Emotional speech synthesis system based on a three-layered model using a dimensional approach

Yawen Xue* Yasuhiro Hamada† and Masato Akagi*
* Japan Advanced Institute of Science and Technology, Ishikawa, Japan
E-mail: xue_yawen@jaist.ac.jp
† Meiji University, Tokyo, Japan

Abstract—**This paper proposes an emotional speech synthesis system based on a three-layered model using a dimensional approach. Most previous studies related to emotional speech synthesis using the dimensional approach focused on the relationship between acoustic features and emotion dimensions (valence and activation) only. However, people do not perceive emotion directly from acoustic features. Hence, the acoustic features have being particularly difficult to predict, and the affectiveness of the synthesized sound is far from that intended. The ultimate goal of this research is to improve the accuracy of acoustic feature estimation and modification rules in order to synthesize affective speech more similar to that intended in the dimensional emotion space. The proposed system is composed by three layers: acoustic features, semantic primitives, and emotion dimensions. Fuzzy Inference System (FIS) is used to connect the three layers. The related acoustic features of each semantic primitive are selected for synthesizing the emotional speech. On the basis of morphing rules, the estimated acoustic features can be applied to synthesize emotional speech. Listening tests were carried out to verify whether the synthesized speech can give the intended impression in the dimensional emotion space. Results show that not only is the accuracy of estimated acoustic features raised but also the modification rules work well for the synthesized speech, resulting in the proposed method improving the quality of synthesized speech.**

I. Introduction

In the field of human-computer-interface (HCI), one of the goals is to improve user experiences by providing genuine human communication. Thus, a speech-to-speech translation (S2ST) system plays a consequential role for converting a spoken utterance from one language into another to enable people who speak different languages to communicate [1]. Conventional S2STs focus on processing linguistic information only, which is deficient in synthesizing affective speech, such as emotional rather than neutral speech. Therefore, a system that can recognize and synthesize emotional speech would be momentous.

To construct a system for synthesizing emotional speech, several studies have already obtained some achievements. Most methods are based on a concatenative approach, like unit selection, or a statistical parametric approach, like the Hidden Markov Model (HMM) with the Gaussian Mixture Model (GMM) [2] [3]. Both methods can synthesize emotional speech with good quality when the emotion is present in a category such as happy, sad, or angry. However, they can only synthesize the emotional speech with the average emotion (not strong or weak emotions) in the emotion category, and both need a huge database for training, although it is difficult to collect many human responses when listening to emotional speech. However, in human speech communication, people sometimes strengthen or weaken emotional expressions depending on the situation [4]. Thus, a small number of discrete categories is not sufficient to mimic the emotional speech in daily life. Therefore, some researchers proposed a multi-dimensional approach to express emotion on a continuous-valued scale instead of categorical methods [5] [6]. By using the rule-based synthesis method, tendencies of the variations can be acquired using a small database. With the tendencies of variation, the synthesized speech can convey all degrees of an emotion.

An emotional synthesis system based on dimensional space previously proposed by the authors, named the two-layered model [7], has already worked in a rule-based emotional speech synthesis scenario. However, two main problems remain. First, the estimated acoustic features are not accurate enough for this kind of method to extract rules just between the acoustic features and emotional space. Second, some synthesized speech cannot give the impression intended, which means that the method for modifying acoustic features still has a problem. The ultimate goal of our work is to improve the conventional dimensional method in order to precisely predict the acoustic features as well as to synthesize affective speech much similar to that intended in dimensional space.

To improve the accuracy of estimated acoustic features, a three-layered model is adopted. According to the Branswikian lens model [8] shown in Figure 1, people's emotion perception is multi-layered. Human beings do not perceive emotion directly from the acoustic features, so semantic primitives such as bright, high, strong, and so on are also of great importance. In these circumstances, this paper basically utilized the three-layered model proposed by Huang and Akagi [9] (acoustic features layer, semantic primitives layer, and emotion layer). For the emotion layer, in this study, a dimensional emotion space is used to model the human emotions. The Valence-Activation (V-A) axes in the two-dimensional emotion space can describe the strength, such as very or slightly happy, which gives a more flexible interpretation of emotional states. An Adaptive-Network-based Fuzzy Inference System (ANFIS) [10] [11] connects the three layers and estimates corresponding values of dimensional axes.

The modification of estimated acoustic features for emo-

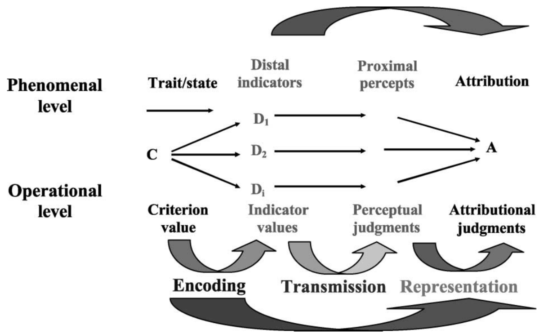


Fig. 1. A Brunswikian lens (1956) model of the vocal communication of emotion. (Scherer, 2003) [8]

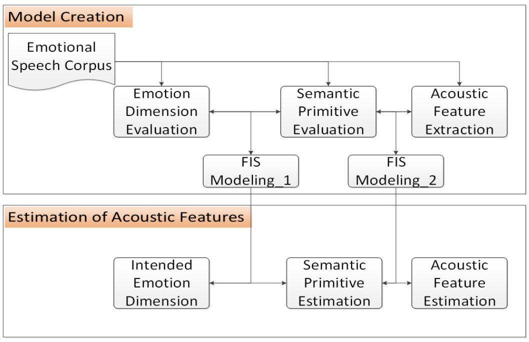

Model Creation
Emotional Speech Corpus
Emotion Dimension Evaluation
Semantic Primitive Evaluation
Acoustic Feature Extraction
FIS Modeling_1
FIS Modeling_2
Estimation of Acoustic Features
Intended Emotion Dimension
Semantic Primitive Estimation
Acoustic Feature Estimation

Fig. 2. Flow chart for estimating acoustic features

tional speech synthesis is revised in this paper, especially for the fundamental frequency (F0) related acoustic features. Hamada et al. [7] separately extracted and modified the F0 related acoustic features such as average F0 pitch, highest F0 pitch, F0 mean value of rising slope, and rising slope of the first accentual phrase in accordance with the extracted rules in the model. However, as there are strong connections among them, these F0 related acoustic features are unsuitable to handle independently. In this paper, the Fujisaki model [14] is adopted to extract the trajectory of the F0 contour from which we can obtain the F0 related acoustic features all at once.

