Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model

Xingfeng Li*, Masato Akagi

Graduate School of Advanced Science and Technology, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, 923-1292, Japan

Abstract

This study presents a scheme for multilingual speech emotion recognition. Determining the emotion of speech in general relies upon specific training data, and a different target speaker or language may present significant challenges. In this regard, we first explore 215 acoustic features from emotional speech. Second, we carry out speaker normalization and feature selection to develop a shared standard acoustic parameter set for multiple languages. Third, we use a three-layer model composed of acoustic features, semantic primitives, and emotion dimensions to map acoustics into emotion dimensions. Finally, we classify the continuous emotion dimensional values into basic categories by using the logistic model trees. The proposed approach was tested on Japanese, German, Chinese, and English emotional speech corpora. The recognition performance was examined and enhanced by cross-speaker and cross-corpus evaluation, and stressed the fact that our strategy is particularly suited for the task of multilingual emotion recognition even with a different speaker or language. The experimental results were found to be reasonably comparable with those of monolingual emotion recognizers as a reference.

Keywords: Multilingual emotion recognition, Human emotional perception, Emotional space, Three-layer model

1. Introduction

Identifying an emotional state from human voices based on speech emotion recognition (SER) has been increasingly turned into a principal focus within the affective computing research for interpreting the semantics of a spoken utterance. The purpose is to enable a machine with sufficient intelligence to recognize human speech not only regarding what is said, but also how it is expressed. SER is promising for many potential applications, one of which refers to a call-center service. The system can provide a user-friendliness response to a customer upon identifying emotions from his or her voice [1]. On the other hand, fatigue can be detected from a driver's voice by a car-board system and the driver can be alerted to ensure a safe driving [2]. Likewise, SER is potential for giving feedback in games and human-robot interactions [3, 4]. Other natural human-computer interactions such as web-movies [5], health care systems [6] and "Affective Mirror" [7] can also be enriched by recognition of emotions in a speaker's voice.

These days, psychology research has proven that human can easily perceive emotions in speech cross-language, even they do not understand the verbal content being spoken [8, 9]. With the gain in accuracies and high generalization abilities of current SER systems to be expected, increasing interest on SER has lately been shifting from monolingual to multilingual scenarios to be able to recognize speech emotions cross-language such as any human listener can easily react to. On the one hand, this interest can incredibly enhance the full development of emotion systems in a real-world-context. On the other hand, it allows for investigating similarities between languages which in turn can remedy the issue of data sparsity in training emotion models by combining more speech instances from different databases.

A number of effort has been done to be able to identify emotional state from speech and report substantial recognition results [10, 11, 12]. Other examples of relevant research include [13, 14]. Nevertheless, most of these have focused on monolingual emotion classification of a specific language, such as English, Chinese, German, and French, and so on. However, it was found that the best vocal features for SER in these works usually differ from one language to another. In such scenario, changing a source language requires reselecting a set of optimal acoustic features and retraining a system, which in turn stresses the fact that adapting these monolingual SER systems into multilingual SER tasks is still a challenge.

Recently, a number of studies have been developed to approach the problem of variations among different languages. Some attempts have been focused on examining many vocal cues [15], or studying different feature normalization and selection algorithms [16, 17], or combining the different acoustic models or classifiers [18, 19]. In [20], authors focused on transferring adaption schemes to make the instances among joint corpora 'similar'. However, an appealing knowledge based on human speech emotion perception has clarified that, as an alternative to these adaption approaches, commonalities and differences can be identified in a valence (how pleasant or

* Corresponding author
Email addresses: lixingfeng@jaist.ac.jp (Xingfeng Li), akagi@jaist.ac.jp (Masato Akagi)

unpleasant an emotional state is) and arousal (how relaxed or aroused an emotional state is) space [21, 22]. As revealed, the directions and distances from a position of a neutral voice to that of another emotional state are common between languages; however, the positions of neutral speech vary from one language to another. This knowledge allows for determining the common perceptual features for different languages, and advancing the task of SER in the multilingual scenario.

This study originally takes one step to incorporate this human-perception-inspired knowledge for speech emotion recognition under an assumption that the proposed system has the capability to estimate emotion dimensions across multiple languages accurately. To this end, this study still faces two challenges: i) which model is appropriate to recognize and predict emotion; ii) what are the best vocal features for SER tasks.

To the first issue, the main focus is how to map acoustic correlates into emotion dimensions. Over the past decade, there has been a great deal of literature on predicting emotion dimensions from different speech feature subsets using diverse classifiers such as fuzzy inference systems, support vector regressions, and k-Nearest Neighbour, etc. [23, 24, 25]. However, the limitation of these works lies in the fact that performance has been poor regarding valence. Besides these aforementioned classical estimators, the current study is focusing on adopting new models to gain the accuracy of estimation on the valence dimension, such as deep neural network [18], and long short-term memory [26]. Likewise, the estimations were promising for arousal and dominance, while the obtained result for valence was required to be improved. These approaches to SER treated human emotion perception as two-layer processes that predict emotion dimensions directly from acoustic correlates. This framework may not match the cognitive processes utilized to judge emotions as humans do.

As Scherer depicted [27], in an adapted version of the Brunswick lens model [28], the emotions are transmitted from a speaker to a listener by multiple modalities, and the listener perceives emotions as a multi-layer process. Other researchers have further reinforced this conclusion, for instance, Huang and Akagi proposed a multi-layered model to approach human expressive speech underlying and demonstrated that human subjects judge emotions by a small set of perceptions that are expressed by semantic primitives instead of directly from acoustic features [29], where low arousal and negative valence speech, (such as sadness) in general easily make an impression on listeners with dark and heavy feelings, but high arousal and positive valence speech (pleasant or happiness) is oftentimes uttered in a bright and well-modulated way. Other examples of literature treat human emotion perception as a multi-layer process also include [30], which aimed to accurately estimate emotion dimensions based on a three-layer model, consisting of acoustic features, semantic primitives, and emotion dimensions. Most interestingly, it was shown that this three-layer model significantly advanced the accuracies of estimation on emotion dimensions, particularly for valence dimension. This human-perceptual-based strategy inspired our study. We originally examined this three-layer model for multilingual SER tasks, and verified that it is well suited for mimicking the processing of human speech emotion perception across languages [31].

