Acoust. Sci. & Tech.

PAPER

Improving Speech Emotion Dimensions Estimation Using a Three-Layer Model of 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*

Abstract: Most previous studies using the dimensional approach mainly focused on the direct relationship between acoustic features and emotion dimensions (valence, activation, and dominance). However, the acoustic features that correlate to valence dimension are very few and very weak. As a result, the valence dimension has been particularly difficult to predict. The purpose of this research is to construct a speech emotion recognition system that has the ability to precisely estimate values of emotion dimensions especially valence. This paper proposes a three-layer model to improve the estimating values of emotion dimensions from acoustic features. 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. First, a top-down acoustic feature selection method based on this model was conducted to select the most relevant acoustic features for each emotion dimension. Then, a button-up method was used to estimate values of emotion dimensions from acoustic features by firstly using fuzzy inference system (FIS) to estimate the degree of each semantic primitive from acoustic features, then using another FIS to estimate values of emotion dimensions from the estimated degrees of semantic primitives. The experimental results reveal that the constructed emotion recognition system based on the proposed three-layer model outperforms the conventional system.

Keywords: Emotion dimensions, Automatic speech emotion recognition, Multi-layer model, Fuzzy Inference Systems (FIS).

PACS number: 43.71.Hw, 43.71.Bp, 43.71.An, 43.72.Ja

1. Introduction

Most previous techniques for automatic speech emotion recognition focus only on the classification of emotional states as discrete categories such as happy, sad, angry, fearful, surprised, and disgusted [1]. 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. In the real-life, an emotional state has different degrees of intensity and may change over time depending on the situation from low to high degree. Therefore, an automatic speech emotion recognition system should be able to detect the degree or the level of the emotional state from the voice [2]. Hence, a number of researchers advocate the use of dimensional descriptions of human emotion, where emotional states are estimated as a point in a multi-dimensional space [3, 4].

In this study, a three-dimensional continuous model is adopted in order to represent the emotional states using the emotion dimensions, i.e. valence, activation, and dominance. These dimensions are a suitable representation, because they are capable of representing low-intensity as well as high-intensity states [2].

However, although the conventional dimensional model for estimating emotions from speech signals allows the representation of the degree of emotional state, it has the following drawbacks: (i) we do not know what acoustic features are related to each emotion dimension, (ii) the acoustic features that correlate to the valence dimension are less numerous, less strong, and more inconsistent [4], and (iii) the values of emotion dimensions are difficult to estimate precisely only on the basis of acoustic information [5]. Due to these limitations, it has been difficult to directly predict the values of the valence dimension using the acoustic features.

The goal of this paper is to improve the conventional dimensional method in order to precisely predict values of the valence dimension as well as improve prediction of those of the activation and dominance. This will be achieved by constructing a speech emotion recognition system which have the ability to accurately estimate emotion dimensions based on the thee-layer model of human perception. The aim of constructing this system

[*] e-mail: elbarougy@jaist.ac.jp
[†] e-mail: akagi@jaist.ac.jp

is to prove the effectiveness of the proposed three-layer model. The following section introduces the proposed emotion recognition approach based on human perception.

2. Emotion Recognition Strategy

Conventional speech emotion recognition methods are mainly based on investigating the relationship between acoustic features and emotion dimensions as a two-layer model, i.e. acoustic feature layer and emotion dimension layer. For instance, Grimm et al. attempted to estimate the emotion dimensions (valence, activation, and dominance) from the acoustic features by using a fuzzy inference system (FIS) [6]. However, they found that activation and dominance were more accurately estimated than valence. Furthermore, many researchers also tried to investigate the most relevant acoustic features for each emotion dimension by using the correlation between a set of acoustic features and emotion dimensions [3–5, 7]. In all these studies, the valence dimension was found to be the most difficult dimension to estimate. Consequently, some other studies focused only on exploring acoustic features related to valence dimension [8, 9]. Some emotions related to valence were found to share similar acoustic features such as happiness and anger, which were characterized by increased levels of fundamental frequency (F0) and intensity. This is one reason why acoustic discrimination on valence dimension is still problematic i.e. no strong discriminative acoustic features are available to discriminate between positive speech (e.g. happiness) and negative speech (e.g. anger) [7]. Therefore, a number of researchers tried to discriminate between the positive and negative emotions by combining acoustic and linguistic features to improve the valence estimation [7, 10]. However, the results on valence estimation remained poor.

Human perception, as described by Scherer [12] who adopted a version of Brunswik's lens model originally proposed in 1956 [13], is a multi-layer process. Huang and Akagi adopted a three-layer model for human perception. They assumed that human perception for emotional speech does not come directly from a change in acoustic features but rather a composite of different types of smaller perceptions that are expressed by semantic primitives or adjectives describing an emotional voice [14].

The two-layer model has limited ability to find the most relevant acoustic features for each emotion dimension, especially valence, or to improve the prediction of emotion dimensions from acoustic features. To overcome these limitations, this paper aims to identify the most relevant acoustic features describing emotion dimension using a novel idea based on human perception. We attempt to use the above human perception model proposed by Huang and Akagi [14] to find the most correlated acoustic features with emotion dimensions through semantic primitives. We assume that the acoustic features that are highly correlated with semantic primitives will have a significant impact for predicting values of emotion dimensions, especially valence. The findings can guide the selection of new acoustic features with better discrimination in the most difficult dimension.

The feasibility of our three-layer model to improve emotion dimensions estimation; for valence, activation, and dominance was investigated. The proposed model consists of three layers: emotion dimensions (valence, activation, and dominance) constitute the top layer, semantic primitives the middle layer, and acoustic features the bottom layer. A semantic primitive layer is added between the two conventional layers acoustic features and emotion dimensions as shown in Fig. 1.

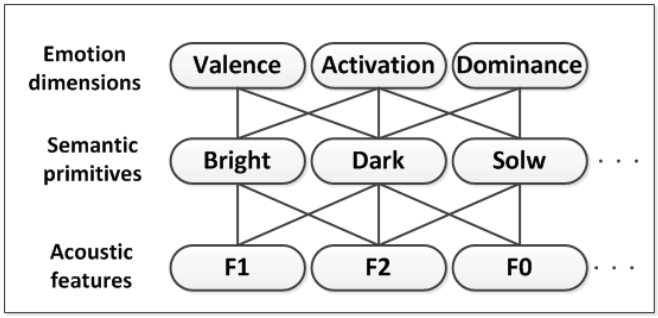


Fig. 1 Three layer model.

