INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Multilingual Speech Emotion Recognition System based on a Three-layer Model

Xingfeng Li, Masato Akagi

Japan Advanced Institute of Science and Technology,
1-1 Asahidai, Nomi, Ishikawa, 923-1292 Japan
lixingfeng@jaist.ac.jp, akagi@jaist.ac.jp

Abstract

Speech Emotion Recognition (SER) systems currently are focusing on classifying emotions on each single language. Since optimal acoustic sets are strongly language dependent, to achieve a generalized SER system working for multiple languages, issues of selection of common features and retraining are still challenging. In this paper, we therefore present a SER system in a multilingual scenario from perspective of human perceptual processing. The goal is twofold. Firstly, to predict multilingual emotion dimensions accurately such as human annotations. To this end, a three layered model consist of acoustic features, semantic primitives, emotion dimensions, along with Fuzzy Inference System (FIS) were studied. Secondly, by knowledge of human perception of emotion among languages in dimensional space, we adopt direction and distance as common features to detect multilingual emotions. Results of estimation performance of emotion dimensions comparable to human evaluation is furnished, and classification rates that are close to monolingual SER system performed are achieved.

Index Terms: emotion recognition in speech, three-layer model, emotion dimension

1. Introduction

Speech processing is widely studied in the area of affective computing to enable computers to possess intelligence to understand human behaviors. Most speech systems are focused on the process of natural-language understanding by capturing the linguistic content of spoken utterances. Such language understanding can be greatly improved if the emotion conveyed by the speaker can be identified [1]. This fact introduces an appealing research topic: speech emotion recognition (SER). On this topic, researchers have recently determined a small set of discrete categories to characterize emotions such as anger, disgust, fear, joy, sadness, and surprise [2]. However, the intensity of a certain emotion varies with time and situation, which may take any arbitrary value such as little anger or much anger. To describe the rich variation in the degree of affective states, dimensional space of valence and activation is generally used to represent emotions [3]. Valence is the quality of being charged with positive or negative significance, which is inherent in emotional appraisal and experience; Activation quantifies how a speech is associated to an intensity level, i.e., particularly strong or weak.

Many studies on SER for different single corpora have been conducted using the language-dependent optimal acoustic sets over the past several decades. Such systems can be simply analyzed in monolingual scenarios; changing the source corpus requires re-selecting the optimal acoustic features and re-training the system. Human-emotion perception, however, has proved to be cross lingual, even without the understanding of the language used [4]. An automatic SER system is expected to recognize emotions such as. We therefore aimed at constructing a multilingual SER system suitable to be used across cultures. To this end, we studied the two essential aspects of estimating emotion dimensions and extracting common features among languages.

The first challenging issue was to accurately estimate emotion dimensions. Traditionally, prediction of emotion dimensions can be directly obtained from acoustic features by using various estimators. Grimm et al., for example, predicted emotion dimensions from acoustic features by using the Fuzzy Inference System (FIS) [5]. The prediction performance of the activation dimension was better than valence. Many researchers later studied new acoustic parameters to improve the accuracy of valence estimation, as analyzed by Wu et al. [6]. However, such estimation remains poor.

Human-emotion perception, as described by Scherer [7] who adopted a version of Brunswiks lens model originally developed in 1956 [8] is a multi-layer process. To precisely estimate emotion dimensions, especially for valence, a three-layer model consisting of acoustic features, semantic primitives and emotion dimensions was adopted by Elbarougy and Akagi in 2012 [9]. The estimation results of valence and activation with this three-layer model outperformed that predicted directly from acoustic features with a traditional system. In this study, we used such a three-layer model to predict emotion dimensions in a multilingual scenario, since it works strongly effective in mimicking human-emotion perception processing in our previous studies [10] [11].

Another challenge in the area of multilingual SER is that it is not clear what common features are efficient in detecting emotions across cultures. Commonalities and differences in human-emotion perception across languages in the valence-activation (V-A) space have recently been studied [4]. It was revealed that direction and distance from neutral to other emotions are similar across languages, and neutral positions of languages are language-dependent. Motivated by this new finding, with the assumption that the proposed multilingual SER can precisely estimate emotion dimensions, we adopted direction and distance from neutral to other emotions as common features to normalize languages for multilingual SER.

