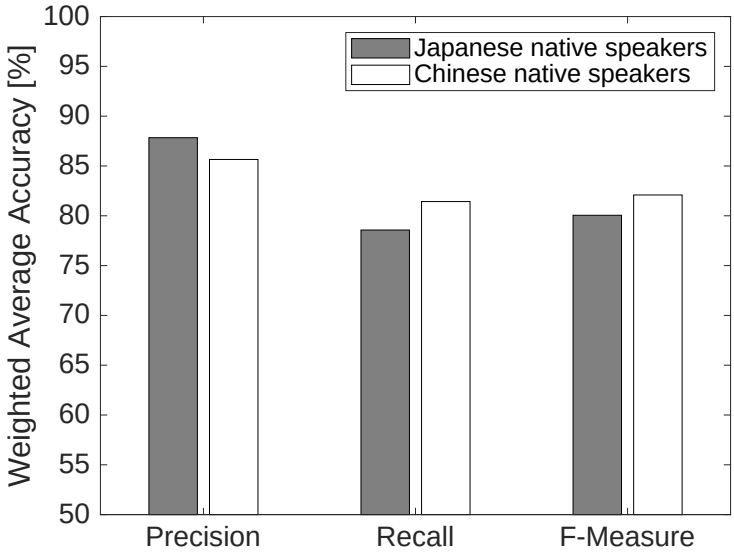


Figure 2.2: Identification results of the weighted average precision, recall, and F-Measure for four emotional categories (neutral, happiness, anger, and sadness) in the Fujitsu database, were reported by human listeners in groups of Japanese native speakers, and Chinese native speakers.

◦ **Results and discussion**

Figure 2.2 showed the weighted average precision, recall, and F-Measure for the four emotional categories in the Fujitsu database, provided by the Japanese native speakers and Chinese native speakers. As shown, the Japanese and Chinese native speakers yielded a small difference within 2% over all those three metrics. This result can be interpreted as evidence that the four emotional states in the Fujitsu database can be perceived independently both relative to mother-tongue and non-native listeners.

2.3 Berlin EmoDB

The German corpus was released by the Institute of Speech and Communication, Technical University of Berlin. Ten professional actors (five males and five females) each uttered ten sentences in German to simulate seven different emotions. The number of utterances of each emotion was as follows: 127 anger, 81 boredom, 46 disgust, 69 fear, 71 joy, 79 neutral, and 62 sadness. This corpus was produced in 16 bit, 16KHz under studio noise

Table 2.3: A list of transcriptions in the Berlin EmoDB in Germany and English

Utterance Id	Germany	Translation in English
b01	Was sind denn das fr Tten, die da unter dem Tisch stehen	What about the bags standing there under the table
b02	Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter	They just carried it upstairs and now they are going down again
b03	An den Wochenenden bin ich jetzt immer nach Hause gefahren undhabe Agnes besucht.	Currently at the weekends I always went home and saw Agnes
b09	Ich will das eben wegbringen und dann mit Karl was trinken gehen	I will just discard this and then go for a drink with Kar
b10	Die wird auf dem Platz sein, wo wir sie immer hinlegen	It will be in the place where we always store it
a01	Der lappen liegt auf dem Eisschrank	The tablecloth is lying on the frigde
a02	Das will sie am Mittwoch abgeben	She will hand it in on Wednesday
a04	Heute abend knnte ich es ihm sagen	Tonight I could tell him
a05	Das schwarze Stck Papier befindet sich da oben neben dem Holzstck	The black sheet of paper is located up there besides the piece of timber
a07	In sieben Stunden wird es soweit sein	In seven hours it will be

conditions. In order to choose similar categories as those in the Fujitsu database (joy, hot anger, sadness, and neutral) for the multilingual SER research. 50 joy, 50 anger, 50 sadness, and 50 neutral, in total 200 utterances were collected from Berlin EmoDB. Five males spoke one hundred utterances, and five females spoke the rest.

The details of this emotional corpus and the chosen stimulus materials were shown in Tables 2.3 and 2.4 corresponding the transcriptions and the number of utterances of each emotional category per speaker.

2.3.1 Human perception of speech emotion in the Berlin EmoDB

This experiment mainly provided insight into the fact that the emotional speech in the Berlin Emo-DB could be identified universally across different mother-tongue listeners. To this end, the Japanese native speakers and Chinese native speakers were recruited to take part in this listening test as non-mother-tongue listeners, none of the listeners can

Table 2.4: Details in the chosen utterances for each speaker per emotional state from the Berlin EmoDB

Speaker Id	Gender	Neutral	Happiness	Anger	Sadness	Total
F08	female	6	9	4	8	27
F09	female	4	1	6	4	15
F13	female	8	7	3	2	20
F14	female	2	6	5	4	17
F16	female	5	2	7	7	21
M03	male	6	7	4	7	24
M10	male	4	3	4	3	14
M11	male	5	7	6	7	25
M12	male	3	2	6	4	15
M15	male	7	6	5	4	22
Total	NULL	50	50	50	50	200

understand the German language. In particular, since it was impractical for us to recruit enough German native speakers for this task, a previous work that has been able to give such results instead was reviewed as a reference [5]. The evidence on the presence of cross-lingual speech perception of emotion was reinforced again, by demonstrating somewhat the same results on the recognition of emotional states in the Berlin Emo-DB by three groups of listeners.

- **Stimulus material**

Table 2.5 details the utterances chosen in each of the four basic emotional states, involving 50 neutral, 50 happiness, 50 anger, and 50 sadness.

Table 2.5: Stimulus material chosen in the listening test

Emotional State	Number of stimulus
Neutral	50
Happiness	50
Anger	50
Sadness	50
Total	200

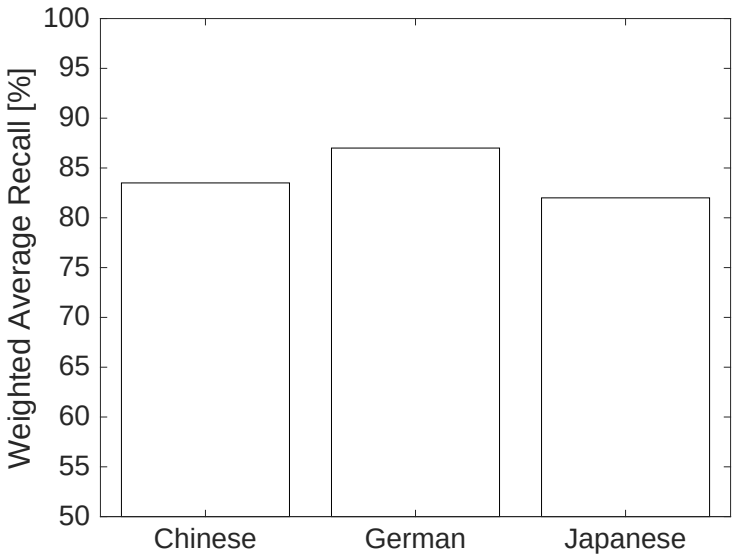


Figure 2.3: Identification results on the weighted mean recall for four basic emotional categories (neutral, happiness, anger, and sadness) in the Berlin EmoDB, obtained by human listeners in groups of Japanese native speakers, Chinese native speakers, and German native speakers in [5].

◦ **Listeners**

They were the same in Section 2.2.1.

◦ **Method**

This experiment was carried out in a soundproof room. Listeners were recruited to label the 200 stimuli to an identified label of each of the four emotional categories. Those stimulus materials were randomly and repeatedly play. Furthermore, in this test, Listeners used the same MATLAB Graphical User Interface as shown in Figure 2.1.

◦ **Metrics**

In light of the fact that the reference [5] did not provide results corresponding precision and F-measure, this test thus showed the result on the recall only.

◦ **Results and discussion**

Figure 2.3 displayed the weighted average recall for the four emotional categories in the Berlin EmoDB individually. They were provided by Japanese native speakers, Chinese native speakers and German native speakers. The highest result was achieved by German native speakers resulting in an averaged recall of 87%; followed by the Chinese native

speakers given an averaged recall of 83.5%, and then was that obtained by Japanese listeners by 82%. It was found that even for the listeners in an Eastern tongue, they can recognize speech emotion from western languages accurately as native speakers did.

2.4 CASIA Corpus

The Chinese emotional database (CASIA) was produced by the Institute of Automation, Chinese Academy of Sciences, containing neutral and five categories of acted emotion: angry, happy, sad, fear, and surprise. It was performed by four professional actors (two males and two females). The data involved dominant and spontaneous parts. The utterances of the dominant part have at least one dominant word, e.g., "anger" or "annoyed" for angry, "pleased" or "joyful" for happiness, and "sad" for sadness, and so on. There were 100 utterances for each emotion. The utterances of the spontaneous part were picked up from news articles, conversations, and essays without emotionally-rich words. There were 300 utterances in this part. Each speaker uttered $(100 + 300) * 6 = 2400$ sentences in total.

We chose 200 sentences from the spontaneous portion among four speakers involving four emotional states: angry, happy, neutral and sad, taking 50 sentences from each category. Different from the Fujitsu database or the Berlin Emo-DB, the spontaneous speech in the CASIA Emotional Corpus did not sufficiently simulate emotions in a natural or clear manner. Four Chinese native speakers (two male and two female) hence were asked to verify the emotional categories in a listening test. The experimental results yielded a mean recognition accuracy of 97, 39, 83, and 93% for neutral, happy, angry, and sad. Compared with the other three well-recognized emotions, happy utterances were recognized with an extremely low accuracy of 39%. The utterances, therefore, were re-annotated by five female and six male Chinese native speakers into the correct categories. The number of utterances of each category was: 68 neutral, 29 happy, 51 angry, and 50 sad. Two spoken utterances could not be identified as any one of the above four emotional categories. Finally, 198 instances were taken from the CASIA corpus.

The uttered transcription in the speech from CASIA corpus was shown in Table 2.6. Moreover, the chosen utterances for each speaker per emotion can be seen in Table 2.7.

Table 2.6: A list of transcriptions in the CASIA Corpus in Chinese and English

Utterance Id	Chinese	Transcription in English
406	shi fen fu za de mi mi dou zheng	A completely complex and secret struggle
418	du shen yi ren zai na biao yan	He/She acts there alone
425	gu niang zai xi ju zhong he bie de nan ren tan hua shi	While the girl talked to the other man in the theater
427	ta jie shu le zhe chang hao wu yi yi de lian ai	He/She ended the meaningless love
431	zhi jie liao dang di dui ta biao bai	Expressing your love to her openly
432	xiao shi hou ting zu mu jiang guo yi ge gu shi	When I was a little girl, my grandma told me a story
440	shan dian de chang du ke neng zhi you shu bai qian mi	The length of lightning maybe few thousand kilometer
441	mai gei na xie chi you zheng shu de fa guo wu qi zhi zao shang	Sell them to the France certificated weapons companies
447	ji zhu zhe ge qi ji bei fa xian de shi jian	Do remember the moment while discovering the wonders
455	dian tai zuo xian chang guang bo	Radio station is doing a live broadcast
466	zhi li yu ke xue yan jiu de ren	The person identified himself strongly with researching
472	mei ge ren dou ying gai jiang yi ge gu shi	Everyone should tell a story
492	dui kang sai jiu zhe yang jie shu le	The battle is over

Table 2.7: Details in the chosen utterances for each speaker per emotional state from the CASIA Corpus

Speaker Id	Gener	Neutral	Happiness	Anger	Sadness	Total
wzh	male	16	8	13	11	48
zzx	male	24	1	13	14	52
lch	female	10	12	13	14	49
zqy	female	18	8	12	11	49
Total	NULL	68	29	51	50	198

2.4.1 Human perception of speech emotion in the CASIA Corpus

The following test particularly provided insight into the fact that the CASIA corpus could be identified universally across different mother-tongue listeners. The experimental results reported in this listening test were given by the Japanese native speakers (who neither speak nor understand Chinese) and Chinese native speakers. Comparable recognition results of the four emotional states were found between Chinese and Japanese listeners, which in turn stressed the fact that the CAISA corpus could be utilized in the study of SER in multilingual scenarios.

- **Stimulus material**

Table 2.8 detailed the utterances chosen in each of the four emotional states, involving 50 neutral, 50 happiness, 50 anger, and 50 sadness.

Table 2.8: Stimulus material chosen in the listening test

Emotional State	Number of stimulus
Neutral	68
Happiness	29
Anger	51
Sadness	50
Total	198

- **Listeners**

They were the same in Section 2.2.1.

- **Method**

This experiment was carried out in a soundproof room. Listeners were asked to rate the 198 stimuli to an identified label of each of the four emotional categories. Those stimulus materials were randomly and repeatedly play. In this test, the same MATLAB Graphical User Interface that shown in Figure 2.1 was used by listeners.

- **Metrics**

They are the same in Section 2.2.1.

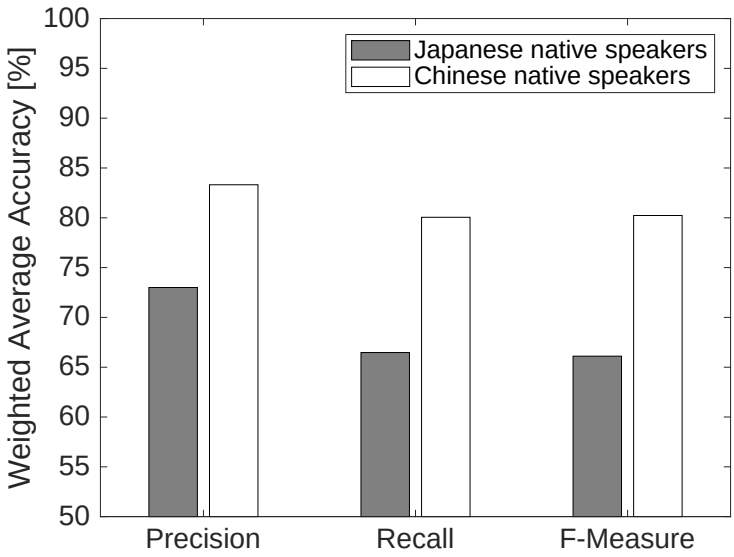


Figure 2.4: Identification results on the weighted mean recall for four basic emotional categories (neutral, happiness, anger, and sadness) in the CASIA corpus, obtained by human listeners in groups of Japanese native speakers and Chinese native speakers.

◦ **Results and discussion**

Figure 2.4 displayed the weighted average recall for the four emotional categories in the CASIA corpus. These results were provided by Japanese native speakers and Chinese native speakers. The perceptual results on identifying emotional states in the Chinese emotional speech by the Japanese native speakers was slightly lower when compared to those given by the Chinese native speakers. Not surprisingly, judging emotional states in spontaneous speech is a challenging task. However, the averaged accuracy over all evaluation metrics given by Japanese listeners was greater than 66% which is much higher than the chance level of 25%. Instead, this Chinese emotional corpus can be studied as an intriguing effort to study the inference rules between acted and spontaneous speech in a cross-lingual scenario.

Table 2.9: Error Analytic: Merged confusion matrix over two Japanese native speakers for the speech perception of emotion in the Fujitsu dataset

	Expressed Emotion			
Identified Emotion	Neutral	Happiness	Anger	Sadness
Neutral	**38 (13.57)**	0 (0)	1 (0.36)	1 (0.36)
Happiness	20 (7.14)	**57 (20.36)**	2 (0.71)	1 (0.36)
Anger	15 (5.36)	1 (0.36)	**63 (22.50)**	1 (0.36)
Sadness	10 (3.57)	1 (0.36)	7 (2.50)	**62 (22.14)**
Total Error	45 (16.07)	2 (0.71)	10 (3.57)	3 (1.07)

Note, values are the amount of incorrect identified samples (with % in the brackets), while the number of the correct identified samples are in bold. Expressed emotions involving four emotional categories that are studied in the multilingual SER. This matrix reports 60 errors and 220 correct identified samples in total

Table 2.10: Error Analytic: Merged confusion matrix over four Chinese native speakers for the speech perception of emotion in the Fujitsu dataset

	Expressed Emotion			
Identified Emotion	Neutral	Happiness	Anger	Sadness
Neutral	**66 (11.79)**	7 (1.25)	3 (0.54)	4 (0.71)
Happiness	10 (1.79)	**144 (25.71)**	2 (0.36)	4 (0.71)
Anger	10 (1.79)	18 (3.21)	**124 (22.14)**	8 (1.43)
Sadness	36 (6.43)	2 (0.36)	0 (0)	**122 (21.79)**
Total Error	56 (10)	27 (4.82)	5 (0.89)	16 (2.86)

2.5 Discussion and Summary

This chapter presented an introduction of the three emotional speech corpora that were used in the research of multilingual SER. This subsection displayed the confusion matrixes that were established by summing the amount of error and the confusion patterns emerging from the listening tests in categorical labeling. It can provide some insights into the universal and dependent aspects on speech perception of emotion among different languages.