The difficulty in this research is to find out whether the extracted rules are suitable for the intended emotion as the relationship between the three layers is nonlinear. The input and output of our system are the dimensional parameter values in V-A space and the corresponding acoustic feature displacements, respectively. ANFIS is used to connect the three layers from the emotion space layer to the semantic primitives layer and from semantic primitives layer to the acoustic features layer. The related acoustic features of every semantic primitive are selected when synthesizing the emotional speech. Listening tests were carried out to verify whether the synthesis speech can give a position similar to that anticipated. On the basis of the listening test, effectiveness is discussed.

II. Outline of the Emotional Speech Synthesis System

This section outlines the emotional speech synthesis system. The system can be divided into two parts. The first is applied to estimate the acoustic features, and the second is used to modify them. The ultimate goal of this paper is to improve the estimation accuracy and the modification of acoustic features. Figure 2 shows the flow chart for estimating the acoustic features. In the model creation part, the evaluated emotion dimensions, evaluated semantic primitives, and the extracted acoustic features first need to be acquired by listening tests and some tools, which are detailed in Chapter IV. To connect the three layers, we use the fuzzy inference system (FIS). FIS Modeling_1 is applied for connecting the emotion dimensions layer and the semantic primitives layer whose input is the value of Valence and Activation and output is the estimated values of the according semantic primitives. The input of FIS Modeling_2 is the values of the semantic primitives, and output is the estimated acoustic features. This is fully explained in Chapter V. After the system has been built, the parameter values can be estimated when given the intended position in the emotion dimensions. To obtain the emotional synthesized speech, the estimated parameter values need to be modified using some tools and models. This will be discussed in Chapter VI.

III. Three-Layered Model

The concept of the three-layered model is explained here. In 2008, Huang and Akagi proposed a three-layered model for expressive speech perception [9], which states that humans perceive the emotion of expressive speech not directly from the academic terms such as F0 contour, power envelop or power spectrum but from a series adjectives such as fast, bright, or strong. On the basis of the three-layered model, the recognition system proposed by Elbarougy and Akagi [12] is adopted with opposite input and output to synthesize emotional speech. However, the recognition system is irreversible as the relationship between the three layers is nonlinear. The difficulty in this research is to find out whether the extracted rules are suitable for the intended emotion as the relationship between the three layers. Opposite to the work of the speech emotion recognition system, the input of the emotional speech synthesis system is the position in valence-activation space, and the output is the according acoustic features as shown in Figure 3. In our system, the emotion dimension is at the bottom, the semantic primitives layer is in the middle, and the acoustic features layer is at the top.

IV. Speech Materials and Experiment

In this section, the database used in this study is explained first. Then, in accordance with our method, the acoustic features extracted from the database, semantic primitives, and emotion dimension acquired by listening tests are presented in the next subsections.

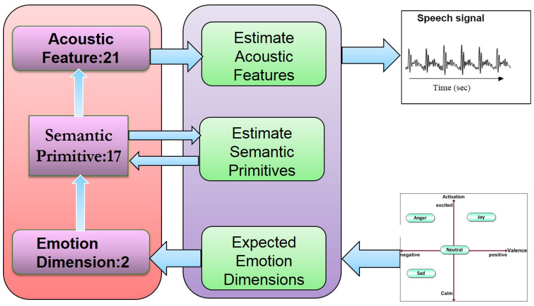


Fig. 3. Structure of three-layered model

A. Speech Materials and Subjects

A Fujitsu database containing 179 utterances spoken by a professional female voice actress is used in this study. The 179 utterances consist of five different emotion states: neutral, happy, sad, cold anger, and hot anger. In addition, there are 20 different Japanese sentences in this database. Each sentence has one for neutral and two for other emotion states. The total number of utterances is 179 because one cold anger utterance is missing from the database.

In the listening test, 11 graduate students, all native Japanese speakers without any hearing impairment, were asked to evaluate the utterances as subjects.

B. Acoustic Feature Extraction

For synthesizing emotional speech, the acoustic features are vital because they can hugely affect the effectiveness of the synthesized speech. In the field of F0, power envelope, power spectrum, and duration, 16 acoustic features are put to use in accordance with the work of Huang and Akagi [9].

Except for the acoustic features related to duration that is extracted by segmentation manually, the rest are obtained by the high quality speech analysis-synthesis system STRAIGHT [13]. Also, the same as in the work of Hamada et al. [7], five acoustic features related to the voice quality were focused on as they are important for perceiving the expressive voice. All together, 21 acoustic features are classified into the following subgroups:

F0 related features: The acoustic features related to F0 were extracted: F0 mean value of a rising slope of the F0 contour (RS), highest F0 (HP), average F0 (AP), and rising slope of the F0 contour for the first accentual phrase (RS1).

Power envelope related features: Mean value of power range in accentual phrase (PRAP), power range (PWR), rising slope of the power for the first accentual phrase (PRS1), the ratio between the average power in high frequency portion (over 3 kHz), and the average power (RHT) were measured.