Therefore, the approach we adopt to estimate values of emotion dimensions includes the following steps:

- Feature selection: The most relevant acoustic features were selected by using a top-down method. First, the semantic primitives which have high correlations with each emotion dimension were selected. Then, the acoustic features which have high correlations with the selected semantic primitives found in the first step were selected.
- Building a three-layer model for each emotion dimension: For example, in the case of valence dimension, the three layers are: valence dimension in the top layer, the highly correlated semantic primitives with valence dimension in the middle layer, all the highly correlated acoustic features with all semantic primitives in the bottom layer.
- Emotion dimensions estimation: By using the constructed three-layer model, a button-up method was used to estimate values of emotion dimensions

from acoustic features as follows. First, FIS was used to estimate the degree of each semantic primitive from acoustic features, and then another FIS was used to estimate values of emotion dimension from the estimated degrees of semantic primitives in the first step.

The achieve the aim of this paper the following investigations are required: (1) whether the selecting acoustic features based on the proposed three-layer model of human perception will help us to find the most related acoustic features for each emotion dimensions, (2) whether using these selected acoustic features as inputs to an automatic emotion recognition system will improve the accuracy of all emotion dimensions especially valence, (3) finally, whether the automatic emotion recognition system is effective in the following cases: speaker-dependent, multi-speaker, and multi-language.

3. Databases and Experimental Evaluation

To construct an emotion recognition system, the elements of the proposed model 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 the below subsections.

3.1. Speech Material and Subjects

In this paper, our aim is to prove a new concept, not to construct a real-life application, consequently, acted emotions are quite adequate as a testing data [15]. Therefore, in order to validate the proposed system, we used two acted databases of emotional speech: one in Japanese (single-speaker) and the other in German (multi-speaker).

The Japanese database is the multi-emotion single-speaker Fujitsu database produced and recorded by Fujitsu Laboratories. A professional actress was asked to produce utterances using five emotional speech categories, i.e., neutral, joy, cold anger, sadness, and hot anger. In the database, there are 20 different Japanese sentences. Each sentence has one utterance in neutral and two utterances in each of the other categories. Thus, there are nine utterances for each sentence and 180 utterances for all 20 sentences. However, one cold anger utterance is missing so, the total number of utterance for Japanese database is 179.

The Japanese database is inadequate for validating our emotion recognition system fully, because it is a single speaker database which is only suitable for speaker-specific task. To investigate the effectiveness of the proposed system for multi-speaker and different languages, a Berlin database [17] was selected. It comprises of seven emotional states: anger, boredom, disgust, anxiety, happiness, sadness, and neutral speech. Ten professional German actors (five female and five male) spoke ten sentences with 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 because: (1) it is an acted-speech database the same as the Fujitsu database, (2) it contains four categories similar to those in the Fujitsu database (happy, angry, sad, and neutral), and (3) it is a multi-speaker and multi-gender database which enable us to investigate the effect of speaker and gender variation in speech emotion recognition. To compare the results of the two databases, we used only the four similar categories. Furthermore, for training proposes, we used sentences equally distributed between 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 five males and the other 100 by five females divided equally between the four emotional states.

To evaluate semantic primitives and emotion dimensions, we used listening tests. The Fujitsu database was evaluated by 11 graduate students, all native Japanese speakers (nine male and two female). While Berlin database was evaluated using nine graduate students, all native Japanese speakers (eight male and one female). No subjects have hearing impairments.

3.2. Acoustic Features

To construct a speech emotion recognition system, acoustic features are needed to be investigated. In this research, the most relevant acoustic features that have been successful in related works and features used for other similar tasks were selected. Therefore, 16 acoustic features that originate from F0, power envelope, power spectrum, and duration were selected from the work by Huang and Akagi [14]. In addition to these 16 acoustic features, five new parameters related to voice quality are added, because voice quality is one of the most important cues for the perception of expressive speech. Acoustic features related to duration are extracted by segmentation, and the rest are extracted by the high quality speech analysis-synthesis system STRAIGHT [18], leading to extraction of a set of 21 acoustic features that can be grouped in several subgroups:

F0 related features: F0 contour and power envelope varied greatly with different expressive speech cat-

egories, both for the accentual phrases as well as for the overall utterance. For each utterance the measurements made were F0 mean value of rising slope of the F0 contour (F0_RS), highest F0 (F0_HP), average F0 (F0_AP), and rising slope of the F0 contour for the first accentual phrase (F0_RS1).

Power envelope related features: in a similar way to that for the F0 contour, for each utterance the measurements were: mean value of power range in accentual phrase (PW_RAP), power range (PW_R), rising slope of the power for the first accentual phrase (PW_RS1), the ratio between the average power in high frequency portion (over 3 kHz), and the average power (PW_RHT);

Power spectrum related features: for spectrum we used formants, spectral tilt, and spectral balance:

- Formants: measures were the mean value of (first formant frequency (SP_F1), second formant frequency (SP_F2), third formant frequency (SP_F3) taken approximately at the midpoint of the vowels /a/,/e/,/i/,/o/, and /u/. The formants frequencies were calculated with LPC-order 12.
- Spectral tilt (SP_TL): is used to measure voice quality, and it was calculated from the following equation

$$SP\_TL = A1 - A3 \tag{1}$$

  where A1 is the level in dB of the first formant, and, A3 is the level of the harmonic whose frequency is closest to the third formant [19].
- Spectral balance (SP_SB): this parameter serves for the description of acoustic consonant reduction [20], and it was calculated according to the following equation

$$SP\_SB = \frac{\Sigma f_i.E_i}{\Sigma E_i} \tag{2}$$

where f_i is the frequency in Hz, and E_i is the spectral power as a function of the frequency [21].

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

Voice quality: Voice quality conveys both linguistic and paralinguistic information, which can be distinguished by acoustic source characteristics. Currently investigation into voice quality has focused on measures of breathiness, such as H1-H2, where H1 and H2, are the amplitudes (dB) of the fundamental frequency and the second harmonic, respectively. As indicated by Menezes et al. in [11], H1-H2 is concerned with glottal opening. In this study, the mean value of H1-H2 for vowel /a/,/e/,/i/,/o/, and /u/ per utterance MH_A, MH_E, MH_I, MH_O, and MH_U are used as an indication for voice quality.

All the 21 acoustic features were extracted for both Fujitsu and Berlin databases. In order to avoid speaker dependency on the acoustic features that are used, we adopt an acoustic feature normalization method, 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.