Overall, Tables 2.9 and 2.10 presented a low error rate of all predictions of 21.43% for the Japanese native speakers and 18.57% for the Chinese native speakers. The emotional state of neutral commonly yielded the highest amount of total error in two groups of listeners. Most interestingly, the neutral speech was mistaken for happy speech (20 out of 45 total incorrect identified samples) as given by the Japanese listeners; however, Chinese listeners confused neutral often with sadness (36 out of 56 total misunderstanding samples). Moreover, Japanese native speakers often confused anger with sadness (7 predictions) in the Fujitsu database, which could be interpreted as an evidence that the arousal-inspired (activation or deactivation) emotional states within mother-tongue speakers seem to be ambiguous. In particular, the Chinese listeners frequently confused with happiness with anger, interpreting the fact that the prediction of valence-inspired emotions could be a challenging task for the non-native speakers.

Concerning the study of cross-lingual speech perception of emotion relative to the Berlin EmoDB, we showed two merged confusion matrix in Tables 2.11 and 2.12. As can be seen from these tables, the error rate was relatively lower in two groups of listeners (18% for all predictions by Japanese listeners; and 16.5% for all predictions by Chinese listeners). The emotional state with the highest amount of total error was neutral. And listeners often confused anger with happiness and vice versa independent to the non-native speakers. These findings were consistent with that obtained from previous Tables 2.9 and 2.10 suggesting a difficulty in predicting valence dimension in a cross-lingual scenario.

Tables 2.13 and 2.14 reports the results of the error analysis for the speech perception of emotion in the CASIA corpus. Making a reference to the previous results demonstrated in Tables 2.9 and 2.10, most interestingly, it indicates that the finding in those cross-culture study is extremely opposite, i.e., consequently, neutral is the most difficult to identify where neutral speech is mistaken for positive (happy) speech in terms of mother-tongue listeners, yet negative (anger/sadness) speech for the non-native listeners. Notably, it further stressed that anger and happiness are easy to be ambiguous to non-native speakers, suggesting that valence is challenging to detect from speech.

Table 2.11: Error Analytic: Merged confusion matrix over two Japanese native speakers for the speech perception of emotion in the Berlin Emo-DB

	Expressed Emotion			
Identified Emotion	Neutral	Happiness	Anger	Sadness
Neutral	**71 (17.75)**	2 (0.5)	6 (1.5)	21 (5.25)
Happiness	11 (2.75)	**79 (19.75)**	7 (1.75)	3 (0.75)
Anger	12 (3)	8 (2)	**80 (20)**	0 (0)
Sadness	1 (0.25)	0 (0)	1 (0.25)	**98 (24.5)**
Total Error	24 (6)	10 (2.5)	14 (3.5)	24 (6)

Table 2.12: Error Analytic: Merged confusion matrix over four Chinese native speakers for the speech perception of emotion in the Berlin EmoDB dataset

	Expressed Emotion			
Identified Emotion	Neutral	Happiness	Anger	Sadness
Neutral	**191 (23.88)**	2 (0.25)	1 (0.13)	6 (0.75)
Happiness	23 (2.88)	**163 (20.38)**	13 (1.63)	1 (0.13)
Anger	16 (2)	17 (2.13)	**165 (20.63)**	2 (0.25)
Sadness	49 (6.13)	1 (0.13)	1 (0.13)	**149 (18.63)**
Total Error	88 (11)	20 (2.50)	15 (1.88)	9 (1.13)

Table 2.13: Error Analytic: Merged confusion matrix over two Japanese native speakers for the speech perception of emotion in the CASIA Corpus

	Expressed Emotion			
Identified Emotion	Neutral	Happiness	Anger	Sadness
Neutral	**118 (29.8)**	11 (2.78)	2 (0.51)	5 (1.26)
Happiness	24 (6.06)	**31 (7.83)**	2 (0.51)	3 (0.76)
Anger	30 (7.58)	19 (4.8)	**52 (13.13)**	0 (0)
Sadness	26 (6.57)	11 (2.78)	0 (0)	**63 (15.91)**
Total Error	80 (20.20)	41 (10.35)	4 (1.01)	8 (2.02)

Table 2.14: Error Analytic: Merged confusion matrix over four Chinese native speakers for the speech perception of emotion in the CASIA Corpus

	Expressed Emotion			
Identified Emotion	Neutral	Happiness	Anger	Sadness
Neutral	**245 (30.93)**	19 (2.4)	1 (0.13)	3 (0.38)
Happiness	45 (5.68)	**74 (9.34)**	0 (0)	1 (0.13)
Anger	33 (4.17)	27 (3.41)	**140 (17.68)**	4 (0.51)
Sadness	19 (2.4)	5 (0.63)	1 (0.13)	**175 (22.1)**
Total Error	97 (12.25)	51 (6.44)	2 (0.25)	8 (1.01)

In closing this chapter, three emotional speech corpora that can be identified universally among mother-tongue speakers and non-native speakers have been examined and determined to be studied in the research of multilingual SER. Notably, the cross-lingual listening tests give two insights into the speech perception of emotion: (1) neutral is the most difficult emotional state to be identified even for the mother-tongue speakers, where native-speakers frequently confused it with positive emotions, yet negative emotions for non-native speakers; (2) Anger and happiness seems to be ambiguous for non-native speakers, suggesting that predicting valence from speech is a challenging task.

Chapter 3

Feature Extraction and Evaluation

This chapter presented the kind of features to implement and validate the three-layer multilingual SER model. These features have been examined over the physiological (acoustic features) and psychological (perceptual judgments, involving evaluation of semantic primitives and emotion dimensions) domains. In the following sections, we discussed in detail on extracting acoustic features, followed by evaluating semantic primitives and emotional dimensions.

3.1 Acoustic features

3.1.1 Prosodic-related features

The set of prosodic features was analysed in our earlier attempt [7], and can be grouped into five categories:

Fundamental frequency (F0): maximum, mean, mean of rising slopes of the speech over all accentual phrases, and rising slope of the first accentual phrase.

Power spectrum: mean value of the first, second, and third formant in dB, spectral tilt, and spectral balance.

Power envelop: range(max-min), ratio of mean power in high frequency domain over 3 kHz and the mean power over whole speech, mean of rising slopes of the speech over all accentual phrases, and rising slope of the first accentual phrase.

Timing: total length of whole speech, length of consonants, ratio of length of consonants to that of vowels.

Voice quality: mean of the difference between the fundamental frequency (H1) and the second harmonic (H2) for each vowel. Since the vowels vary with languages, in this study, we only focused on the common vowels among the languages, namely, /a/, /i/, and /u/.

The above-mentioned acoustic features derived from F0, the power envelop, power spectrum, and voice quality were calculated using STRAIGHT [75]. In addition, the acoustic correlates related to timing were extracted by manual segmentation.

3.1.2 Spectral-related features

The set of modulation spectral features, abbreviated as MSF, was collected and calculated from the modulation spectrogram. We herein referred to a previous attempt on extracting MSF using an auditory-inspired system [76], incorporating a $32-band$ auditory filterbank with centre frequencies scaled by the equivalent rectangular bandwidth (ERB) from 3 to 35 $ERB_{numbers}$ and a $6-band$ modulation filterbank with centre frequencies ranging from 2 to $64Hz$. The modulation spectrogram allows for analysis of the modulation frequency content across different acoustic frequency bands. The MSFs were hence calculated over two different domains:

Acoustic frequency domain: spectral centroid, spectral spread, spectral skewness, spectral kurtosis, spectral flatness, and spectral slope.

Modulation frequency domain: spectral centroid, spectral spread, spectral skewness, spectral kurtosis, and spectral tilt.

As for the acoustic frequency domain, six statistics were calculated over modulation frequency bands providing 36 acoustic correlates; additionally, 160 acoustic features were obtained from the modulation frequency domain over 32 acoustic frequency bands for five statistics. In total, 196 acoustic features were extracted from the modulation spectrogram.

In summary, 215 acoustic features were established as an initial feature pool consists of 19 prosodic features and 196 spectral features from modulation spectrogram.

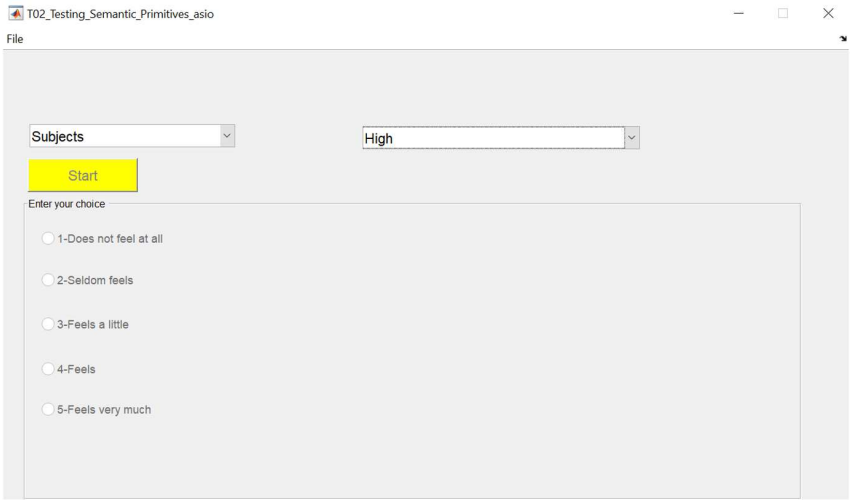


Figure 3.1: MATLAB GUI for evaluating semantic primitives in speech.

3.2 Assessment of semantic primitives

The semantic primitives that chosen for describing emotional multilingual speech in this study were the same in [3], involving seventeen adjectives, i.e., bright, dark, high, low, strong, weak, calm, unstable, well-modulated, monotonous, heavy, clear, noisy, quiet, sharp, fast and slow.

Each semantic primitive was evaluated on the Fujitsu database, Berlin Emo-DB, and CASIA dataset. Eleven Japanese native speakers (nine male and two female, mean age: 26.8 years old) were asked to evaluate the Fujitsu database, and ten Chinese native speakers (five male and five female, mean age: 25.3 years old) were asked to evaluate the CASIA dataset. However, it was impractical for us to recruit enough German native speakers for the listening test. Nonetheless, psychology research has recently shown that speech emotions can be recognized across different languages [31, 77], so we asked nine Japanese native speakers (eight male and one female, mean age: 26.2 years old) to evaluate the Berlin-Emo DB instead. None of these nine participants can understand German. The consistency in the perception of emotions of different nationalities have already confirmed in Chapter 2. Besides, all three groups of participants are from Japan Advanced Institute of Science and Technology under master or doctor course, and no subjects have hearing impairments or mental disorders.

As for the evaluation of semantic primitives, the emotional speech was randomly and once played, and evaluated 17 times by the participants on a whole utterance level, each time on one of 17 semantic primitives for all utterances in per corpus. Each of the these semantic primitives was scored a five-point scale: '1-Does not feel at all', '2-Seldom feels', '3-Feels a little', '4-Feels', and '5-Feels very much', using a MATLAB Graphical User Interface as shown in Figure 3.1.

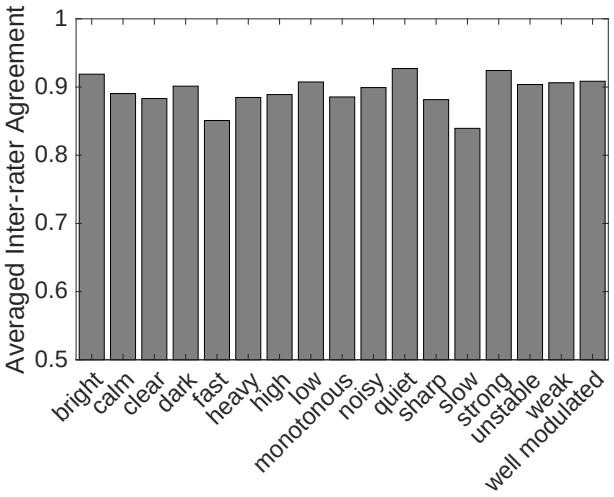
For each instance of speech n in corpus c, where $c \in \{Fujitsu, BerlinDB, CASIA\}$, $1 \leq n \leq N$, the averaged ratings $\overline{x}_{n,c}^{(p)}$ of listeners' responses $\widehat{x}_{n,c}^{e,(p)}$ among all evaluators E were calculated for each semantic primitive, where p refers to one of the semantic primitives from bright, dark, high, low, strong, weak, calm, unstable,well-modulated, monotonous, heavy, clear, noisy, quiet, sharp, fast, and slow.

$$\overline{x}_{n,c}^{(p)} = \frac{1}{E} \sum_{e=1}^{E} \widehat{x}_{n,c}^{e,(p)}, \tag{3.1}$$

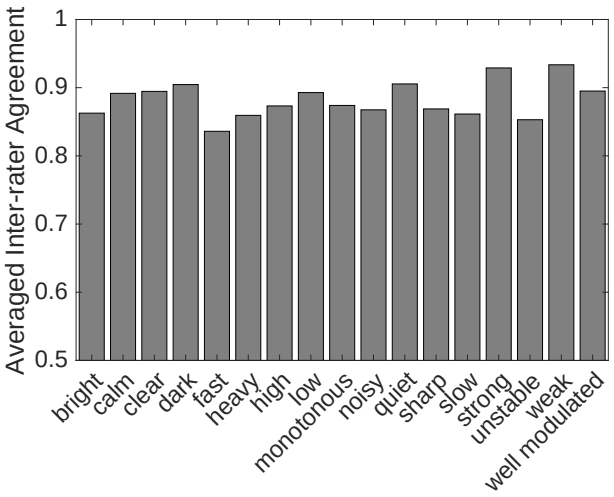
The inter-evaluator agreement was evaluated using Eq. 3.2 following the related study reported in [9].

$$CC_c^{e,(p)} = \frac{\sum_{n=1}^{N} \left(\widehat{x}_{n,c}^{e,(p)} - \frac{1}{N} \sum_{n'=1}^{N} \widehat{x}_{n',c}^{e,(p)} \right) \left(\overline{x}_{n,c}^{(p)} - \frac{1}{N} \sum_{n'=1}^{N} \overline{x}_{n',c}^{(p)} \right)}{\sqrt{\sum_{n=1}^{N} \left(\widehat{x}_{n,c}^{e,(p)} - \frac{1}{N} \sum_{n'=1}^{N} \widehat{x}_{n',c}^{e,(p)} \right)^2} \sqrt{\sum_{n=1}^{N} \left(\overline{x}_{n,c}^{(p)} - \frac{1}{N} \sum_{n'=1}^{N} \overline{x}_{n',c}^{(p)} \right)^2}} \tag{3.2}$$

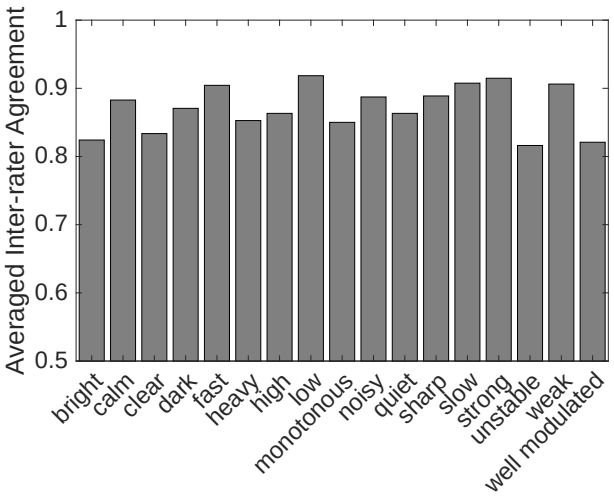
Figure 3.2a , 3.2b and 3.2c demonstrate the results of inter-rater agreement over all human listeners individually on the three emotional corpora. On average, the inter-rater correlation coefficient was moderate to high within the range of 0.84-0.93, 0.84-0.93, and 0.82-0.92 over the Fujitsu database, Berlin EmoDB, and CASIA dataset, indicating good evaluation results and good agreement among listeners. This result can be further interpreted as the evidence that those semantic primitives are proper for describing expressive speech in spite of language/culture, and human listeners can understand them prettily.



(a) Fujitsu database



(b) Berlin EmoDB



(c) CASIA corpus

Figure 3.2: The inter-rater agreement for the semantic primitives evaluation of each of the three corpora for Fujitsu database, Berlin EmoDB, and CASIA corpus over all human listeners.