Power spectrum related features: First formant frequency (F1), second formant frequency (F2), and third formant frequency (F3) were taken approximately at the midpoint of the vowels /a/, /e/, /i/, /o/, and /u/. The formant frequencies were calculated at an LPC-order of 12. Spectral tilt (SPTL) is used to measure voice quality and was calculated using the following equation:

$$SP\_TL = A_1 - A_3 \tag{1}$$

where A_1 is the level in dB of the first formant, and A_3 is the level of the harmonic whose frequency is closest to the third formant [15]. To describe acoustic consonant reduction [16], spectral balance (SB) is adopted. It was calculated in accordance with the following equation:

$$SP\_SB = \frac{\sum f_i \cdot E_i}{\sum E_i} \tag{2}$$

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

Duration related features: Total length (TL), consonant length (CL), and ratio between consonant length and vowel length (RCV).

Voice quality: According to Menezes et al. [17], H1-H2 is concerns glottal opening, which means the mean value of difference between the first and second harmonics for vowels /a/, /e/, /i/, /o/, and /u/ per utterance. MH_A, MH_E, MH_I, MH_O, and MH_U were used as indexes of voice quality.

C. Semantic Primitives Evaluation

As mentioned above, in the three-layered model, the bottom layer is the emotion dimension layer, the middle layer is the semantic primitives layer, and the top layer is the acoustic features layer. Therefore, the value of semantic primitives is essential for building this model. 11 subjects were asked to give subjective values for 17 adjectives: Bright, Dark, High, Low, Strong, Weak, Calm, Unstable, Well-modulated, Monotonous, Heavy, Clear, Noisy, Quiet, Sharp, Fast, and Slow. These words were selected by Huang and Akagi [9] as they can describe emotional speech in a balanced way. The 17 semantic primitives were evaluated on a five-point scale ("1-Does not feel so at all", "2-Seldom feels so", "3-Feels slightly so ", "4-Feels so", "5-Feels very much so"). Separately for each semantic primitives, the inter-rater agreement was measured by pairwise Pearosn's correlation between two subjects' ratings which shows that all subjects agreed from a moderate to a high level.

D. Emotion Dimensions Evaluation

The evaluation of emotion dimension is divided into two parts: valence and activation. The 11 subjects were required to rate the 179 utterances on a five-point scale {-2, -1, 0, 1, 2 }. Valence was from -2 (very negative) to +2 (very positive), and activation was from -2 (very calm) to +2 (very excited). The value of the evaluation has a high inter-rater agreement, which shows that all subjects had similar impressions of the emotional speech. This work expands on the work of Elbarougy and Akagi [12].

V. ESTIMATION OF ACOUSTIC FEATURES

A. Fuzzy Inference System

To obtain the estimated acoustic features, two kinds of fuzzy inference system were used for training. One, called FIS

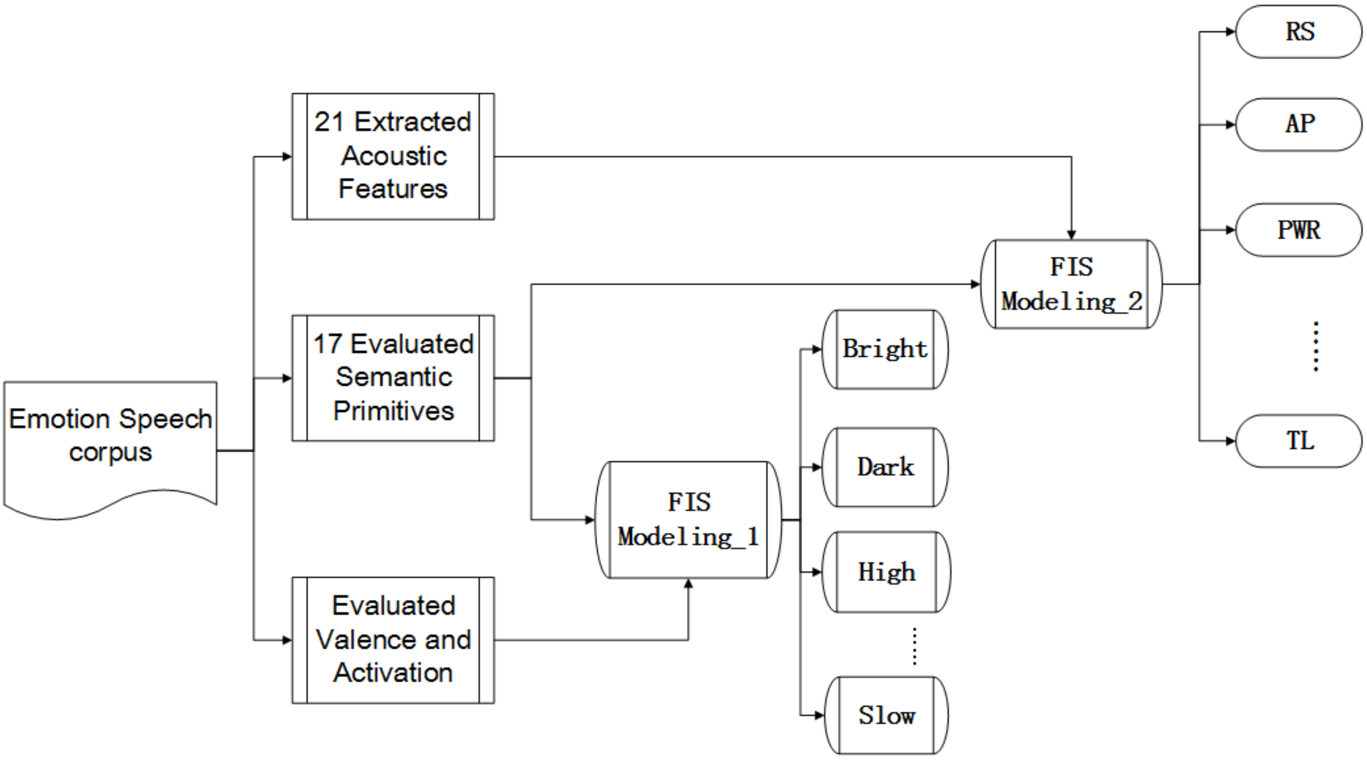

21 Extracted Acoustic Features
17 Evaluated Semantic Primitives
Evaluated Valence and Activation
Emotion Speech corpus
FIS Modeling_1
FIS Modeling_2
Bright
Dark
High
Slow
RS
AP
PWR
TL