3.3. Evaluations of Semantic Primitives

In this study, the human perception model as described by Scherer [12] is adopted. This model assumes that human perception is a multi-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 by Huang and Akagi [14]. In this study 'smaller perception' means an earlier process of perception. These smaller percepts or adjectives are finally used to detect the emotional state of the speaker. These adjectives can be subjectively evaluated by human subjects. Therefore, the following set of adjectives describing the emotional speech 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. These adjectives were selected because they reflect a balanced selection of widely used adjectives that describe emotional speech. They are originally from the work of Huang and Akagi [14].

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 five-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, and 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. For Japanese database, the average of Pearson's correlation among every pairs of two subjects for all semantic primitives evaluation were ranged between 0.68 and 0.85, moreover, for German database, the average of correlations were ranged between 0.66 and 0.86. This result suggests that all subjects agreed from a moderate to a very high degree.

3.4. Emotion Dimensions Evaluation

Most existing emotional speech databases have been annotated using the categorical approach, while, few databases have been annotated using the dimensional approach [22]. 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 by the listening tests along three dimensions: valence, activation, and dominance. For emotion dimension 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 exited), and dominance (from -2 very weak to +2 very strong).

The subjects used a MATLAB GUI to evaluate the stimuli. Repetition was allowed. They 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 the work of Mori et al. [23] for emotion dimension evaluation, the basic theory of emotion dimension was explained to the subjects before the experiment started. Then they took a training session to listen to an example set composed of 15 utterances, which covered the used five-point scale, 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, and then choose score on the five-point scale for each dimension individually. The average of the subjects' rating for each emotion dimension was calculated per utterance.

The average of Pearson's correlation coefficient among every pairs of two subjects were as follows: for Japanese database 0.90, 0.85, and 0.89 for valence, activation, and dominance, respectively, and for German database 0.83, 0.87, and 0.86 for valence, activation, and dominance, respectively. This indicates that all subjects agreed to a high degree for all emotion dimension evaluation.

4. Selection of Acoustic Features and Semantic Primitives

This section describes the proposed acoustic features selection method to identify the most relevant acoustic features for emotion dimensions valence, activation, and dominance. For this purpose, we proposed a thee-layer model that imitates the human perception to understand the relationship between acoustic features and emotion dimensions.

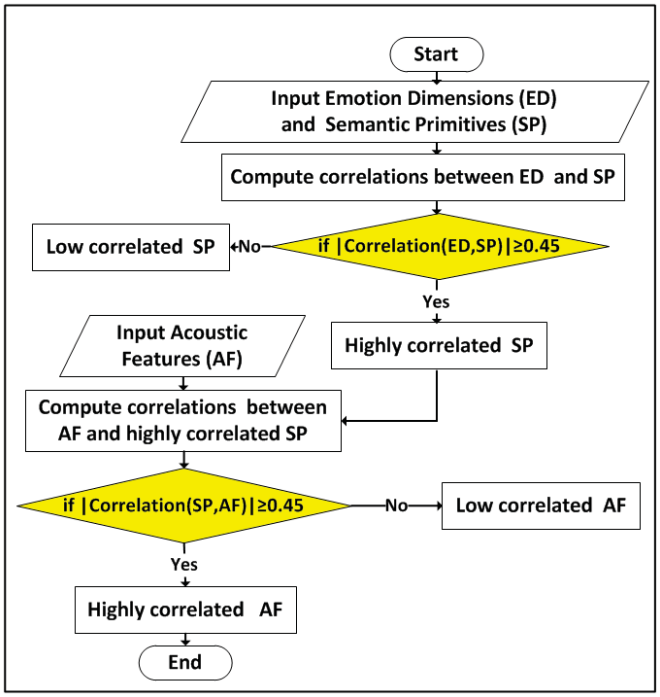


Fig. 2 Process for acoustic feature selection.

4.1. Selection Procedures

Our selection method is based on the following assumptions: 1) semantic primitives which are highly correlated with the emotion dimension are given large impact in the estimation of that dimension, and 2) acoustic features which are highly correlated with the semantic primitive are given large impact in the estimation of that semantic primitive. In this study, we consider the correlation highly correlated if its absolute value is grater than or equal to 0.45. To accomplish this task, the top-down method shown in Fig 2 was used as follows:

- the correlation coefficients between each emotion dimension (top-layer) and each semantic primitives (middle layer) were calculated;
- the highly correlated semantic primitives were selected for each emotion dimension;
- the correlation coefficients between each selected semantic primitive (middle layer) in the second step and each acoustic feature (bottom layer) were calculated,
- the highly correlated acoustic features were selected for each semantic primitive.

For each emotion dimension, the selected acoustic features in the final step are considered as the most relevant features to the dimension in the top-layer.

4.2. Correlation between elements of the three-layer model

First, the correlations between the elements of the top layer and the middle layer were calculated as follows:

Table 1 Correlation coefficients between semantic primitives (SP) and emotion dimensions (ED) (German Database).

ED \ SP	Bright	Dark	High	Low	Strong	Weak	Calm	Unstable	Well-modulated	Monotonous	Heavy	Clear	Noisy	Quiet	Sharp	Fast	Slow	#
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
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
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 2 Correlation coefficients between acoustic features (AF) and semantic primitives (SP) (German Database).

AF \ SP	Bright	Dark	High	Low	Strong	Weak	Calm	Unstable	Well-modulated	Monotonous	Heavy	Clear	Noisy	Quiet	Sharp	Fast	Slow	#
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
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
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
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
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
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
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
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
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
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
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

let $x^{(i)} = \{x_n^{(i)}\}(n = 1, 2, \ldots, N)$ be the sequence of the rated values of the i^{th} emotion dimension by the listening test, $i \in \{valence, activation, dominance\}$. 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 from another listening test, $j \in \{bright, dark, \ldots, slow\}$. Where N is the number of utterances in used database ($N = 179$ for Japanese and $N = 200$ for German). 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:

$$R_j^{(i)} = \frac{\sum_{n=1}^{N}(s_{j,n} - \overline{s_j})(x_n^{(i)} - \overline{x}^{(i)})}{\sqrt{\sum_{n=1}^{N}(s_{j,n} - \overline{s_j})^2}\sqrt{\sum_{n=1}^{N}(x_n^{(i)} - \overline{x}^{(i)})^2}} \quad (3)$$

where $\overline{s_j}$ and $\overline{x}^{(i)}$ are the arithmetic means for the semantic primitive and emotion dimension, respectively. Table 1 lists the correlation coefficients between all semantic primitives and all emotion dimensions for the German database. Where, the numbers in bold represent the higher correlations demonstrated by the absolute value of the correlation, which is ≥0.45. In addition, '#' in the last row and last column represents the number of higher correlations. For example, the number 7 in last column of the valence row indicates that their are seven semantic primitives highly correlated with valence.