3.2.1 Assessment of emotion dimensions

In terms of evaluating emotions in the two-dimension emotional space of valence and arousal, we carried out listening experiments by following the related study reported in [78], where the definition of the emotions of valence and arousal were demonstrated to the listeners, before they listened a small set of demos involving different degrees of a specific emotion. The same participants that evaluated the semantic primitives were asked to score the values for emotion dimensions on a five-point scale (-2, -1, 0, 1, 2) for valence (-2 being very negative and +2 being very positive) and arousal (-2 being very

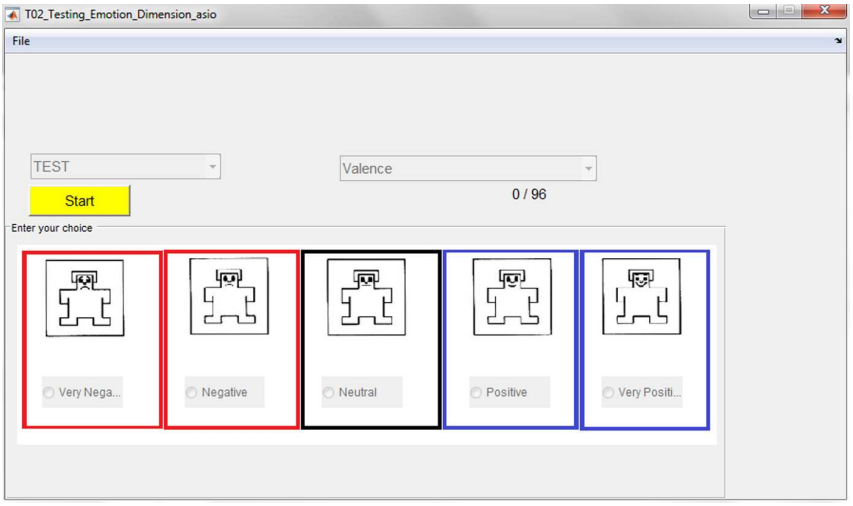


Figure 3.3: MATLAB GUI for evaluating valence dimension in speech.

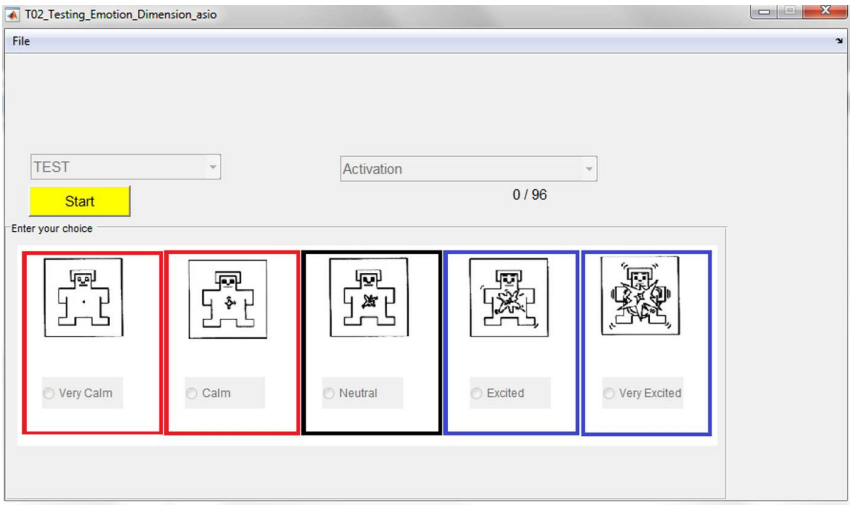
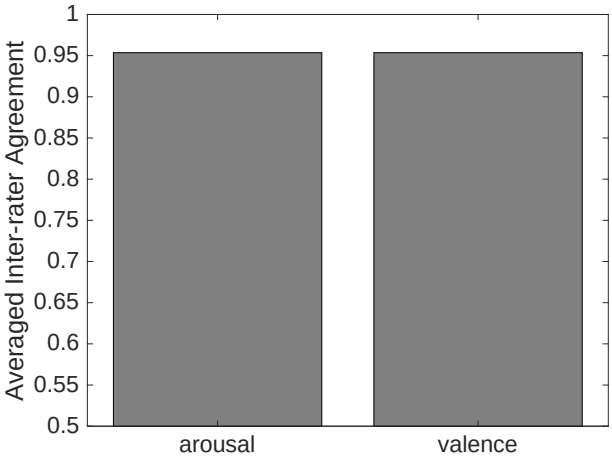


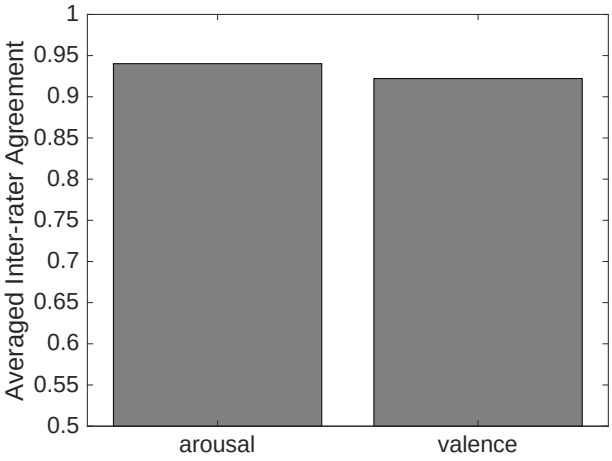
Figure 3.4: MATLAB GUI for evaluating Activation/Arousal dimension in speech.

relaxed and +2 being aroused) using the MATLAB Graphical UserInterfaces as seen in the Figures 3.3and 3.4 respectively. The emotional speech was randomly and once played to each listener in a soundproof room, and was evaluated two times by participants on a whole utterance level, once for each emotion dimension for all utterances in one dataset. The averaged ratings given by the listeners were calculated for each emotion dimension using Eq. 3.1, in such a scenario, p refers to valence or arousal.

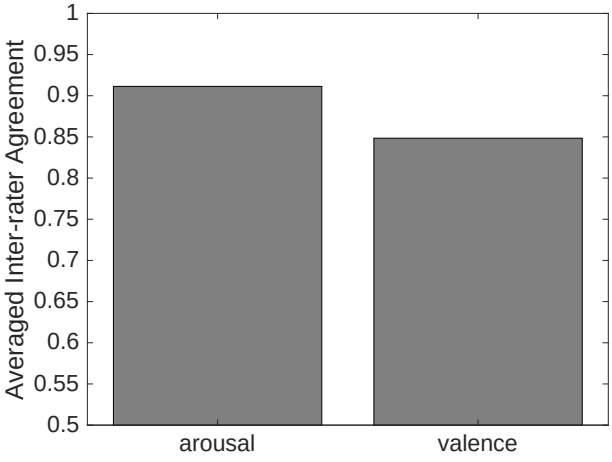
The inter-rater agreement on evaluations over emotion dimensions was also measured by Eq. 3.2, and shown in Figure 3.5a , 3.5b and 3.5c with respect to Fujitsu database, Berlin EmoDB and CASIA corpus respectively. On average, the agreement among listeners was moderate to high within the range of 0.85-0.96 over the three emotional corpora. All listeners reached a high agreement at a significant level where $p < 0.001$. In particular, the highest correlation was 0.96 and 0.96 for valence and arousal on the Fujitsu database; this might have been due to the fact that all emotions in this corpus were clearly produced by one professional actress. The inter-rater agreement of the Berlin EmoDB was moderate. However, the agreement among the participants was relatively low on the CASIA dataset. Although Berlin Emo-DB and CASIA both feature multiple actors, the spontaneous emotional utterances in the CASIA dataset were from news articles, conversations, and essays; the poor performance may be attributed to the fact that the spontaneous speech in the CASIA dataset does not sufficiently simulate emotions in a natural and clear way. In addition, it was found that the valence dimension generally yielded a lower inter-rater agreement than arousal, indicating that human evaluations are more poorly correlated in terms of valence in comparison to arousal.



(a) Fujitsu database



(b) Berlin EmoDB



(c) CASIA corpus

Figure 3.5: The inter-rater agreement for the emotion dimensions evaluation of each of the three corpora for Fujitsu database, Berlin EmoDB, and CASIA corpus over all human listeners.

3.3 Discussion and Summary

This section provided some insights into the speech perceptual judgments of emotion over three emotional speech corpora of Fujitsu database, Berlin EmoDB, and CASIA corpus on the basis of the evaluations of emotion dimensions.

Figure 3.6 showed the emotional space distributions for each corpus for all speakers as a result of eleven (Fujitsu database), nine (Berlin EmoDB), and ten (CASIA corpus) listeners' evaluations. As can been seen, (1) all these three emotional corpora contain a high percentage of emotional utterances in neutral or negative speech with moderate to high arousal values. (2) In particular, compared with Fujitsu database, it has to be noted that Berlin EmoDb and CASIA corpus hold a smaller amount of emotional utterances with extraordinarily high and low intensity. This distribution was probably due to the fact of the utterances acted in the Fujitsu database was quite clear and typical. By contrast, the utterances coming from the Berlin EmoDB, and CASIA corpus were more closer to spontaneous speech. (3) Furthermore, it can be found that Berlin EmoDB and CASIA corpus contained a higher number of utterances in the neutral state.

Particularly, Figures 3.7 and 3.8 demonstrated the details of valence/arousal dimensional evaluation for four emotional categories of neutral, happy, anger, and sad. Emotional categories in the Fujitsu database were quite clear, showing the smallest values of standard deviation. Neutral utterances in Berlin EmoDB were moderately positive, whereas were varied from moderately negative to moderately positive with respect to the CASIA corpus, and suggesting a higher value of standard deviation. This fact can also be interpreted as reasons why human listeners' recognition of category neutral was worse than other categories (c.f. Tables 2.9 to 2.12). Moreover, it has to be noted the fact that the perceived values of valence were only moderately positive with respect to Berlin EmoDB and CASIA corpus or even very negative in Berlin EmoDB, likewise, it might be the reasons to ambiguous between emotions of neutral and happy, and that of happy and anger (c.f. Section 2.5).

This chapter firstly introduced a set of acoustic features to be studied in this research in Section 3.1. Section 3.2 next described a listening experiment to collect the human evaluations/judgment to semantic primitives that used to describing multilingual speech emotion in a three-layer model. Furthermore, the two-dimensional emotion space relative

to valence and arousal dimensions was also subjectively evaluated by the human listening test as described in Section 3.3. Lastly, the estimation and classification approaches were given in Section 3.4. Chapter 4 will introduce the implementation and validation of the proposed SER system in multilingual scenarios.

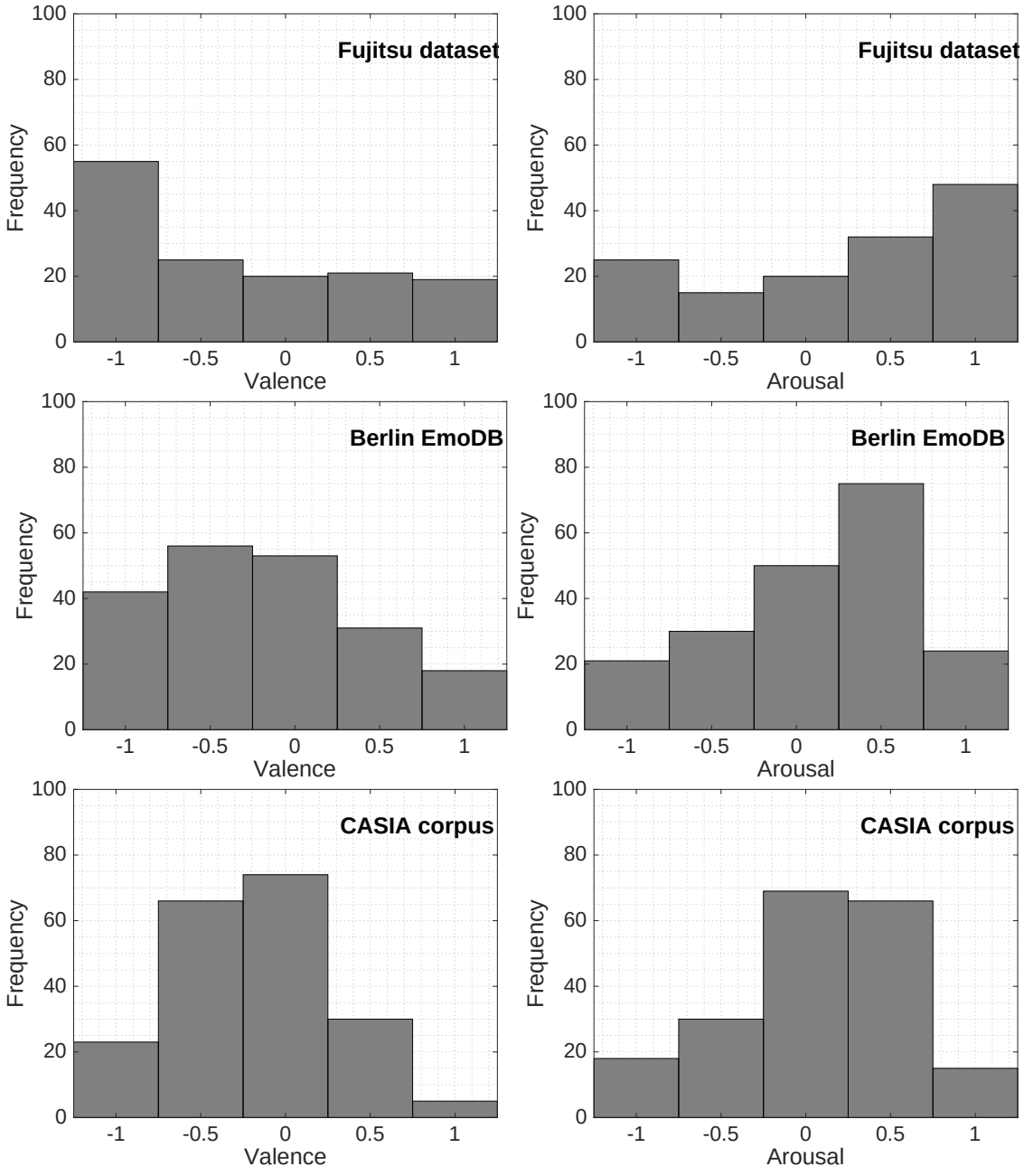


Figure 3.6: Histogram of emotions for each of the three emotional speech corpora for Fujitsu database, Berlin EmoDB, and CASIA corpus regarding valence and arousal dimensions.

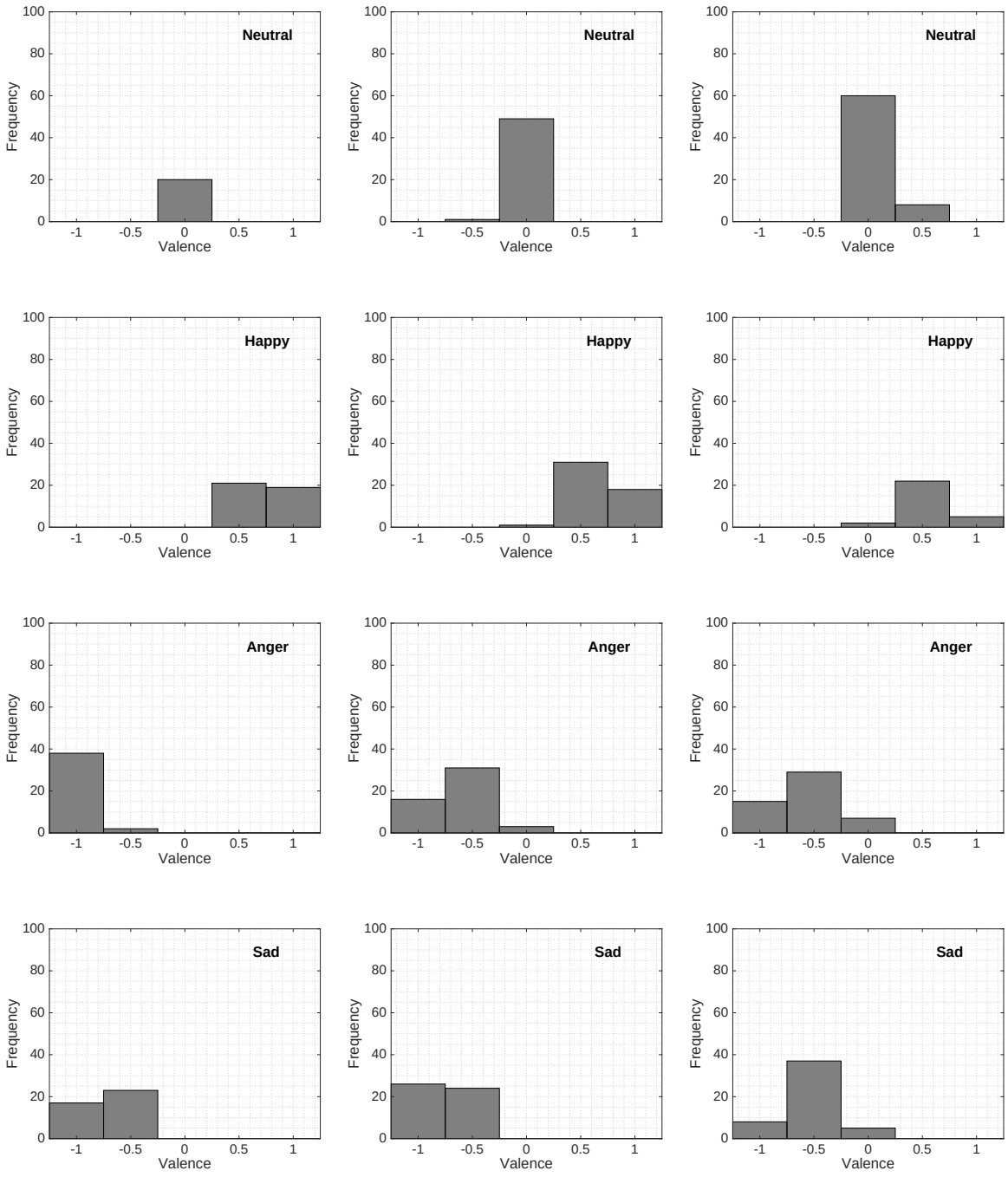


Figure 3.7: Human valence dimensional emotion evaluation of categories, neutral, happy, anger and sad for each of the three emotional speech corpora for Fujitsu database (left column), Berlin EmoDB (middle column), and CASIA corpus (right column).