Fig. 4. Flow chart for training FIS

Modeling_1, is for estimating the semantic primitives from the value of valence and activation. The other, called FIS Modeling_2, is for obtaining the estimated value of acoustic features from the semantic primitives. Fuzzy logic is considered as it turns human knowledge into mathematical models using If-Then rules and is based on the non-linear functions of arbitrary complexity. The relationship between the acoustic features and emotion dimension is also non-linear. Moreover, fuzzy logic is based on natural language, and the natural language used in our system is in the form of semantic primitives. While sometimes it is difficult to transform human knowledge into a rule base, ANFIS overcomes this problem by using artificial neutral networks that can identify fuzzy rules and tune the parameters of membership functions automatically.

There are two reasons that fuzzy logic is considered instead of other methods such Deep Neural Work (DNN). One is that ANFIS has membership function with interpolate method which means that from a few set of database, the tendency of variance in the whole V-A space can be obtained . While DNN do not has the ability of obtain the relationships of all positions in V-A space between acoustic features and dimensional space with a limit database as the function DNN used is point to point. Another one is that fuzzy logic is on the basic of natural language and the natural language used in our system is the form of semantic primitives (the middle layer in three-layered model).

Figure 4 shows the flow chart for estimating the acoustic features. ANFIS is a system with multi-input and single-output. For FIS Modeling_1, 17 ANFISs were trained because 17 semantic primitives whose inputs were the values of valence and activation were used in the middle layer. FIS Modeling_2 has 21 ANFISs for the same reason that 21 acoustic features are modified to synthesize the emotional speech. The input of the FIS Modeling_2 is the semantic primitives evaluated in the listening test when training the model. The values of acoustic features were found to change greatly for emotional speech and neutral speech [18]. Different people have different vocal tracts, which will influence some acoustic features such as formant frequency. For avoiding speaker-dependency and emotion-dependency, all acoustic features were normalized by the neutral speech using (3)

$$\hat{f}_{(i,m)} = \frac{f_{(i,m)}}{\sum_{i=1}^{l} f_{(i,m)}/l} \tag{3}$$

where m is the number of acoustic features ($m = 1, \ldots, 21$) and i is the number of utterances in the database. $f_{(i,m)}(i = 1, 2, \ldots, l, \ldots, 179)$ is a sequence value of the mth acoustic feature which come from the extracted values in the database explained in Section IV(B). The first l represents the value of the neutral acoustic features, and the rest are set to other emotional states. By using (3), $\hat{f}_{(i,m)}$ can be calculated which represents the normalized value from the ith utterances of the mth acoustic feature. The requirement for using ANFIS is that all input and output should be from 0 to 1. Therefore, using the range and minimum value of every variable, all acoustic features, all semantic primitives, and all emotion dimensions were normalized in the range [0,1] when training the model. Using (4), the acoustic features between 0 and 1 can be got

$$\tilde{f}_{(i,m)} = \frac{\hat{f}_{(i,m)} - fmin_m}{fran_m} \tag{4}$$

where m is the number of acoustic features ($m = 1, \ldots, 21$) and i is the number of utterances in the database ($i = 1, \ldots, 179$). $\hat{f}_{(i,m)}$ is the normalized value from (3). $fmin_m$ and $fran_m$ is the minimum value and range of the mth acoustic features which have appeared in (4). Using (4), $\tilde{f}_{(i,m)}$ which means the normalized value in the range [0,1] can be used as the input for training ANFIS. For semantic primitives and emotion dimension, the normalized part into [0,1] is the same as acoustic features which is firstly needed to subtract the minimum value and then divide the range value of the semantic primitives or emotion dimension. All the data sets were divided into the training data (90%) and testing data (10%) in order to avoid the over-fitting of the model being developed. ANFIS is first trained using the training data and then validated using the testing data.

After the process of training the FIS model, equation (5) is used to obtain the estimated semantic primitives when given the input value of valence and activation. In (5), v, a represents the value of valence and activation separately, $F1_n$ means the ANFIS of the nth semantic primitives which is the same as FIS Modeling_1 in Figure 4. And sp_n represents the estimated value of the nth semantic primitives where n is the number of semantic primitive ($n = 1, \ldots, 17$).

$$sp_n(v, a) = F1_n(v, a) \tag{5}$$

The 17 estimated semantic primitives are the input of the FIS Modeling_2. When training the model, the extracted acoustic features were normalized in the range [0,1] so that the denormalized procedure is needed using (6) to get the actual estimated acoustic features.

$$\check{f}_m(sp_n) = F2_m(sp_n) \times fran_m + fmin_m \tag{6}$$

Equation (6) is adopted to obtain the acoustic features for synthesis where m is the number of acoustic features ($m = 1, \ldots, 21$), $fran_m$ is the range of the mth acoustic feature, and $fmin_m$ is the minimum value of the mth acoustic feature which give the same meaning as (4). $F2_m$ represents the ANFIS of the mth acoustic feature which is the same as FIS Modeling_2 in Figure 4. And $\check{f}_m$ is the estimated acoustic features.

B. Related Acoustic Features of Semantic Primitives

Since not all 21 acoustic features have a strong relation with the 17 semantic primitives, the acoustic features with less relation with all semantic primitives are not put into use to simplify modifications of acoustic features without spurious and wrong estimations of them. The related acoustic features of every semantic primitive were selected for synthesizing the emotional speech. The selection procedure is based on the following hypothesis: acoustic features highly related to semantic primitives hugely affect how emotional speech is synthesized. The selection procedure was done in the following four steps:

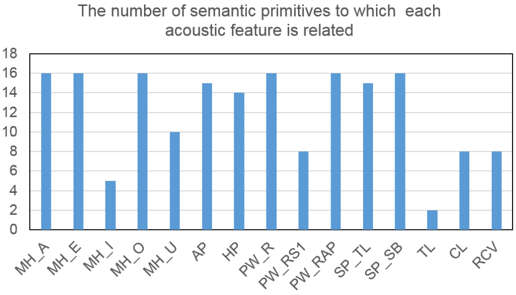


Fig. 5. Number of semantic primitives to which every acoustic feature is related

Step (1): Choosing one utterance with the maximum extent of one semantic primitive, such as Bright, from our database.