The second issue to be considered is the extraction of the best vocal features that can efficiently work for estimation of emotion dimensions. Most literature commonly asserts that prosodic features such as energy, fundamental frequency, and duration often deliver many emotional cues [32, 33, 34, 35]. Our previous effort to multilingual SER has also been put into developing effective acoustic sets from this domain [31]. Despite the substantial achievement reported, three restrictions from previous work still limit the full development of SER in the multilingual scenario.

First, speech emotion requires full representation, that may not be settled yet. Even though prosodic features formed the efficient feature type for predicting arousal dimension, however, associations of these features to valence dimension have generally been observed to be weak and limited. Nonetheless, valence is promising for distinguishing emotions with similar arousal state but differing in emotional categories, such as happiness and anger. However, there is still less evidence for the contribution of vocal parameters to this dimension.

Second, the previous study to SER focused on relations between acoustic features and semantic primitives, and semantic primitives and emotion dimensions by examing a full set of features without taking into consideration selection of best ones. Nevertheless, the functional mapping used in the three-layer model is fuzzy inference system, that can be defined as a nonlinear mapping from an input space into an output space. In such scenario, optimal features can be advantageous to estimation, conversely, a full set of features holding irrelevant and redundant features that may reduce the estimation accuracy.

Third, as an aside, the early work on assessing the performance of emotion systems using a 10-fold cross validation on the same training and test instances in a specific corpus. Other inherent mismatches between training and test data, such as different speakers or languages have not been investigated yet.

To solve each of these three problems that stemmed from our previous study. We first hypothesized that the combination of features from prosodic and spectral domains could improve the estimation performance of valence and arousal. Prosodic features were first decided on the grounds that these features are advantageous for distinguishing low and high arousal emotions in accordance with human perceptions [37, 38]. This study examined prosodic features are the same in our earlier attempt, and have been successfully used in the study of SER [31]. In addition, we examined the spectral features in view of the fact that these features are generally treated as strong correlates of the shape of the vocal tract and the rate of change in articulator movement [39, 40], that varies from one emotional state to another [10]. It was further reported that the valence dimension is also reflected in the acoustic correlates of spectral cues [39, 41]. Beyond the conventional and most

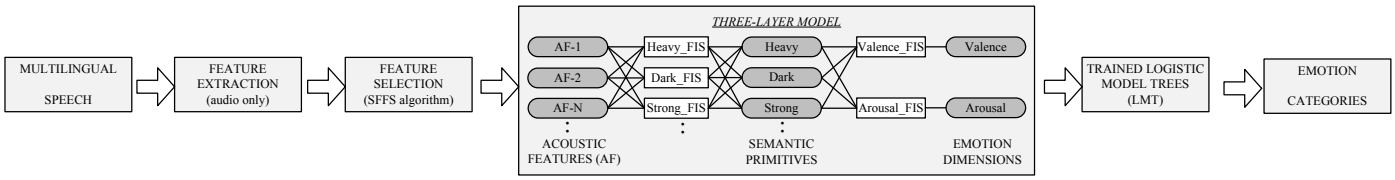


Figure 1: Block diagram of the multilingual SER system, illustrating the methodology proposed in this study, from preprocessing the emotional speech (feature extraction and sequential floating forward selection (SFFS)) to predicting emotional dimensions over recognized emotional categories.

common spectral features, such as Mel-frequency cepstral coefficients and perceptual linear predictive coefficients that were analyzed in terms of only short time frames in [42], our work takes one step on examining modulation spectral features, toward investigating the significant longer temporal-behaviour information used by human listeners and trying to gain insight into perceptual-inspired spectral features. Secondly, we adopted a two-stage feature selection algorithm to select the best features from an original set of acoustic features and semantic primitives, separately. Thirdly, two open-scheme evaluations were conducted over leave-one-speaker-out and cross-corpus validations to demonstrate the gain in accuracies and generalization from our proposed strategies to multilingual SER.

Beyond most of the current SER systems, the main contributions of this paper can be summarized as follows: i) we propose to incorporate the knowledge of human emotion perception to develop an emotion recognition model in multilingual tasks; ii) we define a robust set of combined features to represent emotional information in speech not sensitive to different languages; iii) we performed extensive evaluations from different aspects, taking into account the impacts of acoustic features and speaker normalisation and no-normalisation, besides, the generalization ability of the proposed system is even assessed by conducting a cross-corpus evaluation.

The remainder of this paper is organized as follows: Section 2 describes the block diagram of the proposed multilingual emotion recognition system along with the emotional corpora used in this study. It also describes the collection of acoustic features, semantic primitives, and emotion dimensions used for establishing the three-layer model. Section 3 details the implementation of the system and discuss the speaker-independent and cross-corpus validations. Section 4 compares the achievements of this study and related literature. Finally, conclusions are drawn in Section 5.

2. Methodology of multilingual emotion recognition

An essential aspect of designing the multilingual SER system is the methodology of modeling relations between emotions and speech features among languages. To make machines reach a comparable level of performance as humans, this research uses an emotional perception-inspired three-layer model incorporating acoustic features, semantic primitives, and emotion dimensions that possess the ability to predict emotion dimensions accurately. Figure 1 depicts the block diagram of our multilingual SER system. The input emotional speech is first pre-processed to obtain a set of powerful speech features. The best features are then used to predict emotion dimensions through the semantic primitives in the three-layer model. In the emotion recognition stage, four distinct dimensional-based features, i.e., values of valence and arousal and direction and distance from neutral to other emotions in the dimensional space, are extracted from the speech to recognize emotional categories by using logistic model trees (LMT). The following subsections detail the system as follows: Section 2.1 illustrates the emotional corpora in four different languages that we used. Section 2.2 analyses two different sets of speech features for the task of multilingual emotion recognition. Section 2.3 describes the experimental setup for collecting human evaluations on semantic primitives and emotion dimensions.

2.1. Emotion corpora

This subsection is an overview of the four corpora of acted emotional speech in Japanese, German, Chinese, and English.

2.1.1. Fujitsu database

The Fujitsu database was chosen as the Japanese emotional corpus. It was recorded by Fujitsu Laboratory and acted by a professional actress. The female speaker was asked to express 20 different sentences nine times with five emotions: 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 was lost, this database contains a total of 179 utterances.

