Voice Conversion to Emotional Speech based on Three-layered Model in Dimensional Approach and Parameterization of Dynamic Features in Prosody

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 a system to convert neutral speech to emotional with controlled intensity of emotions. Most of previous researches considering synthesis of emotional voices used statistical or concatenative methods that can synthesize emotions in categorical emotional states such as joy, angry, sad, etc. While humans sometimes enhance or relieve emotional states and intensity during daily life, synthesized emotional speech in categories is not enough to describe these phenomena precisely. A dimensional approach which can represent emotion as a point in a dimensional space can express emotions with continuous intensity. Employing the dimensional approach to describe emotion, we conduct a three-layered model to estimate displacement of the acoustic features of the target emotional speech from that of source (neutral) speech and propose a rule-based conversion method to modify acoustic features of source (neutral) speech to synthesize the target emotional speech. To convert the source speech freely and easily, we introduce two methods to parameterize dynamic features in prosody, that is, Fujisaki model for f0 contour and target prediction model for power envelope. Evaluation results show that subjects can perceive intended emotion with satisfactory order of emotional intensity and naturalness. This fact means that this system not only has the ability to synthesize emotional speech in category but also can control the order of emotional intensity in dimensional space even in the same emotion category.**

I. Introduction

Speech processing is widely studied in the area of human-computer interaction (HCI). Some popular applications such as Speech to Speech Translation system (S2ST) and story teller system have already bring convenience to human daily life. Convensional S2ST system and story teller system take linguistic information into consideration only, which lose the appealing of non-linguistic information such as emotion, and para-linguistic information such as emphasis. This fact brings an interesting research topic: emotional speech synthesis (ESS).

ESS [1] has greatly facilitated the advancement of HCI [2]. Nowadays, there are many techniques to synthesize emotional speech, which can mainly be divided into, those that take the concatenative approach[3] [4] with huge-scale training corpora, such as unit selection, and those that take the statistical approach, such as the hidden Markov model (HMM) [5] [6] and Gaussian mixture model (GMM) [7]. These techniques can synthesize emotional speech with an acceptable quality when emotion is represented in clear categories such as anger, happiness and sadness. In actual daily life, however, humans sometimes relieve or reinforce a variety of emotional expressions depending on the situation [8]. In such cases, a small number of discrete categories cannot sufficiently mimic the actual emotions expressed in real life. When emotional intensity is considered as continuous, both concatenative and statistical approaches require a huge database for training in spite of the difficulty of collecting human response to hear emotional speech. Therefore, our objective with this work is to convert neutral speech into a desired emotional speech with controlled emotional intensity.

In order to represent emotion on a continuously-valued scale rather than categorically, researchers have already proposed a two-dimensional emotion space, spanned by two attributes, valance and activation, to represent emotion as a point in a valence and activation (V-A) space [9] [10]. To obtain synthesized emotional speech from the position on the V-A space, two steps are required: *estimation step*, which is used to estimate displacement of acoustic features from that of the source (neutral) speech according to the position in the V-A space and *modification step*, which is used to convert the source (neutral) speech with the estimated displacement of the acoustic features to any intended emotional speech.

An emotion conversion system has already been proposed by the authors in [11] [12]. Huang and Akagi [14] proposed a three-layered model proposed, with reference to the concept of the Branswikian lens model [13], which is based on the belief that humans perceive emotion not directly from acoustic features, some semantic primitives, such as fast, bright, and so on also play an important role [14]. Following the hypothesis that human production of emotion is completely opposite of human perception, the opposite three-layered model as Huang and Akagi [14] is contucted to *estimation part*. The input of the model is the position in V-A space and the outputs are the estimated displacement of acoustic features that can give the impression represented by the position in V-A space. The accuracy of the estimated displacements of the acoustic features was suitable to use for converting neutral speech. However, F0 and power envelope cannot be controlled continuously in [12] and the modified speech gives only limited impressions, although a *modification step* was implemented for converting acoustic features themselves directly into the target speech.

In this paper, we study methods for controlling the two main dynamic features in prosody. The target prediction model [15] [16] for power envelope and the Fujisaki model [17] for F0 contour are conducted to parameterize the acoustic features of neutral speech. From the relationship between the extracted acoustic values of source (neutral) speech and estimated displacement of the acoustic features in the first step, we convert the parameter values to reproduce the power envelope and F0 contour in order to synthesize emotional speech. The key -point of the method is that both contours are parameterized. Evaluation results from several listening tests show that subjects can perceive emotion with satisfactory order of emotion intensity and naturalness, which means that this system not only can synthesize emotion in category but also has the ability to control the emotion intensity in massive scale.