Step (2): Extracting the values of 17 semantic primitives ($\check{sp}_n(n = 1, \ldots, 17)$) of the utterance with the highest value of Bright from database and putting them in the FIS Modeling_2 using (6) so that the values of 21 acoustic features ($\check{f}_m(m = 1, \ldots, 21)$) can be obtained. On the other hand, the values of 17 semantic primitives ($\overline{sp}_n(n = 1, \ldots, 17)$) of the utterance with the neutral voice as the input of FIS Modeling_2 and the according acoustic features ($\bar{f}_m(m = 1, \ldots, 21)$) can be extracted.

Step (3): The following function is used to calculate percentage variation between the brightest and the neutral speech of one acoustic feature

$$per_m = \frac{\check{f}_m(\check{sp}_n)}{\bar{f}_m(\overline{sp}_n)} \tag{7}$$

Step (4): Selecting the highly correlated acoustic features for synthesizing. Considering the number of acoustic features related to semantic primitives, the percentage (per_m) above 1.4 and below 0.7 is chosen.

After the related acoustic features of every semantic primitive have been obtained, the number of semantic primitives to which every acoustic feature is related is calculated as shown in Figure 5. This figure also shows that the number of acoustic features related to the semantic primitives have some tendencies which is limited to 16 at most. The larger number the acoustic feature has, the closer relationship that the acoustic feature is related to semantic primitives.

C. System Evaluation

The three layers are connected using ANFIS so that by giving the value of activation and valence, the estimated acoustic features can be obtained. However, the accuracy of the estimated acoustic features and semantic primitives has not been explored yet. Correlation coefficient $R^{(j)}$ between the estimated acoustic feature y and the extracted acoustic feature

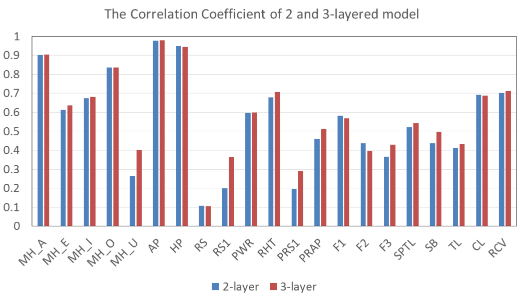


Fig. 6. Correlation coefficient of two- and three-layered models

x can be determined by the following equation:

$$R^{(j)} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}} \tag{8}$$

where $\overline{x}$ and $\overline{y}$ are the average values for $x = \{x_i^{(j)}\}$, $y = \{y_i^{(j)}\}$, respectively. The correlation coefficients of estimated and evaluated semantic primitives use the same equation. The two-layered model of Hamada et al. [7] uses the same method for evaluating the system performance. By using the two synthesis systems separately, the values of valence and activation of 179 utterances are given as the inputs, and acoustic features can be obtained. Figure 6 displays the correlation coefficient results of the two and three-layered models (blue and red columns, respectively). From this figure, we can see that among the performances of estimating the 21 acoustic features, correlation coefficients were higher for 12 acoustic features when using the three-layered model, the same for seven acoustic features when using both models, and higher for two acoustic features when using the two-layered model. Therefore, a conclusion can be made that the three-layered model more accurately estimated the acoustic features than the two-layered model.

VI. Modification of Acoustic Features

A. Modified value of acoustic features

After the ANFIS was used to obtain the estimated acoustic features, voice morphing was done using the estimated acoustic features as shown in Figure 7. All 21 acoustic features obtained from ANFIS were modified in accordance with the following equation:

$$fmod_m(v,a) \quad = \quad f_{(1,m)} \quad \times \quad \frac{\check{f}_m(sp_n(v,a))}{\check{f}_m(sp_n(0,0))} \tag{9}$$

where v means the value of the valence, and a means the value of the activation. m is the number of acoustic features ($m = 1, \ldots, 21$). $f_{(1,m)}$ is the extracted value without any normalization of the mth acoustic feature from the 1st utterance in the database which is the neutral voice. $\check{f}_m(sp_n(v,a))$ is the estimated acoustic feature value using (5) and (6) when input the value of valence and activation. $\check{f}(sp_n(0,0))$ is the estimated acoustic feature value using (5) and (6) when input the value of valence and activation are both 0. $fmod_m$ is the modified value of the mth acoustic feature. The original voice was morphed using the modified values of acoustic features.

B. Modeling of F0 contour

Duration and spectrum parts were the same as those in the work of Huang and Akagi [9], and the spectrum of glottal waveform part is the same as that in the concept of Hamada et al. [7]. In segmentation part, durations of phoneme, phrase, and accent parts were measured manually. By using STRAIGHT [13], power envelope and spectral sequence were extracted. The spectrum of glottal waveform is extracted using the ARX-LF model [19]. This procedure was exactly the same for the two-layered model [7].