Second, the correlations coefficients between elements of the middle layer (semantic primitive), and the bottom layer (acoustic feature) are calculated as follows: Let $f_l = \{f_{l,n}\}(n = 1, 2, \ldots, N)$ be the sequence of values of the m^{th} acoustic feature, $l = 1, 2, \ldots, L$, and L be the number of extracted acoustic features in this study $L = 21$. Then the correlation coefficient $R_l^{(j)}$ between the acoustic parameter f_l and the semantic primitive $s^{(j)}$ can be determined by the following equation:

$$R_l^{(j)} = \frac{\sum_{n=1}^{N}(f_{l,n} - \overline{f_l})(s_n^{(j)} - \overline{s}^{(j)})}{\sqrt{\sum_{n=1}^{N}(f_{l,n} - \overline{f_l})^2}\sqrt{\sum_{n=1}^{N}(s_n^{(j)} - \overline{s}^{(j)})^2}} \quad (4)$$

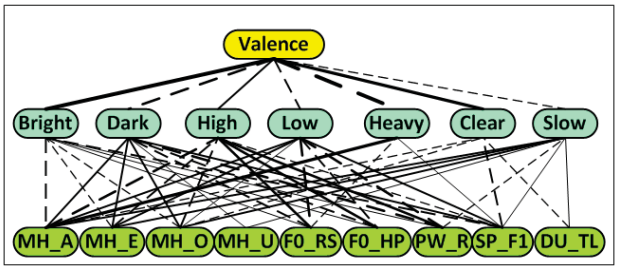
where $\overline{f_l}$, and $\overline{s}^{(j)}$ are the arithmetic means for the acoustic feature and semantic primitive respectively.

Table 2 lists the correlation coefficients between all semantic primitives and 11 acoustic features that has at least two highly correlation with semantic primitives,

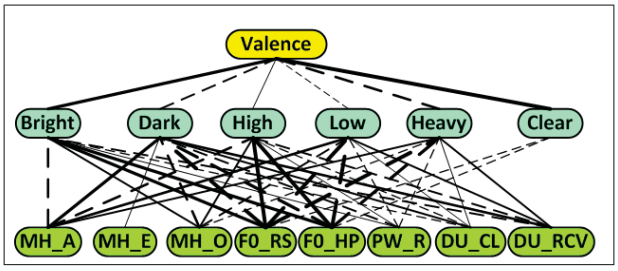
for the German database. Similar analysis was done for the Japanese database in our previous work [16].

4.3. Selection Results

For each emotion dimension, a perceptual three-layer model was constructed as follows: emotion dimension in the top layer, the most relevant semantic primitives for this dimension in the middle layer, and the most relevant acoustic features in the bottom layer. For example, Figs. 3(a) and 3(b) illustrate the valence perceptual model for German and Japanese database, respectively. Where the solid and dashed lines in these figures represent positive and negative correlations, respectively. Also, the thickness of each line indicates the strength of the correlation: the thicker the line, the higher the correlation.



(a) German database.



(b) Japanese database.

Fig. 3 Valence perceptual model.

In case of valence dimension for the German database as shown in Fig. 3(a), it is evident that seven semantic primitives were found highly correlated with valence as shown in the middle layer in Fig. 3(a). These seven semantic primitives are highly correlated with nine acoustic features as shown in the bottom layer in Fig. 3(a).

The valence perceptual model for German and Japanese language are compared as follows: For both languages, the valence dimension is found to be 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 six semantic primitives but also similar correlations between the emotion dimensions and the corresponding semantic primitives.

In addition, comparing the relationship between semantic primitives and acoustic features, it is found that the six semantic primitives that were shared by both German and Japanese have a similar correlations with six common acoustic features (MH_A, MH_E, MH_O, F0_RS, F0_HP, and PW_R). This finding suggests the possibility of some type of universality of acoustic cues associated with semantic primitives. Therefore, the proposed method can be used effectively to select the most relevant acoustic features for each emotion dimension regardless the used language.

4.4. Discussion

Our model mimics the human perception process for understanding emotions on the basis of Brunswick's lens model [13], where the speaker expresses his/her emotional state through some acoustic features. These acoustic features are interpreted by the listener into some adjectives describing the speech signal, and 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 negative valence and very weak activation, resulting in it being detected as a very Sad emotional state in the categorical approach.

On the other hand, the conventional acoustic features selection method was based on the correlations between acoustic features and emotion dimension as a two-layer model. To investigate the effectiveness of the proposed feature selection method, the results were compared with the conventional method. Table 3 lists the correlations coefficients between acoustic features and emotion dimensions directly.

From this table, evidently only one acoustic feature is highly correlated with the valence dimension ($|correlation(SP\_F1, Valence)| = 0.55 > 0.45$), while eight acoustic features are highly correlated with the activation and dominance dimensions. Therefore, valence shows a smaller number of highly correlated acoustic features than the activation and dominance. These results are similar to those of many previous studies [4]. Due to this drawback, most previous studies achieved a very low performance for valence estimation using the conventional approach [6, 24].

The most important result is that, using the proposed three-layer model for feature selection, the number of relevant acoustic features to emotion dimensions increases. For example, the number of relevant features for the most difficult dimension valence increases from one to nine using the proposed method. Moreover, the number of features increased from eight to nine for ac-

Table 3 Correlation coefficients between acoustic features (AF) and emotion dimensions (ED) (German Database).

