Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space

Yawen Xue*, Yasuhiro Hamada, Masato Akagi

School of Information Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, 923-1292, Japan

Abstract

This paper proposes a rule-based voice conversion system for emotion which is capable of converting neutral speech to emotional speech using dimensional space (arousal and valence) to control the degree of emotion on a continuous scale. We propose an inverse three-layered model with acoustic features as output at the top layer, semantic primitives at the middle layer and emotion dimension as input at the bottom layer; an adaptive-based fuzzy inference system acts as connectors to extract the non-linear rules among the three layers. The rules are applied by modifying the acoustic features of neutral speech to create the different types of emotional speech. The prosody-related acoustic features of F0 and power envelope are parameterized using the Fujisaki model and target prediction model separately. Perceptual evaluation results show that the degree of emotion can be perceived well in the dimensional space of valence and arousal.

Keywords: Emotional voice conversion, rule-based speech synthesis, emotion dimension, three-layered model, Fujisaki F0 model, target prediction model.

1. Introduction

In terms of human-computer interaction (HCI), synthesized speech has burgeoned at a rapid rate in recent years to fulfill the demand for daily speech communication. Natural sounding synthetic speech with only linguistic information is currently used in modern applications such as text to speech systems, navigation systems, robotic assistants, story teller systems and speech to speech translation systems. Fujisaki proposed that information conveyed by speech should be summarized through *linguistic information*, which is discrete categorical information explicitly represented by the written language or uniquely inferred from context; but also *paralinguistic information*, discrete and continuous information added by the speaker to modify or supplement the linguistic information, as well as *non-linguistic information*, information not generally controlled by the speaker, such as the speaker's emotion, gender, age, etc [1]. Synthesized speech with only linguistic information cannot encompass all of these factors, thus resulting in unnatural speech sounds. Therefore, affective synthesized speech that allows communication of nonlinguistic information, such as affect and intent, is increasingly required [2] [3] [4]. Affect is not restricted to emotion; for instance [5] [6], there are social affective expressions, such as expression of politeness, sarcasm, irritation, flirtation, etc., which may be more or less controllable. Emotions, ranging from an underlying emotional state to full-blown emotions, contribute substantially to the acoustic manifestation of the spoken language. In order to improve the naturalness of synthetic speech, it is necessary to incorporate the effect of emotion on speech.

Previous methods for emotional voice conversion utilized a categorical approach to express emotional states [7]. One method is the piece-wise linear mapping using a probabilistic model, Gaussian Mixture Models (GMM) [8] [9] [10] [11]. Kawanami [12] first applied GMM for spectrum transformation to emotion voice conversion. Tao [13] tested three different methods for prosody conversion and found that GMM is suitable for a small database while a classification and regression tree model will give better results if a large context-balanced corpus can be obtained. Inanoglu [14] combined a Hidden Markov Model, GMM and F0 segment selection method for transforming F0, duration and short-term spectra in data-driven emotion conversion when large amounts of parallel data are needed. Aihara [15] improved the GMM-based emotional voice conversion for both voice quality and prosody feature conversion.

Former studies [12] [14] [15] [16] [17] [18] [19] considered converting neutral speech to simple categories of emotions such as joy, anger and sad. Tao tried to label the emotion database using four degrees “strong,” “normal,” “weak,” “unlike” to each emotion category [13]. However, daily social emotions conveyed by humans are mild and not purely one emotion or another, but a mixture of emotions, e.g., anger and sad and fearful; they can be described as a continuum of nonextreme states [20] [21]. So synthetic speech with simple categories of emotions is not sufficient. This paper focuses on converting neutral speech

*Corresponding author

Email addresses: xue_yawen@jaist.ac.jp (Yawen Xue), y-hamada@jaist.ac.jp (Yasuhiro Hamada), akagi@jaist.ac.jp (Masato Akagi)

into a continuum of emotional types with varying degrees.

When modeling the emotion, two primary problems are to be considered. The first one is how to describe emotions. In the literature, there are many descriptive systems for emotion. The most straightforward description is the utilization of emotion-denoting words or category labels [12] [13] [14] [15] [16] [17], called emotion category [22]. Also there are other less-well-known methods, prototype descriptions [23], appraisal-based descriptions [24], the circumplex model [25], physiological descriptions [26] and dimensional approaches [20] [27] [28]. Emotion in daily speech communication is highly diverse. Many human-machine dialogues need machines to express mild and nonextreme emotional states. Therefore, an emotion dimensional approach which satisfies the requirement to express a range from low-intensity to high-intensity states is appropriate for representing a continuum of non-extreme emotional states [20] for controlling the degree of emotion.

Another problem is how to model the process of expression and perception of emotion by human beings. Many researchers [29] [30] [31] [32] base their theory and research on a modified version of the Brunswik's functional lens model [33] of perception as shown in Fig.1. Brunswik's model suggests that the process of perception of emotion is multi-layered. Huang and Akagi [34] as shown in Fig.2 proposed a three-layered model for expressive speech perception based on the Brunswik's model with emotion (listener attributions) at the top layer, semantic primitives (proximal percepts) at the middle layer, and acoustic feature (distal indicators) at the bottom layer. The three-layered model has already been applied by some researchers in the emotion recognition area [35] [36]. In this paper, we assume that the human production of emotion follows the opposite direction of human perception. This means the encoding process of the speaker is the inverse process of the decoding of the listener. Hence, an inverse three-layered model is employed as the structure between emotion and acoustic feature.

In this paper, the voice conversion system for emotional speech is built with a single speaker. In order to control the degree of emotion, the emotion dimension is adopted to express the emotional state as a point in dimensional space so the degree can be controlled by changing the position in the emotion dimension. This paper mainly focuses on prosody-related feature conversion. In the emotion conversion system as shown in Fig.3, two inputs (intended position in dimensional space and neutral speech) and two steps (rule extraction and rule application) are necessary. In the first step, the rules between acoustic feature variations of neutral and emotional ones can be extracted using a fuzzy inference system. The inverse three-layered model is set as the structure between emotion dimension and acoustics with emotion dimension as the bottom layer, the semantic primitive layer at the middle and acoustic layer at the top. As the emotional experience is biologically based, these rules have the potential ability to be applied for arbitrary speakers or languages.

The second step is to apply the rule-based voice conversion method to modify the acoustic features of neutral speech to emotional ones following rules extracted from the first step. It is widely understood that emotion is conveyed by means of a number of prosodic parameters such as voice quality and speech rate as well as fundamental frequency [4] [7] [13]. In this step, some essential prosody features such as duration, F0 contour and power envelope are parameterized by an interpolation method, Fujisaki model [37] [38] and target prediction model [39]. Then the modified acoustic features are synthesized using STRAIGHT [40], a VOCODER which can decompose speech signal into parameters so as to precisely control and modify them. Fig.3 will be explained in detail in Section 5.

This paper is structured as follows. Section 2 introduces the conceptual grounds of emotion dimension. Section 3 reviews the three-layered model as the construction between emotion dimension and acoustic features. Section 4 describes the listening tests done to obtain the relation between the acoustic features and each emotion dimension. In Section 5, the structure of the emotional voice conversion system is explained. Section 5.1 illustrates the extraction of the prosody rules for emotional voice conversion using a fuzzy inference system. The prosody conversion method is explained in Section 5.2, and Section 5.3 reports the perceptual evaluation of the resulting emotion conversion system. Lastly, the conclusion is made in Section 6.

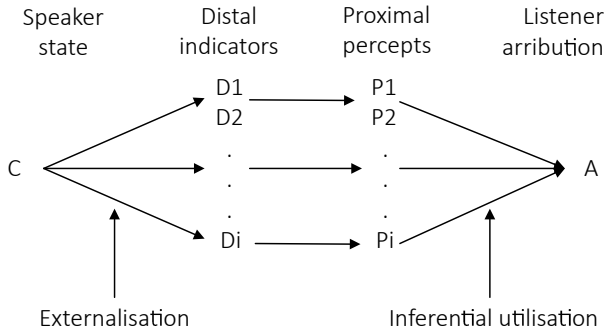

Figure 1: *Scherer's [29] modified Brunswikian lens model adopted for vocal communication of emotion.*

2. Emotion dimension representation

As mentioned above, many frameworks have been proposed already for representing emotion. Among them, categorical approach is the most common way while more and more researchers based their research on dimension representation for emotion [20] [28]. This section aims to explain the two methods in detail.

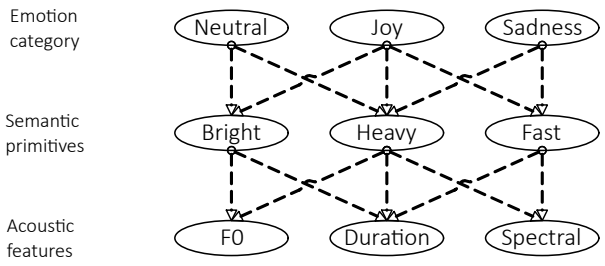

Figure 2: *Three-layered model [34].*

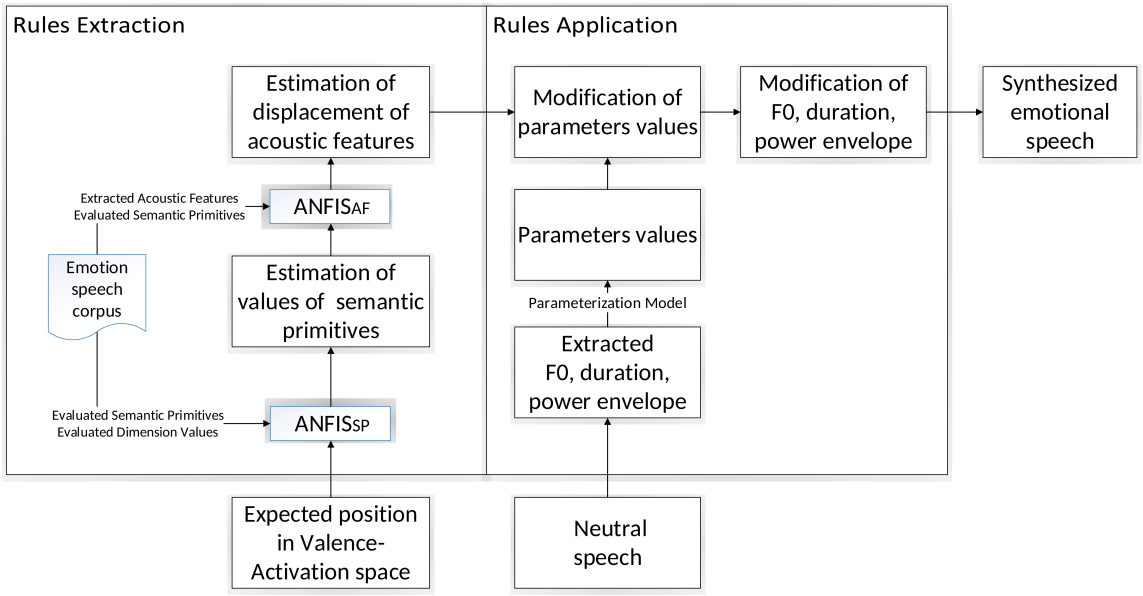

Figure 3: *Scheme of emotion conversion system.*

2.1. Categorical representation

The emotion category approach is the most straightforward method with simple emotion-denoting labels. It has been shown that emotion-denoting labels in human language are extremely powerful. It is reported that there are 107 emotion-denoting labels in English [41] and 235 in German [42]. However, it is difficult to apply all items when concentrating on emotion speech recognition or emotion speech synthesis. According to the research aim, some basic emotions or essential everyday emotion terms are selected. The merit of emotion category representation is that it is the simplest and least costly method for both emotion recognition and emotion synthesis. In the field of emotion synthesis, much previous research has attempted synthesizing affective speech with categorical emotion terms [7] [43]. However, many researchers [13] [20] [44] argued that discrete category representation ignores the diverse and fuzzy peculiarity of emotion and sometimes it is difficult to define a clear-cut boundary among the non-overlapping categories. Therefore, the complexity of emotional states may not be reflected well by categorical representation.

