Speech Emotion Recognition System Based on a Dimensional Approach Using a Three-Layered Model

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*—This paper proposes a three-layer model for estimating the expressed emotions in a speech signal based on a dimensional approach. Most of the 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 less numerous, less strong, and the valence dimension has being particularly difficult to be predicted. The ultimate goal of this study is to improve the dimensional approach in order to precisely predict the valence dimension. The proposed model consists of three layers: acoustic features, semantic primitives, and emotion dimensions. We aimed to construct a three-layer model in imitation of the process of how human perceive and recognize emotions. In this study, we first investigated the correlations between the elements of the two-layered model and elements of the three-layered model. In addition, we compared the two models by applying a fuzzy inference system (FIS) to estimate emotion dimensions. In our model FIS was used to estimate semantic primitives from acoustic features, then to estimate emotion dimensions from the estimated semantic primitives. The experimental results show that the proposed three-layered model outperforms the traditional two-layered model.**

I. Introduction

Affective computing is a new and exciting topic of research that relates to capturing and recognizing emotions through different modalities. It is feasible to build architectures and techniques to facilitate computers in making affective decisions and expressing certain emotional states through various modes of human-computer interaction [1]. In other words, it is an attempt to make a computer capable of observing, interpreting and generating emotional states [1], [2].

In this context, a large number of studies on emotional speech and its classification have been conducted [3], [4]. As for automatic emotion recognition, there are several approaches that exist such as those based on K-nearest neighborhood (KNN)[1], Gaussian mixture model (GMM) [5], and hidden Markov Model (HMM) [6]. However, most of the techniques focus only on the classification of emotional states as discrete categories such as happy, sad, anger, fear, surprise, and disgust [7], [8].

In contrast, we believe it is also important to detect the variability within a certain emotion (e.g., "a little sad" or "very happy") in addition to the emotion categories. This is evidenced by the fact that we often soften or emphasize such emotional expressions flexibly depending on the situation in actual human speech communication. Therefore, a single label or any small number of discrete categories may not accurately reflect the complexity of the emotional states conveyed in everyday interaction [9]. Hence, a number of researchers advocate the use of 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., [10], [11], [12], [13], [14]).

In the categorical approach, where each affective display is classified into a single category, a complex mental or affective state or blended emotions perhaps is too difficult to handle [15]. Contrarily, in the dimensional approach, emotional transitions can be easily captured, the numerical representations are more appropriate to reflect the gradient nature of emotion expressions, in which observers can indicate their impression of moderate (less intense) and authentic emotional expressions on several continuous scales [16], [17].

In this work, the three-dimensional continuous model is adopted in order to represent the emotional states using emotion dimensions i.e. Valence (V), Activation (A) and Dominance (D). This approach is chosen because it exhibits great potential to model the occurrence of emotions in real world as in a realistic scenario, emotions are not generated in a prototypical or pure modality, but rather than in complex emotional states, which are a mixture of emotions with varying degrees of intensity or expressiveness [18]. Therefore, this approach allows a more flexible interpretation of emotional states [19].

The traditional dimensional model for emotion recognition from speech signal allows the representation of any emotional state or blended emotions. However, this model has the following problems: (i) it is difficult to estimate with a high precision the emotion dimensions based only on acoustic information [20]; (ii) the dimensional approach is mostly based on the statistical relationship between the acoustic features and the emotion dimensions [21]; and (iii) the acoustic features that correlate to the valence dimension are less numerous, less strong and inconsistent [11]. Due to these limitations, the valence dimension has been particularly difficult to predict by using the acoustic features directly. The ultimate goal of our work is to improve the traditional dimensional method in

order to precisely predict the valence dimension as well as improve the activation and dominance.

For simplicity we considered the traditional dimensional model as a two-layered model, because this model was based on the relationship between an acoustic features layer and an emotion dimension layer. There are several studies reported based on the two-layered model such as by Grimm et al. 2007, which attempted to estimate the emotion dimensions valence, activation and dominance from the acoustic features by using a fuzzy inference system [22], [23]. However, they found that the estimation was better for activation and dominance than for valence. Furthermore, many researchers also tried to investigate a new acoustic parameters to improve the valence estimation, as reported by [24] Wu et al. 2011, which attempted to estimate the emotion dimensions by combining spectral and prosodic features. However, they found that valence was still poorly estimated.

From the above mentioned studies, we can conclude that, the two-layered model is insufficient to model the relationship between the acoustic features and emotion dimension because this model does not imitate the human perception. Humans usually describe emotions by using semantic primitives (adjectives) and each semantic primitive is conveyed by certain acoustic features [25]. Thus, semantic primitives act as the bridge between emotion dimensions and acoustic features. Therefore, it is not only necessary to consider how acoustic features of speech affect the judgement of emotion dimensions, but also to understand how the vagueness nature of humans affects this judgement. The acquirement of this understanding is the key point to improve emotion dimensions estimation.