2. Multilingual SER Architecture

From perspective of human-emotion perception processing, we studied two sub-processes. Figure 1 illustrates the multilingual SER framework. With extracted and normalized acoustic features, the three-layer model along with the human

http://dx.doi.org/10.21437/Interspeech.2016-645

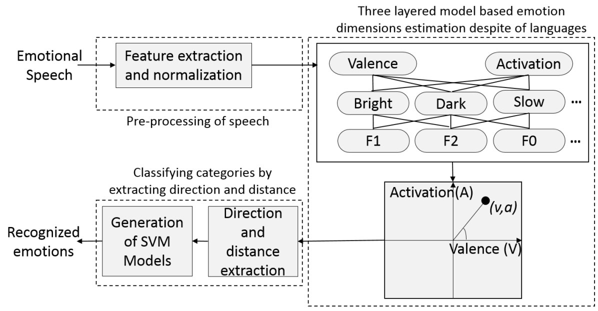

Three layered model based emotion dimensions estimation despite of languages
Emotional Speech
Feature extraction and normalization
Pre-processing of speech
Valence
Activation
Bright
Dark
Slow
...
F1
F2
F0
...
Classifying categories by extracting direction and distance
Activation(A)
(v,a)
Valence (V)
Recognized emotions
Generation of SVM Models
Direction and distance extraction

Figure 1: Block diagram of multilingual SER system

knowledge-based FIS are first used to accurately predict emotion dimensions from acoustic features though semantic primitives. All dimensional predictions are obtained regardless of language. In addition, motivated by the commonalities across languages in the V-A space, a classification method is addressed by adopting direction and distance to classify emotions. The emotional classification results are presented by a support vector machine (SVM) with 10-fold cross validation.

3. Corpora

We used three corpora of acted emotions in Japanese, German, and Chinese to develop the proposed multilingual SER system. In the three corpora, four similar categories: neutral, joy, anger, and sad were selected. The Japanese corpus is Fujitsu Database recorded by Fujitsu Laboratory. It contains 5 emotions, neutral, happy, cold anger, sad, and hot anger. A professional actress was asked to utter 20 different sentences, each sentence has one neutral utterance and two utterances in each of other emotions. The selected utterances from Fujitsu Database are: 20 neutral, 40 happy, 40 hot anger, and 40 sad, totaling 140 utterances.

The German corpus is the well known Berlin Emo-DB. Ten professional actors (five males and five females) each uttered ten sentences in German to simulate seven different emotions. The number of spoken utterances for these 7 emotions in the Berlin Emo-DB are not equally distributed: 127 anger, 81 boredom, 46 disgust, 69 fear, 71 joy, 79 neutral, and 62 sadness. Finally, 200 utterances were selected from this corpus, 50 utterances of 4 similar emotional states as the Japanese corpus.

The Chinese emotional corpus was developed by Institute of Automation, Chinese Academy of Sciences (CASIA), which includes neutral utterances by four professional actors (two males and two females) and five categories of other emotions, anger, happiness, sadness, fear, and surprise. The content 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. There are 100 utterances for each emotion. The utterances of the spontaneous part were selected from news articles, conversations and essays without emotional-rich words. There are 300 utterances in this part. Ultimately 200 utterances of spontaneous content from 4 actors covering 4 emotions (neutral, happiness, sadness, and anger) were selected as the initial step, i.e., 50 utterances of each emotion. Different from Fujitsu Database and Berlin Emo-DB, spontaneous utterances in the CASIA Emotional Corpus do not sufficiently simulate emotions in a natural and clear manner. All 200 selected utterances were re-annotated on categorical labels by 11 Chinese native speakers (five females and six males) to obtain the correct categories. The experimental results showed that 68 utterances were recognized as neutral speech, and 30, 50, and 50 utterances were grouped as happy, angry, and sad. Two spoken utterances cannot be classified into any one of four emotional categories. Hence, we eventually used 198 human-annotated utterances in this study.

