Cross-lingual Speech Emotion Recognition System Based on a Three-Layer Model for Human Perception

Reda Elbarougy [1,2] and Masato Akagi [1]
[1]Japan Advanced Institute of Science and Technology (JAIST), Japan
[2]Department of Mathematics, Faculty of Science, Damietta University, New Damietta, Egypt
E-mail: elbarougy@jaist.ac.jp, akagi@jaist.ac.jp

***Abstract*—The purpose of this study is to investigate whether emotion dimensions valence, activation, and dominance can be estimated cross-lingually. Most of the previous studies for automatic speech emotion recognition were based on detecting the emotional state working on mono-language. However, in order to develop a generalized emotion recognition system, the performance of these systems must be analyzed in mono-language as well as cross-language. The ultimate goal of this study is to build a bilingual emotion recognition system that has the ability to estimate emotion dimensions from one language using a system trained using another language. In this study, we first propose a novel acoustic feature selection method based on a human perception model. The proposed model consists of three layers: emotion dimensions in the top layer, semantic primitives in the middle layer, and acoustic features in the bottom layer. The experimental results reveal that the proposed method is effective for selecting acoustic features representing emotion dimensions, working with two different databases, one in Japanese and the other in German. Finally, the common acoustic features between the two databases are used as the input to the cross-lingual emotion recognition system. Moreover, the proposed cross-lingual system based on the three-layer model performs just as well as the two separate mono-lingual systems for estimating emotion dimensions values.**

I. Introduction

Most of the previous techniques for automatic speech emotion recognition focus only on the classification of emotional states as discrete categories such as happiness, sadness, anger, fear, surprise, and disgust [1], [2]. However, a single label or any small number of discrete categories may not accurately reflect the complexity of the emotional states conveyed in everyday interaction [3]. Hence, a number of researchers advocate the use of a dimensional description of human emotion, where emotional states are not classified into one of the emotion categories but estimated on a continuous-valued scale in a multi-dimensional space (e.g., [4], [5], [6], [7]).

In this study, a three-dimensional continuous model is adopted in order to represent the emotional states using emotion dimensions i.e., valence, activation and dominance. This approach is chosen because it exhibits great potential to model the occurrence of emotions in the real world as well as in a realistic scenario. Emotions are not generated in a prototypical or pure modality, but rather in complex emotional states, which are a mixture of emotions with varying degrees of intensity or expressiveness. Therefore, this approach allows a more flexible interpretation of emotional states [8].

Speech perception plays an important role in human-human communications. Additionally, speech recognition systems that mimic human speech perception mechanisms also come to play an important role in human-machine communications. Thus, we need global evidence for speech perception to obtain knowledge for constructing the models. However, there is little knowledge that could contribute to realize universal communication environments.

Using the dimensional approach, emotion categories are represented by regions in an n-dimensional space, where the neutral category lies near the origin, and other emotions lie in a specific region in the n-dimensional space. For example, in the two-dimensional space valence-activation, happy is represented by a region which lies in the first quarter, in which valence is positive, and activation/arousal is high, as shown in Fig. 1.

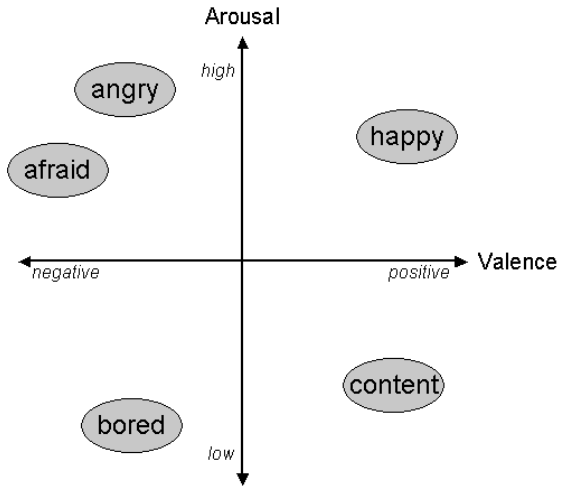


Fig. 1. A two-dimensional emotion space with a valence and an arousal axis. Basic Emotions are marked as areas within the space.

Speech is the most natural and important means of human-human communication in our daily life, when we use the same language. Even without the understanding of one language,

we can still judge the expressive content of a voice, such as emotions. An interesting question to ask is whether emotional states can be recognized universally or not. Culture and society have a considerable weight on the expression of emotions. This, together with the inherent subjectivity among individuals, can make us wonder about the existence of universal emotions. If we consider Darwin's theory of evolution, emotions find their root in biology and therefore can be to some extent considered to be universal [9]. Several studies have indeed shown evidence for certain universal attributes for both speech [10], [11] and music [12], [13], not only among individuals of the same culture, but also across cultures. Dang et al. 2009, for instance, performed an experiment in which humans had to distinguish between 3 and 6 emotions respectively [14]. Their conclusion was that listeners are able to perceive emotion from speech sound without linguistic information with about 60% accuracy in a three-emotion evaluation and about 50% in a six-emotion evaluation.

In this study, we assume that the acoustic features realization of specific emotions is language independent based on the following two assumptions: (1) the positions of neutral voices are different among languages. This may be related to different cultures; (2) the distance and directions from neutral voice to other emotional states are common among languages. Therefore, in order to reduce the speaker and language dependency on acoustic features realizations, we adopt a new acoustic feature normalization process to avoid the speaker and language variation on the used acoustic features.

Several studies have been devoted to the analysis of the most important acoustic features from the point of view of a categorical model, working on mono-lingual [15], [16] and multi-lingual [17] data. However, they have not yet concerned with the same depth the importance of acoustic features from the dimensional model point of view [18]. This paper investigates whether there are common acoustic features between two different languages that allow us to estimate emotion dimensions from a speech voice. Therefore, it is interesting to imitate human perception by building an automatic speech-emotion recognition system that has the ability to detect the emotional state regardless of the input language.

To accomplish this, we work with two databases of emotional speech, one in the Japanese language and the other in the German language. The emotional state in this paper is represented by the dimensional approach. This approach defines emotions as points in a three-dimensional emotion space spanned by the three basic dimensions valence (negative-positive axis), activation (calm-excited axis), and dominance (weak-strong axis).

Firstly, a variety of acoustic features were extracted for each language. Then, a novel feature selection method based on a three-layer model of a human perception model was proposed. The proposed method was used to find the best acoustic feature subsets in the mono-lingual mode for each emotion dimension. Finally, we construct two cross-language emotion recognition systems which can estimate the emotional state by training the system using one language, and testing the system using the other language.

II. Emotion recognition strategy

Scherer [19], in his description of human perception adopted a version of Brunswik's lens model which was originally proposed in 1956 [20]. Based on this model, human perception is a three-layer process.