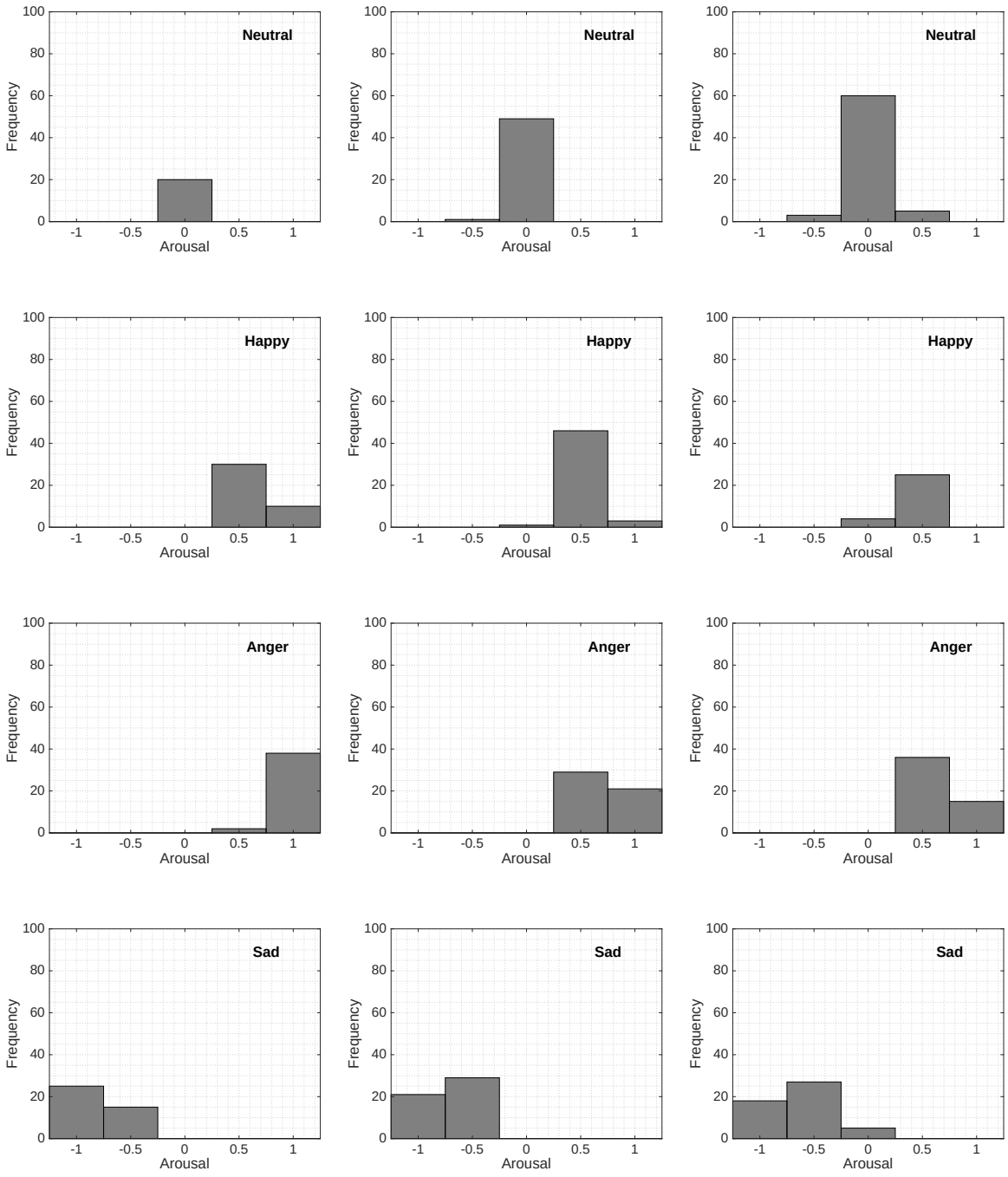


Figure 3.8: Human arousal dimensional emotion evaluation of categories, neutral, happy, anger and sad for each of the three emotional speech corpora for Fujitsu database (left column), Berlin EmoDB (middle column), and CASIA corpus (right column).

Chapter 4

Implementation and Validation to the Proposed Multilingual SER System

4.1 Introduction

This research introduced a framework from a perspective of human speech emotion perception rather than traditions that studied emotional state in the speech directly from acoustic features to approach multilingual SER. Once achieved, this framework could recognize emotional states from speech across multiple languages, and even for a different language without training. With this context, this chapter provided the details in implementing the proposed system and suggesting three promising methodologies to solve the multilingual SER challenge, i.e., (1) bringing together the categorical-based and dimensional-based emotion theories to capture both a discrete label and a gradual degree within an emotional state from multilingual speech; (2) A three-layer model was used to simulate the human perception of emotion based on the judgments of emotion dimensions through semantic primitives; (3) 22 acoustic features and 4 semantic primitives were determined as best patterns with respect to the physiological domain and perceptual domain by using the sequential floating forward selection algorithm. The following sections discussed the details on the implementation and evaluation of each of those three methodologies.

4.2 Implementation to multilingual SER system

This section presented a two-stage estimation scheme for multilingual SER. First, the estimation of emotion dimensions for valence and arousal was addressed by using a three-layer model; Second, the task of emotion categorical classification was furnished by incorporating the human-perception-inspired knowledge [31] based on a classification scheme in logistic model trees.

4.2.1 Feature selection

Large feature sets not only have exorbitant costs in terms of time for system training, but they also involve irrelevant features that reduce recognition accuracy [79]. In this regard, we introduced a two-stage feature selection algorithm to define the best features. In the first stage, we calculated the Fisher discriminant ratio (FDR) to each feature individually that able to eliminate irrelevant features. The normalized multi-class FDR for the uth feature is given as:

$$FDR(u) = \frac{2}{C(C-1)} \sum_{c1} \sum_{c2} \frac{(\mu_{c1,u} - \mu_{c2,u})^2}{(\sigma^2_{c1,u} + \sigma^2_{c2,u})} \tag{4.1}$$

with $1 \leq c_1 < c_2 \leq C$, where $\mu_{c1,u}$ and $\sigma^2_{c2,u}$ are the mean and variance of the uth feature for the c_1th class, and C is the total number of classes. The FDR measure concerns the number of binary comparisons made between two categories, which favors features with well-separated means across classes and small within-class variances. Features with relatively low discrimination ability can then be removed by using the FDR as a threshold. In this simulation, the thresholds for the acoustic features and semantic primitives were empirically set to 0.786 and 840, respectively, in light of the fact that increasing the threshold does not improve performance.

In the second stage, we used the sequential floating forward selection (SFFS) to select the best features from the pre-screened feature set, on the grounds that SFFS is an iterative algorithm to evaluate the selected subset and combined effects of features and k-nearest-neighbor classifier during the evaluation process. Table 4.1 details the 22 acoustic features, four semantic primitives and two emotion dimensions that we used to construct the proposed three-layer model.

Table 4.1: Selected features of each layer for developing the three-layer model based multilingual emotion recognition system

Acoustic Feature		Semantic Primitive	Emotion Dimension
5 IS16 related features	Group	DARK	
maximum; mean rising slopes of the speech over all accentual phrases;	F0		
spectral tilt	Power Spectrum	HEAVY	VALENCE
total length of whole speech	Timing		
harmonic difference H1-H2	Voice Quality		
17 MSF related features	Group		
spectral centroid (SC) cross MF band: 1, 3; spectral slope (SSL) cross MF band: 1; spectral flatness cross MF band: 2;	Acoustic Frequency (AF) Domain	STRONG	AROUSAL
SC cross AF band: 2, 19, 28, 32; SSL cross AF band: 13, 22, 25; spectral skewness cross AF band 13, 23, 30; spectral kurtosis cross AF band:17, 27; spectral spread cross AF band: 25	Modulation Frequency (MF) Domain	WEAK	

4.2.2 Estimation and classification approaches

Adaptive neuro fuzzy inference systems (ANFIS) were first used as bridges over the three layers in order to estimate the emotion dimensions; ANFIS is a neural-fuzzy system based on neural networks and fuzzy systems that can efficiently model non-linear input and output relations by incorporating human knowledge with smaller root mean square errors [80]. Correspondingly, the nature of perception of speech emotion was fuzzy and vague [9]. Furthermore, the proposed three-layer model in this study also incorporated human knowledge from manual evaluations of semantic primitives and emotion dimensions that involve non-linear processing according to human emotion perception. Our earlier effort [7] has proved that ANFIS is an efficient approach for characterizing non-linear relations in this three-layer model that could be a benefit to the estimation of emotion dimensions. Subsequently, with features extracted in the valence and arousal emotional space, the performance of categorical classification was given by the logistic model trees (LMT).

4.2.3 Evaluation Metrics

First, the correlation coefficient(CC) and mean absolute error (MAE), between a system's estimations and human evaluations, are calculated as two metrics, in order to evaluate the performance of estimation of semantic primitives and emotion dimensions. In particular, the CC is merely a preferred metric to evaluate the performance of estimation of semantics primitives in the middle layer, in view of the fact that ANFIS used in a three-layer model captured nonlinear associations between input and output, where smaller MAEs might not definitely result in a good performance in estimation of valence and arousal.

Formally, X_n are the values of an emotion dimension estimated by a system, and the corresponding averaged values of an emotion dimension given by human estimators are Y_n. The CC and MAE are accordingly calculated as:

$$CC = \frac{\sum_1^N \left(X_n - \overline{X}\right)\left(Y_n - \overline{Y}\right)}{\sqrt{\sum_1^N \left(X_n - \overline{X}\right)^2 \sum_1^N \left(Y_n - \overline{Y}\right)^2}} \tag{4.2}$$

$$MAE = \frac{\sum_1^N |X_n - Y_n|}{N} \tag{4.3}$$

where $\overline{X}$ and $\overline{Y}$ are the mean values of X_n and Y_n, respectively. In addition, N is the

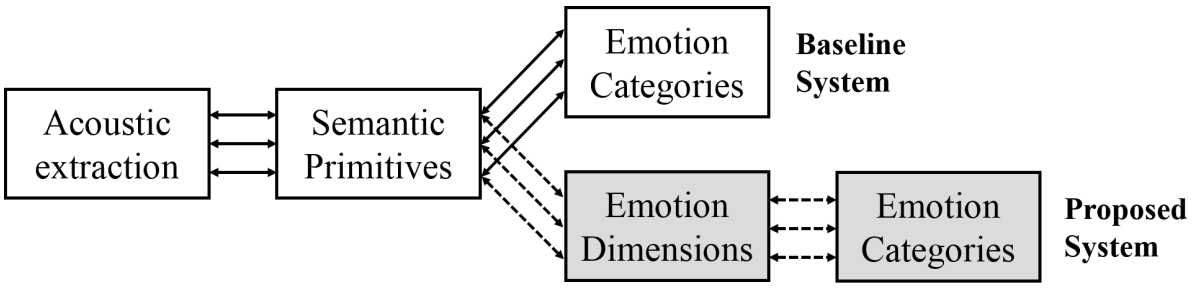


Figure 4.1: Comparison with emotion theories between baseline and proposed approaches.

number of utterances. Notably, CC assigns values that trend to 1 for a closer system's estimation to human evaluations; and MAE assigns values that trend to 0 for a better performance of a system's estimations. Second, the recall (c.f. Eq. 2.1), precision (c.f. Eq. 2.2), and F-measure (c.f. Eq. 2.3) are reported in terms of each emotional state for assessing the performance of categorical classification.

4.3 Experiment1: Comparison between emotion theories

This experiment was an attempt to quantify the performance of the proposed emotion theory in approaching multilingual SER. All results obtained in this section were presented by LOSO cross-validation using a mixed corpus made from the Fujitsu database, Berlin EmoDB and CASIA corpus.

Comparisons were accordingly carried out between the emotion theories of baseline and proposed approaches as demonstrated in Figure 4.1. Each of the two systems mentioned above was the same in acoustic features and semantic primitives as shown Table 4.1. In addition, all these features were transformed to [0,1] by the max-min normalization for training and testing the SER in multilingual scenarios.

Table 4.2: Confusion matrix for multilingual SER using the baseline emotion theory.

Baseline	Neutral	Happy	Anger	Sad
Neu	78	45	11	4
Hap	18	62	39	0
Ang	3	30	107	1
Sad	9	9	14	108

Table 4.3: Confusion matrix for multilingual SER using the proposed emotion theory.

Proposed	Neutral	Happy	Anger	Sad
Neu	119	2	8	9
Hap	23	66	31	0
Ang	5	10	126	0
Sad	16	0	0	124

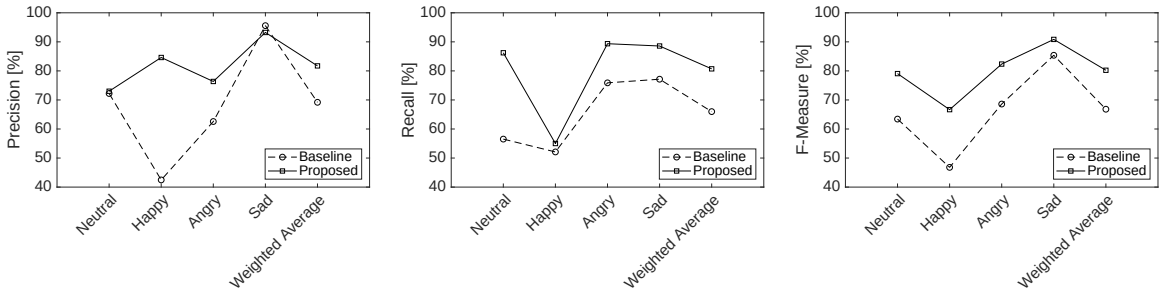


Figure 4.2: Classification results for multilingual SER obtained by different emotion theories of the baseline and proposed approaches.

Figure 4.2 details the results achieved by the baseline and proposed emotion theories, with precision, recall and F-Measure shown for each emotion. It is clear from Figure 4.2 that the proposed emotion theory consistently outperform baseline approach, yielding a better accuracy of 81.72% for averaged precision (baseline:69.19%), 80.71% for averaged recall (baseline: 65.99%), and 80.23% for averaged F-Measure (baseline: 66.81%). The confusion matrices are also demonstrated in Tables 4.2 and 4.3, respectively.

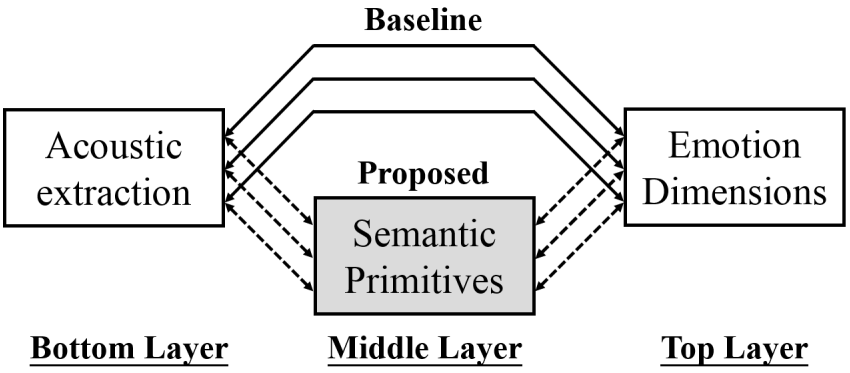


Figure 4.3: Comparison between baseline and proposed computational models.