2.2. Dimensional representation

Humans tend to produce emotion with different degrees of intensity which may change during the course of the speech communication act. Most HCIs require the machine to produce human-like non-extreme emotion. Therefore, in order to build intelligent HCIs, a representation needs to satisfy the requirement that it can express mild emotions rather than full-blown ones.

The dimensional representation method which represents emotion as a point in a multi-dimensional space can scale the emotion intensity from low intensity to high intensity in a continuous way. Despite specifying emotion as an individual emotion category, dimensions used in this representation are gradual in nature and show the essential aspects of emotion concepts.

Through a variety of different methods such as semantic differential ratings and multidimensional scaling, three dimensions [27] (how active or calm, how positive or negative, how powerful or weak) are commonly utilized among researchers. The names of the three dimensions in literature have many versions (eg., pleasure, arousal and dominance; evaluation, activity and potency; and evaluation, activation and power). In this paper, two dimensions, as shown in Fig.4, arousal (synonymous to activation and activity) and valence (synonymous to evaluation and pleasure) are used for representing emotions based on the database we have. In the valence-arousal (V-A) representation as shown in Fig.4, joy is positive and excited while sadness is negative and calm; thus, the position values of joy are all positive and the position values of sadness are all negative in V-A space. On the other hand, anger which is negative but excited shares the negative valence but positive arousal. According to the value of valence and arousal, anger can be divided into hot and cold anger. In psychology, hot anger corresponds to the prototypical full-blown anger emotion; milder and more subtle forms of anger expression exist, including cold anger [45].

3. Three-layered model

Another problem addressed in this paper is that the voice conversion system for emotional speech needs to follow the process of human perception and production of emotion. This section discusses the methods to model the vocal communication of emotion and to apply to a voice conversion system.

3.1. Modified Brunswik's functional lens model for emotion perception

As briefly mentioned above, in the Brunswik's functional lens model as shown in Fig.1, emotion is encoded by means of a number of objective cues, called "distal indicator cues". In

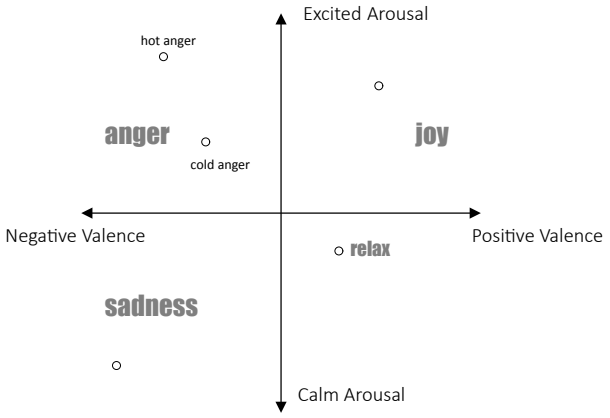

Figure 4: *Dimensional representation.*

the area of speech and emotion, distal indicator cues in principle are related to the objectively measured acoustic features. A listener perceives the distal cues through the transmission channel which are internally viewed as "proximal percepts" in the first perceptual inference process. The listener uses the percepts for "attribution" to judge the speaker's state. According to the Brunswik's lens model, we can see that the perception of emotion is not directly from "distal indicator cues", that is, acoustic to "attribution" emotion, but includes a middle procedure "proximal percepts". This means that the procedure of human emotion perception is a multiple-layered process.

3.2. Three-layered model for emotion perception

Based on the Brunswik's lens model, Huang and Akagi [34] proposed a three-layered model for emotional speech perception with emotion category at the top layer, semantic primitives constitute the middle layer and at the bottom is the acoustic feature layer as shown in Fig.2. Acoustic features refer to the acoustic parameters of voice, e.g., F0, power, duration, and semantic primitive refers to the listener's label of the voice such as bright, fast or hard. The acoustic feature, semantic primitive and emotion category in the three-layered model corresponds to the "distal indicators cues","proximal percepts" and "attribution" respectively in the Brunswik's lens model. They hypothesize that humans perceive emotion not directly from acoustic features but from some descriptors where each descriptor is an adjective for describing the perceived characteristics of the speaker's voice. The combination of descriptors accounts for the decision about which emotion the speech belongs to. The 17 semantic primitives used in the three-layered model are selected by three experiments using multidimensional scaling analysis. Some researchers have utilized this model in the field of emotion recognition [35] [36]. The top layer is modified from an emotion category to an emotion dimension since human beings have the ability to perceive gradual and continuous emotion degrees, not only categorical. They found that applying a three-layered model achieves a better emotion recognition rate compared with a two-layered model with no semantic primitive layer.

3.3. Inverse three-layered model for emotion production

Both Brunswik's functional lens model and the three-layered model are used to account for human emotion perception. According to Juslin who also uses the Brunswik's lens model in [46], two important conclusions can be made: firstly, speakers can communicate emotions successfully to listeners, and secondly, the cue utilization of speakers maps well to the cue utilization of listeners. This indicates that speakers and listeners share the same representation methodology (i.e., coding method) when doing vocal communication. According to this result, we assume that human production of emotion is the mirror effect of human perception of emotions which means the encoding process of the speaker is the inverse process of the decoding process of the listener.

Based on this assumption, the inverse three-layered model is applied to the structure of the voice conversion system for emotional speech. We assume that in order to express the "attribution", i.e., the emotion intended by the speakers, speakers firstly encode the attribution by means of a number of "proximal percepts", that is, semantic primitives. Then the "proximal percepts", are externally expressed by the "distal indicator cues", that is, acoustic features. In the inverse three-layered model, at the top layer is the acoustic feature layer, the middle layer, the semantic primitives and the bottom layer, the emotion dimension representation.

4. Acoustic features related to emotion dimensions

For speech synthesis with different emotional styles in the V-A dimensional space, the related acoustic features to each dimension are explored in this section.

Most previous methods concentrated on related acoustic features within an emotion category [34] [47]. Previous methods such as Schröder [20] [48], applied statistical analysis such as correlation and linear regression analyses to dimension space. According to these results, almost all acoustic variables correlated with the arousal axis. Correlations with the valence axis are less numerous as well as less strong. This leads to confusion when synthesizing the speaking styles related to the valence axis. Statistical methods may make a great contribution to emotion recognition because a combination of acoustic variation may lead to one kind of emotion. However, for emotional voice conversion, even if we modify some acoustic features according to the statistically-derived rules, such as duration which show great differences between emotional and neutral speech, the synthesized speech still is not perceived as a targeted (categorical) emotion.

This section investigates the acoustic features related to each dimension as applied to emotional speech synthesis. Subjects were asked to evaluate the synthesized speech, the specific acoustic features of which, such as F0, have been replaced by the F0 contour from the emotional speech but leaving the other acoustic features of the neutral speech. The idea is that if changing only the F0 contour results in the synthesized speech being rated as similar to the original emotional speech in the arousal dimension, then this means that the F0 contour makes a great

contribution to the arousal axes. If this kind of changing makes results similar to the original neutral speech, this means that F0 contour is not related much to the arousal axes. In this paper, four types of acoustic features relating to emotion are explored: duration, F0 contour, spectral sequence and power envelope.

4.1. Acoustic features replacement procedure

Source and spectral parameters can be extracted flexibly by using the analysis/synthesis method STRAIGHT [40]. Successive refinements on the extraction procedure of source and spectral parameters enable the total system to re-synthesize high-quality speech. The literature on vocal correlates of emotion dimensions, especially with respect to speaking styles, reports the importance of prosodic parameters, such as F0 contour, spectral sequences and power envelopes [13] [20].

In order to determine which particular acoustic features of the emotional speech can be used to convert the neutral speech to emotional speech, it is necessary to keep the linguistic content constant. Thus, our research examined nine sentences with the same linguistic information but different speaking styles/ emotions. These sentences were chosen from the Fujitsu database recorded in the Fujitsu Laboratory by one professional voice actress. One of the 9 sentences is in the neutral speaking style without emotion; the remaining emotions are sadness, joy, hot anger and cold anger, with 2 utterances for each emotion type.

The procedure for replacement of F0 contours shown in Fig.5 is followed. Time information was first modified to keep the speech duration of the neutral and emotional speech constant; this needs to be done before modifying the F0 contour, spectral sequence and power envelope. Time modification was done first by manually segmenting the speech signal at the phoneme level for both neutral and emotional speech; then the time duration of the neutral speech is modified to that of the emotional speech, according to the ratio of the time duration of the neutral and emotional speech. Applying STRAIGHT, the first synthesized speech (neutral speech 2) can be obtained by changing only the time duration to match that of the emotional speech. Then, the F0 contour, spectral sequence, and aperiodic component (Ap) of the neutral speech 2 are extracted using STRAIGHT. At the same time, from the emotional speech, the F0 contour and spectral sequence are also extracted using STRAIGHT. Since the time duration of the neutral speech 2 is the same as that of the emotional speech, the F0 contour of the neutral speech can be directly replaced by that from the emotional speech. The Ap and spectral sequence from neutral speech 2 and the F0 from the emotional speech are combined to be synthesized by STRAIGHT. The synthesized speech with F0 replacement is obtained lastly. By doing this, the spectral sequence and Ap information are kept, but the F0 contour is changed from neutral to emotional.

Fig.5 shows the procedure for replacing of F0 contour. For replacing the spectral sequence, the previous step is the same as F0 replacement. But in the last step, we use the F0 and Ap from the neutral speech, so that the spectral sequence from the emotional speech can be synthesized. This means the spectral sequence from neutral to emotional speech has been changed, but the other information is kept. For power envelope calculation, a Hilbert transform and low-pass filter are used. Synthesized speech with a different power envelope can be obtained by applying the power envelope of the emotional speech to neutral speech 2.

4.2. Experiment

The F0 contour, spectral sequence, power envelope and time duration of the emotional speech are moved one by one to the neutral speech. We obtained 32 samples of synthesized speech (8 utterances with the same linguistic information but different speaking styles, 4 types of acoustic features); plus 9 original utterances. Totally, there were 41 stimuli in the perception test in order to explore the influence of each acoustic feature on each emotion dimension.