In 2008, Huang and Akagi adopted a three-layer model for human perception. They assumed that human perception for emotional speech are not directly realized from a change of acoustic features, but rather from a composite of different types of smaller perceptions that are expressed by semantic primitives or adjectives describing emotional voice [21].

Here we attempt to use the above human perception model proposed in [21] to find the most correlated acoustic features for each emotion dimension through semantic primitives. Our model consists of three layers: emotion dimensions valence, activation and dominance which constitute the top layer, semantic primitives which constitute the middle layer, and acoustic features which form the bottom layer. A semantic primitive layer is added between the two traditional layers of acoustic features and emotion dimensions, as shown in Fig. 2.

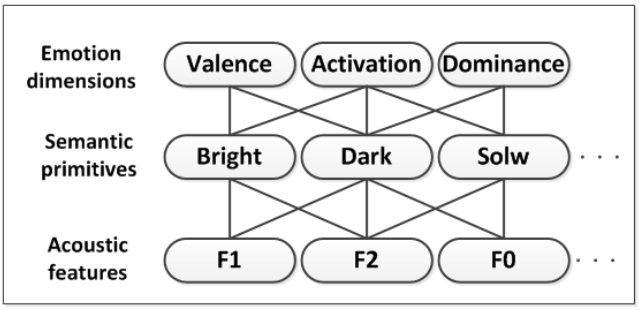


Fig. 2. Three layer model

We assume that the acoustic features which are highly correlated with semantic primitives will have a large impact for predicting values of emotion dimensions. The findings can guide the selection of the most related acoustic features with better discrimination for each emotion dimension.

In our previous paper [22], a three-layer model was proposed for estimating emotion dimensions; valence, activation, and dominance. The proposed model was evaluated using a Japanese database. It was found that this model was effective for selecting acoustic features for each emotion dimension. The prediction for all emotion dimensions was improved using a speech emotion recognition system based on the three-layer model. In this paper, we investigate the effectiveness of the proposed model for another language database, a German database [23]. In addition, we construct a cross-language emotion recognition system in which values of the selected acoustic features were used as the inputs, and values of emotion dimensions are the output of this system.

III. Databases and Experimental Evaluation

In order to construct a cross-lingual emotion recognition system to estimate emotion dimensions valence, activation, and dominance, we need at least two databases in different languages. The elements of the proposed emotion recognition system were collected in this section. The databases and acoustic features used in this study are introduced. Moreover, the semantic primitives and emotion dimensions are evaluated by conducting two listening tests using human subjects as described in next subsections.

A. Speech Material and Subjects

Different types of databases are suitable for different purposes. There are objections against the use of acted emotions. However, acted emotions are quite adequate for testing data. Therefore, it is suitable for a novel method which first requires proof of the concept, rather than construction of a real-life application for the industry [24].

In this paper, in order to validate the proposed system, two emotional speech databases were used, one in the Japanese language and the other in the German language. The Japanese database is the multi-emotion single speaker Fujitsu database produced and recorded by Fujitsu Laboratory. A professional actress was asked to produce utterances using 5 emotional speech categories, i.e., neutral expression, joy, cold anger, sadness, and hot anger. In the database, there are 20 different Japanese sentences. Each sentence has one utterance in the neutral expression and two utterances in each of the other categories. Thus, for each sentence there are 9 utterances and for all 20 sentences there are 180 utterances. One cold anger utterance is missing from this database therefore the total number of utterances for this database is 179 utterances.

The German database is Berlin database [23]. It comprises seven emotional states: anger, boredom, disgust, anxiety, happiness, sadness, and neutral speech. Ten professional German actors (five females and five males) spoke ten sentences with an emotionally neutral content in the seven different emotions. These sentences were not equally distributed between the various emotional states: 69 frightened; 46 disgusted; 71 happy; 81 bored; 79 neutral; 62 sad; 127 angry. This database was selected for the following reasons: it is an acted speech database which is the same as Fujitsu database, since it contains four similar categories:happy, angry, sad, and neutral used in the Fujitsu database, to investigate the effect of multi-speaker and multi-gender on speech emotion recognition. For the purpose of comparing the results of the two databases we used only the four similar categories. Furthermore, for training purpose, we used an equal distribution of the four emotional states, 50 happy, 50 angry, 50 sad, and 50 neutral; in total, 200 utterances were selected from the Berlin database: 100 utterances were uttered by 5 males and the other 100 by 5 females divided equally between the four emotional states.

For evaluating semantic primitives and emotion dimensions, we used listening tests. The Fujitsu database was evaluated by 11 graduate students, native Japanese speakers (9 male and 2 female), while the Berlin database was evaluated using 9 graduate students, native Japanese speakers (8 male and 1 female). None of the subjects have any hearing problems.

B. Acoustic Features

In this research, for constructing a speech emotion recognition system, acoustic features are a very important factor which needs to be investigated. Therefore, the most relevant acoustic features which have been successful in related works and features used for other similar tasks were selected. Those acoustic cues which are considered significant for prosody largely are extracted from fundamental frequency, intensity, and duration. In addition, voice quality is another major factor that researchers have paid much attention to. Therefore, acoustic features which originate from F0, the power envelope, the power spectrum, and voice quality are extracted with the high quality speech analysis-synthesis system STRAIGHT [25]. Moreover, acoustic features which are related to duration are extracted by segmentation. We eventually extracted a set of 21 acoustic features which can be grouped in several subgroups:

F0-related features: f0 mean value of the rising slope (F0_RS), the highest F0 (F0_HP), the average F0 (F0_AP) and the rising slope of the first accentual phrase (F0_RS1).

Power envelope-related features: mean value of the power range in the accentual phrase (PW_RAP), the power range (PW_R), the rising slope of the first accentual phrase (PW_RS1), the ratio between the average power in the high-frequency portion (over 3 kHz) and the average power (PW_RHT);

Power spectrum-related features: the first formant frequency (SP_F1), the second formant frequency (SP_F2), the third formant frequency (SP_F3), spectral tilt (SP_TL), and spectral balance (SP_SB);

Duration related features: total length (DU_TL), consonant length (DU_CL), ratio between consonant length and vowel length (DU_RCV).

These above-mentioned 16 acoustic feature were selected from the work by Huang and Akagi, where they proved that these acoustic features have a significant correlation with semantic primitives [21]. In addition to these 16 acoustic features, 5 new parameters related to voice quality are added, because voice quality is one of the most important cues for the perception of expressive speech.

Voice quality: the mean value of the difference between the first harmonic and the second harmonic H1-H2 for vowels /a/,/e/,/i/,/o/, and /u/ per utterance MH_A, MH_E, MH_I, MH_O, and MH_U.

All the 21 acoustic features were extracted for both the Fujitsu and Berlin databases.

In order to avoid speaker and language dependency on the acoustic features that are used, we adopt the new acoustic feature normalization, in which all acoustic feature values are normalized by those of the neutral speech. This was performed by dividing the values of acoustic features by the mean value of neutral utterances for all acoustic features.