In line with these findings, a three-layered model is proposed in this paper. This model consists of three layers: emotion dimensions valence, activation and dominance constitute the top layer, semantic primitives constitute the middle layer, and acoustic features in the bottom layer. Our model is based on the semantic primitive concept, which plays a large role in the way we perceive emotional speech and measure their similarity. Semantic primitives in our study are adjectives which describe the sounds such as "Dark" or "Slow". To our knowledge, the only previous attempt at using the semantic primitive concept for the purpose of modeling human perception is reported by [25].

In this paper, the feasibility of the three-layered model to improve emotion dimensions estimation for valence, activation, and dominance was investigated. In order to achieve this goal we investigated the correlations between the elements of the three-layered model i.e. (i) the correlations between acoustic features layer and semantic primitive layer, and (ii) the correlations between semantic primitives layer and emotion dimension layer. Moreover, the correlations between the elements of the two-layered model i.e. the correlations between acoustic features layer and emotion dimension layer was also studied. Then, the correlations between the two models can be compared in order to prove that the three-layered model is well suited for estimating the emotion dimensions. Subsequently, an emotion recognition system based on the proposed three-layer model was constructed. In this system, a Fuzzy Inference System (FIS) was applied twice, firstly, to estimate semantic primitives from acoustic features, then, to estimate emotion dimensions values from the estimated semantic primitives in the first step. Finally, the proposed emotion recognition system which is based on the three-layered model was assessed by comparing the results with the conventional system which is based on the two-layered model.

The remainder of this paper is organized as follows. Section II introduces the the used database, acoustic features and experiment setup to evaluate semantic primitives and emotion dimensions. Section III explains a multi-layer emotional speech perception model and compares the relationship between the elements of the two-layered model and the elements of the three-layered model in order to prove that the relation between the elements of the three-layered model are stronger than the relation between the elements of the two-layered model. In Sec. IV we present the details of speech emotion recognition system for estimating the emotion dimensions values from speech using FIS classifier. Finally, Sec. V describes our conclusions.

II. DATA AND EXPERIMENT SETUP

In this section, the database and acoustic features used in this study are introduced. Moreover, the semantic primitive and emotion dimension are evaluated by conducting two listening test using human subjects as described in next subsections.

A. Speech Material and Subjects

For this study, we use 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, 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, for each sentence there are 9 utterances and for all 20 sentences there are 180 utterances. The sampling frequency was 22050 Hz, with 16 bit resolution.

Subjects for listening tests were 11 graduate students, native Japanese speakers (9 men and 2 women) without any hearing troubles.

B. Acoustic Features

In this research, for constructing a speech emotion recognition system, acoustic features are very important factor needed to be investigated. Therefore, the most relevant acoustic features which have been successful in related works and features used for other similar task were selected. Those acoustic cues considered significant for prosody largely are extracted from fundamental frequency, intensity, and duration. In addition, voice quality is another major focus that researchers have paid much attention to. Therefore, acoustic features which originate from F0, power envelope, power spectrum, and voice quality are extracted by the high quality speech analysis-synthesis system STRAIGHT [26]. Moreover, acoustic features which are

related to duration are extracted by segmentation, eventually extracting a set of 21 acoustic features which can be grouped in several subgroups:

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

Power envelope related features: mean value of power range in accentual phrase (PW_RAP), power range (PW_R), rising slope of 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: first formant frequency (SP_F1), second formant frequency (SP_F2), third formant frequency (SP_F3), spectral tilt (SP_TL), 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 features, were selected from the work by Huang and Akagi, where they proved that these acoustic features have a significant correlation with semantic primitives [25]. 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 vowel /a/,/e/,/i/,/o/, and /u/ per utterance MH_A, MH_E, MH_I, MH_O, and MH_U.

C. Semantic Primitives Evaluation

Semantic primitives (adjectives) are required as the bridge between the acoustic features and the emotion dimensions in our emotion recognition system. The used adjectives were (Bright, Dark, High, Low, Strong, Weak, Calm, Unstable, Well-modulated, Monotonous, Heavy, Clear, Noisy, Quiet, Sharp, Fast, and Slow) they are originally from [25]. These adjectives were selected as candidates for semantic primitives because they reflect a balanced selection of widely used adjectives that describe emotional speech. For the evaluation, we used a listening test. In this test, subjects were asked to rate the whole database according to the degree of each 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 individual subject ratings 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

The Fujitsu database was evaluated by 11 subjects along a three dimensions valence, activation, and dominance by using listening test. For each utterance, subjects were asked to choose one out of 5 given degrees depicting the level for each dimension. We used a 5-point scale {-2, -1, 0, 1, 2}: 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 average of the subjects rating for each emotion dimension was calculated per utterance. 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 the correlations were above 0.84, 0.75, and 0.80 respectively.

III. Multi-layer emotional speech perception model

In this section, we investigate the effectiveness of the thee-layered model which imitate the human perception to improve the relationship between acoustic features and emotion dimensions. To accomplish this task, the correlations between elements of the traditional two-layer model were compared with the correlations between elements of the proposed three-layer model. In case of the two-layered model, we investigate the correlations between the acoustic features and emotion dimensions directly as described in Sec III. A. While, in case of the proposed model, we investigate the correlations between the acoustic features and the semantic primitives, moreover, the correlations between the semantic primitives and the emotion dimensions see Sec III. B.