AF \ ED	Valence	Activation	Dominance	#
MH_A	-0.33	**-0.82**	**-0.81**	2
MH_E	-0.18	**-0.70**	**-0.71**	2
MH_I	-0.03	-0.19	-0.24	0
MH_O	-0.28	**-0.67**	**-0.68**	2
MH_U	-0.25	**-0.47**	**-0.47**	2
F0_RS	0.21	**0.69**	**0.65**	2
F0_HP	0.19	**0.59**	**0.54**	2
F0_AP	-0.05	-0.14	-0.13	0
F0_RS1	-0.05	-0.10	-0.09	0
PW_R	0.23	**0.75**	**0.74**	2
PW_RHT	-0.25	0.44	**0.49**	1
PW_RS1	0.08	0.14	0.14	0
PW_RAP	0.08	0.36	0.35	0
SP_F1	**-0.55**	**-0.49**	-0.43	2
SP_F2	-0.03	-0.29	-0.29	0
SP_F3	-0.04	-0.04	0.01	0
SP_TL	0.28	0.26	0.26	0
SP_SB	-0.02	-0.05	-0.02	0
DU_TL	-0.28	-0.38	-0.39	0
DU_CL	-0.24	-0.36	-0.36	0
DU_RCV	-0.14	-0.39	-0.37	0
#	1	8	8	17

tivation and from eight to ten for dominance. The selected acoustic features can be used to improve emotion dimensions estimation as described in detail in the next section.

5. Automatic Emotion Recognition System

The aim of speech emotion recognition system based on the dimensional approach can be viewed as using an estimator to map the acoustic features to real-valued emotion dimensions (valence, activation, and dominance). The selected acoustic features from the previous section are used as an input to the proposed system to predict emotion dimensions. Emotion dimension values can be estimated using any estimator such as K-nearest neighborhood (KNN), Support Vector Regression (SVR), or Fuzzy Inference System FIS. 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 an FIS estimator. Therefore, FIS was used to connect the elements of the three-layer model. Most statistical methodology are mainly based on a linear and precise relationship between the input and the output. However, the relationships among acoustic features, semantic primitives, and emotion dimensions are non-linear. Therefore, fuzzy logic is a more appropriate mathematical tool for describing this non-linear relationship [6, 14, 25].

5.1. System Implementation

Adaptive-Network-based Fuzzy Inference System (ANFIS) [25] was used to construct the FIS models that connect the elements of our recognition system. Each FIS has a structure of multiple inputs and 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. As an example, for the German database, to estimate the valence dimension using the perceptual model in Fig. 3(a), a bottom-up method was used to estimate the values (1 to 5 rated by the listening test) of the seven estimated semantic primitives in the middle layer from the nine acoustic features in the bottom layer as shown in Fig. 4. To accomplish this task, seven FISs were required: one to estimate each semantic primitive. In addition, one FIS was required to estimate the value of valence dimension from the seven semantic primitives. Similarly, the activation and dominance can be estimated using FIS for each semantic primitive and one FIS for the activation and dominance, respectively.

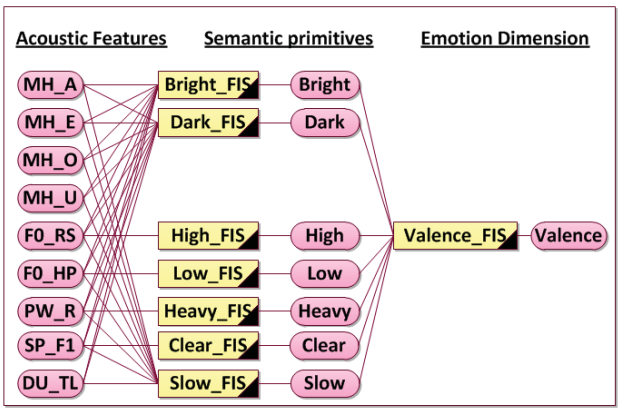
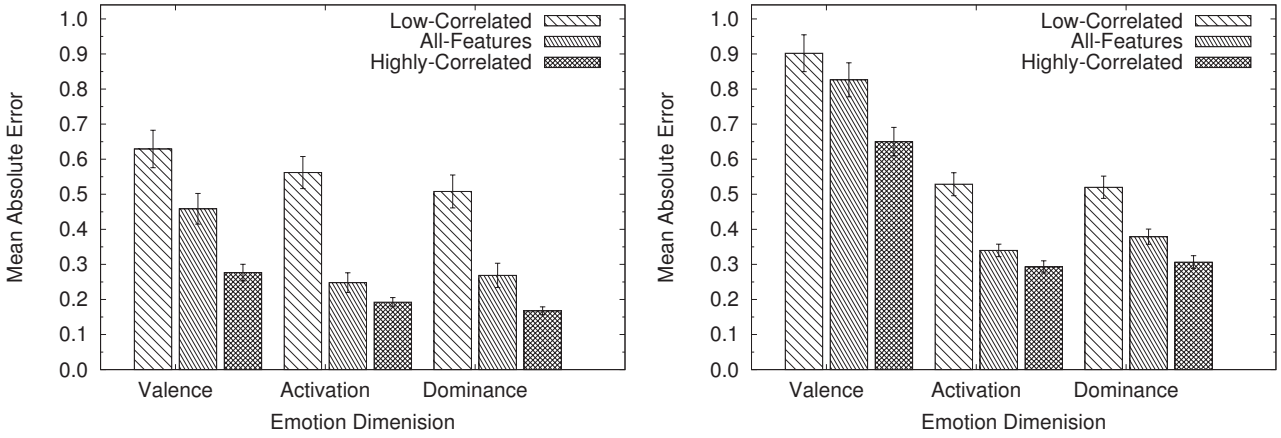


Fig. 4 Block diagram of the proposed approach for estimating valence based on the three-layer model (implementation for German database depicted in Fig. 3(a)).

5.2. Effectiveness of the selected features

This subsection aims to investigate whether the selected acoustic features using the proposed method in Section 4 will improve emotion dimensions estimation. To accomplish this, the proposed automatic emotion recognition system was tested using three different groups of acoustic features, for each emotion dimension: (1) highly correlated acoustic features (absolute values of their correlations with semantic primitives is ≥ 0.45),



(a) Japanese Database (Single Speaker).

(b) German Database (Multi Speaker).

Fig. 5 Mean Absolute Error (MAE) between human evaluation and estimated values of emotion dimensions.

(2) lower correlated acoustic features as shown in Fig. 2, and (3) all the acoustic features.

In order to measure the performance of the proposed system, the mean absolute error (MAE) between the predicted values of emotion dimensions and the corresponding average values given by human subjects is used as a metric of the discrimination associated with each group. The MAE is calculated in accordance with the following equation

$$MAE^{(j)} = \frac{\sum_{i=1}^{N} |\widehat{x}_i^{(j)} - x_i^{(j)}|}{N} \tag{5}$$

where $j \in \{valence, activation, dominance\}$, $\widehat{x}_i^{(j)}$ is the output of the emotion recognition system, and $x_i^{(j)}, -2 \leq x_i^{(j)} \leq 2$ is the evaluated value by human subjects as described in Subsection 3.4.