C. Semantic Primitives Evaluation

In this study, the human perception model as described by Scherer [19] is adopted. This model assumes that human perception is a three-layer process. It was assumed that the acoustic features are perceived by a listener and internally represented by a smaller perception e.g adjectives describing emotional voice as reported in [21]. These smaller percepts or adjectives are finally used for detecting the emotional state of the speaker. These adjectives can be subjectively evaluated by human subjects. Therefore, a set of adjectives describing the emotional speech were selected as candidates for semantic primitives. These adjectives are: Bright, Dark, High, Low, Strong, Weak, Calm, Unstable, Well-modulated, Monotonous, Heavy, Clear, Noisy, Quiet, Sharp, Fast, and Slow. These adjectives were selected because they reflect a balanced selection of widely used adjectives that describe emotional speech, and were used originally in [21].

For the evaluation, we used listening tests. In these tests, the stimuli were presented randomly to each subject through binaural headphones at a comfortable sound pressure level in a soundproof room. Subjects were asked to rate each of the 17 semantic primitives on a 5-point scale ("1-Does not feel at all", "2-Seldom feels" , "3-Feels a little" , "4-feels" , "5-Feels very much"). The 17 semantic primitives were evaluated for the two databases, then ratings of the individual subject were averaged for each semantic primitive per utterance. The inter-rater agreement was measured by means of pairwise Pearson's correlations between two subjects' ratings, separately for each semantic primitive. It was found that all subjects agreed from moderate to a very high degree.

D. Emotion Dimensions Evaluation

Most of the existing emotional speech databases were annotated using the categorical approach. Few databases were annotated using the dimensional approach. The Fujitsu and Berlin databases are categorical databases. Therefore, listening tests are required to annotate each utterance in the used databases using the dimensional approach. Thus, the two databases were evaluated through the listening tests along the three dimensions of valence, activation, and dominance. For the emotion dimensions evaluation, a 5-point scale {-2, -1, 0, 1, 2} was used: valence (from -2 very negative to +2 very positive), activation (from -2 very calm to +2 very excited), and dominance (from -2 very weak to +2 very strong).

The subjects used a MATLAB GUI to evaluate the stimuli. Repeats were allowed. Subjects were asked to evaluate one emotion dimension for the whole database in one session. There were three sessions, one for each emotion dimension. As done in [26] for emotion dimensions evaluation, before starting the experiment, the basic theory of emotion dimension was explained to the subjects. Then, they took a training session to listen to an example set composed of 15 utterances, which covered the used 5- point scale, which are three utterances for each point in the used scale. In the test, the stimuli were presented randomly for each utterance. Subjects were asked to evaluate their perceived impression from the way of speaking, not from the content itself, then to rate each dimension individually using the 5-point scale. The average per utterances of the subjects rating for each emotion dimension was calculated . The subjects show a high inter-rater agreement. It was found that all subjects agreed to a high degree on the valence, activation, and dominance.

IV. Selection of Semantic Primitives and Acoustic Features

A. Selection Procedures

This section describes the proposed acoustic features selection method to identify the most relevant acoustic features for the emotion dimensions of valence, activation and dominance. For this purpose, we investigate the effectiveness of the three-layer model, which imitates human perception, to understand the relationship between acoustic features and emotion dimensions. To accomplish this task, a top-down method shown in Fig. 3 was used as follows:

- the correlation coefficients between evaluated values for each emotion dimension (top-layer) and evaluated values of each semantic primitives (middle layer) were calculated using Eq.(1) as shown in Table 1;
- the highly correlated semantic primitives were selected for each emotion dimension as an adjective that describes this dimension;
- the correlation coefficients between evaluated values for each selected semantic primitive (middle layer) in the second step and extracted values for each acoustic feature (bottom layer) were calculated using Eq.(2), as shown in Table 2,
- and the highly correlated acoustic features were selected for each semantic primitive.

For each emotion dimension, the selected acoustic features are considered to be the features most relevant to the used

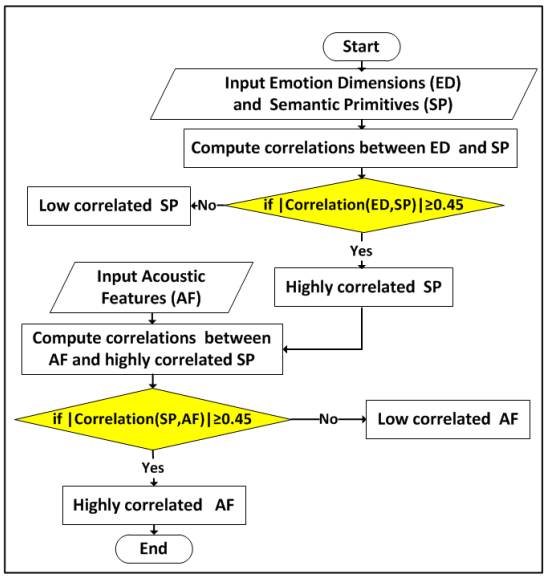


Fig. 3. Process for acoustic feature selection

TABLE I
GERMAN DATABASE: CORRELATION COEFFICIENTS BETWEEN THE SEMANTIC PRIMITIVES AND THE EMOTION DIMENSIONS

m		Bright	Dark	High	Low	Strong	Weak	Calm	Unstable	Well-modulated	Monotonous	Heavy	Clear	Noisy	Quiet	Sharp	Fast	Slow	#
1	Valence	**0.9**	**-0.7**	**0.6**	**-0.6**	0.1	-0.4	-0.2	0.1	0.3	-0.2	**-0.9**	**0.8**	0.1	-0.4	0.3	0.4	**-0.5**	7
2	Activation	**0.7**	**-0.9**	**0.9**	**-0.9**	**0.9**	**-1.0**	**-0.9**	**0.9**	**0.9**	**-0.9**	**-0.6**	**0.7**	**0.9**	**-1.0**	**0.9**	**0.8**	**-0.8**	17
3	Dominance	**0.6**	**-0.9**	**0.8**	**-0.9**	**1.0**	**-1.0**	**-0.9**	**0.9**	**0.9**	**-0.8**	**-0.5**	**0.6**	**0.9**	**-1.0**	**1.0**	**0.8**	**-0.8**	17
	#	3	3	3	3	2	2	2	2	2	2	3	3	2	2	2	2	3	41