A. Using the Two-Layered Model

In order to investigate the relationship between acoustic features and emotion dimensions by using the traditional two-layered model, the correlations coefficients between extracted parameter values for each acoustic feature and evaluated scores of each dimension are calculated. Let $f_m = \{f_{m,n}\}(n =$

TABLE I
THE CORRELATION COEFFICIENTS BETWEEN THE ACOUSTIC FEATURES AND THE EMOTION DIMENSIONS I.E. VALENCE (V), ACTIVATION (A), AND DOMINANCE (D). (**#:** IS THE NUMBER OF SIGNIFICANT CORRELATIONS)

m	Acoustic Feature	V	A	D	#
1	MH_A	-0.23	**-0.85**	**-0.83**	2
2	MH_E	-0.08	**-0.45**	**-0.45**	2
3	MH_I	0.27	-0.03	-0.17	0
4	MH_O	-0.13	**-0.76**	**-0.75**	2
5	MH_U	0.07	-0.17	-0.23	0
6	F0_RS	0.34	**0.78**	**0.65**	2
7	F0_HP	0.29	**0.77**	**0.64**	2
8	F0_AP	-0.08	-0.12	-0.12	0
9	F0_RS1	-0.09	-0.16	-0.17	0
10	PW_RAP	0.24	**0.53**	**0.50**	2
11	PW_R	**-0.47**	0.33	0.37	1
12	PW_RS1	-0.02	-0.22	-0.23	0
13	PW_RHT	0.17	0.39	0.36	0
14	SP_F1	-0.21	0.14	0.13	0
15	SP_F2	0.04	0.13	0.13	0
16	SP_F3	0.00	0.17	0.21	0
17	SP_TL	0.40	0.28	0.27	0
18	SP_SB	0.05	0.40	0.36	0
19	DU_TL	-0.12	-0.31	-0.32	0
20	DU_CL	-0.27	**-0.61**	**-0.59**	2
21	DU_RCV	-0.29	**-0.60**	**-0.57**	2
	#	1	8	8	

TABLE II
THE CORRELATION COEFFICIENTS BETWEEN THE ACOUSTIC FEATURES AND SEMANTIC PRIMITIVES. (**#:** IS THE NUMBER OF SIGNIFICANT CORRELATIONS)

m		Bright	Dark	High	Low	Strong	Weak	Calm	Unstable	Well-modulated	Monotonous	Heavy	Clear	Noisy	Quiet	Sharp	Fast	Slow	#
1	MH_A	**-0.7**	**0.8**	**-0.7**	**0.7**	**-0.8**	**0.8**	**0.8**	**-0.8**	**-0.7**	**0.6**	**0.6**	-0.4	**-0.8**	**0.8**	**-0.7**	**-0.5**	**0.5**	16
2	MH_E	-0.4	**0.5**	-0.4	0.4	**-0.5**	**0.6**	**0.5**	**-0.5**	-0.4	0.3	0.4	-0.2	**-0.5**	**0.6**	**-0.5**	-0.3	0.3	8
3	MH_O	**-0.6**	**0.7**	**-0.6**	**0.6**	**-0.7**	**0.8**	**0.7**	**-0.7**	**-0.6**	**0.5**	**0.5**	-0.3	**-0.7**	**0.8**	**-0.7**	-0.4	0.4	14
4	F0_RS	**0.8**	**-0.9**	**1.0**	**-1.0**	**0.6**	**-0.7**	**-0.8**	**0.7**	**0.7**	**-0.7**	**-0.9**	**0.5**	**0.8**	**-0.9**	**0.6**	**0.5**	**-0.5**	17
5	F0_HP	**0.8**	**-0.8**	**0.9**	**-0.9**	**0.6**	**-0.7**	**-0.8**	**0.8**	**0.8**	**-0.7**	**-0.8**	**0.5**	**0.8**	**-0.9**	**0.6**	**0.5**	**-0.5**	17
6	PW_RAP	**0.5**	**-0.6**	**0.5**	**-0.5**	0.4	**-0.5**	**-0.5**	0.4	0.4	-0.4	**-0.5**	0.3	**0.5**	**-0.5**	0.4	0.3	-0.3	9
7	PW_R	-0.1	0.0	0.1	-0.1	**0.5**	-0.4	**-0.5**	**0.5**	0.4	**-0.5**	0.1	-0.4	**0.5**	-0.4	**0.6**	0.3	-0.3	6
8	DU_TL	-0.1	0.2	-0.1	0.2	-0.3	0.3	0.2	-0.2	-0.2	0.1	0.1	-0.1	-0.2	0.2	-0.2	**-0.5**	**0.5**	2
9	DU_CL	**-0.5**	**0.6**	**-0.5**	**0.5**	**-0.5**	**0.6**	**0.5**	**-0.5**	-0.4	0.3	**0.5**	-0.3	**-0.5**	**0.5**	**-0.5**	**-0.6**	**0.6**	14
10	DU_RCV	**-0.6**	**0.7**	**-0.6**	**0.6**	**-0.5**	**0.6**	**0.6**	**-0.5**	-0.4	0.3	**0.6**	-0.4	**-0.6**	**0.7**	**-0.5**	-0.3	0.3	12
	#	7	8	7	7	8	8	9	8	4	5	7	2	9	8	8	5	5	

