Voice conversion system to emotional speech in multiple languages based on three-layered model for dimensional space

Yawen Xue
Japan Advanced Institute of Science and Technology
Nomi City, Japan 923-1211
Email:xue_yawen@jaist.ac.jp

Ysuhiro Hamada
Meiji University
Tokyo, Japan
Email: y-hamada@jaist.ac.jp

Reda Elbarougy
Damietta University, Egypt
Japan Advanced Institute of Science and Technology
Email: elbarougy@jaist.ac.jp

Masato Akagi
Japan Advanced Institute of Science and Technology
Nomi City, Japan 923-1211
Email: akagi@jaist.ac.jp

***Abstract*—Commonalities and differences of human perception for perceiving emotions in speech among different languages in dimensional space have been investigated in previous work. Results show that human perception for different languages is identical in dimensional space. Directions from neutral voice to other emotional states are common among languages. According to this result, we assume that, given the same direction in dimensional space, we can convert the neutral voices in multiple languages to emotional ones with the same impression of emotion. It means that the emotion conversion system could work for other languages even if it is trained with a databases in one language. We try to convert neutral speech in two different languages, English and Chinese using an emotion conversion system trained with Japanese database. Chinese is a tone language, English is a stress language and Japanese is an accent language. We find that all converted voices can convey the same impression as Japanese voices. On the case, we can make a conclusion that given the same direction in dimensional space, the synthesized speech among multiple language can convey the same impression of emotion. In a word, the Japanese emotion conversion system is compatible to other languages.**

I. Introduction

With developments of technology, people in different countries have more and more aspiration to communicate with each others although they have different mother languages. However, it is impossible to have a conversation if a common language is not shared, that makes a challenge to design a worldwide communication environment. One possible solution to this challenge is to construct a speech-to-speech translation (S2ST) system that can convert a spoken utterance in one language to that of another language [1]. Information in speech mainly can be divided into three parts: linguistic information which represent the content of the utterance inferred from context, paralinguistic information which is added by the speaker to modify the linguistic information and nonlinguistic information which can not be controlled by the speaker, such as emotion, gender, etc [1] [2]. Conventional S2ST systems directly translate one source language to the target one of neutral utterance no matter what the emotion or emphasis conveyed by the source utterance. It means conventional S2ST systems only consider about linguistic information. While non-linguistic information such as, emotional state conveyed by the source language is crucial to be preserved and passed in daily life. On that case, a system that can deal with emotional speech in multi-language is of importance [3].

Many previous studies on emotional speech synthesis have been conducted which can be mainly divided into two parts: concatenative approach such as unit selection [4] [5] and statistical approach such as hidden Markov model (HMM) and Gaussian mixture model (GMM) [6] [7] [8]. Both of them can synthesize emotional speech well when emotion is represented in categories. While the intensity of a certain emotion varies with time and situation, which may take any arbitrary value such as little sad or much sad [9]. To describe rich variations in the degree of affective states, a dimensional space of valence and activation (V-A space) [10] [11] is generally used to represent emotion as a point in dimensional space. As for concatenative approach is based on the concept of concatenating short samples of recorded sound and HMM can only give the average inherent in the statistical approach, both of them need a huge database when represent emotion in a continuous scale. In order to represent emotion in a continuous scale, a rule-based emotion conversion method is conducted in this research. On this case, the synthesized speech can convey all degrees of emotions.

In [12], a rule-based emotional speech conversion system based on the three-layered model on dimensional approach [10] is proposed. Inputs of this system are Japanese neutral speech and a position on the dimensional space representing the emotion category and degree. Through the estimation and modification procedures in this conversion system, converted emotional voices in Japanese can be outputted.

While this system is trained with Japanese database, we attempt to know whether it can be applied to other languages. It is found that human perception of emotions in speech in different languages is identical in the dimensional space [13] [14]. Following this finding, we assume that, given the same direction in dimensional space, the system can convert the neutral voices in multiple languages to emotional ones with the same impression of emotion. It means that even if the

conversion system constructed for one language, it can also work for other languages. In order to explore the validity of this hypothesis, we utilize the emotional speech conversion system in [12] constructed for Japanese to convert other languages such as English and Chinese. Subjects were asked to evaluated the synthesized voices in all three languages in listening tests. As a result, even Chinese is a tone language and English is a stress language, we find that all converted emotional voices in Chinese and English can convey the same impression of emotion categories and similar emotion intensity as those in Japanese. The hypothesis is confirmed that even the emotional conversion system is built in one language, it can be applied for other languages. This means that the emotion conversion system is compatible to other languages.