TABLE II
GERMAN DATABASE: CORRELATION COEFFICIENTS BETWEEN THE ACOUSTIC FEATURES AND SEMANTIC PRIMITIVES

m		Bright	Dark	High	Low	Strong	Weak	Calm	Unstable	Well-modulated	Monotonous	Heavy	Clear	Noisy	Quiet	Sharp	Fast	Slow	#
1	MH_A	**-0.6**	**0.8**	**-0.7**	**0.8**	**-0.8**	**0.8**	**0.7**	**-0.7**	**-0.7**	**0.7**	**0.5**	**-0.6**	**-0.8**	**0.8**	**-0.8**	**-0.7**	**0.7**	17
2	MH_E	**-0.5**	**0.6**	**-0.6**	**0.6**	**-0.7**	**0.7**	**0.7**	**-0.7**	**-0.6**	**0.6**	0.4	-0.4	**-0.7**	**0.7**	**-0.7**	**-0.6**	**0.6**	15
3	MH_O	**-0.5**	**0.6**	**-0.6**	**0.6**	**-0.6**	**0.7**	**0.6**	**-0.6**	**-0.6**	**0.6**	**0.4**	**-0.5**	**-0.6**	**0.7**	**-0.6**	**-0.5**	**0.6**	16
4	MH_U	-0.4	**0.5**	-0.4	**0.5**	-0.4	**0.5**	0.4	-0.4	-0.4	0.3	0.3	-0.4	-0.4	**0.5**	**-0.5**	**-0.5**	**0.5**	7
5	F0_RS	**0.5**	**-0.6**	**0.7**	**-0.7**	**0.7**	**-0.7**	**-0.8**	**0.7**	**0.8**	**-0.8**	**-0.5**	0.4	**0.7**	**-0.7**	**0.7**	0.4	-0.4	14
6	F0_HP	**0.5**	**-0.6**	**0.7**	**-0.6**	**0.6**	**-0.6**	**-0.7**	**0.7**	**0.7**	**-0.7**	-0.4	0.3	**0.6**	**-0.6**	**0.6**	0.3	-0.3	13
7	PW_R	**0.5**	**-0.7**	**0.7**	**-0.7**	**0.7**	**-0.7**	**-0.8**	**0.8**	**0.8**	**-0.8**	-0.4	0.4	**0.8**	**-0.8**	**0.7**	**0.5**	**-0.5**	15
8	PW_RHT	0.1	-0.3	0.3	-0.3	**0.6**	-0.4	**-0.5**	**0.6**	**0.5**	**-0.5**	0.0	0.0	**0.6**	**-0.5**	**0.5**	0.2	-0.2	8
9	PW_RAP	0.3	-0.3	0.4	-0.4	0.4	-0.3	-0.4	0.4	**0.5**	**-0.5**	-0.2	0.2	0.4	-0.4	0.4	0.0	-0.1	2
10	SP_F1	**-0.6**	**0.6**	**-0.5**	**0.6**	-0.3	**0.5**	0.3	-0.3	-0.4	0.3	**0.5**	**-0.6**	-0.3	**0.5**	-0.4	-0.4	**0.5**	9
11	DU_TL	-0.3	0.4	-0.3	0.4	-0.3	0.4	0.2	-0.2	-0.2	0.1	0.3	**-0.5**	-0.2	0.4	-0.3	-0.4	**0.5**	2
	#	7	8	7	8	7	8	7	7	8	8	3	4	7	9	8	5	7	118

dimension in the top layer. Firstly, the correlations between the elements of the top layer and the middle layer were calculated as follow: let $x^{(i)} = \{x_n^{(i)}\}(n = 1, 2, \ldots, N)$ be the sequence of the values of the i^{th} emotion dimension rated with the listening test, $i \in \{Valence, Activation, Dominance\}$, where N is the number of utterances in our database ($N = 179$ for the Japanese database and $N = 200$ for the German database). Moreover, let $s^{(j)} = \{s_n^{(j)}\}(n = 1, 2, \ldots, N)$ be the sequence of the values of the j^{th} semantic primitive rated with another listening test, $j \in \{Bright, Dark, \ldots, Slow\}$, where N is the number of utterances in our database. Then, the correlation coefficient $R_j^{(i)}$ between the semantic primitive $s^{(j)}$ and the emotion dimension $x^{(i)}$ can be determined by the following equation:

$$R_j^{(i)} = \frac{\sum_{n=1}^{N}(s_{j,n} - \overline{s_j})(x_n^{(i)} - \overline{x}^{(i)})}{\sqrt{\sum_{n=1}^{N}(s_{j,n} - \overline{s_j})^2}\sqrt{\sum_{n=1}^{N}(x_n^{(i)} - \overline{x}^{(i)})^2}} \quad (1)$$

where $\overline{s_j}$, and $\overline{x}^{(i)}$ are the arithmetic mean of the semantic primitive and emotion dimension respectively. Correlation coefficients between semantic primitives and emotion dimensions for the German database are shown in Table 1.

The correlation coefficients between elements of the middle layer (semantic primitive), and the bottom layer (acoustic feature) are calculated as follows. Let $a_m = \{a_{m,n}\}(n = 1, 2, \ldots, N)$ be the sequence of values of the m^{th} acoustic feature, $m = 1, 2, \ldots, M$, where M is the number of extracted acoustic features in this study, $M = 21$. Moreover, let $s^{(j)} = \{s_n^{(j)}\}(n = 1, 2, \ldots, N)$ be the sequence of the rated values of the j^{th} semantic primitive, $j \in \{Bright, Dark, \ldots, Slow\}$, where N is the number of utterances in our database. Then the correlation coefficient $R_m^{(j)}$ between the acoustic parameter a_m and the semantic primitive $s^{(j)}$ can be determined by the following equation:

$$R_m^{(j)} = \frac{\sum_{n=1}^{N}(a_{m,n} - \overline{a_m})(s_n^{(j)} - \overline{s}^{(j)})}{\sqrt{\sum_{n=1}^{N}(a_{m,n} - \overline{a_m})^2}\sqrt{\sum_{n=1}^{N}(s_n^{(j)} - \overline{s}^{(j)})^2}} \quad (2)$$

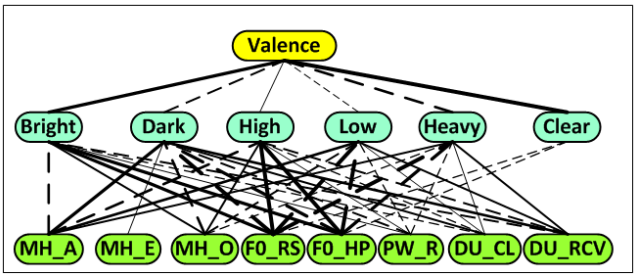
where $\overline{a_m}$, and $\overline{s}^{(j)}$ are the arithmetic mean for the acoustic feature and semantic primitive respectively.

Table 2 lists only 11 acoustic features, which yield a significant correlation with semantic primitives.