II. Scheme of emotion conversion system

The emotion conversion system that converts neutral speech into emotional ones represents emotion in the V-A space. It requires two steps: estimation and modification. As shown in Fig.1, in the estimation step, the inputs are the expected V-A values and the outputs are estimated displacements of the acoustic features from the source (neutral) speech. The estimation step is structured using the three-layered model, which consists of the acoustic features at the top layer, semantic primitives at the middle layer, and V-A space at the bottom layer [14]. An adaptive -network-based fuzzy inference system (ANFIS) [18] based on fuzzy logic is utilized as the connection among the three layers. From an emotion corpus, the two evaluated dimension values, and the 17 evaluated semantic primitives, are collected via listening tests and the 21 acoustic features are extracted to use for training ANFIS1 and ANFIS2. ANFIS1 estimates the values of semantic primitives from the position in V-A space and ANFIS2 estimates the displacement of the acoustic features with the estimated values of semantic primitives. In the modification step, acoustic features are extracted from the source (neutral) speech using STRAIGHT [19] and parameter values are obtained using the proposed parameterization methods. Considering the relationship between the estimated displacements of the acoustic features in the first step and the extracted acoustic features of neutral speech in the second step, the parameterized acoustic features are modified. Applying STRAIGHT, we can resynthesize emotional speech using the modified acoustic features.

III. Estimation of displacement of acoustic features

In order to obtain the estimated displacements of the acoustic features from the given positions in the V-A space, we need to conduct an estimation procedure. The three-layered model is the structure and ANFIS is the connection of the estimation step.

A. The elements of the system

179 utterances from the Fujitsu database are used for training the system. The Fujitsu database contains five different emotional states: neutral, happy, sad, hot anger, and cold anger that are uttered by one professional female speaker.

The three-layered model has a structure of the estimation procedure. The concept of the three-layered model [14] follows the human perception mechanism, which is based on the belief from Brunswik's Lenz Model [13] that humans perceive emotion not directly from academic acoustics such as F0 and power envelope but from a series of adjective words. The three-layered model is constructed within this emotional speech conversion system. At the top of the three-layered model, we extracted 21 acoustic features using STRAIGHT from the neutral speech in [12], which contains acoustic features related to 4 F0, 4 power envelope, 5 spectrum, 3 duration and 5 voice quality. We selected 17 semantic primitives (bright, dark, high, low, strong, weak, calm, unstable, well-modulated, monotonous, heavy, clear, noisy, quiet, sharp, fast, and slow) in [14] because these semantic primitives can express emotion in a balanced way. The V-A space is located at the bottom layer, which consists of activation (from calm to excited) and valence (from negative to positive). The values of the V-A space and semantic primitives are obtained by carrying out listening tests.

B. Estimation procedure

Following the work in [12], ANFIS is used to connect the three-layered model. As shown in Fig. 1, two kinds of ANFISs are built. ANFIS1 is trained when given the evaluated valence and activation values as inputs and evaluated semantic primitives as outputs. The inputs of ANFIS2 are the estimated semantic primitives and the outputs are the displacements of the extracted acoustic features from that of source (neutral) speech. During the estimation procedure, when given the expected values of valence and activation, ANFIS1 gives the estimated values of the semantic primitives and ANFIS2 acquire estimated displacement of acoustic features when given the values of estimated semantic primitives from ANFIS1 as inputs.

IV. Modifying acoustic features

The estimated displacement of the acoustic features are then used to synthesize emotional speech. In [12], the acoustic features related to F0 and power envelope cannot be controlled continuously so that the modified speech gives only limited impressions. In this study, the target prediction model and the Fujisaki model are applied for controlling the power envelope and F0.