4.4 Experiment2: Comparison between computational models

Differ from traditions that estimated emotion dimensions directly from acoustic features; this study introduced a three-layer model to improve the performance of estimation on emotion dimensions, suggesting that human subjects judge expressive speech by a small set of perception that is expressed by a small set of semantic primitives instead of directly from acoustic features. This experiment was hence to study the contribution of the three-layer model by comparing it to an acoustic model that was trained on the basis of acoustic features only, as shown in Figure 4.3. All these two systems conducted in this experiment were the same in collecting the acoustic features for an SER model (c.f. Table 4.1).

Figures 4.4 display the correlation coefficient and mean absolute error between system's output and human listeners' evaluations obtained using the baseline and proposed SER models. As can be seen from this figure, the proposed three-layer model achieved a better performance compared with that obtained by the baseline model, yielding a higher correlation coefficient and lower mean absolute error over both valence and arousal dimensions. Statistical test (one-way ANOVA) indicate that the differences of MAE between the models in terms of baseline and proposed turned out to be statistically significant for both valence ($F(1, 1074) = 63.3098, p < 0.0001$) and arousal ($F(11074) = 39.74, p < 0.0001$). These results could be interpreted as evidence for the effectiveness of the proposed three-layer model for mimicking human speech perception of emotion in multilingual scenarios.

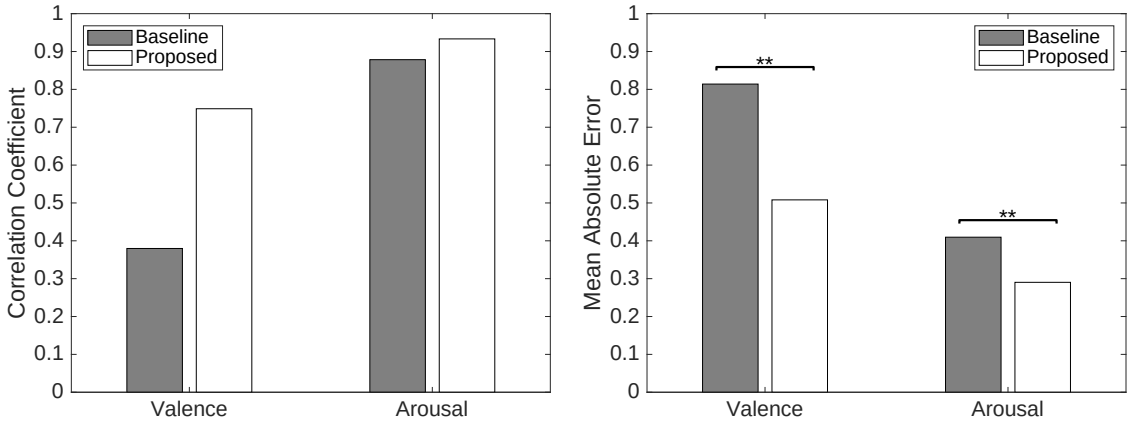


Figure 4.4: Correlation coefficient and mean absolute error between system's output and human evaluations for valence and arousal with respect to Baseline and Proposed SER models. $**$ indicate the estimations differ significantly ($p < 0.0001$) between models of baseline and the proposed.

4.5 Experiment3: Comparison between individual feature sets

This experiment attempted to show that the use of the proposed acoustic features and semantic primitives can benefit SER in the multilingual scenario. All results in this section were obtained by leave-one-speaker-out (LOSO) validation using a mixed corpus made from the Fujitsu database, Berlin Emo-DB, and CASIA dataset.

Comparisons were accordingly carried out in a two-fold manner. The first comparison is carried out on the domain of acoustic features with respect to the feature set of prosodic-related, spectral-related, and their combination, i.e., the proposed set, abbreviated as Proposed. Each of the aforementioned three feature sets was examined under condition no speaker normalization. The second comparison is an effort to study the effectiveness of chosen semantic primitives in the middle layer. In this regard, experiment on the raw set and not chosen semantic primitives were studied for comparison.

4.5.1 Effectiveness to acoustic features

Three systems were conducted by studying the optimal acoustic features individually relative to the prosodic-related, spectral-related and proposed using the SFFS algorithm. Besides, all these three system were the same in the collection of semantic primitives.

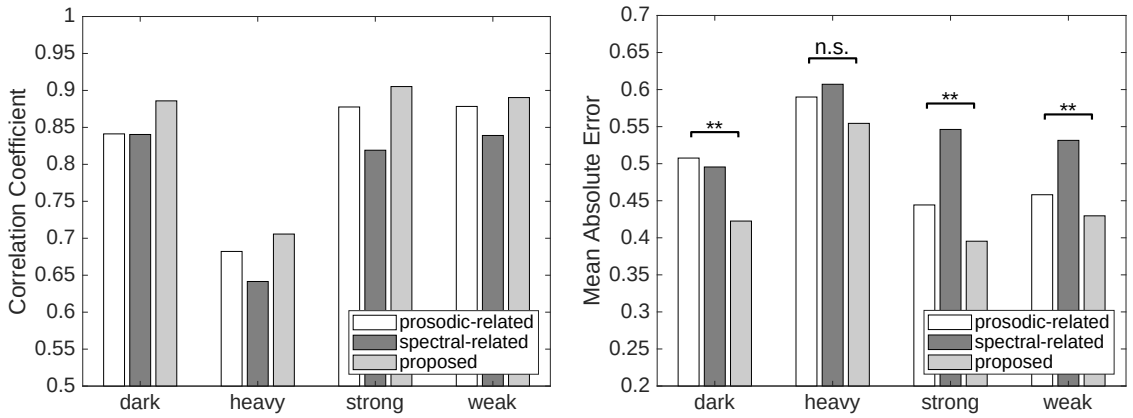


Figure 4.5: Estimation performance of semantic primitives obtained by multilingual emotion recognition systems using different features sets; $**$ indicate the estimations differ significantly ($p < 0.001$) between acoustic features of prosodic-related, spectral-related and the proposed; $n.s.$ indicate there is no significant statistical difference.

We first evaluate the performance obtained during estimation of semantic primitives. Results confirmed that the proposed acoustic features provided a better performance that was closer to human evaluations and outperformed the one obtained with prosodic-related and spectral-related features over four semantic primitives with respect to both the correlation coefficient and mean absolute error as shown in Figure 4.5.

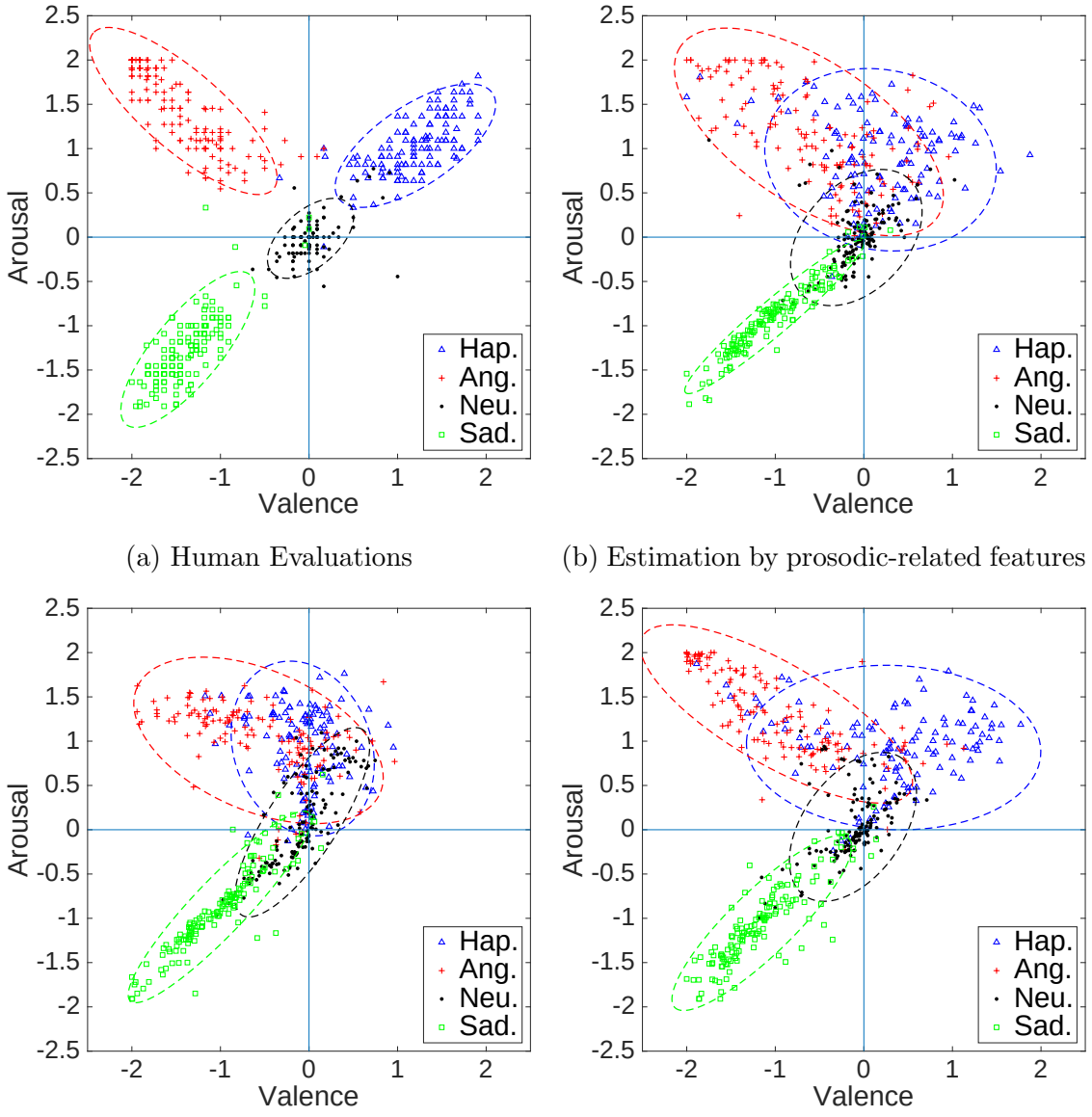
Figures 4.6 further display the scatter plots of manual evaluation and systems' estimation. The performance of valence and arousal estimation was quantified in terms of CC and MAE, and was demonstrated in Table 4.4. As can be seen, the proposed features always furnished the best performance yielding a greater CC and smaller MAE compared with those obtained by prosodic-related and spectral-related features.

Statistical test (one-way ANOVA) was performed between two systems's estimations on all emotional utterances.The proposed features yielded statistically significant improvements over the prosodic-related and spectral-related features individually for both valence: $F(2, 1608) = 19.87$, $p < 0.0001$ and arousal: $F(2, 1608) = 28.23$, $p < 0.0001$. This fact indicated that the estimation of emotion dimensions are benefit from our examined acoustic features.

Table 4.4: Estimation performance of emotion dimensions obtained by multilingual emotioin recognition systems using different features sets.

Feature	Valence			Arousal		
	Prosodic-related	Spectral-related	Proposed	Prosodic-related	Spectral-related	Proposed
CC	0.640	0.568	0.749	0.907	0.865	0.933
MAE	0.64	0.726	0.508 **	0.364	0.435	0.290 **

*Note. ** indicate that the estimation differ significantly between feature set of prosodic-related, spectral-relate, and proposed.*



(a) Human Evaluations

(b) Estimation by prosodic-related features

(c) Estimation by spectral-related features

(d) Estimation by the proposed features

Figure 4.6: Scatter plots of systems estimations of emotional utterances for mixed data (Fujitsu database, Berlin Emo-DB, and CASIA dataset) in 2D emotion space, obtained by a three-layer model incorporating feature set of prosodic-related, spectral-related, and Proposed.

In closing this subsection, Tables 4.5, 4.6 and 4.7 show three confusion matrix, providing the emotional classification performance in terms of recall, precision, and F-measure over LOSO cross-validation for different sets of features. As can be observed that classification results in terms of recall, precision, and F-measure turn out to receive a notable gain from the proposed features. As an aside, a one-way ANOVA analysis was conducted between two systems of classification results on 15 speakers (four speakers in CASIA dataset, ten speakers in Berlin Emo-DB, and one speaker in Fujitsu database) to test whether the proposed features significantly advance the task of multilingual SER.

We conducted two one-way ANOVA analysis between features of prosodic-related and proposed, and spectral-related and proposed, taking into account an effect of feature set. Results showed a significant difference, $F(1, 28) = 12.1594$, $p < 0.05$ for averaged precision, $F(1, 28) = 10.0299$, $p < 0.05$ for averaged recall, $F(1, 28) = 11.5225$, $p < 0.05$ for averaged F-measure, with the proposed features outperformed the prosodic-related features at recognizing multilingual speech emotions. Further, this analysis showed that the proposed features also significantly improved the SER performance compared with spectral-related features, $F(1, 28) = 12.2307$, $p < 0.05$ for averaged precision, and $F(1, 28) = 12.7435$, $p < 0.05$ for averaged recall, and $F(1, 28) = 14.8244$, $p < 0.05$ for averaged F-measure. In line with these findings, it showed that the proposed features outperformed the prosodic-related and spectral-related features at identifying speech emotions.

Table 4.5: Confusion matrix for using prosodic-related features only on the mixed emotional speech corpora of Fujitsu database, Berlin EmoDB amd CASIA corpus

Prosodic-related Features		LABELED EMOTION				
		Neutral	Happy	Anger	Sad	Total
PREDICTED EMOTION	Neutral	120	4	6	8	138
	Happy	35	45	39	0	119
	Anger	25	12	104	0	141
	Sad	24	0	0	116	140
	Total	204	61	149	124	538
	Precison	58.82%	73.77%	69.80%	93.55%	**74.04%**
	Recall	86.96%	37.82%	73.76%	82.86%	**71.56%**
	F-Measure	70.18%	50.00%	71.72%	87.88%	**70.73%**

Table 4.6: Confusion matrix for using spectral-related features only on the mixed emotional speech corpora of Fujitsu database, Berlin EmoDB amd CASIA corpus

Spectral-related Features		LABELED EMOTION				
		Neutral	Happy	Anger	Sad	Total
PREDICTED EMOTION	Neutral	100	21	4	13	138
	Happy	36	16	67	0	119
	Anger	22	16	102	1	141
	Sad	26	0	0	114	140
	Total	184	53	173	128	538
	Precison	54.35%	30.19%	58.96%	89.06%	**59.25%**
	Recall	72.46%	13.45%	72.34%	81.43%	**61.71%**
	F-Measure	62.11%	18.60%	64.97%	85.07%	**59.21%**

Table 4.7: Confusion matrix for using proposed features on the mixed emotional speech corpora of Fujitsu database, Berlin EmoDB amd CASIA corpus

Spectral-related Features		LABELED EMOTION				
		Neutral	Happy	Anger	Sad	Total
PREDICTED EMOTION	Neutral	119	2	8	9	138
	Happy	23	66	31	0	120
	Anger	5	10	126	0	141
	Sad	16	0	0	124	140
	Total	163	78	165	133	539
	Precison	73.01%	84.62%	76.36%	93.23%	**81.72%**
	Recall	86.23%	55.00%	89.36%	88.57%	**80.71%**
	F-Measure	79.07%	66.67%	82.35%	90.84%	**80.23%**

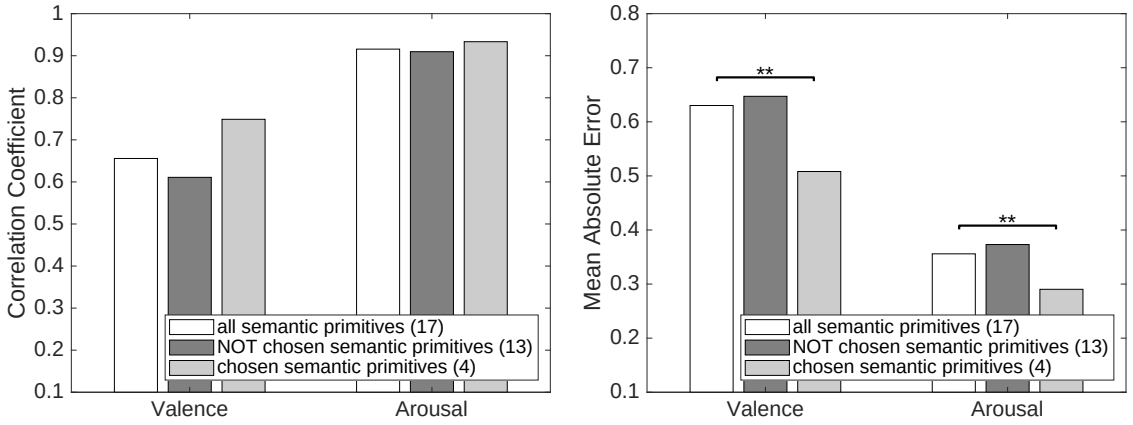


Figure 4.7: Estimation performance of emotion dimensions obtained by multilingual emotion recognition systems using different semantic primitives; ** indicate the estimations differ significantly ($p < 0.001$)between all 17 semantic primitives, not chosen 13 semantic primitives and 4 chosen semantic primitives.