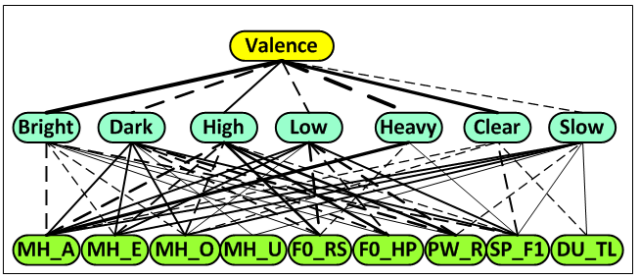
B. Selection Results

Using this method, firstly, the most relevant semantic primitives were selected for each emotion dimension. Secondly, the most relevant acoustic features for each semantic primitive were selected. Finally, a perceptual three-layer model was constructed for each emotion dimension as follows: the emotion

dimensions in the top layer, and the most relevant semantic primitives for this dimension are in the middle layer, while the relevant acoustic features are in the bottom layer. For example, Fig. 4 illustrates the valence perceptual model, where the solid lines in this figure represent a positive correlation, and the dashed ones indicate a negative correlation. The thickness of each line indicates the strength of the correlation; the thicker the line is, the higher the correlation.



(a) Japanese Database



(b) German Database

Fig. 4. Valence perceptual model

Using the acoustic features selection method, which was introduced in the previous subsection for the German database, it was found that there were 7 semantic primitives which are highly correlated with valence. These semantic primitives are the adjectives describing the valence dimension, as shown in the middle layer of Fig. 4(b). These 7 semantic primitives are highly correlated with 9 acoustic features, as shown in the bottom layer of Fig. 4(b), which implies that these acoustic features can be used to improve the estimation of the valence dimension. In a similar way, two perceptual three-layer models were constructed for activation and dominance. For activation, it was found that 9 acoustic features were highly correlated with the semantic primitives which are more correlated to activation, while for dominance, it was found that 10 acoustic features were selected to represent the most relevant acoustic features. The results for the Japanese database were presented in [22].

Here, in order to construct a perceptual three-layer model for each emotion dimension in the case of cross-language: we firstly construct a perceptual three-layer model individually for each dimension for the two databases, then the common acoustic features between the two languages were selected to constitute the bottom layer for the cross-language perceptual models. Moreover, the common semantic primitives between

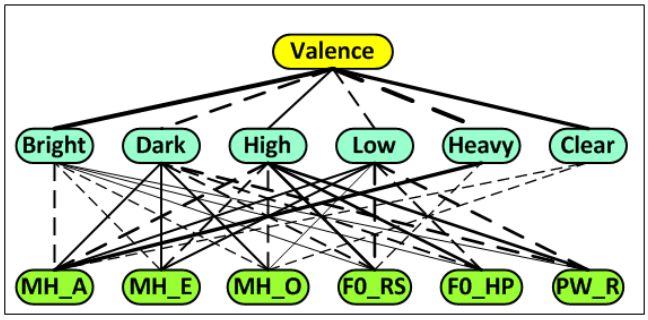


Fig. 5. Cross-lingual: valence perceptual model.

the two-languages were selected as semantic primitives for the cross-language case. For example, the valence perceptual model for the cross-language is shown in Fig. 5.

C. Discussion

Our model mimics the human perception process for understanding emotions based on Brunswick's lens model [20], where the speaker expresses his/her emotional state through some acoustic features. These acoustic features are interpreted by a listener as some adjectives describing the speech signal. From these adjectives, the listener can judge the emotional state. For example, if the adjectives describing the voice are Dark, Slow, Low and Heavy, these make the human listener feel that the emotional state is of negative valence and of very weak activation, so that it can be detected as a Sad emotional state in the categorical approach.

On the other hand, the traditional acoustic features selection method was based on the correlations between acoustic features and the emotion dimension as a two-layer model. In order to investigate the effectiveness of the proposed feature selection method, the results were compared with the traditional method. Therefore, the traditional feature selection method was used by calculating the correlations between acoustic features and emotion dimensions directly, as shown in Table III.

From this investigation, it is obvious that only one acoustic feature is highly correlated with the valence dimension, 8 acoustic features are highly correlated with the activation and dominance dimensions. Therefore, valence shows a smaller number of correlated acoustic features than that of activation and dominance. This result is similar to that described in many previous studies [27], [5]. The poor correlation between the acoustic features and valence is the reason behind the very low performance for valence estimation using the traditional approach. Due to this drawback, most of the previous studies achieved a good performance for the activation and dominance estimation, while a lower performance was obtained for the valence [28], [29].

The most important result is that, using the proposed three-layer model for feature selection, the number of acoustic features correlated to emotion dimensions increases. For example, the number of correlated features for the most challenging

TABLE III
CORRELATION COEFFICIENTS BETWEEN THE ACOUSTIC FEATURES AND THE EMOTION DIMENSIONS FOR THE GERMAN DATABASE

m	AF/ED	V	A	D	#
1	MH_A	-0.33	**-0.82**	**-0.81**	2
2	MH_E	-0.18	**-0.70**	**-0.71**	2
3	MH_I	-0.03	-0.19	-0.24	0
4	MH_O	-0.28	**-0.67**	**-0.68**	2
5	MH_U	-0.25	**-0.47**	**-0.47**	2
6	F0_RS	0.21	**0.69**	**0.65**	2
7	F0_HP	0.19	**0.59**	**0.54**	2
8	F0_AP	-0.05	-0.14	-0.13	0
9	F0_RS1	-0.05	-0.10	-0.09	0
10	PW_R	0.23	**0.75**	**0.74**	2
11	PW_RHT	-0.25	0.44	**0.49**	1
12	PW_RS1	0.08	0.14	0.14	0
13	PW_RAP	0.08	0.36	0.35	0
14	SP_F1	**-0.55**	**-0.49**	-0.43	2
15	SP_F2	-0.03	-0.29	-0.29	0
16	SP_F3	-0.04	-0.04	0.01	0
17	SP_TL	0.28	0.26	0.26	0
18	SP_SB	-0.02	-0.05	-0.02	0
19	DU_TL	-0.28	-0.38	-0.39	0
20	DU_CL	-0.24	-0.36	-0.36	0
21	DU_RCV	-0.14	-0.39	-0.37	0
	#	1	8	8	17

dimension valence increases from one feature using the traditional method to 9 features using the proposed method. Moreover, the number of features for activation increased from 8 to 9, and for dominance from 8 to 10. These selected acoustic features can be used to improve emotion dimension estimation, as described in details in the next section.

The comparison between Fig. 4(a), and Fig. 4(b) helps us to find the common acoustic features and semantic primitives related to each emotion dimension for the Japanese and German languages. For example, in both languages, the valence dimension is usually positively correlated with bright, high and clear semantic primitives, while it is negatively correlated with dark, low, and heavy semantic primitives. Therefore, the two languages not only share the same semantic primitives but also similar correlations between the emotion dimensions and the corresponding semantic primitives. Similarly, in both languages, the 6 semantic primitives were found to be correlated with 6 acoustic features.

V. Automatic Emotion Recognition System

The task of emotion recognition using the dimensional approach can be viewed as using an estimator to map the acoustic features to real-valued emotion dimensions. The desired output is not a classification into one of a finite set of categories but an estimation of real-values for the emotion dimensions of valence, activation, and dominance. However, every point in the dimensional space can be mapped into one emotion category.