A. Target prediction model for power envelope

A target prediction model [15] [16] can predict the stable power target in each short-term interval. When the power envelope is approximated by a 2nd-order critically damped system, the model can estimate the target power envelope using short-term power sequences without being given the onset positions of the power transition. Inputting the original power envelope (the green line in Fig. 5) to the target prediction model, the estimated target of power envelope (the

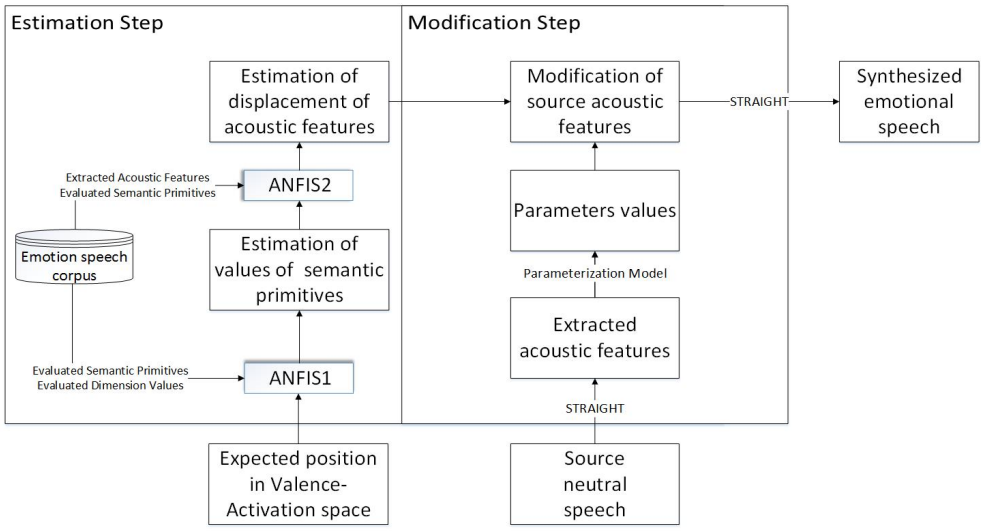

Fig. 1. *Scheme of emotion conversion system.*

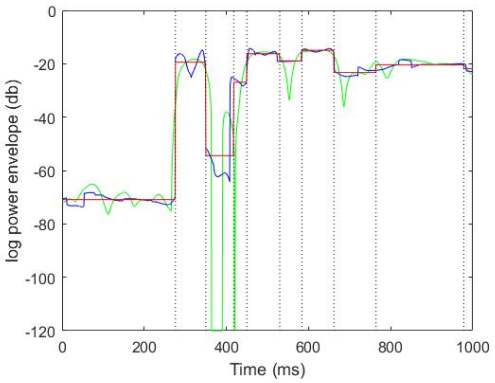

Fig. 2. *The input (green line) and output (blue line) of target prediction model. The red line is the stepwise function for which output of target prediction (blue line) is partially averaged with in the black dashed time points for every phoneme*

blue line in Fig. 5) can be obtained as output. The onset point $T1_i$ and ending point $T2_i$ of the ith phoneme from the original power envelope are segmented manually as shown in Fig. 5 with the black dashed line. From the estimated target of power envelope, we calculate the average amplitude Aq_i of the ith phoneme. It means that we change the estimated power envelope to the stepwise function (the red line in Fig. 5) using $T1_i$, $T2_i$ and Aq_i. They are inputs of Eq. 1 that using 2nd-order critically damped system to reproduce power envelope. By controlling $T1_i$ and $T2_i$, time duration of power envelope can be controlled. Using Aq_i, we can control the magnitude of the power envelope. And $G_b(t)$ represents the step response function.

$$e_y^2(t) = \sum_{i=1}^{I} Aq_i[G_b(t - T1_i) - G_b(t - T2_i)]. \quad (1)$$

B. Fujisaki model for F0 contour

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

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

$$Gp_i(t) = \begin{cases} \alpha_i^2 t \exp(-\alpha_i t), & t \geq 0 \\ 0, & t < 0, \end{cases} \quad (3)$$

$$Ga_j(t) \begin{cases} \min[1 - (1 + \beta_j t)\exp(-\beta_j t), \gamma], & t \geq 0 \\ 0, & t < 0. \end{cases} \quad (4)$$

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,
Aa_j: 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.