TABLE III
THE CORRELATION COEFFICIENTS BETWEEN THE SEMANTIC PRIMITIVES AND THE EMOTION DIMENSIONS I.E. VALENCE (V), ACTIVATION (A), AND DOMINANCE (D). (**#:** IS THE NUMBER OF SIGNIFICANT CORRELATIONS)

m	Semantic Primitive	V	A	D	#
1	Bright	**0.76**	**0.69**	**0.54**	3
2	Dark	**-0.59**	**-0.86**	**-0.76**	3
3	High	0.43	**0.79**	**0.64**	2
4	Low	**-0.45**	**-0.79**	**-0.65**	3
5	Strong	-0.15	**0.91**	**0.96**	2
6	Weak	-0.04	**-0.95**	**-0.98**	2
7	Calm	0.01	**-0.94**	**-0.91**	2
8	Unstable	-0.11	**0.90**	**0.88**	2
9	Well-modulated	0.01	**0.84**	**0.78**	2
10	Monotonous	0.05	**-0.74**	**-0.67**	2
11	Heavy	**-0.72**	**-0.64**	**-0.48**	3
12	Clear	**0.96**	0.35	0.21	1
13	Noisy	-0.17	**0.89**	**0.88**	2
14	Quiet	-0.11	**-0.93**	**-0.89**	2
15	Sharp	-0.23	**0.89**	**0.92**	2
16	Fast	0.04	**0.76**	**0.75**	2
17	Slow	-0.11	**-0.73**	**-0.71**	2
	#	5	16	16	

$1, 2, \ldots, N$) be the sequence of values of the m^{th} acoustic feature, $m = 1, 2, \ldots, M$, M is the number of extracted acoustic features in this study. Moreover, let $x^{(i)} = \{x_n^{(i)}\} (n = 1, 2, \ldots, N)$ be the sequence of values of the i^{th} emotion dimension, $i \in \{V, A, D\}$, where N is the number of utterances in our database. Then the correlation coefficient $R_m^{(i)}$ between the acoustic parameter f_m and the emotion dimension $x^{(i)}$ can be determined by the following equation:

$$R_m^{(i)} = \frac{\sum_{n=1}^{N}(f_{m,n} - \overline{f_m})(x_n^{(i)} - \overline{x}^{(i)})}{\sqrt{\sum_{n=1}^{N}(f_{m,n} - \overline{f_m})^2}\sqrt{\sum_{n=1}^{N}(x_n^{(i)} - \overline{x}^{(i)})^2}} \quad (1)$$

where $\overline{f_m}$, and $\overline{x}^{(i)}$ are the arithmetic mean for the acoustic feature and emotion dimension respectively. Table I shows the correlation coefficients for all acoustic features and all emotion dimensions. From this table, it is evident that eight acoustic features have high correlation with the activation and dominance dimensions as demonstrated by the absolute value of the correlation, which was greater than 0.45 as shown in bold in the table. Furthermore the emotion dimension valence shows a smaller absolute values of correlations than the activation and dominance. This result is consistent with many previous studies [27], [11]. The poor correlation between the acoustic features and valence is the reason behind the very low performance for valence estimation using the traditional approach.

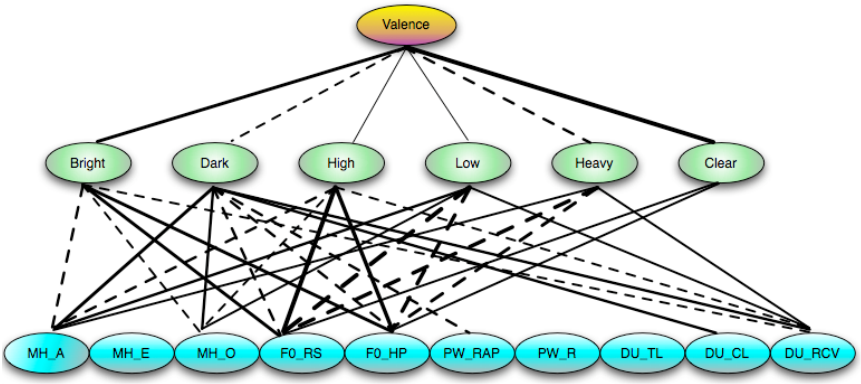
B. Using the Three-Layered Model

1) The relation between acoustic features and semantic primitives: The correlation coefficients were calculated between extracted parameter values for each acoustic feature and evaluated scores of each semantic primitives by using equation similar to Eq. 1. The correlation coefficients for 10 acoustic features which have a significant correlation with semantic primitives are presented in Table II. Most of the semantic primitives gave a high correlation with at least 5 acoustic features except for "Clear" and "Well-modulated" which gave only 2 and 4 significant correlations respectively.