2.1.2. Berlin Emo−DB

The German corpus was the well-known Berlin Database of Emotional Speech (Emo−DB) recorded by five male and five female professional actors [43]. The data contains different numbers of spoken utterances in seven emotions: 127 anger, 38 disgust, 55 fear, 64 happiness, 53 sadness, 79 boredom, and 78 neutral. Overall the database consists of 494 emotional utterances.

2.1.3. CASIA dataset

The Chinese emotional database (CASIA) was produced by the Institute of Automation, Chinese Academy of Sciences; it contains neutral and five categories of acted emotion: angry, happy, sad, fear, and surprise. It was produced by four

professional actors (two males and two females) [44]. The data consists of 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, etc. There are 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 are 300 utterances in this part. Each speaker uttered $(100 + 300) * 6 = 2400$ sentences in total.

We chose 200 sentences from the spontaneous portion for four speakers involving four basic emotions: 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 do not sufficiently simulate emotions in a natural or clear manner. Four Chinese native speakers (2 male and 2 female) hence were asked to verify the emotional categories in a listening test. The experimental results provided 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 utterances were eventually labelled as follows: 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. These 198 instances were taken from the CASIA corpus.

2.1.4. SAVEE database

The Surrey Audio−Visual Emotion (SAVEE) database, which was produced for the purpose of developing an emotion recognition system, consists of 480 British English utterances made by four native English male actors in seven different emotional categories: anger, disgust, fear, happiness, sadness, surprise and neutral.

The above four emotional corpora were used for training and testing the multilingual emotion recognition system. To guarantee an equal contribution from each language to the implementation, subsets on a similar scale and involving the same emotions were taken from the emotional corpora and used. The set of basic emotions to be recognized was: neutral, happiness, anger, and sadness. Table 1 details the utterances chosen from each corpus.

Table 1: Details of four emotional corpora

Corpus	Language	Emotion				
		Neu.	Hap.	Ang.	Sad.	Total
Fujitsu database	Japanese	20	40	40	40	140
Berlin Emo-DB	German	50	50	50	50	200
CASIA database	Chinese	68	29	51	50	198
SAVEE database	English	75	75	75	75	300

Emotional categories: neutral (Neu.), happiness/happy (Hap.), anger/angry (Ang.), sad(ness) (Sad.).

2.2. Extraction of acoustic features

2.2.1. Prosodic related features

The set of prosodic features, abbreviated as IS16, was analysed in our earlier attempt [31], 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 [45]. In addition, the acoustic correlates related to timing were extracted by manual segmentation.

2.2.2. 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 [46], 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 from IS16 and 196 spectral features from MSF.

2.3. Evaluation of semantic primitives and emotion dimensions

To achieve the task on multilingual SER, we adopted a three-layer model after [30], based on the assumption that

Table 2: Coefficients of Pearson's correlation ($\overline{cc}$) for semantic primitive evaluation of Fujitsu database, Berlin Emo-DB, and CASIA dataset by human listeners averaged over all speakers and whole utterances.

	Bright	Dark	High	Low	Strong	Weak	Calm	Unstable	Well-modulated	Monotonous	Heavy	Clear	Noisy	Quiet	Sharp	Fast	Slow
Fujitsu database	0.92	0.90	0.89	0.91	0.92	0.91	0.89	0.90	0.91	0.89	0.88	0.88	0.90	0.93	0.88	0.85	0.84
Berlin Emo-DB	0.86	0.90	0.87	0.89	0.93	0.93	0.89	0.85	0.90	0.87	0.86	0.89	0.87	0.91	0.87	0.84	0.86
CASIA dataset	0.82	0.87	0.86	0.92	0.91	0.91	0.88	0.82	0.82	0.85	0.85	0.83	0.89	0.86	0.89	0.90	0.91

human perception of emotions embedded in speech does not originate directly from a change in acoustic cues, but through an indirect route of small perceptions on semantic primitives. For instance, low arousal and negative valence speech easily make an impression on listeners in the form of dark and heavy feelings, but high arousal and positive valence speech is often uttered in a bright and well-modulated way. The set of semantic primitives was derived from [29], and examined by a multidimensional scaling analysis. These semantic primitives were used in a multi-layer model for describing emotional speech, namely, bright, dark, high, low, strong, weak, calm, unstable, well-modulated, monotonous, heavy, clear, noisy, quiet, sharp, fast, and slow.

To conduct this three-layer model, speech emotions have to be evaluated for semantic primitives and emotion dimensions. Note that we did not carry out listening tests on the SAVEE database because the use of it was not for an implementation of this three-layer model, but only for a test task in a cross-corpora evaluation.

2.3.1. Semantic primitive-based emotion evaluation

Each semantic primitive and emotion dimension 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 [21, 22], 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. To ensure the consistency in the perception of emotions of different nationalities, three extra Japanese native speakers (two male and one female, mean age: 35.0 years old) took part in a categorical perception test to label the Berlin Emo-DB. Experimental result showed a mean recognition rate of 82.0% over four emotional categories. This was somewhat same to that obtained by the German native speakers in [43], reporting a 87.0% recognition accuracy and stressed the fact that speech emotion recognition is a cross-lingual process. 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 evaluating the semantic primitives, the emotional speech was played randomly and evaluated 17 times by the participants on a whole utterance level, once for each semantic primitive for all utterances in one corpus. Each of 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"*.

Table 3: Coefficients of Pearson's correlation ($\overline{cc}$) for emotion dimensions evaluation of Fujitsu database, Berlin Emo-DB, and CASIA dataset by human listeners averaged over all speakers and whole utterances.

	Valence	Arousal
Fujitsu database	0.96	0.96
Berlin Emo-DB	0.92	0.94
CASIA dataset	0.85	0.91

For each instance of speech n in corpus c, where $c \in \{Fujitsu, Berlin, Casia\}$, $1 \leq n \leq N$, 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.

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

The inter-evaluator agreement was evaluated using Eq. 2 following the related study reported in [23].

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

Table 2 demonstrated the average results of inter-rater agreement of the three emotional corpora. 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 Emo−DB, and CASIA dataset, indicating good evaluation results and good agreement among listeners.