4. Pre-Processing

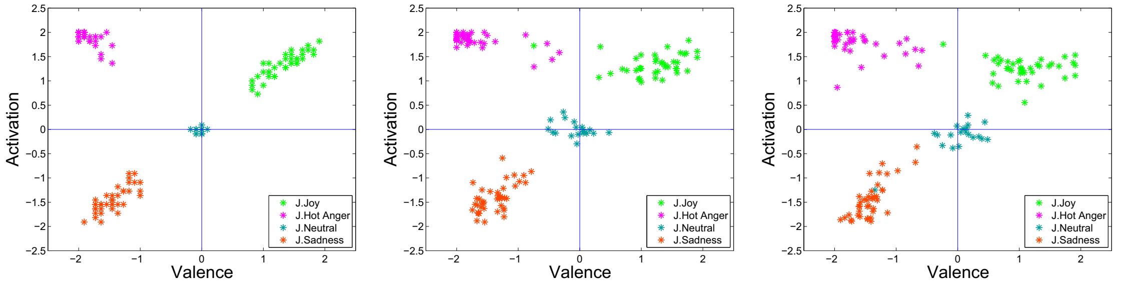
4.1. Acoustic-feature extraction

The selection of acoustic features as inputting parameters is crucial. Therefore, the most relevant acoustic features that have been successfully used in previous studies were selected [10] [11]. These acoustic features can be grouped as five subgroups: four F0-related features, four power envelop-related features, five power spectrum-related features, three duration-related features, and five voice quality-related features. Acoustic features derive from F0, power envelop, power spectrum, and voice quality were collected by STRAIGHT [12]. Besides, acoustic features related to duration were extracted by manual segmentation. A total of 21 acoustic features were extracted. All acoustic features were normalized by dividing the values of acoustic features by the mean value of neutral voices to avoid speaker and language dependency on acoustic cues.

4.2. Semantic primitives and emotion-dimension evaluation

In this study, we chose a three-layer model based on that by Elbarougy and Akagi [9], with which it is assumed that human-emotion perception comes not directly from a change in acoustic cues but from a smaller perception expressed by adjectives. The following set of adjectives describing emotional voice were selected as candidates for semantic primitives: bright, dark, high, low, strong, weak, calm, unstable, well-modulated, monotonous, heavy, clear, noisy, quiet, sharp, fast, and slow. They are originally from [13]. The three emotional corpora were evaluated in terms of each semantic primitive using listening tests. Fujitsu Database and CASIA Emotional Corpus were evaluated by eleven Japanese native speakers (nine males and two females) and ten Chinese native speakers (five males and five females), respectively. However, it was difficult to find German native speakers as listeners; therefore, nine Japanese speakers (eight males and one female) were asked to evaluate the Berlin Emo-DB corpus. Emotional voices were evaluated 17 times by the participants, once for each semantic primitive. The participants were asked to rate each of the 17 semantic primitives on a five-point scale: ”1-Does not feel at all”, ”2-Seldom feels”, ”3-Feels a little”, ”4-feels”, ”5-Feels very much”. Additionally, since emotions were characterized by the V-A space in this study, the three corpora therefore needed to be annotated with dimensional labels along with valence and activation. The same participants were selected to evaluate emotion dimensions on a five-point scale -2, -1, 0, 1, 2, valence (from -2 very negative to +2 very positive), activation (from -2 very calm to +2 very excited).

The basic theory of semantic primitives and emotion dimensions was explained to the participants before the experiment. They then participated in a training session to listen to an example set composed of 20 utterances, which covered the five-point scale. The purpose of this training set was to enable the participants to understand these adjectives. All stimuli were played randomly through binaural headphones at a comfortable sound pressure level in a soundproof room. The participants were asked to evaluate their perceived impression from the manner of speaking not from the content itself then score on the five-point scale for each adjective. Moreover, the inter-rater agreement was measured by means of pairwise Person’s correlation between two participants’ ratings for each semantic primitive. The average rating for each adjective and emotion dimension was calculated per utterance.