The modification method of F0 related acoustic features (such as average pitch, highest pitch, f0 mean value of rising slope, and rising slope of the first accentual phrase) is changed in this paper. Because it is not suitable to extracted the F0 related acoustic features separately as previous work done. The Fujisaki model [14] is adopted to extract the trajectory of F0 contour from which we can obtain the F0 related acoustic features all at once. The Fujisaki model is a mathematical model represented by the sum of phrase components, accentual components, and the base line (Fb). The F0 contour can be expressed by

$$\begin{aligned} \ln F_0(t) = \ln Fb + \sum_{i=1}^{I} Ap_i Gp_i(t - T_{0i}) \\ + \sum_{j=1}^{J} Aa_j\{Ga_j(t - T_{1j}) - Ga_j(t - T_{2j})\} \end{aligned} \tag{10}$$

$$Gp_i(t) = \begin{cases} \alpha_i^2 t \exp(-\alpha_i t), & t \geq 0 \\ 0, & t < 0 \end{cases} \tag{11}$$

$$Ga_j(t) \begin{cases} \min[1 - (1 + \beta_j t)\exp(-\beta_j t), \gamma], & t \geq 0 \\ 0, & t < 0 \end{cases} \tag{12}$$

where $G_{p(t)}$ represents the impulse response function of the phrase control mechanism, and $G_{a(t)}$ represents the step response function of the accent control mechanism. The symbols in these equations forecast

F_b: baseline value of fundamental frequency,

I: number of phrase commands,

J: number of accent commands,

A_{pi}: magnitude of the ith phrase command,

A_{aj}: amplitude of the jth accent command,

T_{0i}: timing of the ith phrase command,

T_{1j}: onset of the jth accent command,

T_{2j}: end of the jth accent command,

α: natural angular frequency of the phrase control mechanism,

β: natural angular frequency of the accent control mechanism,

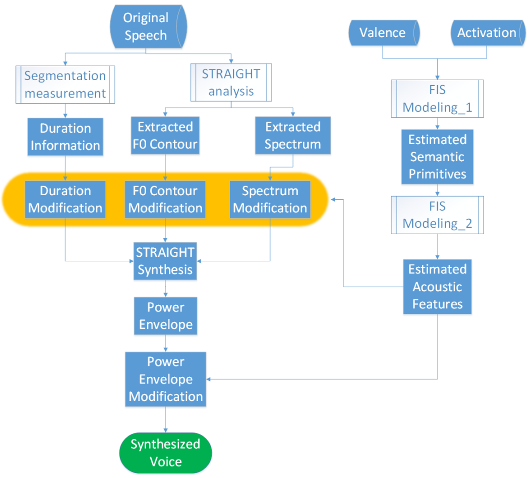


Fig. 7. Process of modifying voice

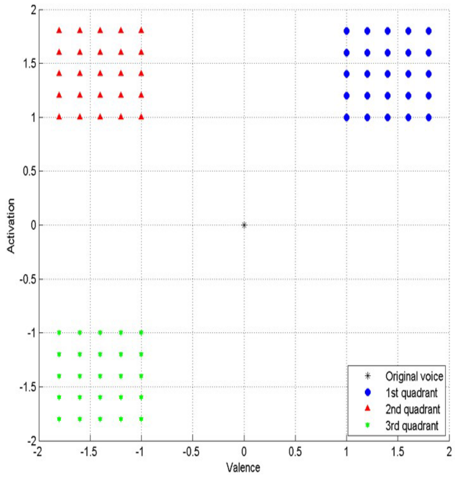


Fig. 8. Stimuli position in valence-activation space

γ: relative ceiling level of accent components.

Many researchers utilize the Fujisaki model, and the work of Mixdorff [20] is adopted in this paper where $\alpha = 1.0/s$ and $\beta = 20/s$. The modification procedure of the parameters of the Fujisaki model (Ap, Aa, and Fb) from estimated acoustic features related to F0 is done by three steps.

Firstly, the gradient of F0 and that of envelope are related to the duration. The related acoustic features of the duration are total length, consonant length, and ratio between consonant length and vowel length. The related parameters of Fujisaki model are T0, T1 and T2. According to the ratio between the extracted values of the acoustic features and the estimated values of the acoustic features, T0, T1 and T2 were modified.

Then the parameter values of Fujisaki model of the original speech was estimated following the work of Mixdorff [20]. Since Fb, Ap and Aa are related to the values of the estimated F0 related acoustic features, Fb, Ap and Aa were modified.

Lastly, by calculating the values of F0 related acoustic features, the optimized values of Fb, Ap and Aa were estimated. By controlling fundamental frequency, neutral speech was converted into emotional speech related to a position on the V-A space. Using modified F0 contour modeled Fujisaki model, target voice was converted.

C. Evaluation

1) Listening Test: Listening tests are required to verify whether synthesized speech can be perceived as the intended position. What's more, the naturalness of synthesized voice is test by doing subject evaluation.

Subjects: Seven Japanese students (six males and one female; mean age: 25 years old) were invited to do the listening tests. The number of subjects is the same for the two-layered model.

Stimuli: In the listening test for the two-layered model [7], 76 synthesized voices were used with the same position in V-A space as shown in Figure 8. A happy voice was situated in the 1st quadrant, an angry voice the 2st quadrant, and a sad voice in the 3rd quadrant. Every quadrant had 25 pieces of synthesis speech, and there was one neutral voice in the center position, which is the original spoken by a professional voice actress in the Fujitsu database. The content of all utterances was

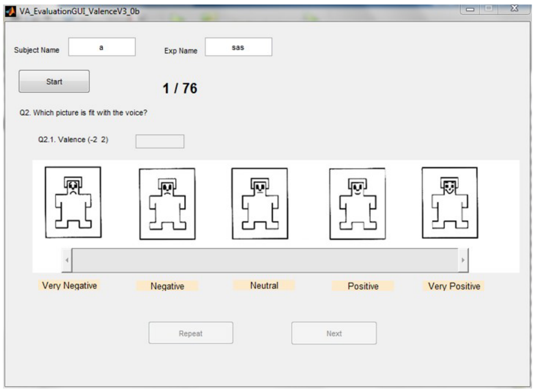
- /Atarashi meru ga todoite imasu./ (Japanese original).
- /You've got a new e-mail./ (English translation).

Procedure: In a soundproof room, subjects were invited to listen to the stimuli, which were presented through an audio interface (FIREFACE UCX, Syntax Japan) and headphones (HDA200, SENNHEISER). The mean sound pressure level of the original voice was 65 dB, and the sound pressure level of all stimuli ranged from 63 dB to 67 dB.