2.3.2. Emotion dimension-based emotion evaluation

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 [47], 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

Table 4: 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	F0		
over all accentual phrases;			
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;	Acoustic Frequency		
spectral slope (SSL) cross MF band: 1;	(AF) Domain	STRONG	AROUSAL
spectral flatness cross MF band: 2;			
SC cross AF band: 2, 19, 28, 32;			
SSL cross AF band: 13, 22, 25;	Modulation Frequency		
spectral skewness cross AF band 13, 23, 30;	(MF) Domain	WEAK	
spectral kurtosis cross AF band:17, 27;			
spectral spread cross AF band: 25			

being very negative and +2 being very positive) and arousal (-2 being very relaxed and +2 being aroused). 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. 1. In this scenario, p refers to valence or arousal.

The inter-rater agreement on evaluations over emotion dimensions was also measured by Eq. 2, and shown in Table 3. As can be seen, the agreement among listeners was moderate to high within the range of $0.85-0.96$ over the three emotional corpora. The highest correlation was 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 Emo−DB 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.

3. Experiments

3.1. Experimental setup

This section presents a two-stage estimation scheme for multilingual SER. First, estimation of emotion dimensions for valence and arousal was addressed by using a three-layer model; Second, the task of emotional classification was furnished by incorporating the human-perception-inspired knowledge [21], namely, mapping four common features of valence, arousal, and the directions and distances from a neutral voice to other emotional states into basic categories based on a classification scheme.

3.1.1. Preprocessing

Most researchers believed that speaker normalization (SN) advances the accuracies of SER [42, 48]. We herein adopted an approach to SN after [40], and compared it to a scenario in no speaker normalization, where this process takes into account the effect of variations among different speakers. 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:

$$f'_{u,v}(n) = \frac{f_{u,v}(n) - \overline{f_{u,v}}}{\sqrt{\frac{1}{N_{u,v}-1}\sum_{m=1}^{N_{u,v}}(f_{u,v}(m) - \overline{f_{u,v}})^2}} \tag{3}$$

$$\overline{f_{u,v}} = \frac{1}{N_{u,v}}\sum_{n=1}^{N_{u,v}} f_{u,v}(n). \tag{4}$$

3.1.2. 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 [50]. 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) for each feature individually to eliminate the irrelevant ones. The normalized multi-class FDR for the uth feature is given as:

$$FDR(u) = \frac{2}{C(C-1)}\sum_{c1}\sum_{c2}\frac{(\mu_{c1,u} - \mu_{c2,u})^2}{(\sigma^2_{c1,u} + \sigma^2_{c2,u})} \tag{5}$$

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 details the 22 acoustic features, four semantic primitives and two emotion dimensions that we used to construct the proposed three-layer model.

3.2. Estimation and Classification

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 [49]. Correspondingly, the nature of perception of speech emotion was fuzzy and vague [23]. Furthermore, our proposed three-layer model also incorporated human knowledge from manual evaluations of semantic primitives and emotion dimensions that involve non-linear processing according to human emotion perception. Our previous effort [31, 36] 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).

3.3. 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:

$$CC = \frac{\sum_1^N \left(X_n - \overline{X}\right)\left(Y_n - \overline{Y}\right)}{\sqrt{\sum_1^N \left(X_n - \overline{X}\right)^2 \sum_1^N \left(Y_n - \overline{Y}\right)^2}} \tag{6}$$

$$MAE = \frac{\sum_1^N |X_n - Y_n|}{N} \tag{7}$$

where $\overline{X}$ and $\overline{Y}$ are the mean values of X_n and Y_n, respectively. In addition, N is the 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, precision, and F-measure are reported in terms of each emotional state for assessing the performance of categorical classification.

Formulate, let C_i stands for an emotional class to be classified, where $i \in \{neutral, happiness, anger, sadness\}$, and N_i is the total number of utterances for class C_i. Supposing a classifier predicts correctly NC_i^T utterances for class C_i, and predicts 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 are defined as:

$$Recall = \frac{NC_i^T}{N_i} \tag{8}$$

$$Precision = \frac{NC_i^T}{NC_i^T + NC_i^F} \tag{9}$$

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{10}$$

3.4. Experiment 1*: Comparison with individual feature sets*

This experiment attempted to show that the use of the proposed combination of IS16 and MSF features can benefit SER in the multilingual scenario. All results in this subsection were obtained by leave-one-speaker-out (LOSO) validation using a mixed corpus made from the Fujitsu database, Berlin Emo-DB, and CASIA dataset, providing 538 utterances in total: 138 neutral, 119 happiness, 141 anger, and 140 sadness (see Table 1 for the details of each corpus).

Comparisons were accordingly carried out on the feature set of IS16, MSF, and their combination that is the proposed set, abbreviated as Proposed. Each of the aforementioned three feature sets was examined under two conditions of speaker normalization (SN) and no speaker normalization (NN). A total of 2 $\star$ 3 systems were trained and tested to study the impact of the selection of acoustic features and processing of SN.

3.4.1. Performance of estimation of semantic primitives and emotional space

We first evaluated the performance obtained during estimation of semantic primitives. The CC values for dark, heavy, strong, and weak were detailed in Table 5. Results confirmed that the proposed acoustic features provides a better