The accuracy of the classifier in terms of five-fold cross validation was calculated for the two databases. Figures 5(a) and 5(b) show the MAE for estimating (valence, activation, and dominance), for Japanese and German database, respectively, using three groups of acoustic features (highly correlated, lower correlated, all). The error bars in these figures represent the standard errors. Analysis of variance (ANOVA) was conducted to test whether the three groups are statistically different with respect to the use of correlated acoustic features for emotion dimensions estimation. For the Japanese database, at level 0.001, a significant discrimination among the three groups was observed: valence ($F[2, 534] = 29.30$, $p \leq 0.001$), activation ($F[2, 534] = 59.28$, $p \leq 0.001$), and dominance ($F[2, 534] = 51.14$, $p \leq 0.001$). For the German database the results were significant for all emotion dimensions at level 0.001, the information of the F-test were as follows: valence $F[2, 597] = 6.95$, $p \leq 0.001$), activation ($F[2, 597] = 30.54$, $p \leq 0.001$) and dominance ($F[2, 597] = 20.28$, $p \leq 0.001$).

For both databases, the results reveal that by using the three-layer model, the MAEs obtained using the selected acoustical features group (highly correlated acoustic features) are the smallest in comparison with that using all the features. This means that our feature selection method is effective for improving emotion dimensions estimation.

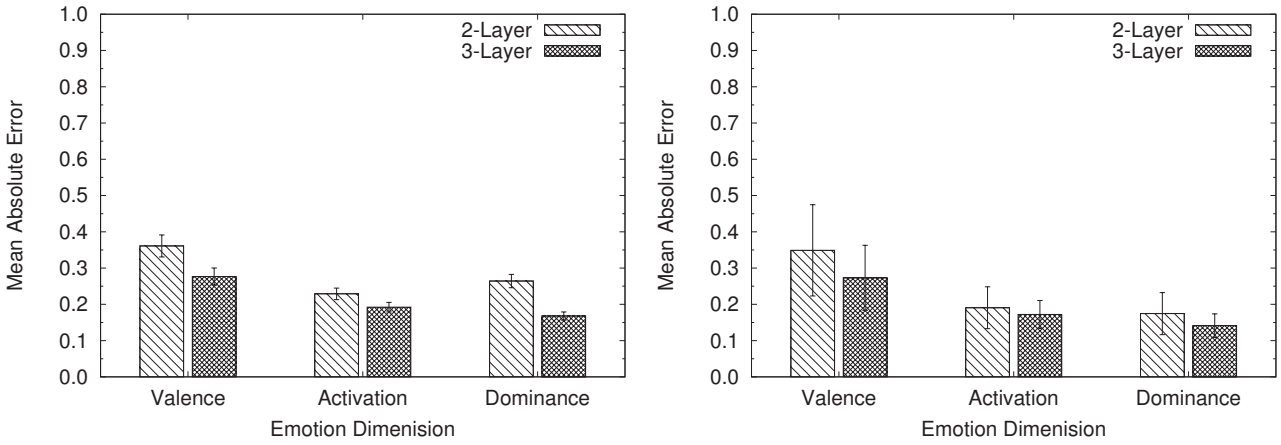
5.3. System Evaluation

In this paper, an automatic speech emotion recognition system based on a three-layer model was implemented. This section presents the evaluation results for the proposed system. To investigate how effectively our system improves emotion dimensions estimation, the performance of the proposed system was compared with that of the conventional two-layer system by using two different languages: Japanese and German, using two different tasks (1) speaker-dependent, and (2) multi-speaker.

The most relevant acoustic features for each emotion dimension were selected using the proposed feature selection method for the two languages as described in Section 4. These selected features were used as the input for the conventional system and the proposed system. The desired output form these systems is the perceived emotion dimensions by listeners, not the emotions intended by speakers.

5.3.1. Evaluation Results for Speaker-dependent Task

In the speaker-dependent task, the automatic emotion recognition system was trained and tested using utterances for one speaker. For a Japanese database, the two automatic systems (the conventional two-layer and proposed three-layer systems) were used to estimate the valence, activation, and dominance from the



(a) Japanese Database (Single speaker).

(b) German Database (Speaker dependent).

Fig. 6 MAE between human evaluation and the two systems output for speaker dependency.

selected acoustic features for 179 utterances included in the Japanese database. The five-fold cross validation was used to evaluate the automatic systems. The MAEs for emotion dimensions (valence, activation, and dominance) between the two systems output and human evaluation are shown in Fig. 6(a). The error bars represent standard errors.

The German database contained ten speakers: five male and five female. Since each speaker made few utterances, the leave-one-out-cross-validation (LOOCV) was used for evaluation. The proposed system and the conventional two-layer system were evaluated using each speaker individually. Finally, the mean value for MAE from all speakers for each emotion dimension was calculated. The results are presented in Fig. 6(b).

Using t-test, at level 0.05, the results for the two databases are as follows: for Japanese database, valence ($t(178)=3.16$, $p\leq 0.05$), activation ($t(178)=2.47$, $p\leq 0.05$), and dominance ($t(178)=4.99$, $p\leq 0.05$). These results are statistically significant for all emotion dimensions. However, for the German database, the results are statistically significant for valence ($t(199)=2.09$, $p\leq 0.05$) and dominance ($t(199)=1.78$, $p\leq 0.05$), but there is no significant differences for activation between the two-layer and the three-layer systems ($t(199)=0.23$, p-value=0.41). As can be seen from Figs. 6(a) and 6(b), the estimation results using the proposed three-layer system outperforms the conventional two-layer system for the two-languages for the speaker-dependent task.

5.3.2. Evaluation Results for Multi-Speaker Task

The German database was used to investigate the effect of multi-speaker on emotion dimension estimation. Thus, the proposed system was validated using the whole database, and all 200 utterances were used to implement this system. Five-fold cross validation was used to evaluate this system. The results for multi-speaker evaluation are shown in Fig. 7. The error bars represent standard errors.