In the listening test, twelve Japanese subjects with normal hearing ability were asked to evaluate the utterances in the V-A space. The stimuli were presented in an individually randomized order per subject over high-quality headphones (type HDA200, SENNHEISER).

Experiments for valence and arousal were done twice, for each dimension for a total of 4 tests. The first time served as a training test to allow the subjects to acquire an impression of all the stimuli. Valence and arousal were evaluated from -2 to 2 with a step of 0.1 (Valence: -2 [Very Negative], -1 [Negative], 0 [Neutral], 1 [Positive], 2 [Very Positive]; Arousal: -2 [Very Calm], -1 [Calm], 0 [Neutral], 1 [Excited], 2 [Very Excited]). Subjects evaluated these scales using a graphic-user interface as shown in Fig.6. During the listening test, subjects were allowed to listen to the stimulus as many times as they wanted.

4.3. Results

The correlation coefficients between subjects are calculated and the average results above 0.7 are chosen for the final analysis. Totally, there were 12 subjects who attended this experiment, but ten subjects were considered for the final analysis. In order to explore the influence of each acoustic feature on each emotion dimension, we assessed the original positions of the original emotional speech. The hollow points in Figs. 7, 8, 9 and 10 show the perceptual position values in V-A space of the original utterances. The neutral speech is almost at the center point which indicates neither positive nor negative, neither active nor calm. The values of joy are in the first quadrant which means positive and active. Hot anger is in the second quadrant which represents negative but active, and cold anger is in the second quadrants although the value of valence and arousal is lower than that of hot anger. For sad emotion, all points are in the third quadrant, which are negative and calm. These findings seem intuitively reasonable, which suggest that our subjects were able to understand the basic meaning of valence and arousal.

In order to investigate the influence of the emotion dimension, the four kinds of acoustic features are replaced separately; thus, the results of the listening tests for the synthesized speech are analyzed in terms of three aspects. Fig.7 shows the results when only the F0 contour is changed to the F0 contours of the

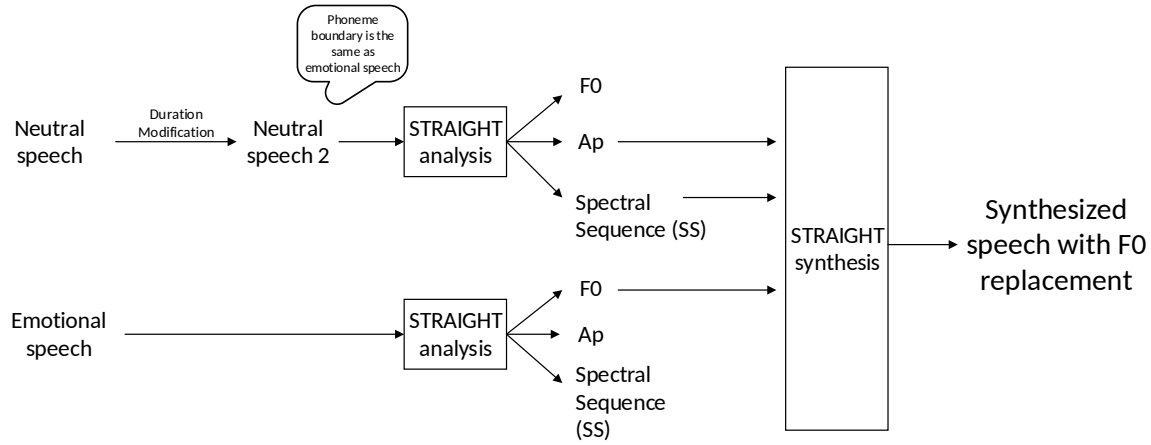

Figure 5: *Procedure of acoustic feature replacement.*

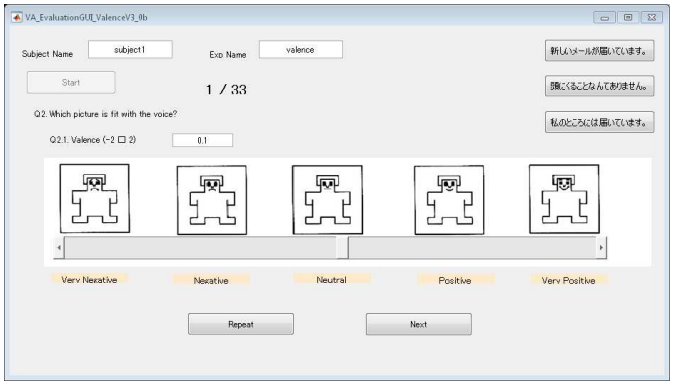

Figure 6: Graphic interface of the perceptual test.

Table 1: *Anova values of each acoustic features to valence and arousal(*** $p < 0.01$, * $p < 0.05$).*

p-value	F0	SS	PW	TM
Valence	**	**	0.53	0.13
Arousal	**	**	0.03*	0.01*

other emotion categories but keeping the other acoustic information such as spectral sequence and power envelope. Figs 8, 9 and 10 show the results when only time duration, power envelope and spectral sequence are changed to those of the other emotions while holding the remaining acoustic values.

4.4. Discussion

Figs.8 and 9 show that by replacing the duration information and power envelope, the synthesized speech is still concentrated at the center point; this means that only modifying the duration or power envelope does not very much change the expressiveness of a neutral utterance. Comparing the results of the original with the replaced ones, shown in Fig.7, we see that if only the F0 contour of the neutral speech is replaced by the F0 contours of joy and hot anger speech, the synthesized speech samples are all evaluated as joyful speech, as the evaluated position values are in the first quadrant. This is an interesting finding because most previous research proposes that

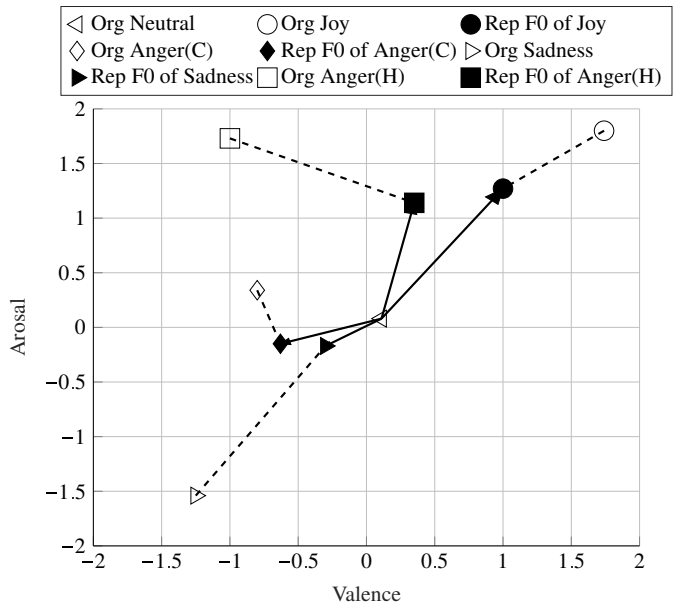

Figure 7: *Perceptual position values of original (Org) emotional utterances and synthesized (Rep) utterances on V-A space when F0 contour (F0) is replaced from neutral to emotional speech. (Anger (C) means cold anger and anger (H) means hot anger.)*

F0-related acoustic features contribute greatly to the emotions of joy and anger. To a certain extent, this is true. But when converting neutral speech to emotional speech, if only F0 information is modified, it is possible to synthesize joyful speech but not angry speech. For sad speech, replacing the F0 of the neutral with that of sad results in the synthesized speech being in the third quadrant; this means sad speech can be synthesized by modifying only the F0-related acoustic features. However, the degrees of valence and arousal are reduced in joyful and sad emotions when only F0 is replaced. For cold anger emotion, by replacing only the F0, it is rated as slightly sad. Our findings show by replacing only F0, sad and joyful speech can be distinguished well, but notice that these emotions have inverse values in terms of both valence and arousal. However, joyful and angry speech cannot be differentiated; note that these differ

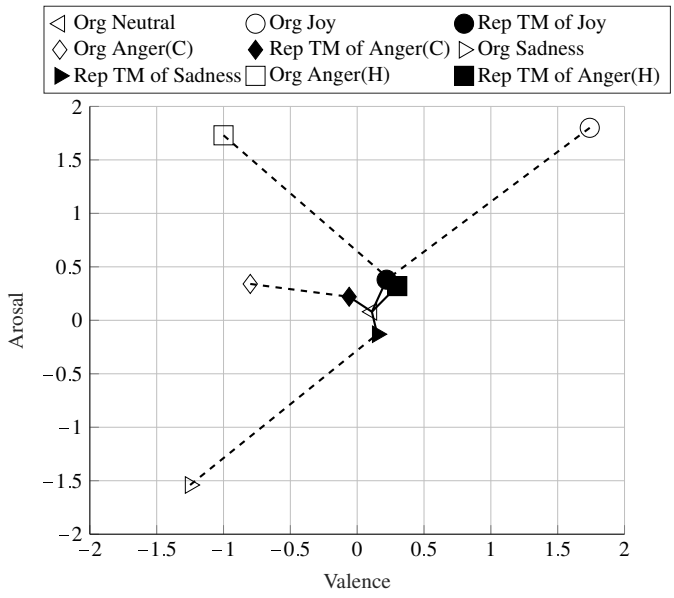

Figure 8: *Perceptual position values of original (Org) and synthesized (Rep) utterances on V-A space when time (TM) duration information is replaced from neutral to emotional speech. (Anger (C) means cold anger and anger (H) means hot anger.)*

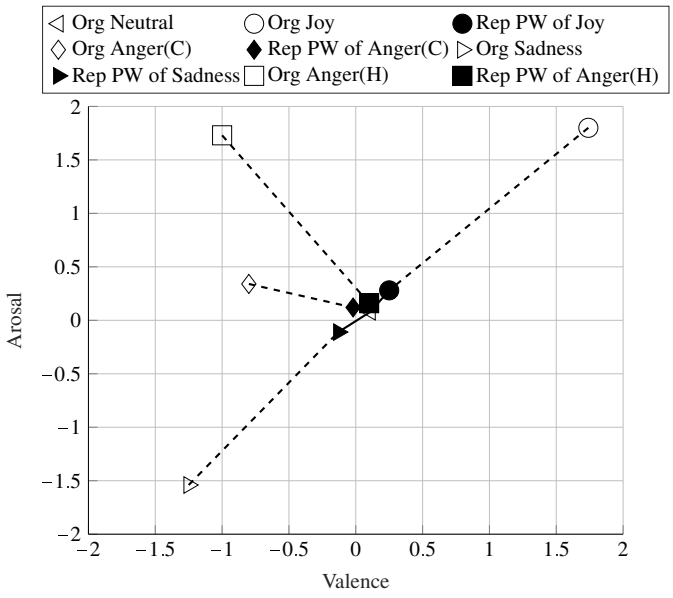

Figure 9: *Perceptual position values of original (Org) and synthesized (Rep) utterances on V-A space when power envelope (PW) is replaced from neutral to emotional speech. (Anger (C) means cold anger and anger (H) means hot anger.)*

only in the valence axis.