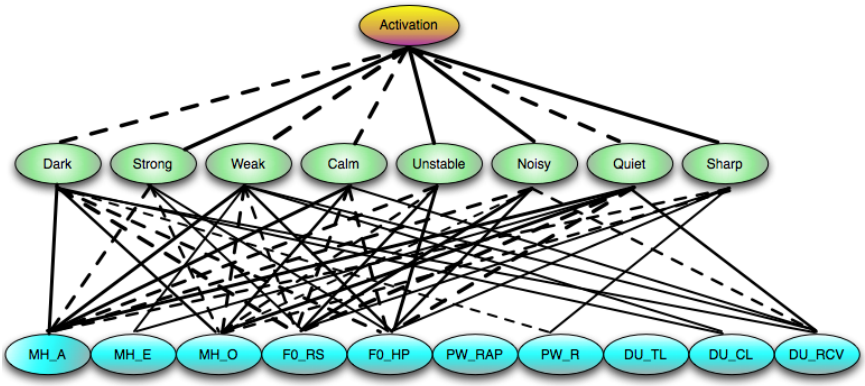
2) The relation between semantic primitives and emotion dimensions: In order to study the relationship between semantic primitives layer and emotion dimensions layer, we calculated the correlation coefficients between the semantic primitives' ratings and the emotion dimensions ratings as shown in Table III by using equation similar to Eq. 1. The most numerous and strongest correlations were found for the activation and dominance. The most important result is that, we found 5 semantic primitives have a significant correlation with valence. This result indicate that the correlations between valence and semantic primitives were better than the correlations between acoustic features and valence.

C. Results and Discussion for the Statistical Analysis

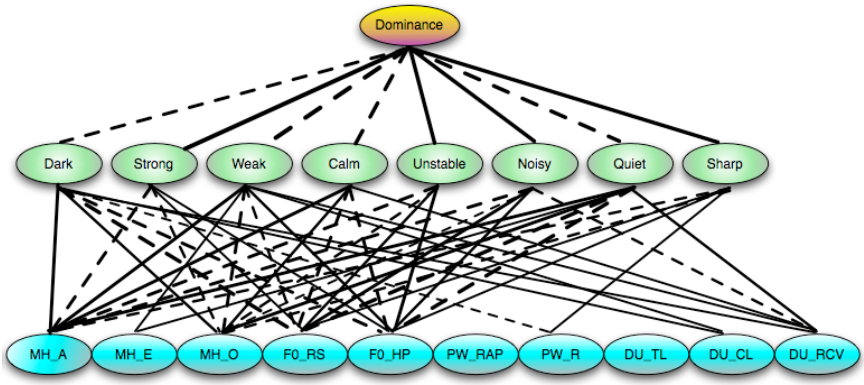
In Subsection IV.A and IV.B we investigated the correlations between the elements of the two-layered model and also between the elements of the three-layered model. The most



(a) The perceptual model of Valence.



(b) The perceptual model of Activation.



(c) The perceptual model of Dominance.

Fig. 1. The perceptual model of Emotion Dimensions.

fundamental result is that, by using the two-layered model, it was observed that the most numerous and strongest correlations were found for the activation and dominance, while the correlations between the acoustic features and the valence were very weak. Although we used a new acoustic features such as voice quality, the correlation with valence is still very weak using the traditional two-layered model, a trend which is also reported in other dimensional emotion recognition studies [11]. Due to this drawback, most of the previous studies achieved a very good performance for the activation and dominance estimation, while a lower performance was obtained for the valence [23], [24].

On the other hand, it was found that 10 acoustic features shows a very good correlation with the semantic primitive as shown in Table II. The semantic primitives in general show higher absolute values of correlations with at least 5 acoustic features. In addition, the correlations between emotion dimension and semantic primitives are stronger than the correlation between emotion dimension and acoustic features. The most important result is that, it was observed that the semantic