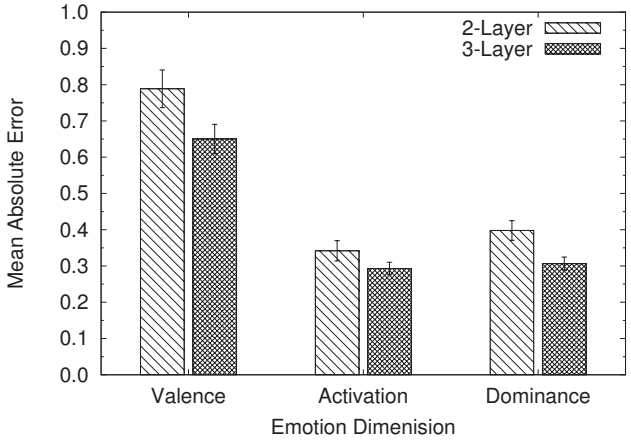


Fig. 7 German Database (multi-speaker): MAE between human evaluation and two systems' output.

The results of the paired t-test at 0.05 significant level were as follows: valence ($t(199)=2.83$, $p\leq 0.05$), activation ($t(199)=1.93$, $p\leq 0.05$), and dominance ($t(199)=3.38$, $p\leq 0.05$). These results are statistically significant for all the emotion dimensions. These results reveal that the proposed system outperforms the conventional one in the multi-speaker task.

5.4. Discussion

Using the acoustic feature selection method described in Section 4, the most relevant acoustic features were selected for each emotion dimensions, for the Japanese and the German databases. To investigate the effectiveness of the selected acoustic features, the proposed system was tested using three different groups of acoustic

features: selected, not selected, and all. The best performance for emotion dimensions estimation were achieved using the selected acoustic features group, for each emotion dimension, as demonstrated by the smallest values of the MAEs, for both German and Japanese databases.

The MAEs for all dimensions, as shown in Figs. 6(a), 6(b), and 7, clearly show that the proposed three-layer system is effective and gives the best results for all emotion dimensions (valence, activation, and dominance) for both speaker-dependent and multi-speaker task. However, the MAEs for the multi-speakers task were higher than those for the speaker-dependent task.

For the German and the Japanese databases, the overall best results were achieved for all emotion dimensions using speaker-dependent task. For both databases, all MAE values were very small; the maximum MAE was 0.28 for valence for the Japanese database as shown in Fig. 6(a). This value indicates that on average the error between human evaluation and system output is 0.28 which means that the output of the proposed system are very close to human evaluation.

From this discussion, it is evident that the valence dimension estimation could be improved by using the proposed model. Therefore, the most important results from this study is that the proposed automatic speech emotion recognition system based on the three-layer model for human perception was superior to the conventional two-layer system.

6. Mapping Values of Emotion Dimensions into Emotion Categories

The categorical and dimensional approaches are closely related, i.e. by detecting the emotional content using one of these two schemes, we can infer its equivalents in the other scheme. For example, if an utterance is estimated with positive valence and high activation we could infer that this is happy, and vice versa. Therefore, any improvement in the dimensional approach will lead to an improvement in the categorical approach and vice versa.

In this section, we want to strengthen our findings in this study by demonstrating that the dimensional approach can actually help us to improve the automatic emotion classification. So, the estimated values of emotion dimensions (valence, activation, and dominance) were used as inputs for Gaussian Mixture Model (GMM) to predict the corresponding emotional category. The classifications results into emotion categories using acoustic features directly is compared with the classification results using the estimated values of emotion dimensions as shown in Tables 4 and 5 for the Japanese and German databases, respectively.

Table 4 Classification results for Japanese database using GMM classifier:

(a) By mapping acoustic features directly to emotion categories (Average recognition rate 53.9%).

Category	Classification rate (%)				
	Neutral	Joy	Cold Anger	Sad	Hot Anger
Neutral	**30.0**	15.0	45.0	5.0	5.0
Joy	2.5	**40.0**	12.5	2.5	42.5
Cold Anger	7.7	12.8	**71.8**	5.1	2.6
Sad	0.0	7.5	12.5	**77.5**	2.5
Hot Anger	2.5	45.0	2.5	0.0	**50.0**

(b) By mapping the estimated emotion dimensions for speaker-dependent task to emotion categories (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
Sad	0.0	0.0	0.0	**100**	0.0
Hot Anger	0.0	2.5	5.0	0.0	**92.5**

6.1. Classification for Japanese Database

For the Japanese database, first, the acoustic features were used as input to train the GMM classifier to classify the Japanese database into five emotion categories: neutral, joy, hot anger, sadness, and cold anger. Moreover, the estimated values of emotion dimensions were used as input to train GMM to classify the values of every point in the space valence-activation-dominance into one emotion category. The confusion matrix of the results is shown in Table 4(a) for mapping acoustic features into categories and in Table 4(b) for mapping values of emotion dimensions into emotion categories. In these tables, the numbers represent the percentages of recognized utterances of the emotion category in the left column versus the number of utterances for emotions in the top line.

6.2. Classification for German Database

The results of classification of the German database into four emotion categories: neutral, happy, angry, and sad are represented by the confusion matrix as follows: Table 5(a) for mapping acoustic feature into categories, Table 5(b) for mapping emotion dimensions into categories for multi-speaker estimation, and Table 5(c) for mapping emotion dimensions into categories for speaker-dependent estimation.

6.3. Discussion

Emotion dimensions values are mapped into the given emotion categories using a GMM classifier. This is a

Table 5 Classification results for German database using GMM classifier:

(a) By mapping acoustic features directly to emotion categories (Average recognition rate 60.0%).

Category	Classification rate (%)			
	Neutral	Happy	Anger	Sad
Neutral	**66.0**	16.0	0.0	18.0
Happy	12.0	**54.0**	32.0	2.0
Anger	2.0	42.0	**54.0**	2.0
Sad	16.0	6.0	12.0	**66.0**

(b) By mapping the estimated emotion dimensions for multi-speaker task to to emotion categories (Average recognition rate 75.0%).

Category	Classification rate (%)			
	Neutral	Happy	Anger	Sad
Neutral	**74.0**	10.0	4.0	12.0
Happy	6.0	**62.0**	32.0	0.0
Anger	2.0	18.0	**80.0**	0.0
Sad	16.0	0.0	0.0	**84.0**

(c) By mapping the estimated emotion dimensions for speaker-dependent task to emotion categories (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**

remarkable improvement on the recognition rate. For the Japanese database, the overall recognition rate was 53.9% for direct classification using acoustic features and 94% using emotion dimensions. For the German database, the rate of direct classification using acoustic features was 60%, which increased to 75% and 95.5% using emotion dimensions for multi-speaker and speaker-dependent tasks, respectively. The result reveals that the recognition rate in speaker-dependent tasks is higher than in multi-speaker tasks. This corresponds with previous studies indicating that speaker-dependent training of the estimator achieves the most accurate emotion classification results [26]. The most important result is that, the classification using emotion dimensions instead of acoustic features improves the recognition rate.