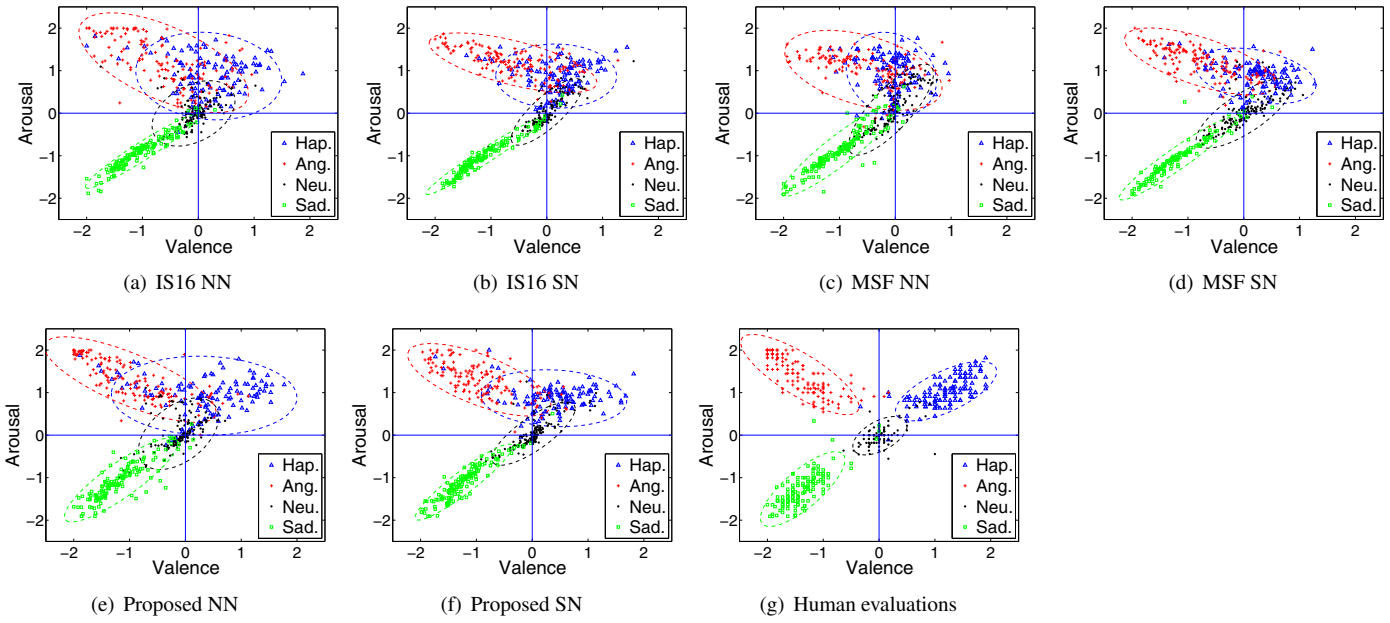
Table 5: Estimation performance of semantic primitives obtained by multilingual emotion recognition systems using different feature sets

Features	Speaker Normalization	DARK			HEAVY			STRONG			WEAK		
		IS16	MSF	Proposed	IS16	MSF	Proposed	IS16	MSF	Proposed	IS16	MSF	Proposed
CC	NN	0.841	0.840	**0.886**	0.682	0.642	**0.706**	0.878	0.819	**0.905**	0.878	0.839	**0.890**
	SN	0.873	0.880	**0.892**	0.727	0.772	**0.784**	0.887	0.905	**0.922**	0.913	0.912	**0.922**

Table 6: Estimation performance of emotion dimensions obtained by multilingual emotion recognition systems using different feature sets

Features	Speaker Normalization	Valence			Arousal		
		IS16	MSF	Proposed	IS16	MSF	Propsed
CC	NN	0.640	0.568	**0.749**	0.907	0.865	**0.933**
	SN	0.654	0.712	**0.792**	0.915	0.919	**0.930**
MAE	NN	0.644	0.726	**0.508★♦**	0.364	0.435	**0.290★♦**
	SN	0.630	0.600	**0.497n.s.**	0.356	0.330	**0.295n.s.**

★ *and* ♦ *indicate that the estimations differ significantly between feature set of IS16 and Proposed, and MSF and Proposed under condition of no speaker normalisation (NN) (*$p < 0.001$*);* ***n.s.*** *indicate that estimations not differ statistically significant between NN and speaker normalisation (SN) for the Proposed features in terms of valence (*$p = 0.7285$*) and arousal (*$p = 0.7649$*).*



(a) IS16 NN (b) IS16 SN (c) MSF NN (d) MSF SN

(e) Proposed NN (f) Proposed SN (g) Human evaluations

Figure 2: 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 IS16, MSF, and Proposed under conditions of no speaker normalisation (NN) and speaker normalisation (SN).

performance that is closer to human evaluations and outperformed that obtained with IS16 and MSF features that trained with and without speaker normalization.

Figures 2 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 6. As can be seen, the Proposed features (IS16 combined with MSF) always furnished the best performance yielding a greater CC and smaller MAE compared with those obtained by IS16 and MSF features, both under scenarios of NN and SN.

Statistical test (one-way ANOVA) was performed between two systems's estimations on all emotional utterances. Mostly interesting, the differences of MAE between the sets of proposed features in term of NN and SN turned out to be not statistically significant for both valence ($F(1, 1074) = 0.1206$, $p = 0.7285$) and arousal ($F(1, 1074) = 0.0895$, $p = 0.7649$). On the other hand, under no speaker normalisation condition, the proposed features yielded statistically significant improvements over the IS16 and MSF features individually for both valence: $F(1, 1074) = 15.4174$, $p < 0.001$, and $F(1, 1074) = 39.6463$, $p < 0.001$; and arousal: $F(1, 1074) = 17.0243$, $p < 0.001$, and $F(1, 1074) = 54.8106$, $p < 0.001$.

This fact indicated that our system could better cope with the variation of acoustic correlates among speakers without speaker normalization, which in turn benefit the full development of SER in the real-life applications, especially in a scenario without knowledge for an unknown speaker.

Table 7: Classification results for multilingual SER (All), obtained by three types of features on IS16, MSF, and Proposed with speaker normalisation (SN), and no speaker normalisation (NN); and detailed classification performance of each of the multilingual corpora for Fujitsu databse, Berlin Emo-DB, and CASIA dataset.