When replacing spectral sequences, as shown in Fig.10, synthesized speech was evaluated as the original emotion, especially for hot and cold anger. This means that if there is a suitable method for modifying the spectral sequence of neutral speech, all emotions can be synthesized, although the degrees of valence and arousal are reduced compared to the original speech. What's more, for joyful and sad speech, by replacing only the F0, we can get closer to the original position in the V-A space than by replacing only the spectral sequence. This indicates that F0 is more related to the arousal axis than the valence axis. However, for the valence axis, the spectral sequence is more important.

The results of the ANOVA (analysis of variance) are shown in Table.1. From this table, we can see that F0 and spectral sequence have significant contributions to valence and arousal axes ($p < 0.01$); power envelope and duration are much related to the arousal dimension ($p < 0.05$) but show no significance with the valence dimension ($p > 0.05$).

We conclude that both the F0 contour and spectral sequence are important to voice conversion for emotional speech. The power envelope and duration show little influence on the valence axes. In this paper, we focused on the prosody-related features such as duration, F0 and power envelope. The controlling of spectral sequence will be researched in the future. Since the utterances examined in this experiment are from a single speaker, and speakers have individuality differences when encoding emotion [53], future research is necessary to examine the commonalities among speakers for a better understanding of synthesizing different speaking styles.

5. The voice conversion system for emotional speech

A rule-based voice conversion technique is utilized for modifying the acoustic parameters of the neutral speech in order to convey the target emotion. Previous methods on emotional voice conversion systems mainly focused on applying a statistical approach, GMM, Deep neural network (DNN) or neural network (NN) [12] [14] [15] [16] [17]. GMM often suffers from over-smoothing problems and the non-linear mappings such as DNN or NN need large databases for training using categorical emotion representations. However, for the dimensional approach, it is difficult to collect a sufficiently large enough database with continuous emotional degrees. A rule-based strategy is applied with a limited database in this paper to obtain tendencies of variation between emotion dimensions and semantic primitives, and then to extract rules between semantic primitives and acoustic features.

In the rule-based emotional voice conversion system, the two-dimensional space of valence and arousal is used for representing the emotion; and the inverse three-layered model is used as the structure relating the acoustic features and emotion dimensions, as shown in Fig.3. The emotional voice conversion system needs two inputs and two steps. Firstly, we need to input the position in the V-A space, which represents the desired emotion degree, and this step is referred to as the rule extraction step. It is this step which allows us to estimate the acoustic values of the desired emotion through the inverse three-layered model. This then allows us to calculate the differences in acoustic features between the emotional and neutral speech. In the next step, the rule application step, the ratios of difference between the estimated acoustic features of the desired emotion and the acoustic features of neutral speech are applied to the extracted parameter values of the neutral speech. In order to

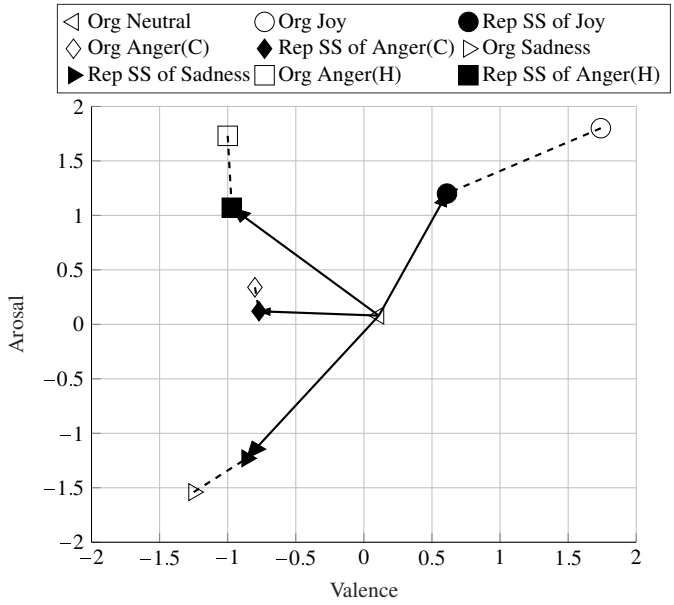

Figure 10: *Perceptual position values of original (Org) and synthesized (Rep) utterances in V-A space when spectral sequence (SS) is replaced from neutral to emotional speech. (Anger (C) means cold anger and anger (H) means hot anger.)*

modify the differences for the neutral speech, we concentrated on the prosody-related features, duration, F0 contour and power envelope. In order to control these features, the F0 contour and power envelope are parameterized using the Fujisaki model and target prediction model. After the modifications, applying the synthesis tool STRAIGHT to the modified acoustic features, the converted utterances with desired emotional degrees can be synthesized.

5.1. Rule extraction

In this section, we illustrate how the inverse three-layered model is applied to the database to obtain the value of the various elements; how the fuzzy inference system connects the three layers to output rules relating the emotional dimensions to the semantic primitives; how the rules are extracted from semantic primitives to acoustic features and finally, how the effectiveness of the inverse three-layered model is evaluated by means of calculating mean absolute errors [49].

5.1.1. Database

We used the multi-emotional single speaker Japanese Fujitsu Database, recorded at Fujitsu Laboratories. A professional voice actress uttered 179 utterances in 5 speaking styles, joy, cold anger, hot anger, sad and neutral; 20 sentences spoken in 5 speaking styles, including one instance of neutral speech and two repetitions of each of the other speaking styles. One instance of cold anger is missing which makes a total of 179 sentences.

5.1.2. Acoustic feature extraction

Except for duration-related features which are extracted by manual segmentation, the other acoustic features are obtained by the high-quality speech analysis-synthesis system STRAIGHT [40]. Based on the work by Huang and Akagi [34], 16 acoustic features are classified into the following subgroups.

F0 related features: F0 mean value of average F0 (AP), highest F0 (HP), a rising slope of the F0 contour (RS) and rising slope of the F0 contour for the first accentual phrase (RS1st).

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 (SP_TL) was 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. To describe acoustic consonant reduction, spectral balance (SP_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.

Power envelope related features: 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), the average power (PW_RHT) and the mean value of power range in accentual phrase (PW_RAP) were measured.

Duration related features: Total length (TL), consonant length (CL), the ratio between consonant length and vowel length (RCV) were considered related to duration.

All the acoustic features are used for building the inverse three-layered model in the rule extraction step. In the rule application step, the prosody related features such as F0, duration and power envelope are parameterized. The conversion of spectral sequence features will be performed in the future work.

5.1.3. Semantic primitives evaluation

Based on the work by Huang and Akagi [34], 17 semantic primitives were selected to describe the perception of emotional vocalization. The 17 semantic primitives are bright, dark, high, low, strong, weak, calm, unstable, well-modulated, monotonous, heavy, clear, noisy, quiet, sharp, fast, and slow. 11 Japanese subjects were asked to give subjective values 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") for each semantic primitive for each 179 utterances. For each semantic primitive, the inter-rater agreement is measured by pairwise Pearson's correlation between two subjects' ratings. All subjects showed from moderate to a high level agreement.

5.1.4. Emotion dimensions evaluation

The evaluation of emotion dimension is divided into two parts: valence and arousal [35]. The 11 Japanese subjects rated the 179 utterances on a five-point scale {-2, -1, 0, 1, 2 }. Valence

was from -2 (very negative) to +2 (very positive), and arousal was from -2 (very calm) to +2 (very excited). The correlation coefficient between subjects rating for valence is about 0.9 and for arousal, about 0.85, which means subjects showed a high inter-rater agreement.

5.1.5. Applying a fuzzy inference system for extracting rules

The fuzzy logic system using If-Then rules to turn human knowledge into a mathematical model is utilized as the connector for the three layers. The fuzzy logic system is based not only on the non-linear functions of arbitrary complexity but also on natural language. Contrary to the conventional fuzzy logic system which does not have a learning ability, the adaptive neuro-fuzzy inference system (ANFIS) combines the merit of fuzzy inference systems and neural networks as its own structure [50]. ANFIS not only has an inference ability but also a strong learning mechanism. ANFIS is considered instead of other popular methods such as DNN, or NN for two reasons. One is that ANFIS has a membership function with an interpolating method which means that the tendency of the variance in the whole V-A space can be obtained from a small database. A second reason is that fuzzy logic is based on natural language; the natural language in our system is in the form of semantic primitives (the middle layer in the three-layered model).

Fig.11 shows the flow chart for training the ANFIS to extract rules. Firstly, from the emotion speech corpus as introduced in Sections 6.2, 6.3, 6.4, 16 acoustic features ($AF_1, ..., AF_{16}$) are extracted by STRAIGHT; the 17 semantic primitives values ($SP_1, SP_2, ..., SP_{17}$) and the two emotion dimensions (D_1, D_2) are evaluated by subjects' ratings. To avoid any emotion dependency, all acoustic features are normalized by the mean value of neutral speech. For the ANFIS, all input and output need to range from 0 to 1. We then normalized the acoustic features, semantic primitives and emotion dimensions using the range and minimum value of each parameter using the following Eq.3.

$$\tilde{f}_{(i,m)} = \frac{\hat{f}_{(i,m)} - fmin_m}{fran_m} \tag{3}$$

where m is the number of acoustic features ($m = 1, \ldots, 16$) and i is the number of utterances in the database ($i = 1, \ldots, 179$). $\hat{f}_{(i,m)}$ is the normalized value of the neutral speech. $fmin_m$ and $fran_m$ is the minimum value and range of the mth acoustic features. For semantic primitives and emotion dimensions, the normalized part in $[0, 1]$ are the same as the acoustic features.

ANFIS is a system with multi-inputs and a single-output. In the training phase as shown in Fig.11, from the bottom to the middle layer, for each semantic primitive (SP_1, SP_2,..., SP_{17}), we train the appropriate ANFIS ($ANFIS_{SP1}$, ... , $ANFIS_{SP17}$) whose input is the same, that is, the evaluated value of valence and arousal in the emotion dimension (D_1, D_2). From the middle to the top layer, 17 semantic primitives (SP_1, SP_2, ... , SP_{17}) are the input of each ANFIS ($ANFIS_{AF1}$, ... , $ANFIS_{AF16}$), whose outputs are the acoustic features (AF_1, AF_2, ... , AF_{16}).

After the training step, 17 semantic primitives and 16 acoustic features are used in this system to generate 17 ANFISs for estimating semantic primitives and 16 ANFISs for estimating acoustic features. When given the intended position in the V-A space to each of the 17 ANFIS ($ANFIS_{SP1}$, ... , $ANFIS_{SP17}$) for estimating SP , the estimated semantic primitive ($estSP_1$, $estSP_2$, ... , $estSP_{17}$) is obtained. Applying the 17 estimated semantic primitives ($estSP_1$, $estSP_2$, ... , $estSP_{17}$) as the input to each ANFIS ($ANFIS_{AF1}$, $ANFIS_{AF2}$, ... , $ANFIS_{AF16}$) for estimating acoustic features, the estimated acoustic features ($estAF_1$, $estAF_2$, ... , $estAF_{16}$) are acquired as shown in Fig.12.