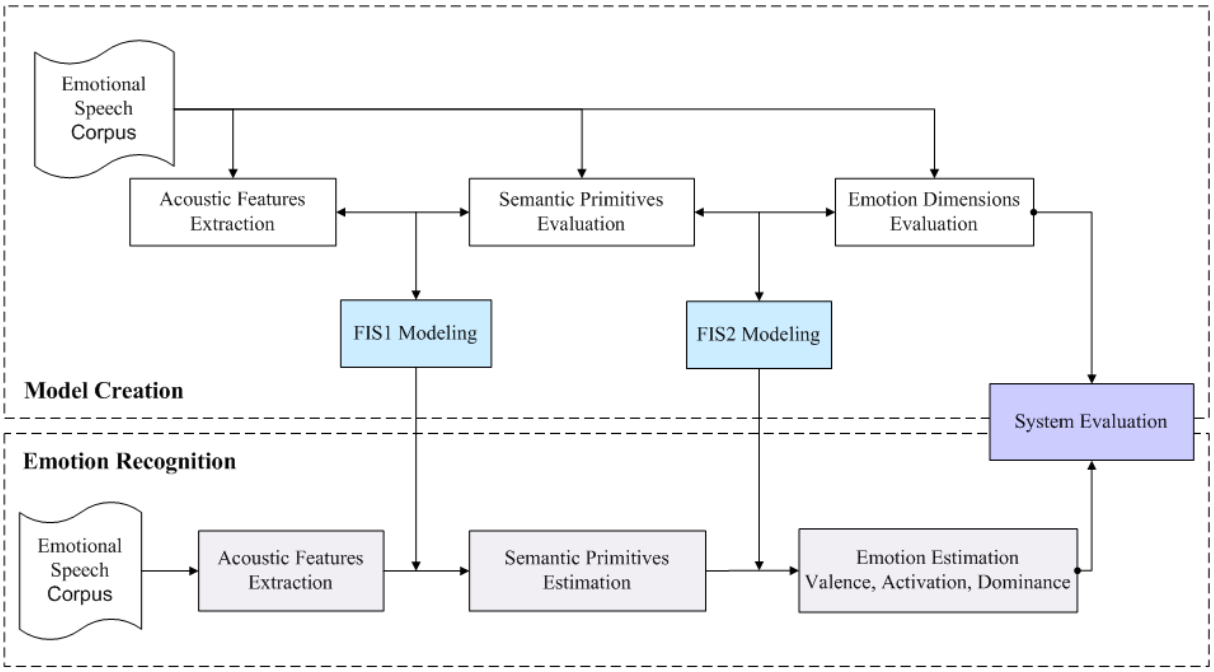


Fig. 2. Block diagram of the proposed emotion recognition system based on the three-layered model.

primitives gave higher correlations with valence dimension than the acoustic features. Therefore, the strong correlations between the semantic primitives and the valence will improve the prediction of this dimension.

Based on the results described in this section, a perceptual model for each emotion dimension was built. Fig. 1(a),(b), and (c) illustrate the perceptual model for valence, activation, and dominance, respectively. In this figure, the solid lines indicate a positive correlation, and the dashed ones, a negative correlation. For the relationship between emotion dimensions and semantic primitives, the highest values are shown in the thick lines, others are shown in thin lines. For the relationship between semantic primitives and acoustic features, the thicker the line is, the higher the correlation. For example, the model in Fig. 1(b) describes that an active speech utterance will sound strong, unstable, noisy and sharp but not dark, weak, calm or quiet. Moreover this figure shows which acoustic features are most related to which semantic primitives of each emotion dimension.

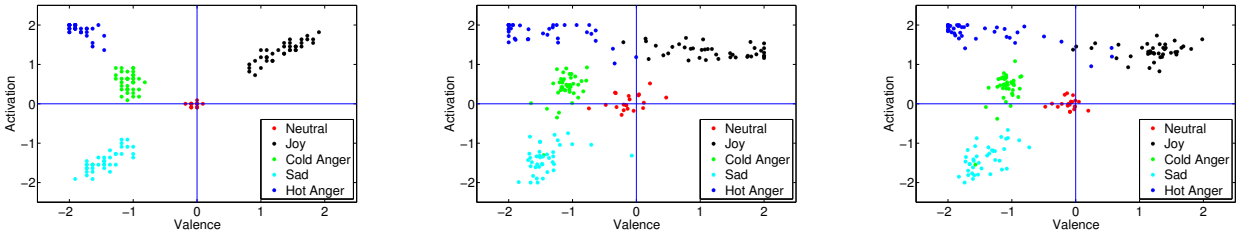
IV. Speech Emotion Recognition System

In the previous section we proved that the correlations between elements of the three-layered model were stronger than the correlations between the elements of the two-layered model. The results from the previous section reveals that the correlation between acoustic features and emotion dimensions become stronger by adding the semantic primitives between them, which mean that we can improve the emotion dimension estimation by building an emotion recognition system based on the three-layered model. The task of the emotion estimator is to map the acoustic features to a real-valued emotion dimensions. The novelty of this study is the use of the three-layered model in order to estimate the emotion dimension simply by mapping the acoustic features into a real-valued semantic primitives followed by mapping the semantic primitives to a real-valued emotion dimensions. We analyzed using Fuzzy Inference System estimator as briefly described in the following subsection.

Figure 2. shows a block diagram of the proposed recognition system based on the three-layered model, this system consists of two main stages. The first stage is model creation which is employed for training the model, and the second stage is applying emotion recognition to test the model. Our system was constructed by using FIS to build the mathematical relationship between the elements of the three-layered model as follows: (1) FIS1 is used to map the acoustic features onto semantic primitives, (2) and also FIS2 is used to map the semantic primitives onto emotion dimensions. The desired output is not a classification into one of a finite set of categories but an estimation of a continuous-valued for emotion dimensions: Valence, Activation, and Dominance.