4.5.2 Effectiveness to semantic primitives

Three systems were further conducted in this experiment on the basis of different semantic primitives, namely the full set of seventeen semantic primitives, the chosen semantic primitives and the not chosen ones. The chosen acoustic features were same in Table 4.1 for all these three systems. The four chosen semantic primitives achieved the best results that are closer to human evaluations, providing a higher CC and lower MAE over both valence and arousal dimensions. Statistically, the ANOVA found a significant main effect of MAE for feature selection on semantic primitives: Valence $F(2, 1608) = 9.02$, $p < 0.001$; Arousal $F(2, 1608) = 11.81$, $p < 0.001$ (see Figure 4.7). This results could be interpreted as evidence that the select semantic primitives advances the estimation of emotion dimensions in multilingual scenarios.

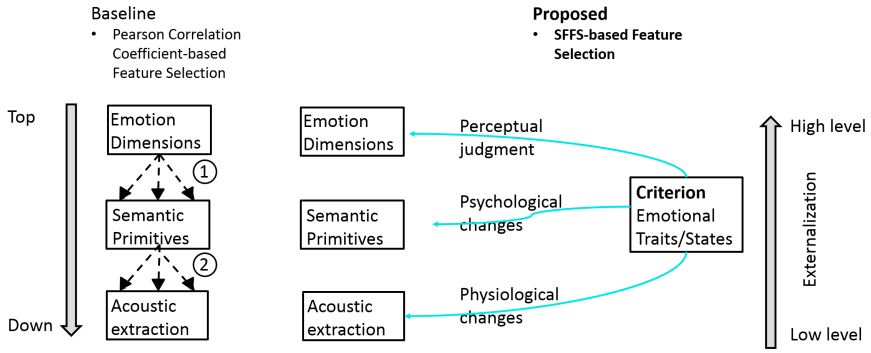


Figure 4.8: Comparison with methodologies to implement the three-layer model-based multilingual SER system.

4.6 Comparison between methodologies to implement a three-layer model

Selection of features to implementation of a three-layer model is one of the most important issue to mimic the human speech perception of emotion. As can seen from Figure 4.8, tradition define the optimal features in a three-layer model by a two-fold scheme, incorporating a feature selection method on the basis of the Pearson correlation coefficient (PCC) [4]. In spite of the promising results reported by the previous efforts, the limitation of PCC, however, has to be noted due to the fact that the PCC can naively capture linear relationship between features and target, but can not capture correlations that are not linear in nature. Human emotion perception, for instance, is vague, complex, and has multi-processes; it does not suffice to always use linear correlation to capture the association between acoustic feature and semantic primitives.

This study assumes that the emotional states in speech are externalized in the form of physiological changes that producing patterns of acoustic features; and accompanied by psychological changes that will affect perceptual judgments in such a way as to provide patterns of proximal percepts. Beyond the filter-based feature selection approach, this study hence used a wrapper-based method SFFS along with k-nearest-neighbor classifier during the evaluation process to define the best features taking into account its ability to study an optimal subset and combined effects of features, and can potentially achieve a better learning performance.

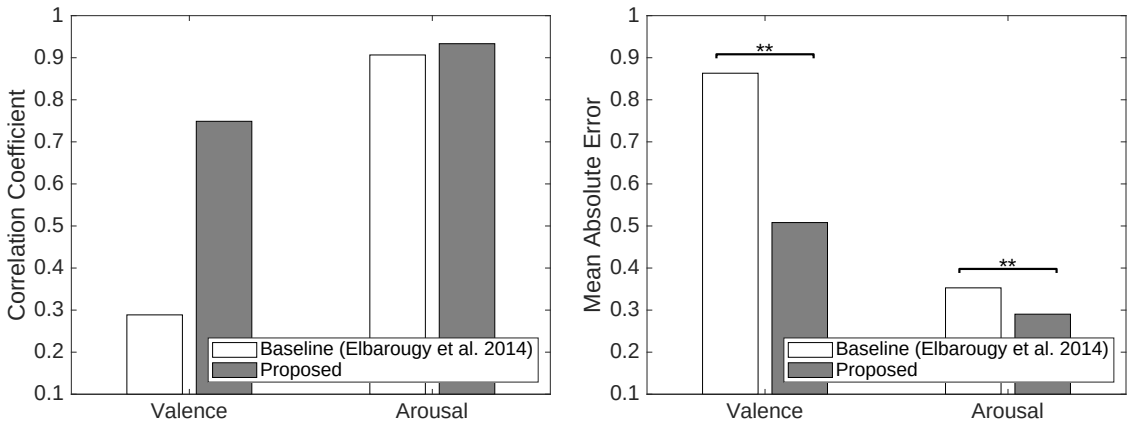


Figure 4.9: Estimation performance of emotion dimensions obtained by multilingual emotion recognition systems on the basis of different implementation methodologies with respect to a baseline and the proposed approach; ** indicate the estimations differ significantly ($p < 0.001$)between the feature selection algorithms of PCC and SFFS in multilingual scenarios.

The raw sets of 215 acoustic features and 17 semantic primitives to be investigated for PCC and SFFS are the same; the threshold in PCC was same in [4], settled to be 0.45. The results of estimation on emotion dimensions are compared with respect the the feature selection methods on the basis of PCC and SFFS, and as given in Figures 4.9. As can be seen, the SFFS achieved the higher values of CC either for valence or arousal. There are also significant estimation accuracy differences between PCC and SFFS with respect to MAE, Valence $F(1, 1072) = 74.18$, $p < 0.001$; and Arousal $F(1, 1072) = 11.64$, $p < 0.001$. This result showed the potential of the SFFS to define optimal features within a three-layer model for mimicking human speech perception of emotion.

Chapter 5

Multilingual SER System Evaluation

5.1 Introduction

The previous chapter detailed the implementation and validation of the three-layer model based multilingual SER system. In this chapter, we will study the question as how well the proposed proposed model perform on identifying multilingual speech emotional states. To this end, we first provide two investigations to study the ability to handle speaker variability and language variability. To study the first question, we adopted an approach to speaker normalization and compared it to the scenario in no speaker normalization as presented in Chapter 4. Regarding the second question, we performed an open data evaluation, i.e., cross-corpus evaluation , training in one corpus and test on a completely new database. Obtained results in this evaluation were compared with those of LOSO validation. Furthermore we compared our strategies with some relevant literature that using the same emotional speech corpus for SER as a reference.

5.2 Effectiveness to speaker variability

Speech emotion recognition is a very challenging task, one of the reason is that the acoustic features in emotional speech generally vary from one speaker to another. Recently, most researchers believed that speaker normalization (SN) advances the accuracies of SER [23, 62]. However, most of today's SER systems still suffer this difficulty due to the fact that it is not always possible to do speaker normalization for an acoustic SER model,

i.e., while there is no data available from the same speaker, those systems may appear to sharp fall off the recognition accuracy.

To quantify the effectiveness of the proposed multilingual SER system in this study to handle speaker variability in the acoustic domain, we herein adopted an approach to SN after [57], and compared it to the proposed approach in a scenario of no speaker normalization. In such stage, the features were mean and variance normalized within the scope of each speaker to compensate for speaker variations. Let $f_{u,v}(n)$ $(1 \leq k \leq N_{u,v})$ stand for the uth feature from speaker v with $N_{u,v}$ denoting its sample size, which in our case is the number of all available samples in the database from that speaker. The normalized feature $f'_{u,v}(n)$ processed by SN is defined as:

$$f'_{u,v}(n) = \frac{f_{u,v}(n) - \overline{f_{u,v}}}{\sqrt{\frac{1}{N_{u,v}-1}\sum_{m=1}^{N_{u,v}}(f_{u,v}(m) - \overline{f_{u,v}})^2}} \tag{5.1}$$

$$\overline{f_{u,v}} = \frac{1}{N_{u,v}} \sum_{n=1}^{N_{u,v}} f_{u,v}(n). \tag{5.2}$$

All results in this subsection were obtained by leave-one-speaker-out validation using mixed corpus made from Fujitsu database, Berlin EmoDB and CASIA corpus. Comparisons were accordingly carried out on the proposed acoustic set (c.f. Table 4.1) under two conditions of SN and NN.

We first evaluated the performance obtained during estimation of semantic primitives. The CC and MAE values for dark, heavy, strong, and weak were detailed in Figure 5.1. We conducted an ANOVA on two estimation results on semantic primitives with respect to the proposed acoustic features in scenarios of SN and NN, among 15 speakers from three emotional speech corpora. The results revealed no significant effect of speaker normalization on both CC and MAE over four semantics primitives (see Figure 5.1), suggesting that the proposed system in this study can hand in predicting semantic primitives speaker-independently, even for those who speak different languages. Furthermore, in the emotion dimensions estimation task, we also found that the effect of speaker normalization for valence in terms of CC, $F(1, 28) = 1.11$, $p = 0.30$, and MAE, $F(1, 28) = 0.03$, $p = 0.86$; and for arousal in terms of CC, $F(1, 28) = 0.39$, $p = 0.54$, and MAE, $F(1, 28) = 0.36$, $p = 0.55$, were also no significant. Note again that without

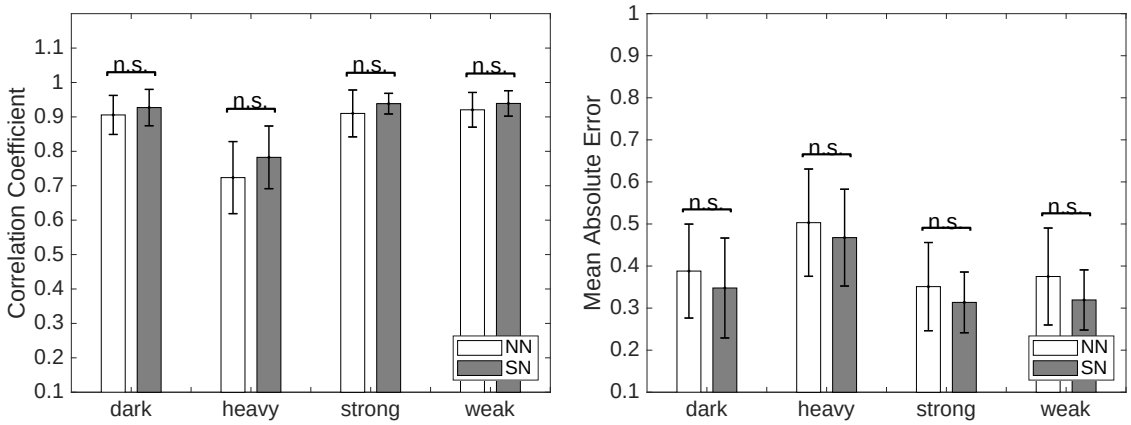


Figure 5.1: Estimation performance of semantic primitives by multilingual emotion recognition system using the proposed acoustic features with respect to SN and NN.

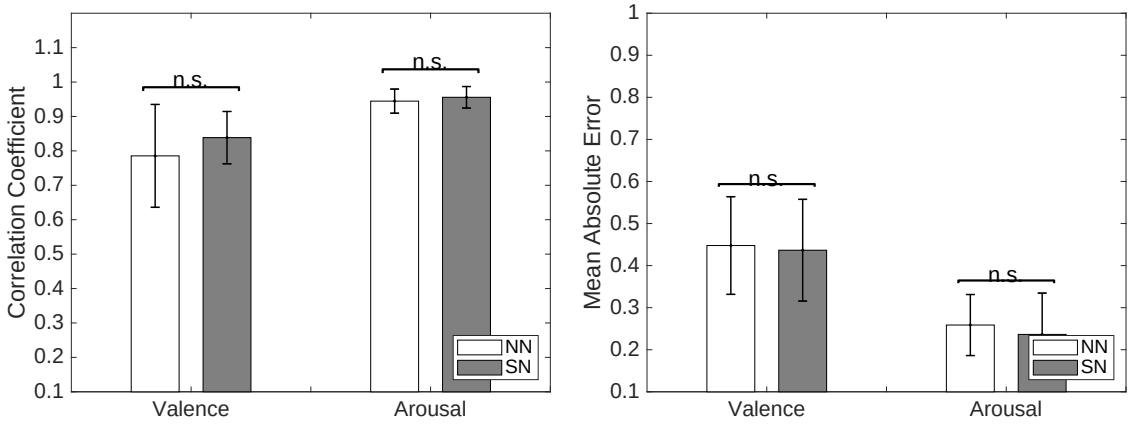


Figure 5.2: Estimation performance of emotion dimensions by multilingual emotion recognition system using the proposed acoustic features with respect to SN and NN.

speaker normalization, the estimation of emotion dimensions are comparable to those obtained in a scenario of SN (c.f. Figure 5.2)

In addition, we conducted one more one-way ANOVA analysis in the scenario of classification performance regarding the F-Measure, between SN and NN, taking into account an effect of speaker normalization. Figure 5.3 showed that for each emotional category within neutral, happy, anger, and sadness, the F-Measure achieved comparable and equal performance in both scenarios of SN and NN, indicating no significant difference. These results could be interpreted as evidence that the proposed system can work well speaker independently, in spite of the language being spoken.

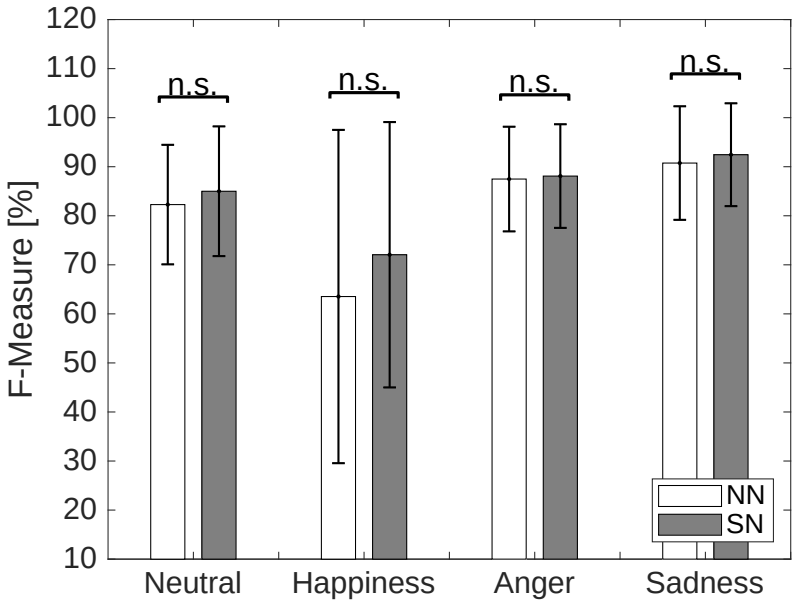


Figure 5.3: F-measure results on classification for each emotional category by multilingual emotion recognition system using the proposed acoustic features with respect to SN and NN.

5.3 Effectiveness to language variability

To further quantify the performance of our proposed strategies to multilingual SER, we also performed an open data evaluation, namely, training in one corpus and test on a completely new database. To this end, the three emotional corpora of Fujitsu database, Berlin Emo-DB, and CASIA dataset were used. All systems conducted in this experiment are same in collecting the acoustic features and semantic primitives for a three-layer model (see Table 4.1). Table 5.1 summarises the permutations over pairs of training and test data and corresponding performance, where three different combinations of the remaining corpora were used for testing each emotional corpus.

As can be seen, the performance obtained using Fujitsu database for testing is slightly higher than the two other corpora, this result was probably beneficial from the more natural and clear recordings in this database. The system trained on the Berlin Emo-DB achieved the highest average F-measure relative to the systems trained on CASIA dataset and their combination, reaching up to 85.5%. This performance might be on the grounds that Berlin Emo-DB and Fujitsu database have very similar

[1]Fujitsu database is a single speaker corpus

Table 5.1: Classification performance for three-layer model based emotion recognition systems over permutations in cross-corpus evaluation.