7. Conclusion

The aim of this paper is to improve the conventional dimensional method in order to accurately estimate emotion dimensions, especially the valence dimension. Therefore, we first proposed a novel acoustic features selection method based on a three-layer model of human perception, for selecting the most relevant acoustic features to each emotion dimensions. This method was successfully applied for two different language databases (Japanese and German), many acoustic features were found to be relevant for the valence dimension as well as for the activation, and dominance.

We then proposed a speech emotion recognition system based on the three-layer model to estimate emotion dimensions (valence, activation, and dominance) from most related acoustic features. The proposed system was evaluated using two different languages (Japanese and German) in two different cases (speaker-dependent and multi-speaker). It was found that the proposed system outperforms the conventional two-layer system in both languages, for speaker-dependent, and multi-speaker tasks.

Finally, the estimated values of emotion dimensions were mapped into the given emotion categories using a GMM classifier for the Japanese and German databases. For the Japanese database, an overall recognition rate was 94% using emotion dimensions. For the German database, the recognition rate was 95.5% for speaker-dependent tasks.

In the future, in order to obtain a much more reliable and rich annotation results for emotion dimension and semantic primitives using a listening test, we will study the effect of using a balanced number of subjects in terms of gender and age. Moreover, we will investigate the effectiveness of the three-layer model for constructing a cross-language emotion recognition system which has the ability to detect emotion regardless of the language used for training.

ACKNOWLEDGMENTS

A part of this study was supported by the Grant-in-Aid for Scientific Research (A) (No. 25240026), 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] C. M. Lee, and S. Narayanan, "Toward Detecting Emotions in Spoken Dialogs," *IEEE Transactions on Speech and Audio Processing*, **13(2)**, 293-303 (2005).

[2] 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*, **8(4)**, 201-212 (2005).

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

[4] 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).

[5] H. P. Espinosa, C. A. Reyes-Garca, L. V. Pineda, "Acoustic feature selection and classification of emotions in speech using a 3D continuous emotion model," *Biomedical Signal Processing and Control*, **7(1)**, 79-87 (2012).

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

[7] K. P. Truong, and S. Raaijmakers, "Automatic Recognition of Spontaneous Emotions in Speech Using Acoustic and Lexical Features," *In: Popescu-Belis, A., Stiefelhagen, R. (eds.) MLMI 2008, LNCS 5237*, Springer, Heidelberg, pp. 161-172 (2008).

[8] G. K. Shashidhar and R. Sreenivasa, "Exploring Speech Features for Classifying Emotions along Valence Dimension," *In Proc. of Pattern Recognition and Machine Intelligence*, pp. 537-542 (India, 2009).

[9] C. Busso, T. Rahman "Unveiling the Acoustic Properties that Describe the Valence Dimension," *In Proc. of Interspeech* (2012).

[10] S. G. Karadogan, and J. Larsen, "Combining semantic and acoustic features for valence and arousal recognition in speech," *Proc. of Cognitive Information Processing* (2012).

[11] C. Menezes, K. Maekawa, and H. Kawahara, "Perception of voice quality in paralinguistic information types," *Proc. of the 20th General meeting of the Phonetic Society of Japan*, Tokyo, Japan, pp. 153-158 (2006).

[12] K. R. Scherer, "Personality inference from voice quality: The loud voice of extroversion," *European Journal of Social Psychology*, **8**, 467-487 (1978).

[13] E. Brunswik, "Historical and thematic relations of psychology to other sciences," *Scientific Monthly*, **83**, 151-161 (1956).

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

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

[16] R. Elbarougy and M. Akagi, "Speech Emotion Recognition System Based on a Dimensional Approach Using a Three-Layered Model," *Proc. of APSIPA ASC* (2012).

[17] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," *Proc. of Interspeech*, (2005).

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

[19] K. Maekawa, "Production and perception of paralinguistic information," *In: Proceedings of Speech Prosody, Nara*, pp. 367374 (2004).

[20] R. J. J. H. van Son, and L. C. W. Pols, "An acoustic description of consonant reduction," *Speech communication* **28**, 125-140 (1999).

[21] M. Kienast, and W. F. Sendlmeier, "Acoustical analysis of spectral and temporal changes in expressive speech," *In: ISCA Workshop on Speech and Emotion, Belfast* (2000).

[22] T. Vogt, E. Andre, and J. Wagner, "Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realisation," *Affect and Emotion in HCI, LNCS 4868*, pp. 75-91 (2008).

[23] 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*, **53**, 36-50 (2011).

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

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

[26] T. Iliou, and C. N. Anagnostopoulos, "Classification on Speech Emotion Recognition-A Comparative Study," *International Journal on Advances in Life Sciences*, **2(1 and 2)** 18-28 (2010).

Reda Elbarougy received his B.Sc., and M.Sc., degrees from Damietta University, Egypt, in May 1997, and February 2006, respectively. Both were in computer science. He was with the Faculty of Science, Damietta University from 1999 to 2009. In July 2009, he joined the Japan Advanced Institute of Science and Technology (JAIST), Japan, as a Ph.D. student. His current research interests include speech analysis, speech emotion recognition, and synthesis.

Masato Akagi received his B.E. from Nagoya Institute of Technology in 1979, and his M.E. and Ph.D. Eng. from the Tokyo Institute of Technology in 1981 and 1984. He joined the Electrical Communication Laboratories of Nippon Telegraph and Telephone Corporation (NTT) in 1984. From 1986 to 1990, he worked at the ATR Auditory and Visual Perception Research Laboratories. Since 1992 he has been on the faculty of the School of Information Science of JAIST and is now a full professor. His research interests include speech perception, the modeling of speech perception mechanisms in human beings, and the signal processing of speech. During 1998, he was associated with the Research Laboratories of Electronics at MIT as a visiting researcher, and in 1993 he studied at the Institute of Phonetics Science at the University of Amsterdam. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, the ASJ, the IEEE, the Acoustical Society of America (ASA), and the International Speech Communication Association (ISCA). Dr. Akagi received the IEICE Excellent Paper Award from the IEICE in 1987, the Best Paper Award from the Research Institute of Signal Processing in 2009, and the Sato Prize for Outstanding Papers from the ASJ in 1998, 2005, 2010 and 2011.