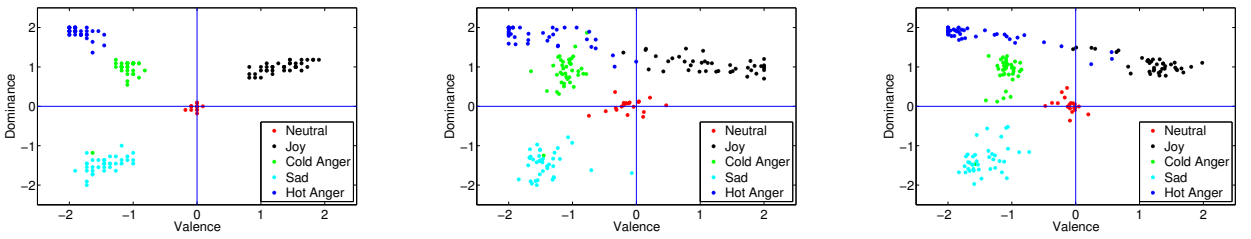
A. *Fuzzy Inference System*

A FIS implements a nonlinear mapping from an input space to an output space by a number of fuzzy if-then rules constructed from human knowledge. The success of a FIS depends on the identification of the fuzzy rules and membership functions tuned to a particular application. It is usually difficult in terms of time and cost, and sometimes impossible, however, to transform human knowledge into a rule base [28]. Even if a rule base is provided, there remains a need to tune the membership functions to enhance the performance of the



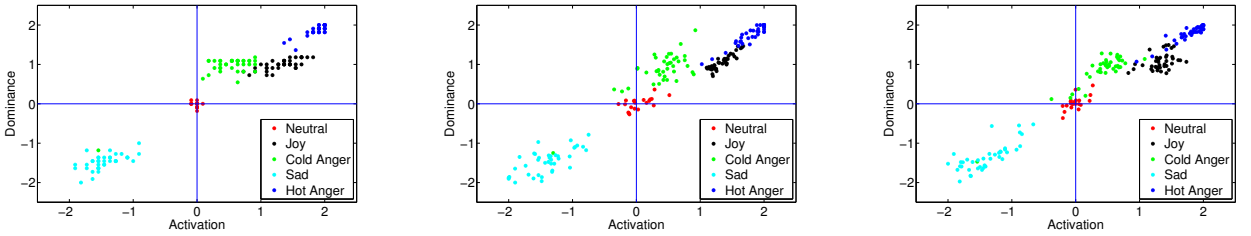
(a) Manually labeled by Human. (b) Estimated by the Two-Layer system. (c) Estimated by the Three-Layer system.

Fig. 3. The distribution of the speech utterances in the Valence-Activation space.



(a) Manually labeled by Human. (b) Estimated by the Two-Layer system. (c) Estimated by the Three-Layer system.

Fig. 4. The distribution of the speech utterances in the Valence-Dominance space.



(a) Manually labeled by Human. (b) Estimated by the Two-Layer system. (c) Estimated by the Three-Layer system.

Fig. 5. The distribution of the speech utterances in the Activation-Dominance space.

mapping. Neuro-fuzzy systems overcome these limitations by using artificial neural networks to identify fuzzy rules and tune the parameters of membership functions in FIS automatically. In this way, the need for the expert knowledge usually required to design a standard FIS is eliminated. A specific approach in neuro-fuzzy systems is ANFIS, which is a Sugeno type FIS implemented in the framework of adaptive neural networks [29].

We use ANFIS to construct FIS models which connect the elements of our recognition system. Each FIS has the structure of multiple inputs and one output. Therefore, in order to construct our recognition system 20 FIS are needed. For semantic primitives estimation, 17 FIS are needed to map the acoustic features to semantic primitives, one for each semantic primitives. Moreover, 3 FIS also needed to map the semantic primitives to emotion dimensions, one for each emotion dimension.

B. Results and Discussion

In order to explicitly demonstrate the emotion recognition improvement, we firstly, constructed two different automatic emotion recognition systems:the first system was based on the traditional two-layer model, in this system FIS was used to map the acoustic features to emotion dimensions directly. While the second system was based on the proposed three-layer model as described in Sec. IV. A. Then, we compared the rated emotions by human subjects and the estimated emotions using the two-layer system and the three-layer system respectively. Fig. 3 shows emotion category distribution in the Valence-Activation space. Where Fig. 3(a) shows the results of experimental evaluation by human subjects for all utterances in our database. Fig. 3(b) and (c) show that by imitating the human perception using the three-layer model as shown in Fig. 3 (c), the system develops its capability to understand human emotion more accurately. As we can see, the the two-layer system in Fig 3. (b) is not able to accurately estimate the valence dimension, while using the three-layer system the valence estimation is improved and become very close to human ratings. Similar results were obtained for the Valence-Dominance space as shown in Fig. 4(a)-(c). However, the estimation results in the Activation-Dominance space are slightly improved as seen from Fig.

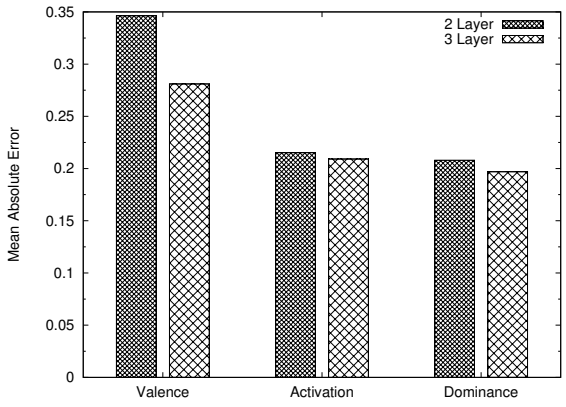


Fig. 6. Comparison between the two-layered model and the three-layered model using FIS.