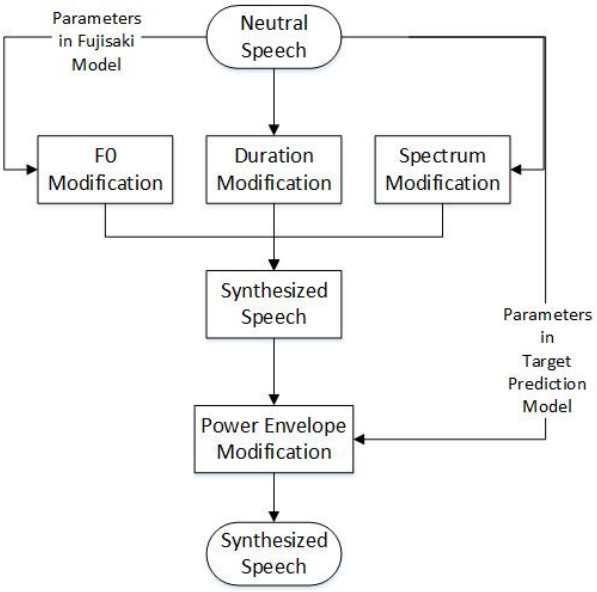

Fig. 3. *Procedure of the modification step.*

Many researchers have utilized the Fujisaki model, and the work of Mixdorff [20] is adopted in this paper, where $\alpha = 1.0/s$ and $\beta = 20/s$.

Through controlling five parameters, F_b, A_{pi}, A_{aj}, T_{0i}, T_{1j}, T_{2j}, we can reproduce the F0 contour into the desired shape.

C. Modification procedure

Extracted acoustic features including the F0 and the power envelope from the source (neutral) speech were modified according to the estimated displacement of the 21 acoustic features in step 1.

The modification procedure is shown in Fig. 3. The modification procedure can be divided into three steps. First, the duration information of every phoneme is segmented manually. We modify the duration related acoustic features. In this step, we modify $T1_i$ and $T2_i$ for power envelope modification in the target prediction model and modify T_{0i}, T_{1j} and T_{2j} for the F0 contour modification in the Fujisaki model. Second, we obtain the modified F_b, A_{pi}, A_{aj} using the estimated displacements of the acoustic features from that of the neutral speech. We apply the Fujisaki model to obtain the modified F0 contour. Then we modify the spectral tilt by formant shift. STRAIGHT is used to synthesize speech utilizing the modified duration, F0 contour and spectral sequence. Third, by modifying the amplitude Aq_i in each phoneme, we acquire the modified power envelope using a 2nd-order critically damped system. The modified power envelope is applied to the speech synthesized in the second step. Then, the final synthesized speech can be obtained. In Fig. 4, the original power envelope (red line) and modified power envelope (blue line) with the position (VA:-0.6,0.6), anger voice are shown. And the original (dashed) and modified (solid) F0 trajectory of happy voice are present in Fig. 4.

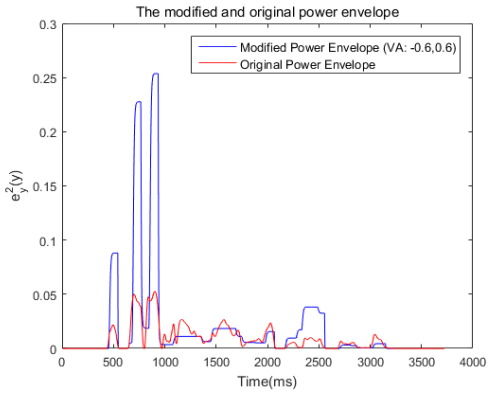

Fig. 4. *The original power envelope (red) and modified power envelope (blue)*

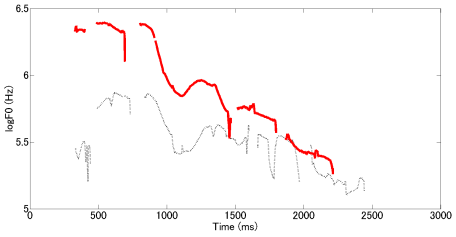

Fig. 5. *Original F0 trajectory of a neutral speech (dashed) and modified F0 of synthesized speech (solid).*

V. Evaluation

A. Listening Test

In order to verify whether the synthesized speech can be well perceived by humans, we carried out subjective listening tests in which subjects evaluated the synthesized speech in the V-A space.

1) Subjects and Stimuli: In the listening test, four Japanese subjects (four males, mean 26 years old) with normal hearing ability gave evaluations on three aspects: activation, valence and naturalness. On the basic of the listening test in [12], 76 stimuli were synthesized in valence and activation space represented by dashed line in Fig. 6. The 76 stimuli contains 25 voices of joy, anger and sad in the 1st, 2nd and 3rd quadrant in V-A space and one neutral voice.

2) Procedure: Subjects were asked to listen to the stimuli presented through an audio interface (FIREFACE UCX, Syntax Japan) and headphones (HDA200, SENNHEISER) in a soundproof room. The original sound pressure level was 64 dB.