In the previous section, a perceptual three-layer model was constructed for each emotion dimension. Emotion dimension values can be estimated using any estimator such as K-nearest neighborhood (KNN), Support Vector Regression (SVR), a Fuzzy Inference System (FIS) or any other estimator. In this study, for selecting the best estimator among KNN, SVR and FIS, pre-experiments not included here indicated that our best results were achieved using the FIS estimator. Therefore, FIS was used to connect the elements of the three-layer model. Most of the statistical methodology is mainly based on linear and precise relationships between the input and the output, while the relationship between acoustic features, semantic primitives, and emotion dimensions is non-linear. Therefore, fuzzy logic is a more appropriate mathematical tool for describing this non-linear relationship [28], [21], [30].

A. System Implementation

An Adaptive-Network-based Fuzzy Inference System (ANFIS) was used to construct FIS models which connect the elements of our recognition system. Each FIS has the structure of multiple inputs and of one output. Having identified the best acoustic features set, we constructed an individual estimator to predict the values (-2 to 2 rated by the listening test) of each emotion dimension. For example, in the case of cross-language, in order to estimate the valence dimension using the perceptual model in Fig. 5, a bottom-up method was used to estimate the values (1 to 5 rated by the listening test) of the 6 semantic primitives in the middle layer from the 6 acoustic features in the bottom layer, as shown in Fig. 6. In order to accomplish this task, 6 FISs were needed, one for estimating each semantic primitive. In addition, one FIS was needed to estimate the value of the valence dimension from the 6 semantic primitives. In a similar way, the activation and dominance can be estimated using FIS for each semantic primitive, and one FIS for the activation and dominance respectively.

VI. System Evaluation

The aim of this study, is to investigate whether an automatic emotion recognition system trained using one language has

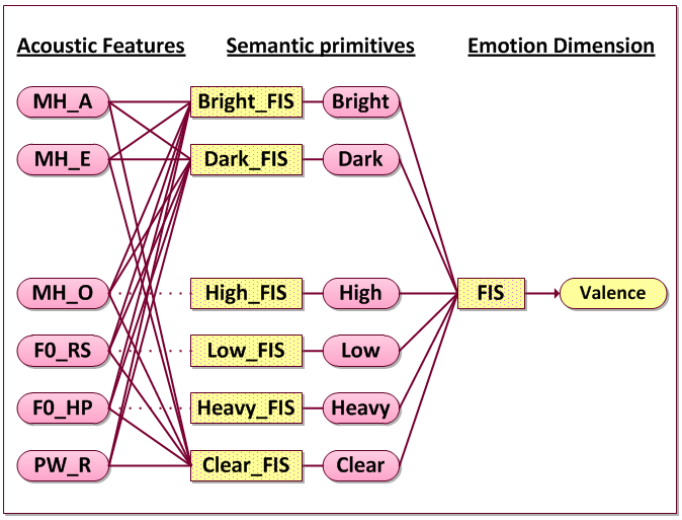
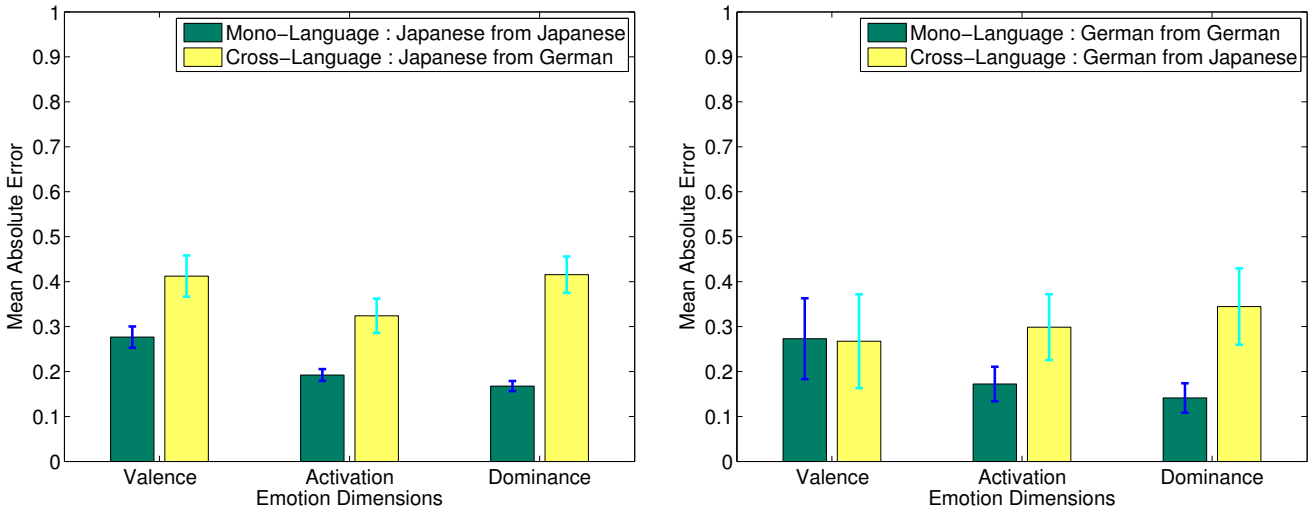


Fig. 6. Block diagram of the proposed approach for estimating valence based on the three-layer model



(a) Japanese Database

(b) German Database

Fig. 7. Mean Absolute Error (MAE) between human evaluation and the estimated values of emotion dimensions, in the case of mono-language and cross-language

the ability to detect the emotion dimension from different languages. To accomplish this task, the most important acoustic features for the two-languages were investigated. As explained in section IV, it was found that 6 acoustic features were common between the two Japanese and German databases. For example, in the case of the valence dimension, 6 acoustic features can be used as the input of the proposed system, as shown in Fig. 6. The features found in German were used to estimate emotion dimensions for the Japanese, and vice-versa.

The mean absolute error MAE between the predicted values of emotion dimensions and the corresponding average value given by human subjects is used as a metric of the discrimination associated with each case. The MAE is calculated according to the following equation

$$MAE^{(j)} = \frac{\sum_{i=1}^{N} |\widehat{x}_i^{(j)} - x_i^{(j)}|}{N} \tag{3}$$

where $j \in \{valence, activation, dominance\}$, $\widehat{x}_i^{(j)}$ is output of the emotion recognition system, and $x_i^{(j)}, -2 \leq x_i^{(j)} \leq 2$ is the values evaluated by the human subjects, as described in Subsection III-D.

The estimations of emotion dimensions using the acoustic features of Japanese utterances from a speech emotion recognition system trained using the information from the German database will be explained in section VI-A, while in section VI-B, the estimations of emotion dimensions for German utterances from another speech emotion recognition system trained using the Japanese database will be presented. In order to avoid the multi-speaker variation, the evaluation of cross-language emotion recognition systems was conducted by training the system using one speaker from one language, and by testing the system using one speaker from the other language.