5 which can be confirmed from Fig. 6. This result showed that the activation and dominance dimensions can be easily estimated with acoustic features, whereas the estimation of the valence dimension yields the best results when the semantic primitives are used.

In order to assess the performance of our system, the output of the proposed system is compared with the output of the traditional one. For each emotion dimension, and for each system, we calculated the mean absolute error between the estimated output using the system, and the evaluated scores by listeners. The mean absolute error is calculated according to the following equation

$$E^{(i)} = \frac{\sum_{i=1}^{n} |\widehat{x}_n^{(i)} - x_n^{(i)}|}{N} \quad (2)$$

where $i \in \{V, A, D\}$, $\widehat{x}_n^{(i)}$ is output of the emotion recognition system, and $x_n^{(i)}$ is the evaluated scores by using human subjects as described in section II.

All results were achieved using 5-fold cross-validation. These results are presented in Fig. 6, we compared between the two-layer model and the three-layered model using FIS estimator. Form this figure, it is clear that the valence prediction is improved by using the proposed model. It is clear to see that the dimensional approach for emotion dimension estimation based on three-layered model was superior to the traditional dimensional approach.

V. CONCLUSION

In this paper, we proposed a three-layered model for emotion speech recognition based on the dimensional approach, in order to precisely predict the valence dimension from the acoustic features. To imitate human perception, a semantic primitive layer was added between the acoustic features and the emotion dimension. For the estimation of the emotion dimensions (valence, activation, and dominance) fuzzy inference system was used. We compared the estimation based on the three-layered model and the two-layered model. Our results demonstrate that the proposed method can be used to accurately predict the valence dimension from the acoustic features through semantic primitives.

ACKNOWLEDGMENTS

This study was supported by a Grant-in-Aid for Scientific Research (22650032) from the Japan Society for the Promotion of Science (JSPS).

REFERENCES

[1] T.L. Pao, Y.T. Chen, and J.H. Yeh "Comparison of Classification Methods for Detecting Emotion from Mandarin Speech ," *IEICE Transactions on Information and Systems*, vol. E91-D(4), pp. 1074-1081, 2008.

[2] T. Nose, and T. Kobayashi, "A Technique for Estimating Intensity of Emotional Expressions and Speaking Styles in Speech Based on Multiple-Regression HSMM ," *IEICE Transactions on Information and Systems*, vol. E93-D(1), pp. 116-124, 2010.

[3] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, "Emotion recognition in human- computer interaction," *IEEE Signal Process.*, vol. 18(1), pp. 3280, January 2001.

[4] D. Erickson, "Expressive speech: Production, perception and application to speech synthesis," *Acoustical Science and Technology*, vol. 26(4), pp. 317325, July 2005.

[5] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using GMMs," *Proc. INTERSPEECH 2006*, pp. 809812, September 2006.

[6] T.L. Nwe, S.W. Foo, and L.C.D. Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41(4), pp. 603623, November 2003.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

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

[14] R. Cowie, "Describing the emotional states that are expressed in speech," *Proc. ISCA Workshop on Speech and Emotion*, pp. 11-18, 2000.

[15] C. Yu, P.M. Aoki, and A. Woodruff, "Detecting User Engagement in Everyday Conversations," *Proc. Eighth Intl Conf. Spoken Language Processing*, 2004.

[16] M. Nicolaou, and H. Gunes, "Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space," *IEEE Transactions on Affective Computing*, vol. 2(2), pp. 92-105, 2011.

[17] D. Wu, T. Parsons, E. Mower and S. Narayanan, "Speech Emotion Estimation in 3D Space," *Proc. ICME 2010*, 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] Q. Zhang, S. Jeong, M. Lee, "Autonomous emotion development using incremental modified adaptive neuro-fuzzy inference system," *Neurocomputing*, vol. 86, pp. 33-44, 2012.

[20] 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*, vol. 7(1), pp. 79-87, January 2012.

[21] S. Patel, and R. Shrivastav, "A Preliminary Model of Emotional Prosody using Multidimensional Scaling," *Proc. InterSpeech 2011*, pp. 2957–2960, 2011.

[22] M. Grimm, E. Mower , K. Kroschel, and S. Narayanan, "Combining categorical and primitives-based emotion recognition," *Proc. EUSIPCO 2006*, 2006.
Grimm, Mower, Kroschel, and Narayanan.
[23] 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.
[24] 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.
[25] C. Huang, and M. Akagi, "A three-layered model for expressive speech perception," *Speech Communication*, vol. 50(10), pp. 810–828, October 2008.
[26] H. Kawahara, and I.M.-katsuse, and A.D. Cheveign, "Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction : Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.
[27] M. Grimm, and K. Kroschel, "Rule-Based Emotion Classification Using Acoustic Features," *Proc. Int. Conf. on Telemedicine and Multimedia Communication*, 2005.
[28] D. Nauck, F. Klawonn, and R. Kruse, "Foundations of neuro fuzzy systems, " *New York:Wiley*, 1997.
[29] 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.