For valence and activation, subjects listened to all stimuli twice. This was done so that they could acquire an impression of the whole stimulus the first time and then evaluate one dimension from -2 to 2. Valence and activation needed to be done separately in order to avoid conceptual confusion. Valence and activation were evaluated using 40 scales (Valence: Left [Very Negative], Right [Very Positive]; Activation: Left [Very Calm], Right [Very Excited]: range $-2 \sim 2$ in increments of 0.1). Subjects evaluated these scales using a

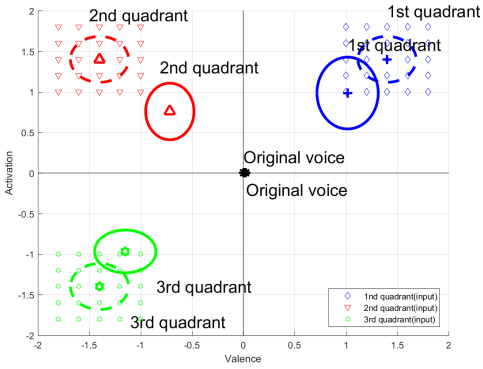

Fig. 6. *The intended with dashed line and obtained results from listening test with solid line in valence and activation space.*

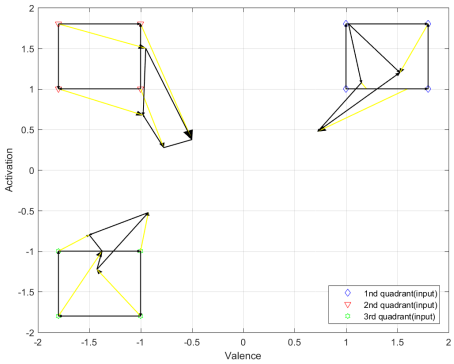

Fig. 7. *The intended intensity and the obtained intensity in valence and activation space (the rectangles are the intended intensity and the quadrilaterals are the obtained intensity).*

graphic user interface. During the listening test, subjects could listen to the stimulus as many times as they wanted.

For naturalness, all synthesized voices were presented once before subjects gave evaluations. The scale of evaluations was divided into five levels from bad to excellent ($1 \sim 5$). Subjects gave evaluations according to original speech spoken by a human whose naturalness is excellent.

B. Results

1) Emotion Perception: The evaluated positions in the valence and activation space are shown in Fig. 6 with solid line in each quadrant. Here, the dashed line are the inputs of the system and the solid lines are the evaluated results from subjects in the listening test (LT). The dashed lines are what we want, and the solid lines are what we actually obtained from the listening test. In Fig. 6, the oval is calculated using average and standard deviation in each quadrant. In Fig. 6, it shows that the category of emotion can be perceived by subjects especially for happy and sad emotion. In Fig. 7, 12 stimuli with either largest or smallest value of valence or activation are chosen to represent the intended and obtained intensity of emotion in each quadrant. The yellow line shows directions from the intended positions to obtained positions and the

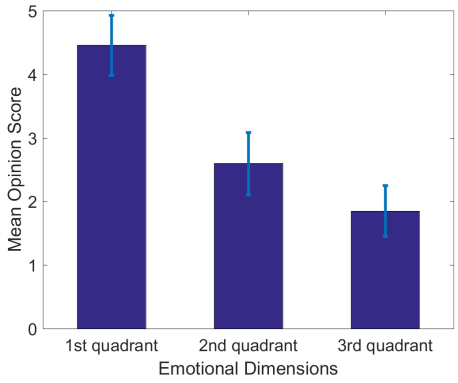

Fig. 8. *The mean opinion score for each quadrant.*

rectangles are the intended intensity and the quadrilaterals are the obtained intensity from the listening test in the V-A space. Among the 12 stimuli, 9 stimuli give the same tendency of emotion intensity as intended, the larger intended valence and activation, the larger perceived valence and activation. This fact confirms this system can not only synthesize emotional speech with intended category but also can convert neutral speech to emotional ones with the same tendency of emotion intensity. However, the value of valence from the rest of the 3 points are not perceived as intended order which influences the whole shape of each emotion intensity. Therefore, controlling valence still need more work in the future.

2) Naturalness: The evaluation of the naturalness of synthesized speech is shown in Fig. 8. The mean opinion score of each quadrant is calculated separately. From these results, we can see that all naturalness scores are above or near 2, which means not bad. The excellent synthesized speech in terms of naturalness was joy, with anger as second. According to the subjects, the reason why sadness was not good is because the duration of sad speech was long but the interval in each phrase was not obvious. Therefore, the synthesized speech seemed like machine-like. Therefore, more precise control of duration ratios between voiced and unvoiced periods is needed to be researched.