		Precision						Recall						F-Measure					
		IS16		MSF		Proposed		IS16		MSF		Proposed		IS16		MSF		Proposed	
		NN	SN	NN	SN	NN	SN	NN	SN	NN	SN	NN	SN	NN	SN	NN	SN	NN	SN
All	Neutral	58.52	64.06	53.35	72.26	73.01	74.51	86.69	89.13	72.46	81.16	86.23	82.61	70.18	74.55	62.11	76.45	79.07	78.35
	Happpiness	73.77	61.33	30.19	70.24	84.62	77.89	37.82	38.66	13.45	49.58	55.00	62.18	50.00	47.42	18.60	58.13	66.67	69.16
	Anger	69.80	74.66	58.96	73.75	76.36	84.11	73.76	77.30	72.34	83.69	89.36	90.07	71.72	75.96	64.97	78.41	82.35	86.99
	Sadness	93.55	97.60	89.06	90.65	93.23	94.24	82.86	87.14	81.43	90.00	88.57	93.57	87.88	92.08	85.07	90.32	90.84	93.91
	Weighted Avg.	**74.04**	**74.96**	**59.25**	**76.99**	**81.72★♦**	**82.91n.s.**	**71.56**	**74.35**	**61.71**	**77.14**	**80.71★♦**	**82.90n.s.**	**70.73**	**73.48**	**59.21**	**76.52**	**80.23★♦**	**82.63n.s.**
Fujitsu database	Neutral	72.00	48.78	66.67	82.61	73.68	55.60	90.00	100.00	20.00	95.00	70.00	100.00	80.00	65.57	30.77	88.37	71.79	71.40
	Happpiness	94.74	72.73	0.00	94.74	100.00	92.90	45.00	20.00	0.00	45.00	50.00	32.50	61.02	31.37	0.00	61.02	66.67	48.10
	Anger	63.93	70.59	50.00	68.97	60.61	74.50	97.50	90.00	100.00	100.00	100.00	95.00	77.23	79.12	66.67	81.63	75.47	83.50
	Sadness	100.00	100.00	100.00	100.00	100.00	100.00	87.50	92.50	95.00	100.00	87.50	97.50	93.33	96.10	97.44	100.00	93.33	98.70
	Weighted Avg.	84.19	76.49	52.38	87.14	84.99	84.30	78.57	72.14	58.57	83.57	77.86	78.60	77.59	68.40	51.28	81.95	77.53	76.00
Berlin Emo-DB	Neutral	64.18	80.00	62.12	80.36	78.57	93.75	86.00	96.00	82.00	90.00	88.00	90.00	73.50	87.27	70.69	84.91	83.02	91.84
	Happpiness	75.00	76.74	56.00	86.11	92.11	88.00	48.00	66.00	28.00	62.00	68.63	88.00	58.54	70.97	37.33	72.09	78.65	88.00
	Anger	77.08	79.59	75.47	78.18	90.38	91.49	74.00	78.00	80.00	86.00	94.00	86.00	75.51	78.79	77.67	81.90	92.16	88.66
	Sadness	88.68	97.92	87.50	90.57	89.09	90.91	94.00	94.00	98.00	96.00	98.00	100.00	91.26	95.92	92.45	93.20	93.33	95.24
	Weighted Avg.	76.24	83.56	70.27	83.80	87.56	91.04	75.50	83.50	72.00	83.50	87.06	91.00	74.70	83.24	69.54	83.03	86.75	90.93
CASIA dataset	Neutral	52.68	60.44	49.11	63.16	69.32	72.06	86.76	80.88	80.88	70.59	89.71	72.06	65.56	69.18	61.11	66.67	78.21	72.06
	Happpiness	30.00	23.81	16.67	34.38	55.00	56.25	10.34	17.24	6.90	34.48	37.93	62.07	15.38	20.00	9.76	34.38	44.90	59.02
	Anger	70.00	73.91	55.00	74.47	82.98	86.79	54.90	66.67	43.14	68.63	76.47	90.20	61.54	70.10	48.35	71.43	79.59	88.46
	Sadness	94.44	95.00	79.41	82.61	93.02	93.33	68.00	76.00	54.00	76.00	80.00	84.00	79.07	84.44	64.29	79.17	86.02	88.42
	Weighted Avg.	64.37	67.27	53.53	66.78	76.73	78.91	62.63	66.67	53.54	66.16	76.26	78.28	60.59	66.07	51.10	66.34	75.66	78.51

★ *and* ♦ *indicate that the classification results differ significantly between feature set of IS16 and Proposed, and MSF and Proposed under condition of no speaker normalisation (NN) (*$p < 0.05$*);* ***n.s.*** *indicate that classification results not differ statistically significant between NN and speaker normalisation (SN) for the Proposed features.*

3.4.2. Performance of categorical classification

Given the estimations of valence and arousal in the emotional space, we extracted four human-perception-inspired common features of the values of valence, arousal, and the directions and distances from a position of a neutral voice to that of another emotional state, to perform categorical-based classification using the logistic model trees. Table 7 details the emotional classification performance in terms of recall, precision, and F-measure over LOSO cross-validation for different sets of features with SN and NN. 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, irrespective of NN and SN.

As an aside, a one-way ANOVA analysis was conducted between two systems of classification results on 15 speakers (4 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 first performed a statistical test between the two classification results on the Proposed features with no speaker normalization and speaker normalization, taking into account an effect of speaker normalization. Results showed no significant improvement, $F(1, 28) = 0.7498$, $p = 0.3939$ for an averaged precision, $F(1, 28) = 0.3722$, $p = 0.5468$ for an averaged recall, and $F(1, 28) = 0.5118$, $p = 0.4803$ for an averaged F-measure, with the Proposed features with speaker normalisation yielded no statistically significant different performance to that achieved by the Proposed features with no speaker normalisation.

In addition, we conducted two more one-way ANOVA analysis in the scenario of no speaker normalisation, between features of IS16 and Proposed, and MSF 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 IS16 features at recognizing multilingual speech emotions. Further, this analysis showed that the Proposed features also significantly improved the SER performance compared with MSF 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 IS16 and MSF features at identifying speech emotions. Notably, the Proposed features can also deal with speaker-independent SER tasks irrespective to languages, even without speaker normalization that is more suitable for real-life applications.

3.5. Experiment 2: Cross-corpus evaluation

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). Table 8 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 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 [23], 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 in the event that the training and test speech delivered different degrees in a specific emotion.

[1] Fujitsu database is a single speaker corpus

Table 8: Classification performance for three-layer model based emotion recognition systems over permutations in cross-corpus evaluation.

Test	Train	Classification Performance						
		Neu.	Hap.	Ang.	Sad.	Rec.	Pre.	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**

Table 9: Classification performance of each language by monolingual SER systems, multilingual systems, and approaches used in [31] for Fujitsu database, Berlin Emo-DB, and CASIA dataset.

		F-Measure			
		Monolingual SER		Multilingual SER	
		[31]	Proposed	[31]	Proposed
Fujitsu database	Neutral	93.02	100.00	65.31	71.40
	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 Emo-DB	Neutral	82.69	96.97	82.00	91.84
	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 dataset	Neutral	52.63	64.66	67.78	72.06
	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**

4. Comparison with related literature

4.1. comparison within our studies

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 9, and compared with that performed in multilingual scenarios. Furthermore, the experimental results given in a previous attempt [31] were also included in Table 9 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 9, our proposed approach advanced the performance of categorical classification on Fujitsu database for both monolingual and multilingual SER systems, and outperformed those obtained by [31]; 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 SER system after [31] ($p < 0.05$). 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.