In the estimation step, the neutral position and the intended position in V-A are given separately to the ANFIS to obtain the acoustic value of neutral speech and the intended speech. We then use the estimated acoustic feature of the intended position in V-A space to divide the estimated AF of the neutral position in V-A space. In this system, we assume that (0,0), the center point in V-A space, is the neutral position. The ratio differences, i.e., the rules between intended and neutral acoustic features, are calculated using the following equation:

$$rule_n = estAF_n / \overline{estAF_n} \tag{4}$$

where $estAF_n$ shows the estimated nth acoustic feature value from the ANFIS of the intended emotional state in V-A space and $\overline{estAF_n}$ shows the estimated nth acoustic feature value from the ANFIS of the neutral speech (0,0) in V-A space. Then $rule_n$ represents the rule for the nth acoustic features($n = 1, 2, ..., 16$) which is applied for modifying the neutral speech in the next step.

5.1.6. System evaluation

All data sets are divided into training data (90%) and testing data (10%). ANFIS is first trained using the training data and then validated using the testing data. By giving the value of arousal and valence to the $ANFIS_{sp1}, ANFIS_{sp2}, ..., ANFIS_{sp17}$, firstly, the estimated semantic primitives, $estSP_1, estSP_2, ..., estSP_{17}$ can be obtained and then we input the estimated semantic primitives to $ANFIS_{AF1}, ANFIS_{AF2}, ..., ANFIS_{AF16}$, and after that, the estimated acoustic features $estAF_1, estAF_2..., estAF_{16}$ can be obtained. The accuracy of the estimated acoustic features and estimated semantic primitives are evaluated by mean absolute error (MAE); this can measure the distance between the estimated values by the proposed system and the annotated values from listeners evaluations.

$$MAE = \frac{\sum_{i=1}^{N} |x_i - y_i|}{N} \tag{5}$$

where $x_i (i = 1, 2, ..., N)$ is the sequence of estimated values of one semantic primitive or one acoustic feature. $y_i (i = 1, 2, ..., N)$ is the sequence of annotated values by listeners for the corresponding semantic primitive and acoustic feature. N is the number of utterances in the database.

Figs 13 displays the MAE results of semantic primitives between the training data and testing data from the three-layered model. Fig.14 shows the MAE of acoustic features from three-layered model and two-layered model. The two-layered model

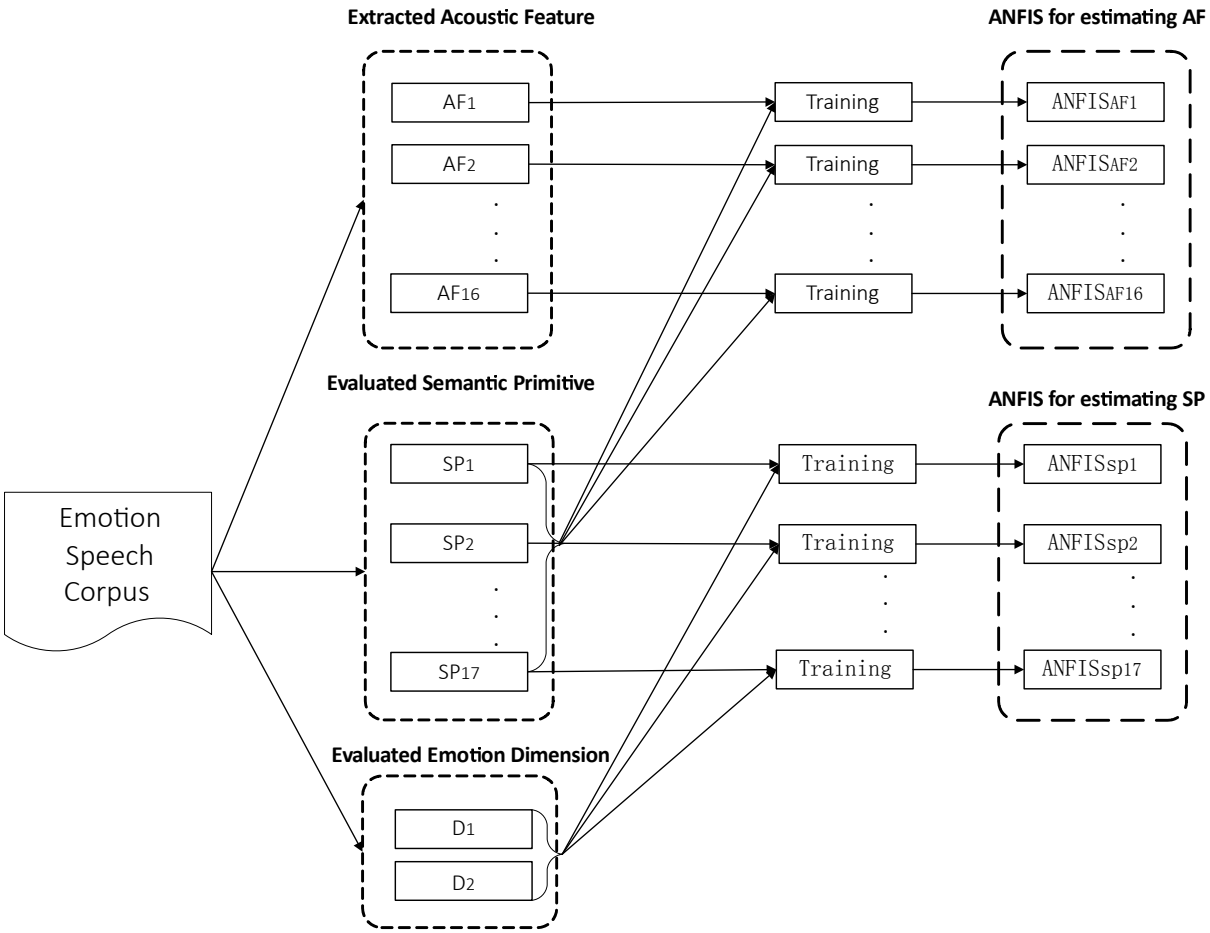

Figure 11: *Procedure for training ANFIS.*

utilized the same methodology of applying the emotion dimension for representing emotion but without considering using the semantic primitive layers. The MAE of 15 semantic primitives are all below 10%, and for the fast and slow semantic primitives, the MAE is somewhat higher, near 10% which means that the estimation accuracy of semantic primitives is very high. From Fig.14, comparing the results from the two-layered and three-layered model, it is found that among 16 acoustic features, the MAE values of 10 acoustic features from the three-layered model are lower than the two-layered model which means that the three-layered model can provide higher estimation accuracy than the two-layered model. Among the 16 acoustic features, all are below 20% and only the MAE of *PW_RAP* is higher than 15% using the three-layered model.

5.2. Rule application

For the emotional voice conversion system, the acoustic parameters of neutral speech need to be modified in order to synthesize the emotional speech. The ratios, rules of the relationships between acoustic features between neutral and intended emotion, are calculated by ANFIS through the inverse three-layered model. In this section, the modification method based on the extracted rules is explained.

As shown in Fig.15, first, the phoneme boundaries of the vowels and consonants are extracted manually from the neutral speech. Then the ratios between the neutral and target emotional speech of the acoustic features TL, CL, RCV are used to modify the phoneme boundaries. The F0 contour is extracted by STRAIGHT at the same time and interpolated using the duration information. The F0 contour is parameterized by a modified version of the Fujisaki model to modify the F0 contour. After F0 modification, STRAIGHT is applied to obtain the modified speech. Lastly, the power envelope modification is done by using the target prediction model. After the power envelope modification, the final converted emotional speech can be acquired.

5.2.1. Fujisaki model for parameterizing F0 contour

Previous work separately modified the F0 related acoustic features, such as average F0 (AP), highest F0 (HP), the mean value of F0 in the rising slope (RS) and rising slope of the first accentual phrase (RS_1st). In our case, separately modifying the acoustic features is not suitable, because modifying one acoustic feature such as RS may influence other acoustic features such as AP and HP and there is no appropriate order for modification. We parameterized the F0 contour to control the entire contour using only a limited set of parameters.

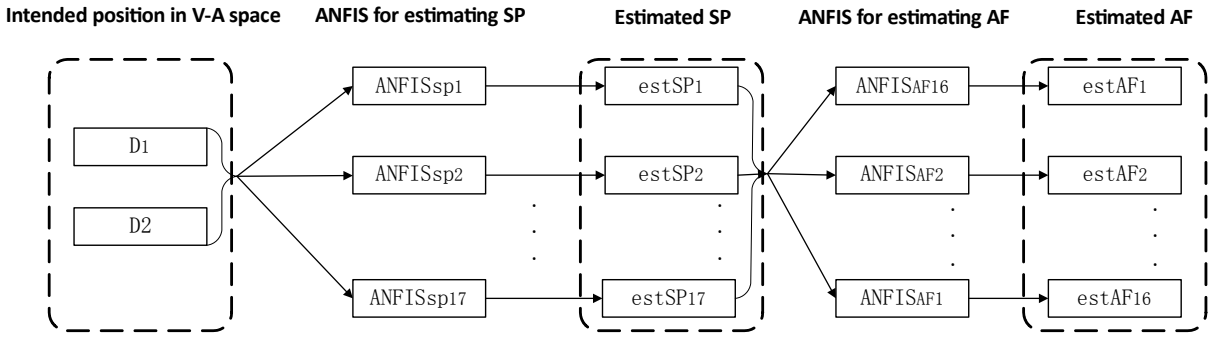

Figure 12: *Applying ANFIS for estimating acoustic features (AF) and semantic primitives (SP).*

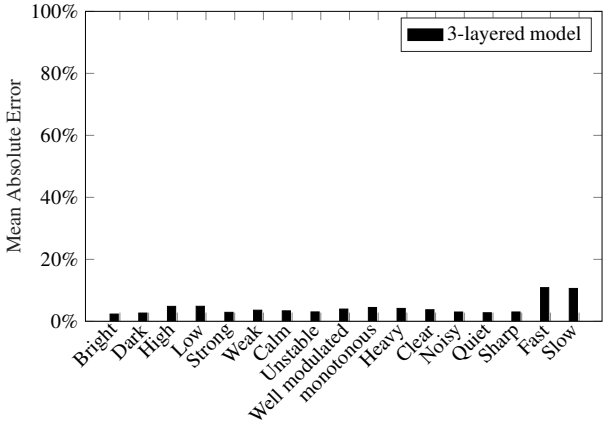

Figure 13: Mean absolute error of semantic primitives.

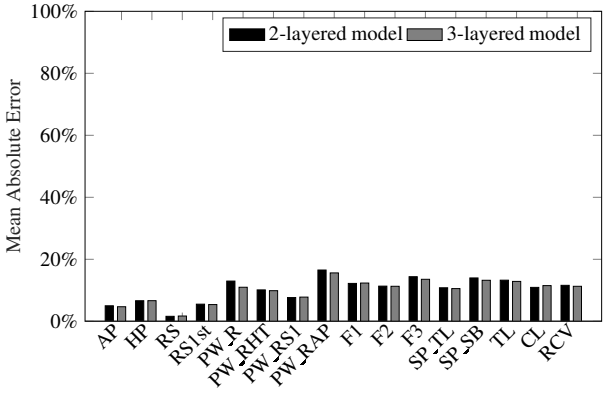

Figure 14: Mean absolute error of acoustic features from three- and two-layered model.