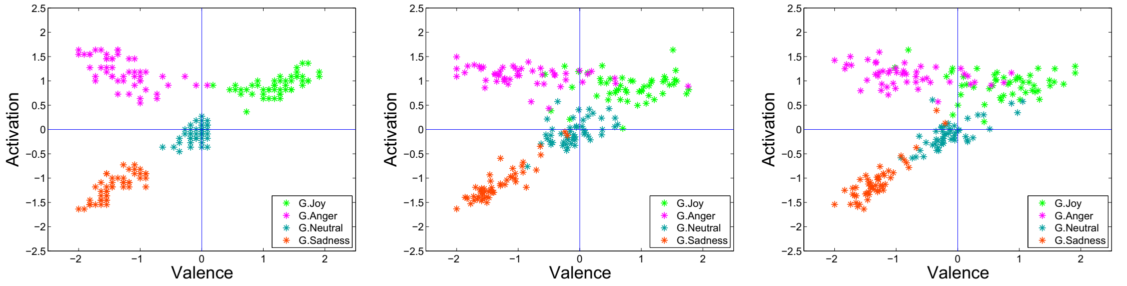


(a) Manually labeled

(b) Estimated using monolingual SER system (Mono SER)

(c) Estimated using proposed multilingual SER system (Multi SER)

Figure 2: Distribution of Japanese corpus in valence-activation (V-A) space

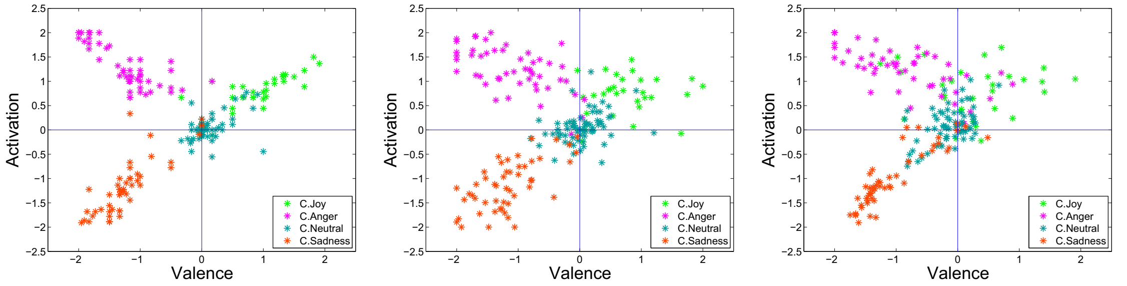


(a) Manually labeled

(b) Estimated using Mono SER

(c) Estimated using proposed Multi SER

Figure 3: Distribution of German corpus in V-A space



(a) Manually labeled

(b) Estimated using Mono SER

(c) Estimated using proposed Multi SER

Figure 4: Distribution of Chinese corpus in V-A space

5. Emotion-dimension estimation

As mentioned above, our goal for this study consisted of two parts. In this section, we discuss the first sub-process which aims at precisely estimating emotion dimensions similar to human annotations. The human-perception-based three-layer model along with FIS are investigated.

5.1. Emotion-dimension estimation

The proposed multilingual SER system was constructed with 21 acoustic features in the bottom layer, 17 semantic primitives in the middle layer, and 2 emotion dimensions in the top layer. This three-layer model based multilingual SER system was trained and tested using the three corpora simultaneously, regardless of language. The FIS was used to establish the mapping from acoustic features to emotion dimensions through semantic primitives. To obtain estimated emotion dimensions, each of the 17 semantic primitives in the middle of three-layer model should be first predicted separately from the 21 acoustic features using 17 FISs. Beyond that, the estimation of emotion dimension can be done from the 17 estimated adjectives in the previous part with another 2 FISs.