4.2. comparison with other studies using the same corpora

As was reviewed in Table 10, 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 the Fujitsu database by exploiting 21 acoustic features in a three-layer model [30]. 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 [30]. We can see from these results that exploring efficient vocal features contributes to advancing the recognition and accuracies of all emotional categories.

A number of effort has been done to be able to recognize emotional state in the 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 [42]. Furthermore, Vlasenko et al. [40] 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 [42, 40]. 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 studies that were able to recognize speech emotional state in the CASIA dataset, [52] 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.

Table 10: 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	[30]	92.50
		Ours	**100.00**
	Multilingual	Ours	**98.10**
Berlin Emo-DB (LOSO)	Monolingual	[40]	89.90
		[42]	85.80
		Ours	**93.00**
	Multilingual	Ours	**91.00**
CASIA dataset (LOSO)	Monolingual	[52]	58.53
		Ours	**69.70**
	Multilingual	Ours	**78.28**

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 [53], training and testing under a 70-30% cross-validation, and reporting a 48.4% average recognition accuracy.

5. Conclusion

We presented a system for recognizing emotions expressed in multilingual speech. We analyzed three crucial issues, the question of which speech features to use to recognize emotions in multiple languages; the effects of speaker normalization and no speaker normalization; and the ability of the system to deal with a completely new language without a training phase for verifying its generalization ability.

The proposed features were evaluated on mixed emotional corpora to classify four emotional categories. The individual set of IS16 and MSF features were included for reference. Experiments were conducted under conditions of speaker normalization and no speaker normalization. The proposed features provided a comparable performance under conditions of speaker normalization and no speaker normalization, yielding an average F-measure of 82.63%, and 80.23% individually for a mixture of three different language corpora, and the results showed the benefit of using combined features.

We carried out two evaluations involving LOSO cross-validation and cross-corpus validation. The average weighted F-measure was within the range of 58.6 to 95.8% in the LOSO cross-validation depending on the speaker (cf. Section 3.4.2), indicating that the inter-speaker differences in expression of emotions might be the main reason for mismatches. In the cross-corpus validation, the highest recognition accuracy was obtained on the Fujitsu Database, followed by Berlin Emo-DB and CASIA, indicating that stereotypical emotions are slightly easier to recognize than authentic ones.

Moreover, we reviewed the literature as general benchmarks. In particular, our previous attempt presented a multilingual emotion recognition system by analyzing the IS16 features only. The proposed features outperformed the IS16 features at identifying multilingual speech emotions, yielding an error reduction rate of 34.50% and 32.46% respectively in relative to conditions of speaker normalization and no speaker normalization. Furthermore, in comparison with the related work targeting recognition of the same emotional corpus, the proposed system demonstrated a promising performance for multilingual SER tasks. In particular, in a classification task for a new target language without training, its performance was comparable to that obtained in a monolingual emotion recognition system.

Numerous studies conducted over the past few decades have tried to recognize emotions in speech. Promising performance has been achieved in monolingual scenarios in various languages. However, implementation of multilingual emotion recognition is still a challenging task. With respect to the processing of human emotional perception, this study achieved a multilingual emotion recognizer by combining prosodic and spectral features in a three-layer model. In the future, we would like to apply it to human-machine interactions, such as in an affective speech-to-speech translation system.

6. Acknowledgments

This study was supported by a Grant-in-Aid for Scientific Research (A) (No. 25240026) and a CSC Scholarship made available by the China Scholarship Council.

References

[1] J. Ma, H. Jin, L. Yang, and J. Tsai, "Ubiquitous Intelligence and Computing: Third International Conference, UIC 2006, Wuhan, China, September 3-6, 2006," *Proceedings (Lecture Notes in Computer Science)*, Springer-Verlag, New York, Inc., Secaucus, NJ, USA, 2006.

[2] S. Chandaka, A. Chatterjee, and S. Munshi, " Support vector machines employing cross-correlation for emotional speech recognition," *Measurement*, vol. 42, no. 4, pp. 611-618, 2009.

[3] C. Jones, and J. Sutherland, "Acoustic emotion recognition for affective computer gaming," *Affect and emotion in human-computer interaction*, vol. 4868, pp. 209-219, Springer, Berlin, Heidelberg, 2008.

[4] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "An adaptive framework for acoustic monitoring of potential hazards," *EURASIP Journal on Audio, Speech, and Music Processing*, 2009.

[5] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," *Proceedings of Acoustics, Speech, and Signal Processing, 2004*, vol. 1, 2004, pp. 577-580.

[6] J. Hirschberg, S. Benus, J.M. Brenier, F. Enos, S. Friedman, and S. Gilman, "Distinguishing deceptive from non-deceptive speech," *Proceedings of the Ninth European Conference on Speech Communication and Technology*, pp. 1833-1836, 2005.

[7] R.W. Picard, "Affective Computing," *MIT Press*, Cambridge, 1998.

[8] R. Huang, and C. Ma, "Towards a speaker-independent real-time affect detection system," *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)*, vol. 1, pp. 1204–1207, 2006.

[9] C. Huang, D. Erickson, and M. Akagi, "Comparison of Japanese expressive speech perception by Japanese and Taiwanese listeners," *Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3323, 2008.

[10] T. L. Nwe, S. W. Foo, and Liyanage C De Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication* vol. 41, no. 4, pp. 603–623, 2003.

[11] Y. Zhou, J. Li, Y. Sun, J. Zhang, Y. Yan, and M.Akagi, "A hybrid speech emotion recognition system based on spectral and prosodic features," *IEICE Trans. on Information and Systems,* vol. 93, no. 10, pp. 2813–2821, 2010.

[12] P. Shen, C. Zhou, and X. Chen, "Automatic speech emotion recognition using support vector machine," *Proc. International Conference on Electronic and Mechanical Engineering and Information Technology (EMEIT)*, vol. 2, pp. 621–625, 2011.

[13] M. Shami, and M. Verhelst, "Automatic classification of expressiveness in speech: a multi-corpus study," *Speaker classification II: Selected Projects*, Springer-Verlag, Berlin, pp. 43–56, 2007.

[14] P. Y. Oudeyer, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, vol.59, no. 1–2, pp. 157–183, 2003.

[15] O. Kalinli, "Analysis of multi-lingual emotion recognition using auditory attention features," *Proc. Interspeech 2016*, pp. 3613–3617, 2016.