The Fujisaki model [1], a mathematical model represented by the sum of phrase components, accentual components, and the baseline Fb, is adopted to parameterize the F0 contour. The F0 contour can be expressed as follows.

$$\ln F_0(t) = \ln Fb + \sum_{i=1}^{I} Ap_i Gp_i(t - T_{0i}) + \sum_{j=1}^{J} Aa_j\{Ga_j(t - T_{1j}) - Ga_j(t - T_{2j})\} \quad (6)$$

$$Gp_i(t) = \begin{cases} \alpha_i^2 t \exp(-\alpha_i t), & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (7)$$

$$Ga_j(t) \begin{cases} \min[1 - (1 + \beta_j t)\exp(-\beta_j t), \gamma], & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (8)$$

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,

γ: relative ceiling level of accent components.

Many researchers utilize the Fujisaki model; the work of Mixdorff [51] is adopted in this paper where $\alpha = 1.0/s$ and $\beta = 20/s$. By using Mixdorff's method the parameters ($T0$, $T1$, $T2$, Ap, Aa, and Fb) in the Fujisaki model are extracted. We then modify the parameters to obtain a modified F0 contour using Equations 6, 7 and 8. We can extract the AP, HP, RS, and RS_{1st} of the modified F0 contour. The root mean square error (RMSE) between the desired acoustic features and extracted one from the modified F0 contour is calculated using the following equation:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(AF_i - \hat{AF}_i)^2}{N}} \quad (9)$$

where AF_1, ... ,AF_n is the desired acoustic feature value which

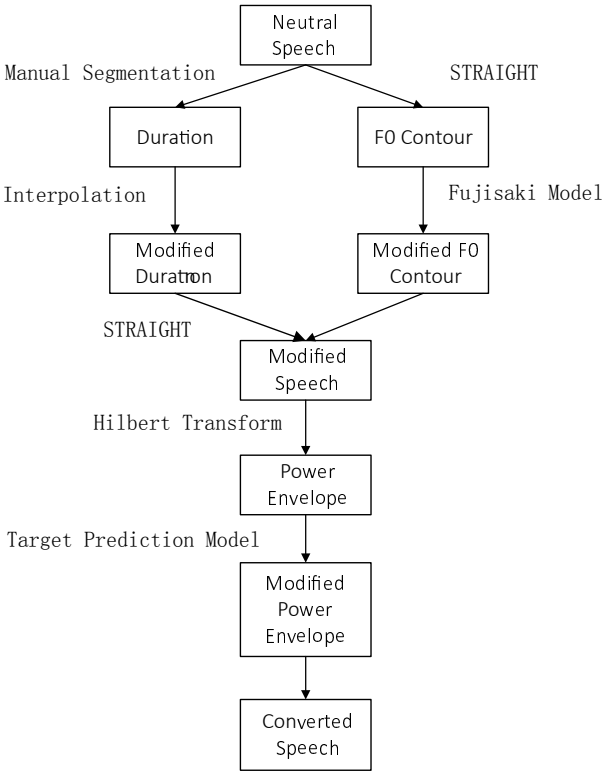

Figure 15: The procedure of modifying the neutral speech to emotional one.

is denormalized after being estimated from ANFIS. And the $A\hat{F}_1$, ... ,$A\hat{F}_n$ is the extracted value from the modified F0 contour. The F0 contour with the smallest RMSE is selected as the final F0 contour.

5.2.2. Target prediction model for parameterizing the power envelope

In order to parameterize the power envelope target, a prediction model which predicts the stable power target in short-term intervals is used to estimate the targets of the power envelope [52]. We then change the targets to a stepwise function by using the segmentation information, the starting and ending points of each phoneme. By modifying the magnitude of the stepwise targets of the power envelope, a modified power envelope is reproduced by a 2nd-order critically damped model.

The power envelope of the neutral speech signal is firstly extracted by

$$e_y(t) = \mathbf{LPF}\left[\left|y(t) + j\text{Hilbert}\left[y(t)\right]\right|^2\right] \tag{10}$$

where LPF[·] is a low-pass filtering and Hilbert[·] is the Hilbert transform. Then we used Eq.11 to change the power envelope in the log power envelope domain.

$$\log e_y(t) = 10\log_{10}(e_y(t)) \tag{11}$$

Then the power envelope is approximated by a 2nd-order critically damped system which can estimate the target power envelope using short-term power sequences without being given the onset positions of the power transition.

A 2nd-order critically damped model is generally represented as follows

$$\left(\Delta^2 - 2\lambda\Delta + \lambda^2\right) y_n = \lambda^2 b \tag{12}$$

where Δ is a differenctial operator in time, λ is a reciprocal time constant, time $n = 0$ is the onset position of the transition and b is a target to which y_n converges in the past if $\lambda > 0$ and $n \leq 0$, or in the future if $\lambda < 0$ and $n \geq 0$. The solution of Eq.12 is

$$y_n = (a + cn)\exp(\lambda n) + b \tag{13}$$

where a and c are constants obtained from the boundary condition. Previous methods that estimated the parameters of 2nd-order critically damped models have predicted all parameters directly by using Eq.13 and the following measure,

$$e\left(n_0 \text{ or } n_1, \lambda\right) = \sum_{n=n_0}^{n_1} \left|y_n^i - y_n\right|^2, \quad n_0 < n_1 \tag{14}$$

where y_n^i is an unknown input sequence. For these methods, a long-term sequence sufficient to start at the onset position of the transition $n_0 = 0$ when $\lambda < 0$ or $n_1 = 0$ when $\lambda > 0$ is essentially required. Then, non-linear optimization under two values, n_0 and λ or n_1 and λ is needed. However, the purpose of our target prediction model is to estimate b only.

Divide Eq.12 such that;

$$(\Delta - \lambda)\left\{(\Delta - \lambda)\, y_n\right\} = \lambda^2 b \tag{15}$$

and assume that

$$x_n = (\Delta - \lambda)\, y_n \tag{16}$$

$$(\Delta - \lambda)\, x_n = \lambda^2 b \tag{17}$$

By substituting Eq.13 into Eq.16,

$$x_n = c\exp(\lambda n) - \lambda b \tag{18}$$

and Equation 18 is a first-order equation.

Assuming that

$$c_m = c\exp(\lambda m) \tag{19}$$

at time $n = m$, the neighborhood x_{m+t} of x_m is represented by

$$x_{m+t} = c_m \exp(\lambda t) - \lambda b \tag{20}$$

Thus, if the measure

$$\begin{aligned} e(\lambda) &= \sum_{t=n_0}^{n_1} \left|(\Delta - \lambda) y_{m+t}^i - x_{m+t}\right|^2 \\ &= \sum_{t=n_0}^{n_1} \left|x_{m+t}^i - x_{m+t}\right|^2 \end{aligned}$$

can be used, non-linear optimization under only λ is needed and it does not require any knowledge of the onset position of the transition estimating the target b, because x_{m+t} is an exponential function. In this prediction, if $\lambda \geq 0$, it is the backward prediction (target in the past). If $\lambda < 0$, it is the forward prediction (target in the future). We use forward prediction (target in the future) to reproduce the power envelope.

In Fig.16, the blue line shows the estimated target of the power envelope using the target prediction model.

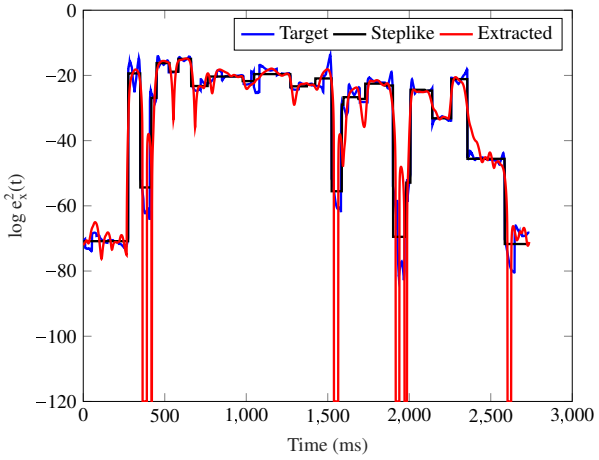

Figure 16: The original extracted power envelope, the estimated target of power envelope and the stepwise target of power envelope.

The onset point $T1_j$ and ending point $T2_j$ of each phoneme was segmented manually. After obtaining the estimated power envelope, we calculated the average value, Au_j of the jth step in each period of one phoneme which consisted of the stepwise function shown in Fig.16, black line. These are the inputs of the Eq.21 that follow the accent mechanism of the Fujisaki model. The stepwise input signals to the power control mechanism are defined by their amplitude Au_j, onset time $T1_j$ and offset time $T2_j$ using Eq.

$$\log e_y(t) = \sum_{j=1}^{J} Au_j[Gu(t - T1_j) - Gu(t - T2j)] \tag{21}$$

where $\log e_y(t)$ is the reproduced power envelope. And the step-response $Gu_{(t)}$ is calculated using the following equation

$$Gu_j(t) = 1 - (1 + \delta t)\exp(-\delta t) \quad t \geq 0 \tag{22}$$

The symbols in these equations forecast

- Au_j: amplitude of the jth step, Au_j is the average value of b in each segmentation,
- $T1_j$: onset of the jth step,
- $T2_j$: offset of the jth step,
- δ: time constant.
- δ is the absolute value of the sum of the negative parts of λ as we use a forward prediction, $\lambda < 0$ (target in the future), to reproduce the power envelope.

In Fig.17, the reproduced power envelope and extracted log power envelope are shown. Signal/Error Ratio (SER) in Eq.23 and Mean Absolute Error (MAE) in Eq.24 are used to evaluate the difference between the extracted and reproduced power envelope. As the voiced signal is more important than the unvoiced parts in this research, SER is calculated only during the voiced part.

$$SER = 10\log_{10}\frac{\sum_{i=1}^{N}(x_i)^2}{\sum_{i=1}^{N}(x_i - y_i)^2} \tag{23}$$

$$MAE = \frac{\sum_{i=1}^{N}|x_i - y_i|}{N} \tag{24}$$

where x_i is the extracted power envelope and y_i is the reproducing power envelope. N is the number of bits in the voiced part.

The value of SER is 18.01dB and the MAE is about 1.82dB which means that the reproduced power envelope is almost the same as the original extracted power envelope. Therefore, we can conclude that this method works well for parameterizing the power envelope. After this, we modified the power envelope by controlling A_{aj} to fit the estimated acoustic features.