For valence and activation, subjects listened to all stimuli twice. The reason is that they were supposed to acquire an impression of the whole stimulus the first time and then evaluate one dimension from -2 to 2 in 40 scales. What is more, valence and activation needed to be done separately. The interval was at least one day so that they would not mistake the conception of valence and activation. Valence and activation were evaluated using forty scales (Valence: Left [Very Negative], Right [Very Positive]; Activation: Left [Very Calm], Right [Very Excited]: range $-2 \sim 2$ by 0.1 step). Subjects evaluated these scales using the graphic user interface (Figure 9). In each evaluation task, subjects could listen to the stimulus repeatedly.

For naturalness, all synthesized speeches are presented once before subjects give evaluations. The scale of evaluations is divided into 5 levels from bad to excellent (1 $\sim$ 5). Subjects give evaluations according to original speech spoken by human whose naturalness is excellent. In addition, all synthesized speeches can be listened repeatedly.

2) Results of listening test: The evaluated positions in V-A space are shown in Figure 10. As the position of stimuli, the number of subjects, and the evaluation in the listening test were the same as for the two- and three-layered models [7],



VA_EvaluationGUI_ValenceV3_0b

Subject Name: a

Exp Name: sas

1 / 76

Q2. Which picture is fit with the voice?

Q2.1. Valence (-2 2)

Fig. 9. Graphic user interface for evaluation

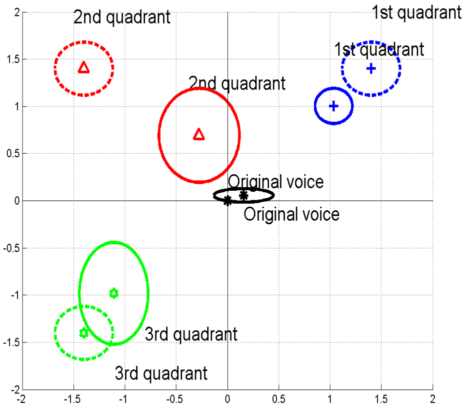


Fig. 10. Evaluated positions in valence-activation space using three-layered model. Blue, red, and green points are the average values of the 1st, 2nd, and 3rd quadrants, respectively. Each circle describes the standard deviation (solid: evaluated value for synthesize voice; dashed: stimulus value for intended emotional voice)

we can compare the results of positions in V-A space between both models [7]. The results of the two-layered model for evaluated position are shown in Figure 11.

To investigate the distance between the intended position of stimuli and listeners' evaluation, we calculated the mean absolute error (MAE) (How much error is there between the stimuli's positions and the evaluated positions?). The values of distance between the intended values and evaluated values of valence and activation were calculated separately in each quadrant. The MAE is calculated in accordance with the following equation:

$$MAE^{(j)} = \frac{\sum_{i=1}^{N} |\hat{x}_i^{(j)} - x_i^{(j)}|}{N} \tag{13}$$

where $j \in \{V, A\}$, $\hat{x}_i^{(j)}$ is the evaluated value for the synthesis voice and $x_i^{(j)}$ is the value of the intended stimulus. The MAEs for each quadrant by two- and three-layered models [7] are shown in Figure 12.

Considering the evaluation results of naturalness, the values of mean opinion score in each quadrant are shown in Figure 13. From this figure, we can see that the naturalness of the 1st quadrant synthesized voice is about 3.8. For the 2nd and 3rd quadrant, the mean opinion score is 2.8 and 2.7. The score of the 1st quadrant is the highest.

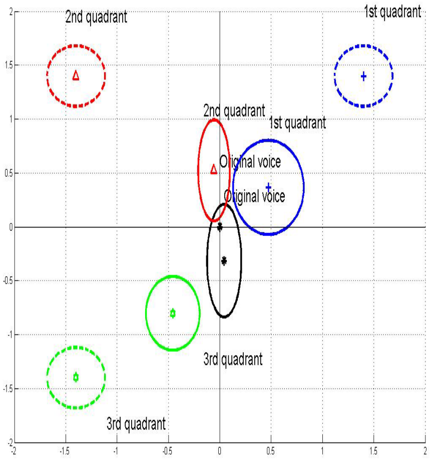


Fig. 11. Evaluated positions in valence-activation space using two-layered model [7]

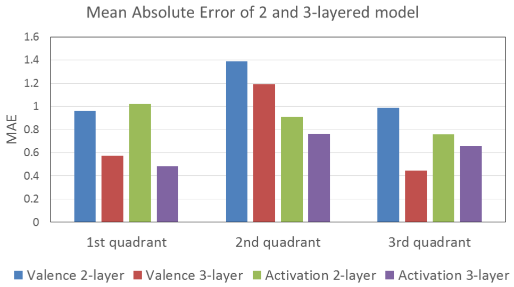


Fig. 12. MAEs of each quadrant using two- and three-layered models

D. Discussion

Comparing Figures 10 and 11, we can find that the evaluated positions from three-layered model in the three quadrants are closer to the intended positions than those from the two-layered model. The position evaluated by two-layered model is close to the center point, which means that the synthesized speech may not express the strong intensity of emotion. In contrast, the position evaluated by the three-layered model is much closer to the intended position. The results of MAEs in Figure 12 reveal that the mean absolute value between stimuli position and evaluated position for the three-layered model is about 0.6, which improves on that for the two-layered model, 1.0. However, from Figure 10, we can see that anger is not as well-perceived as the other two emotions, resulting in the MAEs of valence and activation in the second quadrant being higher than in the others. In the future, the improvement of

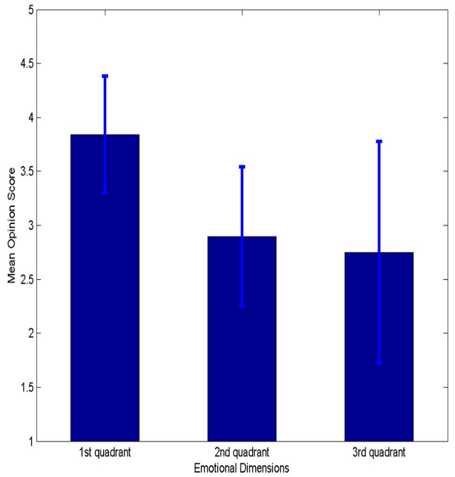


Fig. 13. Naturalness of synthesized voices of each quadrant

anger voice needs to be investigated. From Figure 12, the naturalness of the 1st quadrant is the highest comparing to the two other quadrants. All scores are above 2.5, which means the naturalness of all three quadrants synthesized voices can achieved a acceptable level.