Test	Train	Classification Performance						
		Neutral	Happy	Anger	Sad	Recall	Precision	F-Measure
Fujitsu[1]	Berlin	95.00	67.50	90.00	95.00	85.70	87.40	**85.50**
	CASIA	85.00	45.00	92.50	100.00	80.00	82.30	**79.40**
	Berlin+CASIA	100.00	32.50	95.00	97.50	78.60	84.30	**76.00**
Berlin	Fujitsu	74.00	74.00	68.00	90.00	76.50	77.50	**76.50**
	CASIA	86.00	50.00	66.00	90.00	73.00	73.60	**72.50**
	Fujitsu+CASIA	76.00	64.00	78.00	96.00	78.50	78.50	**78.10**
CASIA	Fujitsu	57.40	37.90	56.90	84.00	61.10	61.70	**61.00**
	Berlin	66.20	48.30	35.30	60.00	54.00	59.50	**54.50**
	Fujitsu+Berlin	66.20	31.00	45.10	78.00	58.60	63.50	**60.50**

characteristics. They are all prototypical, acted corpora of emotional speech with a strong degree or intensity. Conversely, the CASIA dataset contains authentic emotions with an intensity varies gradually from weaker to stronger. On the other hand, the accuracy is in general lower for happiness than for the three other emotions. This result is consistent with Grimm's previous finding [9], and could be due to that the expression of happiness was significantly different for individual speakers. Additionally, we found no significant difference on averaged F-measure over three permutations while testing emotional speech from ten German speakers data in Berlin Emo-DB, $F(2, 27) = 0.3625$, $p = 0.6993$; and from four Chinese speakers in CASIA dataset, $F(2, 9) = 0.4536$, $p = 0.6491$. Together, these results suggested that our proposed system might not be modulated by different target languages.

As an aside, the classification results achieved with cross-corpus validation were somewhat decreased compared to those of LOSO validation. Whereas these results are within the relative error tolerance that can be expected in everyday situations, in view of the fact that emotional corpora usually vary with speakers, recording conditions, languages, or even labeled annotations of acted emotions. For instance, the degrees of each acted emotion in the Fujitsu, Berlin Emo-DB, and CASIA corpora are different; while the Japanese emotional speech is generally coloured by high intensities and the German emotional speech is moderately to highly coloured, the Chinese acted speech is closer to spontaneous speech whose degree changes from the lowest to highest smoothly. This is the main reason why the emotional speech from the CASIA dataset is slightly harder to recognize even by native Chinese speakers. In addition, it may not be sufficient to identify multilingual speech emotions merely using acoustic correlates from

speech, taking into consideration that emotions can also be transmitted from a speaker to a listener by other modalities, such as gestures, facial expressions, and so on.

The principal conclusion that can be drawn from this cross-corpus validation is that our proposed strategies may advance the task to SER, even for a scheme of training on one language and testing on a completely different one under an open speaker and language conditions. It can yield comparable results even the training and test speech vary in degrees in a certain emotion.

5.4 Comparison with related literature

5.4.1 comparison within our previous effort

The proposed system might be beneficial to multilingual speech emotion recognition only if it could reach a comparable performance to a monolingual recognizer. To facilitate this comparison, we also constructed three language-dependent monolingual emotion recognition systems following our proposed strategies. The classification performance of each proposed monolingual system is demonstrated in Table 5.2, and compared with that performed in multilingual scenarios. Furthermore, the experimental results given in a previous attempt [7] were also included in Table 5.2 for reference. All results presented were obtained by the LOSO cross-validation, apart from that of Fujitsu database, which was examined by 10-fold cross-validation on the grounds that it only involves one female speaker.

As shown in Table 5.2, our proposed approach advanced the performance of categorical classification on Fujitsu database for both monolingual and multilingual SER systems, and outperformed those obtained by [7]; Whereas, the averaged F-measure fell from 100% in monolingual scenario to 75% in multilingual scenarios, this is due to the fact classification in a monolingual case was performed by a 10-fold cross validation, training, and testing on close dataset. Conversely, the results obtained by the multilingual SER system were performed on an open data scheme in light of the fact that Fujitsu database has only one speaker.

Regarding the performance of categorical classification on the Berlin Emo-DB, obtainable results are significantly higher than that achieved by the referred multilingual

Table 5.2: Classification performance of each language by monolingual SER systems, multilingual systems, and approaches used in [7] for Fujitsu database, Berlin Emo-DB, and CASIA dataset.

		F-Measure			
		Monolingual SER		Multilingual SER	
		[7]	Proposed	[7]	Proposed
Fujitsu	Neutral	93.02	100.00	65.31	71.40
database	Happpiness	96.30	100.00	39.29	48.10
	Anger	94.87	100.00	75.51	83.50
	Sadness	97.44	100.00	90.91	98.70
	Weighted Avg.	95.75	**100.00**	68.10	**76.00**
Berlin	Neutral	82.69	96.97	82.00	91.84
Emo-DB	Happpiness	76.92	88.00	75.27	88.00
	Anger	84.91	89.11	87.85	88.66
	Sadness	90.91	98.00	92.00	95.24
	Weighted Avg.	83.86	**93.02**	84.28	**90.93**
CASIA	Neutral	52.63	64.66	67.78	72.06
dataset	Happpiness	0.00	36.73	11.43	59.02
	Anger	59.26	80.81	70.71	88.46
	Sadness	56.07	80.00	73.17	88.42
	Weighted Avg.	47.50	**68.60**	61.64	**78.51**

SER system after [7] ($p < 0.05$). Notably, it is interesting that we achieved a better performance on CASIA dataset in multilingual than monolingual scenario, since acoustic features in different languages generally varied from one to another.

For further analysis, the difference between our proposed multilingual and monolingual SER systems is not statistically significant, besides that of Fujitsu database which is not a fair condition for comparison as mentioned above. These findings stressed the fact that the proposed multilingual SER system could perform comparable results to those obtained by the language-dependent speech emotion recognizers.

5.4.2 comparison with other studies using the same corpora

As was reviewed in Table 5.3, the other studies targeting speech emotion recognition have produced substantial results. This subsection aims to demonstrate, discuss, and compare these results obtained in the state-of-the-art approaches to those of our strategy.

In light of the fact that Fujitsu database is a single speaker corpus, all results were shown using 10-fold cross-validation. A 92.5% overall recognition rate was obtained on

Table 5.3: Comparisons of classification performance with state-of-the-art works on Fujitsu database, Berlin Emo-DB, and CASIA dataset

Datasets (Validation Methods)	Tasks	Refs	Unweighted Accuracies
Fujitsu database (10-fold)	Monolingual	[4]	92.50
		Ours	**100.00**
	Multilingual	Ours	**98.10**
Berlin Emo-DB (LOSO)	Monolingual	[57]	89.90
		[62]	85.80
		Ours	**93.00**
	Multilingual	Ours	**91.00**
CASIA dataset (LOSO)	Monolingual	[81]	58.53
		Ours	**69.70**
	Multilingual	Ours	**78.28**

the Fujitsu database by exploiting 21 acoustic features in a three-layer model [4]. By comparison, a monolingual SER system conducted by our proposed approach substantially improved the classification performance, yielding a recognition accuracy up to 100%. On the other hand, a positive result of ours is that an overall recognition rate reached up to 98.1% in a multilingual scenario, resulting in an error reduction rate of 74.67% over the previous attempt [4]. We can see from these results that exploring efficient vocal features contributes to advancing the recognition and accuracies of all emotional categories.

There is numerous effort has been able to recognize emotion in Berlin Emo-DB. Regarding attempts that used combinations of different vocal features to improve the SER performance on speaker-independent tasks, 85.80% accuracy is achieved by exploring prosodic and spectral features in [62]. Furthermore, Vlasenko et al. [57] reported a comparatively improved accuracy of 89.9% by combining utterance-level and frame-level speech features. In contrast, our monolingual SER system presented in this paper showed an average recognition rate of 93.00% using 22 speech features, which is higher compared to the literature as mentioned earlier. More specifically, our proposed multilingual SER system can even furnish a better performance compared to the monolingual recognizers developed in [57, 62]. This might be due to the fact that three-layer model could be more suitable to model the process of human emotion

perception than the conventional models.

Among the works that able to recognize speech emotions in CASIA dataset, [81] once reported a recognition rate of 58.53% by LOSO validation in a monolingual scenario, using 384 acoustic features with speaker normalization, that is an absolute deterioration of 11.17% while comparing it to our proposed monolingual SER system. It should be noted that the multilingual system outperformed the monolingual one in CASIA case. On the one hand, this might be caused by the fact that the number of utterances for each emotional category in this corpus is not equally distributed, which in turn might limit the accuracy of SER. On the other hand, CASIA dataset turned out to receive better performance gain from a combination of Fujitsu database and Berlin Emo-DB, which again indicate that the proposed strategy provides a reasonable means of dealing speaker-independent SER tasks regardless of languages.

To stress the well-established ability of generalization, we carried out a further classification task for a new target language in English. We analyzed the SAVEE corpus using our multilingual emotion recognition system without training, and resulting in an average recognition rate of 43.5%. This was a significant achievement and somewhat comparable to that obtained by a monolingual SER system [82], training and testing under a 70-30% cross-validation, and reporting a 48.4% average recognition accuracy.

5.5 Summary

The purpose of this chapter is to provide an assessment of the proposed multilingual SER system in this study. To this end, three different aspects were presented, i.e., (1) how well can the proposed system perform on speaker variability; (2) how well can the system handing in identifying speech emotion in cross-lingual scenarios; (3) how well the proposed system outperformed the state of the art systems in SER.

First, the selected acoustic features in the Chapter 4 were adopted by two scenarios of speaker normalization and no speaker normalization. These two set of features were examined in a three-layer model by a LOSO cross-validation. The results reveal that the proposed multilingual SER in this study can yielded an comparable results to those obtained by a multilingual SER system trained with speaker normalization, indicating no significantly different by ANOVA analytic.

Second, we conducted an experiment by cross-lingual validation, where training a system using one languages, and testing on a completely new corpus. The results showed that the proposed system was not well modulated by languages, as training languages varies. Particularly, it was found that it can deal with recognition of emotion in speech with changing degrees/intensities.

Third, we recall some relevant studies on SER using the same emotional speech corpora. The proposed strategies showed promising performance from monolingual over multilingual scenarios. Most interestingly, beyond the studied three languages in this study, the proposed multilingual SER system even yielded a comparable results on a English dataset while comparing with a system that was trained in monolingual scenario.

Chapter 6

Conclusion

6.1 Summary

This research focused on designing a computational model for recognition of emotional states in the multilingual speech, by studying a multi-layered process of human speech perception of emotion, rather than attempting direct recognition of multilingual speech emotion by reducing the differences between different languages in most of the traditions. Following the concept of human emotion perception, this research successfully implemented a computational model that can handle multilingual SER independent of speakers and languages where three sub-goals were addressed involving (i) the recognition of discrete labels as well as gradual degrees of emotional speech; (ii) the clarification of appropriate features that can generalize well across multiple languages; (iii) the determination of a framework to explore relevant features to the implementation of the proposed computational model.

More specifically, this research studied common features in a three-layer model relative to acoustic features, semantic primitives, and emotional space. Three different languages in Japanese, German, and Chinese were analyzed. The proposed systems were validated in four scenarios, namely, cross-speaker SER, speaker normalization versus no speaker normalization, multilingual versus monolingual SER, and system cross-lingual SER versus human cross-lingual emotion evaluation. Results of principles relative to each of the three subgoals as mentioned above were followed.

(i) Estimating emotional state relative to the discrete labels as well as the changing degrees within spoken utterances. In this regard, a combined emotion theory brought together dimensional with categorical theories was determined to characterize speech emotion across multiple languages. The proposed theory was found to be beneficial for multilingual SER, providing 40.43% reduction of error rate when compared with obtained by the traditional categorical emotion theory. Besides, it yielded accurate estimation of emotion dimensions, with the high correlation coefficient of 0.75 and 0.93 relative to valence and arousal dimension between systems' estimations and human evaluations. The improvement in categorical classification was attributed to the fact of accurate estimation of emotion dimensions.

(ii) Studying appropriate features that can generalize well across languages to an accurate estimation of emotion dimensions. On the one hand, 22 combined acoustic features derived from prosodic and spectral domains were found to be universal among three languages. These features were compared to each of the two domains above individually, resulting in higher values of correlation coefficient and lower values of mean absolute error of estimation of valence and arousal dimensions. It was further found that combined features improved the SER accuracy yielding 32.46% and 51.53% reduction of error rate individually compared with that provided merely prosodic features, and spectral features. As expected, the modulation spectral features notably advanced the SER in terms of classification between positive and negative emotional state along the valence dimension where exploring prosodic features alone cannot perform well. On the other hand, four chosen semantic primitives were also compared with two sets of semantic primitives, i.e., a full set of 17 semantic primitives and thirteen semantic primitives that were not chosen. Again, the proposed four semantic primitives achieved the highest results on the estimation of emotion dimensions, indicating that they were well suited to handle multilingual SER than others.

(iii) Presenting a framework to determine relevant features to implement the proposed computation SER model. To this end, this research evaluated a wrapper-based feature selection algorithm, namely the SFFS, in capturing the associations within the process of human emotion perception. The filter-based Pearson correlation coefficient feature selection was referred to as a baseline that was commonly used in previous

studies. The proposed approach was demonstrated to result in significant improvement on the estimation of both the valence and arousal dimensions indicating that the effects of combined subsets were essential in the process of human speech emotion perception.

This research designed a computational model for recognition of emotional states independent of speakers and languages. In order to quantify the ability in handling speaker variability, speaker normalization was studied in the acoustic domain and compared to a scenario with no speaker normalization. It was found that there no significant difference exists between the obtained results of categorical classification over four emotional states in two conditions. Nonetheless, the recognition of happiness was found to be the most challenging task that was consistent with the previous study (Grimm, 2007) suggesting that the expression of happiness is highly speaker-dependent, long-term mood or other affective influences might cause that. Most interestingly, this research reported a comparable performance between monolingual and multilingual SER, from which most of the traditional studies cannot recognize well. It can be summarized that the three-layer model on the basis of the process of human speech emotion perception provides an excellent framework to SER in multilingual scenarios. Moreover, even in the event that without training, the recognition results of a new target language were still in the range of performance of human cross-lingual emotion evaluations, indicating the proposed three-layer multilingual SER system could mimic the process of human speech perception naturally.

6.2 Contribution

The most important contribution of this research is to present a framework for SER that has the capability to recognize emotional states independent of speakers and languages, even in a scenario without training for a new target language. The relevant features that can generalize well across different languages were demonstrated in this research in terms of 22 combined features in the acoustic domain, four semantic primitives in the perceptual domain, and the direction and distance from a neutral position to that of an emotional state in the emotional space domain. The commonalities among multiple languages were a general problem that most of the traditional studies suffered. More specifically, the contributions of this research can be summarized as follows:

◦ This study introduced a combined emotion theory, advancing the recognition of an emotional state in multilingual speech not only concerning categorical labels, but also the gradual transmitted degrees of a specific emotion. The changing degree within an emotional state is essential especially for health-care applications like recognition of anxiety, etc.

◦ Besides acoustic features extensively explored in the prosodic domain, this research presented powerful features from the spectral domain in terms of modulation spectral features that can be beneficial for different languages. The universal perceptual features in terms of semantic primitives were also demonstrated. These findings were also essential to other acoustic models in multilingual SER like DNN, providing insight into the universal underlying of the mechanism of human emotion recognition. Once DNN studied relevant features from those domains instead of raw observation in speech, the recognition accuracy is expected to be improved.

◦ This research reinforced the potential in evaluating the effect of combined subsets to capture the associations within the process of human speech emotion perception, advancing implementation of a perceptual model to emotion study formally.

6.3 Future works

1. **Automatic acoustic feature extraction** Acoustic feature extraction and selection is one the most important studied to be explored in the are od speech emotion recognition. This study has introduced effective features both from prosodic and spectral domains. However, acoustic features extraction in this work is semi-automatic, automated acoustic feature extraction algorithm need to be studied and promising for real-time speech emotion recognition.

2. **Study of universal proximal percepts in human judgment** On the one hand, in this study an attempt was made to apply a three-layer model to the speech perception of emotion in multilingual scenarios. The results showed a significant effect of interpretation over proximal percepts represented by semantic primitives. However, compared with the acoustic features from physiological domain, the amount of psychological-based features of semantic primitives is really

smaller. In this regard, one important study can be explored in future is to examine more universal perceptual features in this domain.