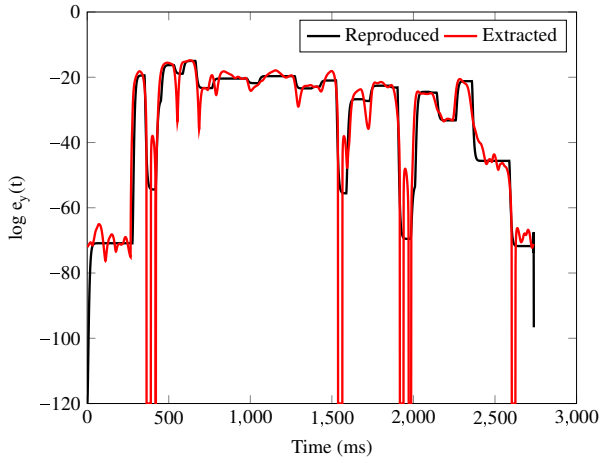

Figure 17: Reproducing power envelope using 2nd-order critically damped model and the extracted power envelope from original speech.

5.3. Perceptual evaluation

The voice conversion system for emotional speech aims to control the degree of emotion in dimensional space. We hypothesize that the system can convert any utterance from any speaker by a given point in dimensional space using a limited database. This is the procedure for the synthesized utterances for the evaluation phase, so there is no reference to the desired

position of the corresponding emotional utterance in the corpus. Hence, the objective measures such as Mel-cepstral distortion or mean squared error between converted and target are not suitable. The inputs of the conversion system are the intended position value in V-A space and the neutral speech. We utilized the distance between the intended position and the evaluated position obtained from the perception experiment to evaluate the category and the degree of emotion.

5.3.1. Stimuli

In the following subjective evaluation experiments, the inputs of the system for emotional voice conversion are three different neutral statements spoken by the single speaker from the Japanese Fujitsu database. The English meaning of the three statements are the following:

1. You have new mail.

2. Nothing new has come to mind.

3. I am already home.

The input positions to the system in V-A space are shown in Fig.18 with solid points. The range of valence and arousal is from -2 to 2 in increments of 0.1. The position values among the three utterances are the same. In the 1st and 3rd quadrants, there are 3 positions. Since there are two kinds of anger emotion, hot anger and cold, there are two positions for each in the 2nd quadrant. One position in the V-A space represents one synthesized utterance with different degrees of emotion. Including the neutral original speech, there are 11 stimuli for each utterance with a total number of 33 synthesized speech utterances.

5.3.2. Experiment procedure

16 Japanese subjects (7 females and 9 males) with normal hearing, average age about 23.3 years old, participated in the experiment. Subjects listened to the stimuli in a random order presented through an audio interface (FIREFACE UCX, Syntax Japan) and headphones (HDA200, SENNHEISER) in a soundproof room. The original sound pressure level was about 64 dB. The subjects evaluated the stimuli with regard to three aspects, valence, arousal and naturalness. Each aspect was evaluated as a separate test in order to avoid the conceptual confusion between valence and arousal, with at least a 3 hour time interval between tests. Subjects evaluated these scales using a graphic user interface as shown in Fig.6. The ranges, scale steps, and other rules are the same as explained in Section 4.

5.3.3. Subjective evaluation result in V-A space

Analysis of the evaluated results mainly focuses on two parts: perception of the emotion category and the degree of emotion. The evaluated results (perceived positions) analyzed in terms of emotion category in the valence and arousal spaces are shown in Fig.19. The oval is calculated using average and standard deviation of valence and arousal values. The central point of each oval is the mean value of each emotion. The radius of the oval shows the standard deviation related to valence and arousal of each emotion. Fig.19 shows that the mean value of evaluated joy, cold anger and sadness can be obtained in the intended quadrant and the standard deviation is acceptable for each emotion; this means that the category of emotion can be perceived well by subjects for joyful, cold anger, and sad emotional speech. But for hot anger, the intended position is the second quadrant while the evaluated line of hot anger is in the first quadrant, so subjects perceived synthesized hot anger as a joyful emotion. The reason for this misunderstanding is that, as we mentioned in Section 4, only by replacing the spectral sequence of hot anger can the neutral speech be perceived as hot anger. For now, our modification method only controls for duration, F0, and power envelope. Therefore, hot anger emotion cannot be well obtained.

The degree of emotion perception is shown in Fig.18. As the input positions of the three different linguistic utterances are the same, the average evaluated values for each position among the three utterances are calculated. In Fig.18, the solid circles represent the intended position and the hollow circles are the positions evaluated from the perceptual experiment in V-A space. The dashed lines show the distance of the two pairs: intended and evaluated. From Fig.18, we can see that the tendencies of the degrees of valence and arousal for the intended and evaluated emotions are the same, except for cold anger. Moreover, we note that the degree of the synthesized speech is more mild than intended. This phenomenon is in line with the results reported in Section 4. It is found that if only the F0 or spectral sequence of neutral speech is replaced by those from the emotional speech, the degree of perceived emotion is decreased.

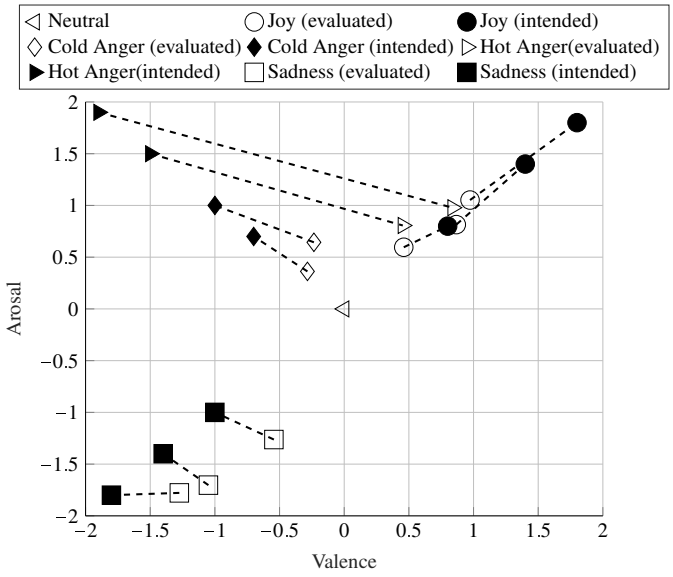

Figure 18: The evaluated and intended positions in V-A space. (the dashed lines are the intended position and the solid lines are the obtained position.)

5.3.4. Subjective evaluation result of naturalness

The naturalness quality of the converted utterances was rated on a 1-to-5 scale [1-bad, 2-poor, 3-fair, 4-good, 5-excellent] using the neutral sentence as the reference. The Mean Opinion Score (MOS) is shown in Fig.20. The MOS of each emotion is calculated separately. From these results, we see that

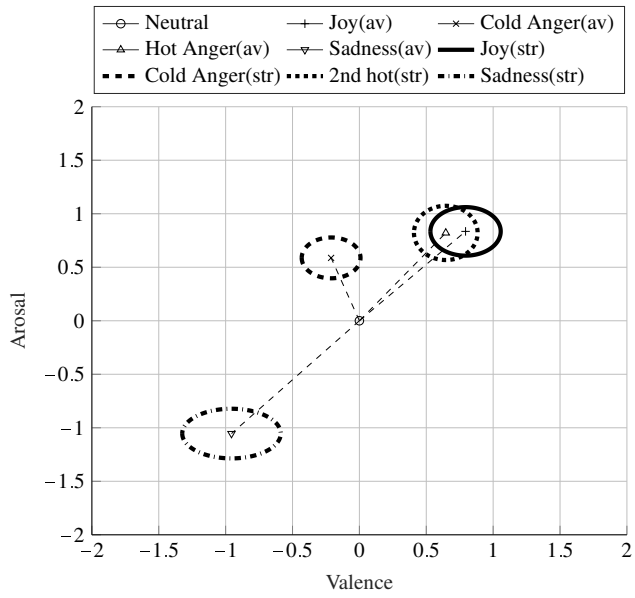

Figure 19: The average and standard deviation of the evaluated results. (*1st, 2nd, 3rd stand for the intended quadrants.* Cold *and* hot *represent the intended cold anger and hot anger.* av *and* str *mean the average and standard deviation values of each quadrant.)*

all naturalness scores are fair, i.e., above 2.5. Joyful speech was rated best (MOS about 3.38), with cold anger as a second (MOS about 3.1). The MOS of hot anger and sad are about 2.98 and 2.27. The reason that the quality of sadness is the lowest is because that the duration of sad speech is long but the pauses between phrases were not markedly obvious. We treated the ratio of modification to voice and unvoiced part the same. Therefore, the synthesized speech seemed machine-like. More precise control of duration ratios between voiced and unvoiced periods is needed in order to improve the quality of sadness; this is a topic that will be researched in the future.

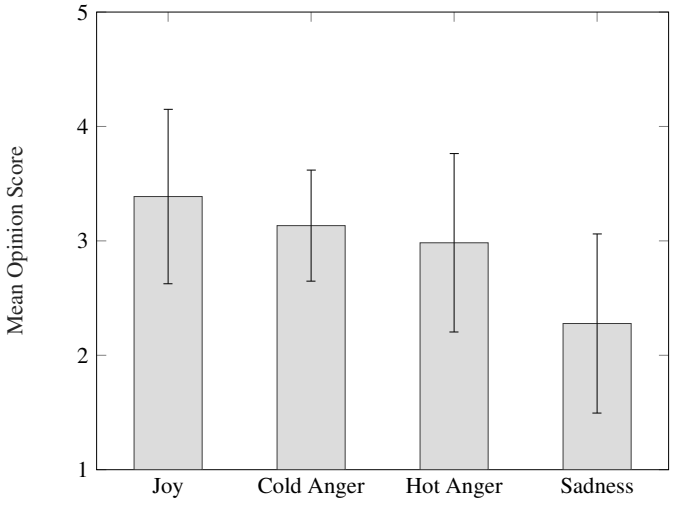

Figure 20: Mean opinion scores for converted speech in each quadrant.

6. Discussion and conclusion

A voice conversion system for emotional speech which utilized dimensional space to represent emotion in order to control the degree of emotion is proposed in this paper. Two dimensions, valence(from positive to negative) and arousal (from excited to calm), are considered to represent emotion. Following Brunswik's functional lens model which assumes that the perception of emotion by humans is multi-layered, the inverse three-layered model is proposed as the structure between emotion dimensions and acoustics. The significant acoustic features related to each dimension are explored by synthesizing speech, certain acoustic features of which are from emotional speech and others, from neutral speech. Perceptual evaluations in V-A space show that F0 and spectral information are the most important factors related to arousal and valence. By replacing the F0 and spectral information of neutral speech from joyful, cold anger and sad emotional speech, the synthesized speech can be perceived as having the same original emotional category, although the degree is decreased by replacing either of them. But by replacing only the F0 of the neutral utterance to the F0 from the hot anger utterance, the synthesized utterance is perceived as a joyful emotion. If only spectral information is replaced by that from the hot anger utterance, the synthesized voice can be perceived as hot anger while the degrees in both valence and arousal dimensions are decreased. These results support the previous studies that voice quality and F0 contribute much to emotions [53] [54]. However, these findings are based on one female voice actress database. Yet, speakers encode their affective states using various acoustic features. In future work, the database will be extend ed to multiple speakers in order to explore speaker individuality for affectiveness.