[16] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," *Automatic Speech Recognition and Understanding* , pp. 523–528, 2011.

[17] B. C. Chiou, and C.P. Chen, "Speech emotion recognition with cross-lingual databases," *Fifteenth Annual Conference of the International Speech Communication Association*, pp. 558–561, 2014.

[18] S. Parthasarathy, and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," *INTERSPEECH, Stockholm, Sweden (2017)*, pp. 1103–1107, 2017.

[19] I. Lefter, J. M. Rothkrantz, P. Wiggers, and D. A. Van Leeuwen, "Emotion recognition from speech by combining databases and fusion of classifiers," *International Conference on Text, Speech and Dialogue*, Springer, Berlin, Heidelberg, pp. 353–360, 2010.

[20] J. Deng, Z. Zhang, and B. Schuller, "Linked Source and Target Domain Subspace Feature Transfer Learning–Exemplified by Speech Emotion Recognition," *22nd International Conference on Pattern Recognition*, Stockholm, Sweden, IAPR, 2014, pp. 761–766, 2010.

[21] X. Han, R. Elbarougy, M. Akagi, J. Li, T. D. Ngo, and T. D. Bui, "A study on perception of emotional states in multiple languages on Valence-Activation approach," *Proc. NCSP2015, Kuala Lumpur, Malaysia (2015).*

[22] M. Akagi, X. Han, R. Elbarougy, R. Hamada, and J. Li, "Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages," *Proc. Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific.*, pp.1-10.

[23] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49,no. 10, pp. 787–800, 2007.

[24] M. Grimm, K. Kroschel, and S. Narayanan, " Support vector regression for automatic recognition of spontaneous emotions in speech," *Acoustics, Speech and Signal Processing, 2007*, vol. 4, pp. 1085–1088, 2007.

[25] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, " A dimensional approach to emotion recognition of speech from movies," *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009*, pp. 65–68, 2009.

[26] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, " LSTM-Modeling of continuous emotions in an audiovisual affect recognition

framework," *Image and Vision Computing*, vol. 31, no. 2, pp. 153–163, 2013.
[27] K. R. Scherer, "Personality inference from voice quality: The loud voice of extroversion," *European Journal of Social Psychology*, vol. 8, pp. 467–487, 1978.
[28] E. Brunswik, "Historical and Thematic Relations of Psychology to Other Sciences," *The Scientific Monthly*, vol. 83, no. 3, pp. 151-161, 1956.
[29] C. Huang, and M. Akagi, "A three-layered model for expressive speech perception," *Speech Communication*, vol. 50, no. 10, pp. 810–828, 2008.
[30] R. Elbarougy, and M. Akagi, "Improving speech emotion dimensions estimation using a three-layer model of human perception," *Acoustical science and technology*, vol. 35, no. 2, pp. 86–98, 2014.
[31] X. Li, and M. Akagi "Multilingual speech emotion recognition using a three-layer model," *Proc. Interspeech 2016*, pp. 3608–3612, 2016.
[32] R. Cowie, E. Douglas−Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–225, 2001.
[33] L. Ten Bosch, "Emotions, speech and the ASR framework," *Speech Communication*, vol. 40, no. 1, pp. 213–225, 2003.
[34] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," *Proc. Spoken Language 1996*, vol. 3, pp. 1970–1973, 1996.
[35] M. Schröder, R. Cowie, "Issues in emotion-oriented computing-towards a shared understanding," *Workshop on Emotion and Computing at KI*, 2006.
[36] X. Li, and M. Akagi "Toward improving estimation accuracy of emotion dimensions in bilingual scenario based on three-layered model," *Proc. O-COCOSDA/CASLRE 2015*, pp. 21–26, 2015.
[37] T. Jahnstone, and K. Scherer, "Vocal communicatioin of emotion," *Handbook of Emotions*, vol.2, pp. 220–235, 2000.
[38] R. Cowie, and R. Cornelius, "Describing the emotional states that expressed in speech," *Speech Communication*, vol. 40, no.1, pp. 5–32, 2003.
[39] J. Benesty, M. M. Sondhi, and Y. Huang, "Springer handbook of speech processing," *Springer Science and Business Media*, 2007.
[40] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll "Combining frame and turn-level information for robust recognition of emotions within speech," *Proc. Interspeech 2007*, pp. 2225–2228, 2007.
[41] M. Goudbeek, and K. Scherer, "Beyond arousal: Valence and potency/contral cues in the vocal expression of emotion," *The Journal of the Acoustical Society of America*, vol. 128, no. 3, pp. 1322–1336, 2010.
[42] S. Wu, T. H.Falk, and W. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, no. 5, pp. 768–785, 2011.
[43] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," *Proc. Interspeech 2005*, pp. 1517–1520, 2005.
[44] "Mandarin emotional speech corpus," *http* : *//www.chineseldc.org/doc/CLDC − SPC − 2005 − 010/intro.htm* Institute of Automation, Chinese Academy of Sciences, 2005.
[45] H. Kawahara, "Straight exploitation of the other aspect of vocoder: perceptually isomorphic decomposition of speech sounds," *Acoustical Scinece and Technology*, vol. 27, no. 6, pp. 349–353, 2006.
[46] Z. Zhu, R. Miyauchi, Y. Araki, and M. Unoki, "Modulation spectral features for predicting vocal emotion recognition by simulated cochlear implants," *Proc. Interspeech 2016*, pp. 262–266, 2016.
[47] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," *Speech Communication*, vol. 53, no. 1, pp. 36–50, 2011.
[48] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2013.
[49] J. S. Jang, "ANFIS: adaptive-network-based fuzzy inference system," *IEEE transactions on systems*, vol. 23, no. 3, pp. 665–685, 1993.
[50] M. Kotti, and F. Paternó, "Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema," *International journal of speech technology*, vol. 15, no. 2, pp. 131–150, 2012.
[51] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2017.
[52] L. Chao, J. Tao, M. Yang, and Y. Li, "Improving generation performance of speech emotion recognition by denoising autoencoders," *Proc. ISCSLP 2014*, pp. 341–344, 2014.
[53] M. Sidorov, C. Brester, W. Minker, and E. Semenkin, "Speech-Based Emotion Recognition: Feature Selection by Self-Adaptive Multi-Criteria Genetic Algorithm," *Proc. LREC 2014*, pp. 3481–3485, 2014.