A. Emotion dimension estimation for the Japanese database from the German database

For detecting Japanese from German, we build 10 automatic emotion recognition systems, one for each German speaker. The 10 systems were trained using the German utterances and tested using Japanese utterances. The Japanese database was tested 10 times using German speakers.

Finally, for each utterance in the Japanese database, we have 10 estimations from 10 different speakers for the three dimensions: valence, activation, and dominance. The average value for each dimension was calculated for each utterance.

For a comparative analysis of the performance of the proposed cross-language emotion recognition system, the results were compared with those of the mono-language emotion recognition system, which was trained and tested using the Japanese database. The MAE for all emotion dimensions, for the mono-language (Japanese-from-Japanese) system and the cross-language (Japanese-from-German) system are illustrated in Fig. 7(a).

From this figure, the MAE for estimating Japanese emotion dimensions from the German database is as follows: the valence increased from 0.28 in the mono-language case to 0.41 in the cross-language case, the activation increased from 0.19 to 0.32, and the dominance increased from 0.17 to 0.42. In all cases, the mean absolute error of emotion dimensions increased, however these increments do not constitute a large difference.

Another interesting evaluation method for the proposed system can be performed by comparing the estimation of MAE for the dimensional approach with the performance achieved based on the categorical approach. Therefor, the dimensional space is mapped into emotion categories using Gaussian Mixture Model GMM classifier. Using GMM every point in the dimensional space is mapped into one emotion

category. Thus, the estimated values of emotion dimensions valence, activation, dominance were used as an input features to train GMM classifier to classify emotional state into emotion categories. Therefore, every point in the dimensional space is mapped into one emotion category. The confusion matrix of the results is shown in Table IV in the mono-language case, and in Table V for the cross-language case.

TABLE IV
MONO-LANGUAGE: CONFUSION MATRIX FOR AUTOMATICALLY CLASSIFYING EMOTION CATEGORIES FROM EMOTION DIMENSIONS USING A GMM CLASSIFIER, FOR JAPANESE-FROM-JAPANESE DATABASE, (AVERAGE RECOGNITION RATE 94.0%)

Category	Classification rate (%)				
	Neutral	Joy	Cold Anger	Sad	Hot Anger
Neutral	**80.0**	10.0	5.0	5.0	0.0
Joy	0.0	**97.5**	2.5	0.0	0.0
Cold Anger	0.0	0.0	**100.0**	0.0	0.0
Sad	0.0	0.0	0.0	**100.0**	0.0
Hot Anger	0.0	2.5	5.0	0.0	**92.5**

TABLE V
CROSS-LANGUAGE: CONFUSION MATRIX FOR AUTOMATICALLY CLASSIFYING EMOTION CATEGORIES FROM EMOTION DIMENSIONS USING A GMM CLASSIFIER, FOR JAPANESE-FROM-GERMAN DATABASE, (AVERAGE RECOGNITION RATE 92.7%)

Category	Classification rate (%)				
	Neutral	Joy	Cold Anger	Sad	Hot Anger
Neutral	**95.0**	0.0	5.0	0.0	0.0
Joy	0.0	**100.0**	0.0	0.0	0.0
Cold Anger	0.0	0.0	**100.0**	0.0	0.0
Sad	0.0	0.0	0.0	**100.0**	0.0
Hot Anger	0.0	0.0	30.0	0.0	**70.0**

The emotion classification accuracies listed in the above tables correspond to the MAEs for the dimensional approach. It is clearly seen that the recognition rate in the mono-language case is 94.0% which decreased to 92.7% for the cross-language system for detecting Japanese-from-German. Therefore, we can conclude that emotion dimensions for the Japanese database can be detected from a speech emotion recognition system trained with the German database with a small error.

B. Emotion dimensions estimation for the German database from the Japanese database

On the other hand, in order to estimate German from Japanese, we construct one cross-language emotion recognition system trained using the Japanese database. This system was tested using the utterances from German, for each German speaker individually. The MAE for the estimation of the whole database using the cross-language emotion recognition system was calculated and compared with the emotion dimensions estimation in the case of the mono-language, as shown in Fig. 7(b).

From this figure, the MAE for estimating German speakers emotion dimensions from a Japanese database is as follows: The estimation for the valence is unchanged, the activation increases from 0.17 in the mono-language case to 0.30 in the cross-language case, and the dominance increases from 0.14 to 0.34. In the cases of the activation and dominance, the mean absolute error of the emotion dimension increases, however these increments do not constitute a large difference. Therefore, we can conclude that the emotion dimension for the German database can be detected from a speech emotion recognition system trained with the Japanese database with a small error.

Moreover, the results of classification for German database into 4 categories Neutral, Happy, Anger, and Sad are as follow: the confusion matrix of the results is shown in Table VI, for mono-language German-from-German emotion recognition system, Table VII for cross-language German-from-Japanese system.

TABLE VI
MONO-LANGUAGE: CONFUSION MATRIX FOR AUTOMATICALLY CLASSIFYING EMOTION CATEGORIES FROM EMOTION DIMENSIONS USING A GMM CLASSIFIER, FOR GERMAN-FROM-GERMAN IN CASE OF SPEAKER-DEPENDENT RESULTS (AVERAGE RECOGNITION RATE 95.5%)

Category	Classification rate (%)			
	Neutral	Happy	Anger	Sad
Neutral	**98.0**	0.0	2.0	0.0
Happy	0.0	**94.0**	6.0	0.0
Anger	0.0	8.0	**92.0**	0.0
Sad	2.0	0.0	0.0	**98.0**

TABLE VII
CROSS-LANGUAGE: CONFUSION MATRIX FOR AUTOMATICALLY CLASSIFYING EMOTION CATEGORIES FROM EMOTION DIMENSIONS USING A GMM CLASSIFIER, FOR GERMAN-FROM-JAPANESE IN CASE OF SPEAKER-DEPENDENT RESULTS (AVERAGE RECOGNITION RATE 89.0%)

Category	Classification rate (%)			
	Neutral	Happy	Anger	Sad
Neutral	**100.0**	0.0	0.0	0.0
Happy	0.0	**76.0**	24.0	0.0
Anger	0.0	4.0	**96.0**	0.0
Sad	12.0	2.0	2.0	**84.0**

The results in the above tables reveal that there is small degradation for detecting emotion using cross-language recognition system, where the average recognition rate decreased from 95.5% in case of mono-language to 89.0% in case of cross-language.

VII. CONCLUSION

In this paper, we investigate whether the following assumption is satisfied or not: an automatic speech emotion recognition system can detect the emotional state regardless of the used language. We adopt a three-layer model of human perception, in order to precisely predict the values of the emotion dimensions from the acoustic features.