VI. Conclusions

In this paper, we proposed a method of modeling acoustic features for an emotional voice conversion system. The three-layered model and ANFIS are utilized as the structure and connection in the estimation procedure. The target prediction model and the Fujisaki model in the modification procedure can satisfactorily parameterize the power envelope and F0 contours continuously. Results of the listening test show that applying target prediction model and Fujisaki model, the conversion system for emotion can convert neutral speech to emotional ones not only with the same category, but also can controlling the order of emotion intensity in a large scale.

Acknowledgements

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

program.

REFERENCES

[1] M. Schrder. “Emotional speech synthesis: a review”. *Proc: INTERSPEECH* pp. 561-564, 2001.

[2] M. Akagi, X. Han, R. Elbarougy, Y. Hamada, and J. Li. “Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages,” *APSIPA 2014 - Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference, December 912 Siem Reap, Cambodia Proceedings*, pp.1–10, 2014.

[3] R. Barra-Chicote, J. Yamagishi, S. King, J. M. Montero, and J. Macias-Guarasa, “Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech”*Speech Communication*, vol. 52, no. 5, pp.394–404, 2010.

[4] A. Iida, N. Campbell, F. Higuchi, and M. Yasumura, “A corpus-based speech synthesis system with emotion,” *Speech Communication*, vol. 40, pp.161–187, 2003.

[5] J. Yamagishi, T. Nose, H. Zen, Z. H. Ling, T. Toda, K. Tokuda, and S, Renals. “Robust speaker-adaptive HMM-based text-to-speech synthesis,” *IEEE Transactions on, Audio, Speech, and Language Processing*, vol. 17, no. 6, pp.1208–1230, 2009.

[6] T. Nose, and T. Kobayashi. “An intuitive style control technique in HMM-based expressive speech synthesis using subjective style intensity and multiple-regression global variance model,” *Speech Communication*, vol. 55.2, pp.347–357, 2013.

[7] J. Jia, S. Zhang, F. Meng, Y. Wang, and L. Cai, “Emotional audio-visual speech synthesis based on PAD,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19(3), pp.570–582, 2011.

[8] I. Albrecht, M. Schrder, J. Haber, and H. P. Seidel, “Mixed feelings: expression of non-basic emotions in a muscle-based talking head”. *Virtual Reality*, vol. 8, no. 4, pp.201–212, 2005.

[9] M. Schrder, et al. “Acoustic correlates of emotion dimensions in view of speech synthesis”. *Proc INTERSPEECH*. 2001.

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

[11] Y. Hamada, R. Elbarougy, and M. Akagi, “A method for emotional speech synthesis based on the position of emotional state in Valence-Activation space”. *APSIPA 2014 - Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference, December 912 Siem Reap, Cambodia Proceedings*, pp.1–7, 2014.

[12] Y. Xue, Y. Hamada, and M. Akagi, “ Emotional speech synthesis system based on a three-layered model using a dimensional approach”. *APSIPA 2015 - Asia-Pacific Signal and Information Processing Association, 2015 Annual Summit and Conference, December 1518 HongKong, China Proceedings*, pp. 505-514, 2015.

[13] K. R. Scherer, “Personality Inference from Voice Quality: The Loud Voice of Extroversion”. *European Journal of Social Psychology*, no. 8, pp.467–487, 1978.

[14] C-F. Huang, and M. Akagi, “A three-layered model for expressive speech perception”. *Speech Communication* , no. 50, pp.810-828, 2008.

[15] M. Akagi, “ Evaluation of a spectrum target prediction model in speech perception”. *J. of Acoust. Society of America*, vol. 87(2), pp.858-865, 1990.

[16] M. Akagi, and Y. Tohkura, “Spectrum target prediction model and its application to speech recognition”. *Computer Speech and Language*, vol. 4, Academic Press, pp.325-344, 1990.

[17] H. Fujisaki, “Information, prosody, and modeling-with emphasis on tonal features of speech,” *Proc. Speech Prosody, Nara, Japan*, pp.1–10, 2004.

[18] J-SR. Jang, “ANFIS: adaptive-network-based fuzzy inference system”. *IEEE Transactions on Systems, Man and Cybernetics*, vol. 23, no. 3, pp.665–685, 1993.

[19] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne,(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*, vol. 27, no. 3, pp.187–207, 1999.

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