Three-layered model provide a high estimation method and Fujisaki model improve the modification method. Both give contributions to the final improvement. The reason why three-layered model achieve higher accuracy than two-layered model is that the relationship between acoustic features and dimensional space is non-linear and so complicated. The two-layered model only consider the relationship once while the three-layered model try to build this system in two steps. Therefore the higher accuracy can be obtained. On the other hand, three-layered model following the process of human perception which assume that human perceive emotion not from acoustic features but can tell the intensity of some adjectives.

All in all, the new estimation method outperforms the two-layered model. The synthesized speech using the revised model with the added modification method can give the intended impression. What's more, the distance between the synthesized and intended speech is smaller for the three-layered model than for the two-layered model, which is a great improvement.

VII. Conclusion

This paper proposed an emotional speech synthesis system using a three-layered model in a dimensional approach. ANFIS was used to connect the three layers for estimating the semantic primitives and acoustic features. The related acoustic features were used for synthesizing the emotional speech by morphing rules. The higher correlation coefficient comparing to the two-layered model [7]shows that three-layered model estimates acoustic features more accurately than the previous two-layered model. Results of subjective evaluations revealed that emotional speeches converted by three-layered model using new modification method, Fujisaki method can give the intended impression to a much similar degree as than the previous two-layered model in the emotion dimension. And the mean opinion score of naturalness is about 3.2 above the average score 2.5 by subject evaluations which is acceptable. Above all, a conclusion can be made that an emotional conversion system utilizing three-layered model in dimensional approach can achieve better quality synthesized emotional speech than previous method.

Acknowledgment

This study was supported by the Grant-in-Aid for Scientific Research (A) (No. 25240026) and the JSPS A3 Foresight program.

References

[1] Akagi, M., Han, X., Elbarougy, R., Hamada, Y., & Li, J. "Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages". Proc. APSIPA2014, CD-ROM, Siem Reap, Cambodia, 2014.

[2] Barra-Chicote, R., Yamagishi, J., King, S., Montero, J. M., & Macias-Guarasa, J. "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech".Speech Communication, 52(5): 394-404, 2010.

[3] Yamagishi, J., Nose, T., Zen, H., Ling, Z. H., Toda, T., Tokuda, K.. & Renals, S. "Robust speaker-adaptive HMM-based text-to-speech synthesis". IEEE Transactions on, Audio, Speech, and Language Processing, 17(6): 1208-1230, 2009.

[4] Albrecht, I., Schrder, M., Haber, J., & Seidel, H. P. "Mixed feelings: expression of non-basic emotions in a muscle-based talking head". Virtual Reality, 8(4), 201-212, 2005.

[5] Schrder, M, et al. "Acoustic correlates of emotion dimensions in view of speech synthesis". Proc INTERSPEECH. 2001.

[6] Grimm, Michael, and Kristian K. "Emotion estimation in speech using a 3d emotion space concept". INTECH Open Access Publisher, 2007.

[7] Hamada, Y., Elbarougy, R., & Akagi, M. "A method for emotional speech synthesis based on the position of emotional state in Valence-Activation space". Proc. APSIPA2014, CD-ROM, Siem Reap, Cambodia, 2014.

[8] Scherer, K.R., "Personality Inference from Voice Quality: The Loud Voice of Extroversion". European Journal of Social Psychology, 8, 467-487, 1978

[9] Huang, C. and Akagi, M. "The building and verification of a three-layered model for expressive speech perception". Proc. JCA2007,CD-ROM, 2007.

[10] Jang, J-SR. "ANFIS: adaptive-network-based fuzzy inference system". IEEE Transactions on Systems, Man and Cybernetics, 23.3, 665-685, 1993.

[11] Nauck, Detlef, Frank K, and Rudolf K. "Foundations of neuro-fuzzy systems". John Wiley & Sons, Inc., 1997.

[12] Elbarougy R, Akagi, M. "Speech Emotion Recognition System Based on a Dimensional Approach Using a Three-Layered Model". Proc. of APSIPA CD-ROM, Los Angers, USA, 2012.

[13] Kawahara, H., Masuda-Katsuse, I., & De Cheveigne, A. (1999). "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, 27(3), 187-207, 1999.

[14] Fujisaki, H. "Information, prosody, and modeling-with emphasis on tonal features of speech". Proc. Speech Prosody, Nara, Japan, 1-10, 2004.

[15] Maekawa,H. "Production and perception of paralinguistic information". Proc of Speech Prosody, Nara. 367-374, 2004.

[16] Van Son, R. J. J. H., and Pols, L. "An acoustic description of consonant reduction." Speech communication 28.2, 125-140, 1990.

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

[18] Vlasenko, Bogdan, et al. “Vowels formants analysis allows straightforward detection of high arousal emotions.” Multimedia and Expo (ICME), 1-6, 2011.

[19] Agiomyrgiannakis, Yannis, and Olivier R. “ARX-LF-based source-filter methods for voice modification and transformation.” Proc. ICASSP, Taipei, Taiwan, 3589-3592, 2009.

[20] Mixdorff, H. “A novel approach to the fully automatic extraction of Fujisaki model parameters”. Proc. ICASSP, Istanbul, Turkey, 1281-1284, 2000.