In this study, a novel feature selection method based on the three-layer model was successfully used to find many acoustic features related to each emotion dimension. These acoustic features were used as the inputs to the speech

emotion recognition system, the output of our system are the estimated values of emotion dimensions: valence, activation, and dominance. Through identification of the best acoustic features, the estimation performance of the proposed system is improved. Therefore, using the proposed model for building an automatic speech emotion recognition system allows us to find many acoustic features, which allow us to investigate the cross-language mode.

For estimating emotion dimensions for the German database from the Japanese database, the proposed system was trained using the Japanese database, which contains one speaker, and tested using 10 German speakers individually. To detect the emotional state for the Japanese database from German database, we trained 10 speech emotion recognition systems using one German speaker at a time. Finally, the results reveal that values of emotion dimensions for the Japanese database can be detected from a cross-language speech emotion recognition system trained with German database with a small error, and and vice-versa.

The most important result is that, using our normalization method, we found that emotion recognition is language independent, which means that our assumption that the positions of neutral voices are different among languages, and that the distance and directions from the neutral voice to other emotional ones are common among languages is confirmed.

ACKNOWLEDGMENTS

This study was supported by the Grant-in-Aid for Scientific Research (A) (No. 25240026), SCOPE (No. 131205001) by Ministry of Internal Affairs and Communications, Grant-in-Aid for Exploratory Research (No. 22650032), and the A3 Foresight Program made available by the Japan Society for the Promotion of Science (JSPS).

REFERENCES

[1] O. Pierre-Yves, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, vol. 59, pp. 157–183, July 2003.

[2] C.M. Lee, and S. Narayanan, "Toward Detecting Emotions in Spoken Dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13(2), pp. 293–303, 2005.

[3] I. Albrecht, and M. Schroder, and J. Haber, and H.-P. Seidel, "Mixed feelings: Expression of non-basic emotions in a muscle-based talking head," *Virtual Reality*, vol. 8(4), pp. 201-212, 2005.

[4] D. Wu, and T.D. Parsons, and S. Narayanan, "Acoustic Feature Analysis in Speech Emotion Primitives Estimation," *Proc. InterSpeech 2010*, pp. 785–788, 2010.

[5] M. Schroder, and R. Cowie, and E.D.-cowie, M. Westerdijk, and S. Gielen, "Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis," *Proc. Eurospeech 2001*, pp. 87–90, 2001.

[6] M. Grimm, and K. Kroschel, "Emotion Estimation in Speech Using a 3D Emotion Space Concept," *in Robust Speech Recognition and Understanding, M. Grimm and K. Kroschel (Eds.)*, June 2007.

[7] I. Kanluan, M. Grimm, and K. Kroschel, "Audio-Visual Emotion Recognition Using An Emotion Space Concept," *Proc. EUSIPCO 2008*, 2008.

[8] Q. Zhang, S. Jeong, M. Lee, "Autonomous emotion development using incremental modified adaptive neuro-fuzzy inference system," *Neurocomputing*, vol. 86, pp. 33-44, 2012.

[9] J.-A. Bachorowski and M.J. Owren, "Sounds of Emotion," Annals of the New York Academy of Sciences, 1000(1), pp. 244-265, January 2006.

[10] R. Banse and K.R. Scherer. "Acoustic profiles in vocal emotion expression," Journal of personality and social psychology, 70(3), pp. 614-636, March 1996.

[11] K.R. Scherer, R. Banse, H.G. Wallbott, and T. Goldbeck. Vocal cues in emotion encoding and decoding. Motivation and Emotion," 15(2), pp. 123-148, June 1991.

[12] T. Fritz, S. Jentschke, N. Gosselin, D. Sammler, I. Peretz, R. Turner, A.D. Friederici, and S. Koelsch, "Universal recognition of three basic emotions in music," Current biology : CB, 19(7) pp. 573-576, April 2009.

[13] P.G. Hunter, E.G. Schellenberg, and U. Schimmack, "Feelings and perceptions of happiness and sadness induced by music: Similarities, differences, and mixed emotions," Psychology of Aesthetics, Creativity, and the Arts, 4(1) pp. 47-56, 2010.

[14] J. Dang, A. Li, D. Erickson, A. Suemitsu, M. Akagi, K. Sakuraba, N. Minematsu, K. Hirose, "Comparison of Emotion Perception among Different Cultures," APSIPA,Sapporo, Japan, 2009.

[15] B. Xie, L. Chen, G.-C. Chen, and C. Chen, "Statistical Feature Selection for Mandarin Speech Emotion Recognition," Springer Berlin / Heidelberg, 2005.

[16] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, and N. Amir, Whodunnit - searching for the most important feature types signalling emotion-related user states in speech, Comput. Speech Lang., vol. 25, no. 1, pp. 4-28, 2010.

[17] T. Polzehl, A. Schmitt, and F. Metze, Approaching multilingual emotion recognition from speech - on language dependency of acoustic/ prosodic features for anger detection, in Proc. of the Fifth International Conference on Speech Prosody, 2010.

[18] H.P. Espinosa, C.A.R. Garcia, L.V. Pineda, "Bilingual Acoustic Feature Selection for Emotion Estimation Using a 3D Continuous Model," *Proc. Automatic Face and Gesture Recognition (FG 2011)*, pp. 786-791, 2011.

[19] K.R. Scherer, "Personality inference from voice quality: The loud voice of extroversion," European Journal of Social Psychology, Vol. 8, pp. 467-487, 1978.

[20] E. Brunswik, "Historical and thematic relations of psychology to other sciences," Scientific Monthly, Vol. 83, pp. 151-161, 1956.

[21] C. Huang, and M. Akagi, "A three-layered model for expressive speech perception," *Speech Communication*, vol. 50(10), pp. 810–828, October 2008.

[22] R. Elbarougy and M. Akagi, "Speech Emotion Recognition System Based on a Dimensional Approach Using a Three-Layered Model," Proc. Int. Conf. APSIPA ASC, 2012.

[23] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," Proceedings of Interspeech, Lissabon, Portugal, 2005.

[24] S. Ramakrishnan, and I. El-Emary, "Speech emotion recognition approaches in human computer interaction," Telecommunication Systems, pp. 1-12, (2011).

[25] H. Kawahara, "STRAIGHT, Exploration of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," Acoust. Sci & Tech., 27(6), pp. 349-353, 2006.

[26] 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, pp.36-50, 2011.

[27] M. Grimm, and K. Kroschel, "Rule-Based Emotion Classification Using Acoustic Features," *Proc. Int. Conf. on Telemedicine and Multimedia Communication*, 2005.

[28] M. Grimm, and K. Kroschel, and E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, pp. 787–800, 2007.

[29] S. Wu, T.H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53(5), pp. 768–785, May 2011.

[30] J.-S.R. Jang, "ANFIS: Adaptive network-based fuzzy inference system," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 23(3), pp. 665–685, 1993.