3. **Improvement to estimation on valence dimension** One the other hand, this study has suggested that accurate estimation of emotion dimensions exactly advances the performance of categorical classification. However, the accuracy of estimation on valence dimension is somehow lower. Whereas, the valence dimension is potential to study positive and negative emotions, such as happiness and anger. Some of the relevant literature has suggested that the valence dimension is a challenging task to be studied from speech, yet could be benefit from facial expression. Another attempt to be studied is to combine the audio and visual features to improve the recognition accuracy of emotional states for human-machine based friendly interactions in the real-life.

Bibliography

[1] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[2] Klaus R Scherer. Personality inference from voice quality: The loud voice of extroversion. *European Journal of Social Psychology*, 8(4):467–487, 1978.

[3] ChunFang Huang and Masato Akagi. A three-layered model for expressive speech perception. *Speech Communication*, 50(10):810–828, 2008.

[4] Reda Elbarougy and Masato Akagi. Improving speech emotion dimensions estimation using a three-layer model of human perception. *Acoustical science and technology*, 35(2):86–98, 2014.

[5] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[6] Andrew Ortony and Terence J Turner. What's basic about basic emotions? *Psychological review*, 97(3):315, 1990.

[7] Xingfeng Li and Masato Akagi. Multilingual speech emotion recognition system based on a three-layer model. In *INTERSPEECH*, pages 3608–3612, 2016.

[8] Hiroya Fujisaki. *Prosody, Models and Spontaneous Speech*, pages 27–42. Computing Prosody, Springer, 1996.

[9] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49(10-11):787–800, 2007.

[10] Rosalind Picard. *Affective Computing*. MA: MIT Press, 1997.

[11] J. Ma, H. Jin, L. Yang, and J. Tsai. Ubiquitous intelligence and computing: Third international conference. Proceedings (Lecture Notes in Computer Science), Springer-Verlag, New York, Inc., Secaucus, NJ, USA, 2006.

[12] Suryannarayana Chandaka, Amitava Chatterjee, and Sugata Munshi. Support vector machines employing cross-correlation for emotional speech recognition. *Measurement*, 42(4):611–618, 2009.

[13] C. Jones and J. Sutherland. Acoustic emotion recognition for affective computer gaming. In *Affect and emotion in human-computer interaction 2008*, volume 4868, pages 209–219, 2009.

[14] Stavros Ntalampiras, Ilyas Potamitis, and Nikos Fakotakis. An adaptive framework for acoustic monitoring of potential hazards. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009:13, 2009.

[15] Björn Schuller, Gerhard Rigoll, and Manfred Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 1, pages I–577. IEEE, 2004.

[16] Julia Hirschberg, Stefan Benus, Jason M Brenier, Frank Enos, Sarah Friedman, Sarah Gilman, Cynthia Girand, Martin Graciarena, Andreas Kathol, Laura Michaelis, et al. Distinguishing deceptive from non-deceptive speech. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[17] Rongqing Huang and Changxue Ma. Toward a speaker-independent real-time affect detection system. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 1, pages 1204–1207. IEEE, 2006.

[18] Chun-Fang Huang, Donna Erickson, and Masato Akagi. Comparison of japanese expressive speech perception by japanese and taiwanese listeners. *Journal of the Acoustical Society of America*, 123(5):3323, 2008.

[19] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[20] Klaus R Scherer. Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1-2):227–256, 2003.

[21] Mohit Shah, Chaitali Chakrabarti, and Andreas Spanias. Within and cross-corpus speech emotion recognition using latent topic model-based features. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):4, 2015.

[22] Peng Song, Wenming Zheng, Shifeng Ou, Xinran Zhang, Yun Jin, Jinglei Liu, and Yanwei Yu. Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization. *Speech Communication*, 83:34–41, 2016.

[23] Bjorn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wollmer, Andre Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll. Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing*, 1(2):119–131, 2010.

[24] Yuan Zong, Wenming Zheng, Tong Zhang, and Xiaohua Huang. Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression. *IEEE Signal Processing Letters*, 23(5):585–589, 2016.

[25] Silvia Monica Feraru, Dagmar Schuller, et al. Cross-language acoustic emotion recognition: An overview and some tendencies. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 125–131. IEEE, 2015.

[26] Kun Han, Dong Yu, and Ivan Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In *Fifteenth annual conference of the international speech communication association*, 2014.

[27] Mohamed R Amer, Behjat Siddiquie, Colleen Richey, and Ajay Divakaran. Emotion detection in speech using deep networks. In *ICASSP*, pages 3724–3728. Citeseer, 2014.

[28] WQ Zheng, JS Yu, and YX Zou. An experimental study of speech emotion recognition based on deep convolutional neural networks. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 827–831. IEEE, 2015.

[29] Jun Deng, Zixing Zhang, Erik Marchi, and Bjorn Schuller. Sparse autoencoder-based feature transfer learning for speech emotion recognition. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 511–516. IEEE, 2013.

[30] Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[31] Xiao Han, Reda Elbarougy, Masato Akagi, Junfeng Li, and Thi Duyen Ngo. A study on perception of emotional states in multiple languages on valence-activation approach. In *2015 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP'15)*. 2015 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP'15), 2015.

[32] Carroll E Izard. *The face of emotion.* Appleton-Century-Crofts, 1971.

[33] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[34] Robert W Frick. Communicating emotion: The role of prosodic features. *Psychological Bulletin*, 97(3):412, 1985.

[35] Renée Van Bezooijen, Stanley A Otto, and Thomas A Heenan. Recognition of vocal expressions of emotion: A three-nation study to identify universal characteristics. *Journal of Cross-Cultural Psychology*, 14(4):387–406, 1983.

[36] Klaus R Scherer, Rainer Banse, Harald G Wallbott, and Thomas Goldbeck. Vocal cues in emotion encoding and decoding. *Motivation and emotion*, 15(2):123–148, 1991.

[37] KR Scherer. Universality of emotional expression. *Encyclopedia of human emotions*, 2:669–674, 1999.

[38] Renee Van Bezooijen. *Characteristics and recognizability of vocal expressions of emotion*, volume 5. Walter de Gruyter, 2011.

[39] Tuomas Eerola, Olivier Lartillot, and Petri Toiviainen. Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In *ISMIR*, pages 621–626, 2009.

[40] Harold Schlosberg. Three dimensions of emotion. *Psychological review*, 61(2):81, 1954.

[41] Hugo Lövheim. A new three-dimensional model for emotions and monoamine neurotransmitters. *Medical hypotheses*, 78(2):341–348, 2012.

[42] Hao Hu, Ming-Xing Xu, and Wei Wu. Gmm supervector based svm with spectral features for speech emotion recognition. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–413. IEEE, 2007.

[43] Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan. Support vector regression for automatic recognition of spontaneous emotions in speech. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–1085. IEEE, 2007.

[44] Theodoros Giannakopoulos, Aggelos Pikrakis, and Sergios Theodoridis. A dimensional approach to emotion recognition of speech from movies. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 65–68. IEEE, 2009.

[45] Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech communication*, 48(9):1162–1181, 2006.

[46] André Stuhlsatz, Christine Meyer, Florian Eyben, Thomas Zielke, Günter Meier, and Björn Schuller. Deep neural networks for acoustic emotion recognition: raising

the benchmarks. In *Acoustics, speech and signal processing (ICASSP), 2011 IEEE international conference on*, pages 5688–5691. IEEE, 2011.

[47] Martin Wöllmer, Moritz Kaiser, Florian Eyben, BjöRn Schuller, and Gerhard Rigoll. Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2):153–163, 2013.

[48] Samuel Kim, Panayiotis G Georgiou, Sungbok Lee, and Shrikanth Narayanan. Real-time emotion detection system using speech: Multi-modal fusion of different timescale features. In *Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on*, pages 48–51. IEEE, 2007.

[49] Narichika Nomoto, Masafumi Tamoto, Hirokazu Masataki, Osamu Yoshioka, and Satoshi Takahashi. Anger recognition in spoken dialog using linguistic and para-linguistic information. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[50] H Teager. Some observations on oral air flow during phonation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(5):599–601, 1980.

[51] James F Kaiser. On a simple algorithm to calculate the'energy'of a signal. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 381–384. IEEE, 1990.

[52] Mohammed Abdelwahab and Carlos Busso. Supervised domain adaptation for emotion recognition from speech. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5058–5062. IEEE, 2015.

[53] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001.

[54] Louis Ten Bosch. Emotions, speech and the asr framework. *Speech Communication*, 40(1-2):213–225, 2003.

[55] Frank Dellaert, Thomas Polzin, and Alex Waibel. Recognizing emotion in speech. In *Fourth International Conference on Spoken Language Processing*, 1996.

[56] Marc Schröder and Roddy Cowie. Issues in emotion-oriented computing-towards a shared understanding. In *Workshop on Emotion and Computing at KI*. Citeseer, 2006.

[57] Bogdan Vlasenko, Björn Schuller, Andreas Wendemuth, and Gerhard Rigoll. Combining frame and turn-level information for robust recognition of emotions within speech. In *Proc. INTERSPEECH 2007, Antwerp, Belgium*, 2007.

[58] Ricardo A Calix, Mehdi A Khazaeli, Leili Javadpour, and Gerald M Knapp. Dimensionality reduction and classification analysis on the audio section of the semaine database. In *Affective computing and intelligent interaction*, pages 323–331. Springer, 2011.

[59] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623, 2003.

[60] Jacob Benesty, M Mohan Sondhi, and Yiteng Huang. *Springer handbook of speech processing*. springer, 2007.

[61] Martijn Goudbeek and Klaus Scherer. Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America*, 128(3):1322–1336, 2010.

[62] Siqing Wu, Tiago H Falk, and Wai-Yip Chan. Automatic speech emotion recognition using modulation spectral features. *Speech communication*, 53(5):768–785, 2011.

[63] Joel Robert Davitz. *The communication of emotional meaning*. McGraw Hill, 1964.

[64] Christer Gobl and Ailbhe Nı Chasaide. The role of voice quality in communicating emotion, mood and attitude. *Speech communication*, 40(1-2):189–212, 2003.

[65] Yongwei Li, Junfeng Li, and Masato Akagi. Contributions of the glottal source and vocal tract cues to emotional vowel perception in the valence-arousal space. *The Journal of the Acoustical Society of America*, 144(2):908–916, 2018.

[66] Zhi Zhu, Ryota Miyauchi, Yukiko Araki, and Masashi Unoki. Contribution of modulation spectral features on the perception of vocal-emotion using noise-vocoded speech. *Acoustical Science and Technology*, 39(6):379–386, 2018.

[67] Humberto Pérez-Espinosa, Carlos A Reyes-García, and Luis Villaseñor-Pineda. Acoustic feature selection and classification of emotions in speech using a 3d continuous emotion model. *Biomedical Signal Processing and Control*, 7(1):79–87, 2012.

[68] Dongrui Wu, Thomas D Parsons, and Shrikanth S Narayanan. Acoustic feature analysis in speech emotion primitives estimation. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[69] Marc Schröder, Roddy Cowie, Ellen Douglas-Cowie, Machiel Westerdijk, and Stan Gielen. Acoustic correlates of emotion dimensions in view of speech synthesis. In *Seventh European Conference on Speech Communication and Technology*, 2001.

[70] Khiet P Truong and Stephan Raaijmakers. Automatic recognition of spontaneous emotions in speech using acoustic and lexical features. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 161–172. Springer, 2008.

[71] Special Eurobarometer. Europeans and their languages. *European Commission*, 2006.

[72] Kurt Braunmüller and Christoph Gabriel. *Multilingual individuals and multilingual societies*, volume 13. John Benjamins Publishing, 2012.

[73] Angela De Bruin, Barbara Treccani, and Sergio Della Sala. Cognitive advantage in bilingualism: An example of publication bias? *Psychological science*, 26(1):99–107, 2015.

[74] Klaus R Scherer. *On social representations of emotional experience: Stereotypes, prototypes, or archetypes?* Hogrefe & Huber Publishers, 1992.

[75] Hideki Kawahara. Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoustical science and technology*, 27(6):349–353, 2006.

[76] Zhi Zhu, Ryota Miyauchi, Yukiko Araki, and Masashi Unoki. Modulation spectral features for predicting vocal emotion recognition by simulated cochlear implants. In *INTERSPEECH*, pages 262–266, 2016.

[77] Masato Akagi, Xiao Han, Reda Elbarougy, Yasuhiro Hamada, and Junfeng Li. Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pages 1–10. IEEE, 2014.

[78] Hiroki Mori, Tomoyuki Satake, Makoto Nakamura, and Hideki Kasuya. Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics. *Speech Communication*, 53(1):36–50, 2011.

[79] Margarita Kotti and Fabio Paternò. Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema. *International journal of speech technology*, 15(2):131–150, 2012.

[80] J-SR Jang. Anfis: adaptive-network-based fuzzy inference system. *IEEE transactions on systems, man, and cybernetics*, 23(3):665–685, 1993.

[81] Linlin Chao, Jianhua Tao, Minghao Yang, and Ya Li. Improving generation performance of speech emotion recognition by denoising autoencoders. In *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*, pages 341–344. IEEE, 2014.

[82] Maxim Sidorov, Christina Brester, Wolfgang Minker, and Eugene Semenkin. Speech-based emotion recognition: Feature selection by self-adaptive multi-criteria genetic algorithm. In *LREC*, pages 3481–3485, 2014.

Publications

Journal Paper

[1] Xingfeng LI, Masato AKAGI, "Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model, Speech Communication, Vol. 110, July 2019, pp.1–12.

[2] Zhizhao PENG, Xingfeng LI, Zhi ZHU, Masashi UNOKI, Jianwu DANG, Masato AKAGI, "Auditory Model as Front-ends for Speech Emotion Recognition Using 3D Convolution and Attention-based Sliding Recurrent Network," Journal of Transactions on Affective Computing. (under review)

International Conference

[3] Xingfeng LI, Masato AKAGI, "The Contribution of Acoustic Features Analysis to Model Emotion Perceptual Process for Language Diversity," Interspeech 2019 (to appear).

[4] Xingfeng LI, Masato AKAGI, "A Three-Layer Emotion Perception Model for Valence and Arousal-Based Detection from Multilingual Speech," Interspeech 2018, pp. 3643–3647, 2018.

[5] Xingfeng LI, Masato AKAGI,"Multilingual Speech Emotion Recognition Using A Three-Layer Model," Interspeech 2016, pp. 3608–3612, 2016.

[6] Xingfeng LI, Masato AKAGI, "Maximal Information Coefficient and Predominant Correlation-Based Feature Selection Toward A Three-Layer Model for Speech

Emotion Recognition," APSIPA ASC 2018, pp. 1428-1434.

[7] Xingfeng LI, Masato AKAGI, "Automatic Speech Emotion Recognition in Chinese Using A Three-Layered Model in Dimensional Approach," NCSP 2016, pp. 17–320, 2016.

[8] Xingfeng LI, Masato AKAGI, "Toward Improving Estimation Accuracy of Emotion Dimensions in Bilingual Scenario Based on Three-Layered Model," OCOCOSDA 2015, pp. 21–26, 2015.

[9] Xingfeng LI, Masato AKAGI, "Improving Estimation Accuracy of Dimension Values for Speech Emotion in Bilingual Cases Using a Three-layered Model,"Proceedings of the auditory research meeting 2015, pp. 577–581, 2015.

Domestic Conference

[10] Xingfeng LI, Masato AKAGI, "Study on Estimation of Bilingual Speech Emotion Dimensions Using A Three-Layered Model," ASJ Autumn 2015, 1-Q-39, pp. 305–308, 2015.

[11] Xingfeng LI, Zhi ZHU, Masato AKAGI, "Acoustic feature selection for improving estimation of emotions using a three layer model," ASJ Spring 2017, 1-Q-14, pp. 117–120, 2017.

[12] Xingfeng LI, Masato AKAGI, "Multilingual emotion recognition from speech using a three layer model," ASJ Autumn 2017, 1-R-33, pp. 235–238, 2017.

[13] Xingfeng LI, Masato AKAGI, "Toward Automatic Multilingual Emotion Detection in 2D Space Using A Three-Layer Model," ASJ Spring 2018, 2-Q-9, pp. 159–162, 2018.

[14] Xingfeng LI, Masato AKAGI, "Acoustic Feature Selection Toward A Three-Layer Emotion Perception Model," ASJ Autumn 2018, 2-Q-10, pp. 1061–1064, 2018.

[15] Xingfeng LI, Masato AKAGI, "Acoustic Features Analysis to Model Emotion Perceptual Process for Language Diversity," ASJ Autumn 2019 (to appear).