The voice conversion system has two parts: rule extraction and rule application. ANFIS, which embraces the concept of human perception of emotion as fuzzy logic, connects the three layers as a non-linear mapping. The low mean absolute error between the estimated value from ANFIS and the reference shows that ANFIS and the inverse three-layered model has the ability to build the non-linear relationship between acoustics and the emotion dimensions. The rules of acoustic features for modifying the neutral speech are extracted using the estimated acoustic features from ANFIS and the extracted acoustic features from neutral speech. In order to convert the neutral speech to the desired emotional speech in dimensional space, the Fujisaki model and target prediction model for parameterizing F0 and power envelope separately are conducted. STRAIGHT is used as the analysis-synthesis tool in this system.

Perceptual evaluation results in V-A space show that the synthesized speech of joyful, sad and cold anger emotion can be perceived well, including the category and the degree, although the perceived degree is decreased compared to the desired values. For hot anger emotion, since spectral modification was not conducted, the synthesized speech of hot anger is perceived as a joyful emotion. In the future, the method for controlling spectral sequences will be researched in order to convert neutral speech to any kind of emotion. Also, previous research has already revealed the commonality and difference in cross-cultural

emotion perception [55] [56] [57]. Since this system does not have a restriction on linguistic information, this will be a good approach for exploring the applications to multiple languages and multiple speakers in the future.

7. Acknowledgments

This study was supported by a Grant-in-Aid for Scientific Research (A) (No.25240026) and JSPS KAKENHI Grant. We sincerely thank Donna Erickson for her valuable comments.

References

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

[2] J. Tao and T. Tan, "Affective computing: A review," *International Conference on Affective computing and intelligent interaction. Springer Berlin Heidelberg,* pp. 981-995, 2005.

[3] R.W. Picard, "Affective computing," MIT Press, Cambridge, 1997.

[4] D. Erickson, "Expressive speech: Production, perception and application to speech synthesis," *Acoustical Science and Technology,* vol. 26, no.4, pp. 317-325, 2005.

[5] A. Rilliard, T. Shochi, JC. Martin, D. Erickson and V. Aubergé, "Multimodal indices to Japanese and French prosodically expressed social affects," *Language and speech,* vol. 52, no. 2-3, pp. 223-243, 2009.

[6] T. Shochi, A. Rilliard, V. Aubergé and D. Erickson, "Intercultural Perception of English, French and Japanese Social Affective Prosody," *The role of prosody in Affective Speech*, pp. 31, 2009.

[7] M. Schröder, "Expressive speech synthesis: Past, present, and possible futures," *Affective information processing, Springer*, London, pp. 11-126, 2009.

[8] T. Toda, L. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu and J. Yamagishi, "The voice conversion challenge 2016," in *Proc. Interspeech2016*, pp. 1632-1636, 2016.

[9] T. Toda, A. Black and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, no. 8, pp. 2222-2235, 2007.

[10] S. Takamichi, T. Toda, A. Black, G. Neubig, S. Salto. and S. nakamura, "Post-filter to modify the modulation spectrum for statistical parametric speech synthesis," *Audio, Speech and Language Processing, IEEE/ACM Transactions,* vol. 18, no. 5, pp. 1006-1010, 2015.

[11] D. Erro, A. Moreno and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 18, no. 5, pp. 922-931, 2010.

[12] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari and K. Shikano, "Gmm-based voice conversion applied to emotional speech synthesis," *IEEE Trans Speech Audio Proc*, vol. 7, pp. 2401-2404, 2003.

[13] J. Tao, Y. Kang and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 14, no.4, pp. 1145-1154, 2006.

[14] Z. Inanoglu and S. Young, "Data-driven emotion conversion in spoken English," *Speech Communication,* vol. 51, no. 3, pp. 268-283, 2009.

[15] R. Aihara, R. Takashima, T. Takiguchi and Y. Ariki, "GMM-based emotional voice conversion using spectrum and prosody features," *American Journal of Signal Processing,* vol.2, no.5, pp. 134-138, 2012.

[16] J. Yadav and K. Rao, "Prosodic mapping using neural networks for emotion conversion in Hindi language," *Circuits, Systems, and Signal Processing,* vol. 35, no.1, pp. 139-162, 2016.

[17] Z. Luo, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using deep neural networks with MCC and F0 features," *Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on. IEEE,* pp.1-5, 2016.

[18] D. Erro, E. Navas and I. Hernez, "Emotion conversion based on prosodic unit selection," *IEEE Transactions on Audio, Speech, and Language Processing,* vol.18, no.5, pp. 974-983, 2010.

[19] A. Iida, N. Campbell, F. Higuchi and M. Yasumura, "A corpus-based speech synthesis system with emotion," *Speech Communication,* vol. 40,no.1, pp. 161-187, 2003.

[20] M. Schröder, "Expressing degree of activation in synthetic speech," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 14, no. 4, pp. 1128-1136, 2006.

[21] J. Dang, et al."Comparison of emotion perception among different cultures," *Acoustical Science and Technology,* vol,31, no. 6, pp. 394-402, 2010.

[22] R. Cowie and R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech communication,* vol. 40, no.1 pp. 5-32, 2003.

[23] B.Fehr, J. Russell, "Concept of emotion viewed from a prototype perspective," *Journal of experimental psychology: General,* vol. 113, no.3, pp. 464-486, 1984.

[24] K. Scherer, P. Ekman, "On the nature and function of emotion: A component process approach," *Approaches to emotion,* vol. 2293, pp. 317, 1984.

[25] RE. Plutchik, HR. Conte, "Circumplex models of personality and emotions," American Psychological Association, 1997.

[26] R. Banse, KR. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of personality and social psychology,* vol.70, no.3, pp. 614, 1996.

[27] H. Schlossberg, "Three dimensions of emotion," *Psychological review,* vol.61, no.2, pp. 81-88, 1954.

[28] M. Grimm and K. Kroschel, "Emotion estimation in speech using a 3d emotion space concept," *Robust Speech Recognition and Understanding. InTech,* 2007.

[29] K. Scherer, "Vocal communication of emotion: A review of research paradigms, " *Speech communication,* vol. 40, no.1, pp. 227-256, 2003.

[30] K. Scherer and R. Klaus, "Methods of research on vocal communication: Paradigms and parameters," *Handbook of methods in nonverbal behavior research,* pp. 136-198, 1982.

[31] A. Kappas, U. Hess and K. Scherer, "Voice and emotion," *Fundamentals of nonverbal behavior,* vol. 200, 1991.

[32] T. Bänziger, G. Hosoya and K. Scherer,"Path Models of Vocal Emotion Communication," *PLoS ONE,* vol. 10, no. 9: e0136675. pone.0136675.

[33] E. Brunswik, "Historical and thematic relations of psychology to other sciences," *Scientific Monthly,* vol. 83, pp: 151161, 1956.

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

[35] R. Elbarougy and M. Akagi, "Improving speech emotion dimensions estimation using a three-layer model of human perception," *Acoustical science and technology,* vol.35, no.2 pp. 86-98, 2014.

[36] X. Li and M. Akagi, "Multilingual Speech Emotion Recognition System based on a Three-layer Model," *Prof. Interspeech2016*, pp. 3608-3612, 2016.

[37] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of the Acoustical Society of Japan (E),* vol. 5, no. 4, pp. 233-242, 1984.

[38] M. O'Reilly and A. Chasaide, "Analysis of intonation contours in portrayed emotions using the Fujisaki model," *The Second International Conference on Affective Computing and Intelligent Interaction. Proceedings of the Doctoral Consortium,* 2007.

[39] M. Akagi and Y. Tohkura, "Spectrum target prediction model and its application to speech recognition," *Computer Speech & Language,* vol. 4, no. 4, pp. 325-344, 1990.

[40] H. Kawahara, I. Masuda-Katsuse and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sound," *Speech communication,* vol. 27, no. 3, pp. 187-207, 1999.

[41] C.M. Whissell, "A dictionary of affect in language," *Perceptual and Motor Skills,* vol. 62, no.1 pp. 127-132, 1986.

[42] K. Scherer, "Emotion as a multicomponent process: A model and some cross-cultural data," *Review of Personality & Social Psychology (1984).*

[43] R. Barra-Chicote, J. Yamagishi, S. King and JM. Montero, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," *Speech Communication,* vol. 52, no.5, pp. 394-404, 2010.

[44] N. Takashi, J. Yamagishi and T. Masuko, "A style control technique for HMM-based expressive speech synthesis," *IEICE TRANSACTIONS on Information and Systems,* vol.90, no. 9, pp. 1406-1413, 2007.

[45] F. Biassoni, S. Balzarotti and M. Giamporcaro, "Hot or cold anger? Verbal and vocal expression of anger while driving in a simulated anger-provoking scenario," *SAGE Open,* vol.6, no.3, 2016.

[46] P. Juslin, "Cue utilization in communication of emotion in music performance: Relating performance to perception," *Journal of Experimental Psychology: Human perception and performance,* vol. 26, no.6, pp. 1797, 2000.

[47] M. Belyk, S. Brown, "The acoustic correlates of valence depend on emotion family," *Journal of Voice,* vol.28, no.4, pp.523.e9-523.e18, 2014

[48] M. Schröder, R. Cowie and E. Douglas-Cowie, "Acoustic correlates of emotion dimensions in view of speech synthesis," *Seventh European Conference on Speech Communication and Technology. 2001.*

[49] Y. Xue, Y. Hamada, and M. Akagi, "Emotional speech synthesis system based on a three-layered model using a dimensional approach," *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific,* IEEE, pp.505-514, Hongkong, 2015.

[50] Jang, J-SR, "ANFIS: adaptive-network-based fuzzy inference system," *IEEE transactions on systems, man, and cybernetics,* vol. 23, no.3, pp. 665-685, 1993.

[51] H. Mixdorff, "A novel approach to the fully automatic extraction of Fujisaki model parameters," *Proc. ICASSP*, Istanbul, Turkey, pp. 1281-1284, 2000.

[52] Y. Xue, Y. Hamada, and M. Akagi, "Voice conversion to emotional speech based on three-layered model in dimensional approach and parameterization of dynamic features in prosody," *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific,* IEEE, pp.1-6, 2016.

[53] I. Grichkovtsova, M. Morel and A. Lacheret, "The role of voice quality and prosodic contour in affective speech perception," *Speech Communication,* vol. 54, no. 3, pp. 414-429, 2012.

[54] C. Gobl and A. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication,* vol. 40, no. 1, pp. 189-212, 2003.

[55] X. Han, E. Reda and M. Akage, "A study on perception of emotional states in multiple languages on Valence-Activation approach," *2015 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP 2015)*, pp. 86-89, 2015.

[56] N. Lim, "Cultural differences in emotion: differences in emotional arousal level between the East and the West," *Integrate Medicine Research,* vol. 5, no. 2, pp. 105-109 2016.

[57] A. Chen, C. Gussenhoven, and T. Rietveld, "Language-specificity in the perception of paralinguistic intonational meaning," *Language and Speech,* vol. 47, no.4, pp. 311-349, 2004.