5.2. Estimation performance regarding emotion dimensions

The estimated results of emotion dimensions by the system were respectively compared to the human evaluation that derived from the subjective listening experiments in the V-A space for Japanese, German, and Chinese. The comparison was conducted by calculating the mean absolute error (MAE) between human responses and system's estimations from Equation (1).

$$MAE^{(j)} = \frac{\sum_{i=1}^{N} \left| \widehat{x}_i^{(j)} - x_i^{(j)} \right|}{N} \qquad (1)$$

where $j \in \{Valence, Activation\}$, $\widehat{x}_i^{(j)}$ is the output of the proposed system, $x_i^{(j)}$ is the human responses by human participants, and N is the number of utterances in each emotion corpus. Simultaneously, to obtain a comparative baseline to evaluate the estimated precision of the proposed system, the mean standard deviation (MSTDEV) of emotion dimensions among the human participants in the listening experiments was used, which was calculated from Equation (2).

$$MSTDEV^{(j)} = \frac{\sum_{i=1}^{N} \sqrt{\frac{\sum_{m=1}^{N1} (x_{m,i}^{(j)} - \overline{\mu}_i^{(j)})^2}{N1}}}{N} \qquad (2)$$

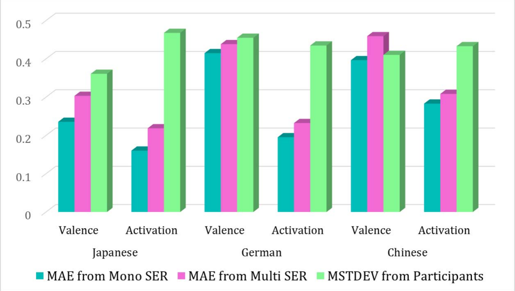


Figure 5: Comparison of MAE of both SER systems and MSTDEV of human evaluation

where $j \in \{Valence, Activation\}$ and N have the same definitions as in Eq. (1), $N1$ is the number of participants of the listening tests ($N1$=11 in Japanese, $N1$=9 in German, $N1$=10 in Chinese), and $\overline{\mu}_i$ is the average value of one utterance from all participants. Particularly, to compare the performance of the proposed multilingual SER system (Multi SER), it was compared with the monolingual SER system (Mono SER), which was trained and tested using only one language. The distribution of all emotional utterances in the V-A space are shown in Figs. 2 to 4, corresponding to Japanese, German, and Chinese corpora. For each space, there are three panels (a), (b), (c); the left panel shows the distribution of human evaluation, the middle panel shows the estimation of the Mono SER, and the right panel shows the estimation of our Multi SER.

The MAE and MSTDEV of all emotion dimensions of the Japanese, German, and Chinese corpora, for Mono SER and Multi SER are illustrated in Figure 5. The MAE of all emotion dimensions of the three corpora from Multi SER were completely within the scope of the MSTDEV of human evaluation as well as that of valence in the Chinese scenario with insignificant increments by 0.05. Additionally, comparisons of the MAEs of Mono SER and Multi SER show that the largest increments were achieved by valence in Japanese (by 0.07). Such increments do not constitute a large difference. In summary, the estimation performance of emotion dimensions of Multi SER was comparable to human evaluation. Furthermore, the performance of Multi SER was close to that of Mono SER for this task.

6. Classification

This section introduces the second sub-process to classify emotional categories using the accurate estimations in the V-A space. The knowledge of commonalities and differences in human-emotion perception among languages in the V-A space illustrates that direction and distance from neutral to other emotions are similar across languages. According to Eqs. (3) and (4), we adopted direction and degree as common features to normalize languages in order to recognize emotional states in multiple languages.

$$angle = \arctan(\frac{y_E - y_N}{x_E - x_N}) \tag{3}$$

$$d(E, N) = \sqrt{(x_E - x_N)^2 + (y_E - y_N)^2} \tag{4}$$

Table 1: Classification rates of each language by Monolingual SER (Mono), Multilingual SER (Multi), and Referenced studies ([15] [16] [17]) for Fujitsu Database, Berlin Emo-DB, and CASIA Emotional Corpus.

[%]	Mono	Multi	Referenced
	Fujitsu Database		
Neutral	100	95	80
Joy	97.5	93	97.5
Anger	95	95	92.5
Sad	95	100	100
Average	96.88	95.75	92.5
	Berlin Emo-DB		
Neutral	98	90	88.5
Joy	86	82	69
Anger	86	82	93.7
Sad	90	96	94.3
Average	90	87.5	86.38
	CASIA emotional corpus		
Neutral	96	89	98
Joy	97	77	82.25
Anger	88	73	90.25
Sad	92	92	94.5
Average	93.25	82.75	91.25

where (x_E, y_E) is the position of the emotional state E, and (x_N, y_N) is the center position of the neutral state N. For this part of the study, we adopted the SVM as a classifier to present classification results. All results were obtained using 10-fold cross-validation. Table 1 summarizes the classification rates for each language. The important fact is that we achieved an overall classification rate of 95.75, 87.50, and 82.75% respectively for the Japanese, German, and Chinese corpora, which are close to those analyzed from Mono SER (96.88, 90, and 93.25%). The main contributions to this comparable performance are the three-layer model and the classification method by extracting direction and distance in the V-A space. In particular, the largest decrement of 10.5% was achieved by comparing Mono SER and Multi SER in the Chinese corpus. This is due to the fact that the selected emotional speeches were selected from news articles, conversations, and essays that do not sufficiently simulate emotions in a natural and clear manner such as with Fujitsu Database and Berlin Emo-DB. Additionally, we attained an error reduction rate of 58.4, 27.5, and 23% for the Japanese, German, and Chinese corpora, respectively, in comparisons to the referenced studies [14] [15] [16] using the same corpus for monolingual SER task, in which the combination of the three-layer model and human-knowledge-based classification method was found to give the best results and be well suited for emotion recognition yielding small classification errors.

7. Conclusion

We developed a multilingual SER from the perspective of human-emotion perception processing by using a three-layer model and studied the estimation of emotion dimensions, regardless of language. Motivated by the knowledge of human-emotion perception across languages in the V-A space, we adopted direction and distance as common features to recognize emotions in multilingual scenarios. The experimental results show that the proposed multilingual SER system can precisely estimate emotion dimensions as well as that through human annotations, and the classification performance of the proposed system is close to that achieved with the monolingual SER system.

8. Acknowledgment

This study was supported by the Grant-in-Aid for Scientific Research (A) (No. 25240026) and the A3 Foresight Program made available by the Japan Society for the Promotion of Science (JSPS).

9. References

[1] M. Akagi, X. Han, and R. Elbarougy, "Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages," in *APSIPA*. IEEE, 2014, pp. 1–10.

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

[3] E. D.-C. M. M. Schrder, R. Cowie and S. C. Gielen., "Acoustic correlates of emotion dimensions in view of speech synthesis," 2001, pp. 87–90.

[4] 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).

[5] 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.

[6] D. Wu, T. D. Parsons, and S. S. Narayanan, "Acoustic feature analysis in speech emotion primitives estimation." in *INTERSPEECH*, 2010, pp. 785–788.

[7] K. R. Scherer, "Personality inference from voice quality: The loud voice of extroversion," *European Journal of Social Psychology*, vol. 8, no. 4, pp. 467–487, 1978.

[8] E. Brunswik, "Historical and thematic relations of psychology to other sciences," *The Scientific Monthly*, vol. 83, pp. 151–161, 1956.

[9] R. Elbarougy and M. Akagi, "Cross-lingual speech emotion recognition system based on a three-layer model for human perception." Proc. APSIPA2013, Kaohsiung, Taiwan (2013).

[10] X. Li and M. Akagi, "Toward improving estimation accuracy of emotion dimensions in bilingual scenario based on three-layered model," in *O-COCOSDA/CASLRE*. IEEE, 2015, pp. 21–26.

[11] X. Li and M. Akagi., "Automatic speech emotion recognition in chinese using a three-layered model in dimensional approach." Proc NCSP2016, Honolulu, Hawaii (2016).

[12] H. Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.

[13] C. Huang and M. Akagi, "A three-layered model for expressive speech perception," *Speech Communication*, vol. 50, no. 10, pp. 810–828, 2008.

[14] 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.

[15] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll, "Emotion recognition in the noise applying large acoustic feature sets," *Speech Prosody, Dresden*, pp. 276–289, 2006.

[16] Z. Yu, L. Junfeng, S. Yanqing, J. Zhang, Y. Yonghong, and M. Akagi, "A hybrid speech emotion recognition system based on spectral and prosodic features," *IEICE TRANSACTIONS on Information and Systems*, vol. 93, no. 10, pp. 2813–2821, 2010.