II. Commonalities of human perception for emotional speech among multi-languages.

In the previous study [13], five emotional speech databases in five different languages, Japanese, German, English, Vietnamese and Chinese, were analyzed in valence-activation space (V-A space). Valence represents speech from positive to negative and activation takes values from high to low. Four emotional states, happy, angry, neutral, and sad, were selected from the five databases. Thirty subjects from three different countries, Japan, China and Vietnam, evaluated the three databases in terms of valence and activation. It is found that human perception for different languages is identical in the dimensional space i.e. the directions from neutral voice to other emotional states are common among languages. Based on this result, we hypothesize that, given the same direction in V-A space from neutral voice to other emotional states, the emotion conversion system can also convert other languages with the same impression of emotion. In order to confirm this hypothesis, we apply the emotion conversion system build for Japanese to two other languages, English and Chinese, without training using the two languages. In the following section, the outline of the emotion conversion system and the procedure for applying the emotion conversion system to other languages are illustrated.

III. Outline of conversion system to emotional speech

The conversion system proposed by the authors, in which neutral speech is converted into emotional ones, represents emotions 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 to identify the position in the V-A space and the outputs are estimated displacements of the acoustic features from the source (neutral) speech. The estimation step is structured using a 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 [15]. The concept of the three-layered model follows the human perception mechanism, which is based on the belief from Brunswiks Lenz Model [17] that humans perceive emotion not directly from academic acoustics such as F0 and power envelope but from a series of adjective words. An adaptive-network-based fuzzy inference system (ANFIS) [16] based on fuzzy logic is utilized to connect the three layers. From a Japanese emotion corpus, evaluated dimension values for the two dimensions and the 17 evaluated primitives are collected via listening tests and the 21 acoustic features are extracted. These values are used 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 from 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 [12]. Considering relationships 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.

A. Database

179 Japanese 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 actor.

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 [15] 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 step

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.

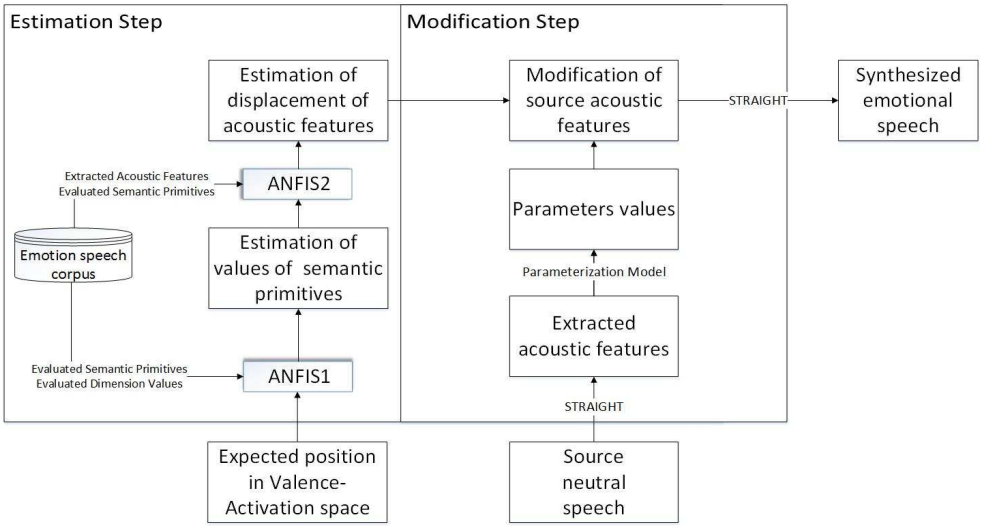


Fig. 1. *Scheme of emotion conversion system.*

C. Modification step

Extracted acoustic features including F0, power envelope, spectrum and duration from the source (neutral) speech were modified according to the estimated displacements of the 21 acoustic features in the estimation step.

The modification procedure is shown in Fig. 2. The modification procedure can be divided into three steps. First, the duration information of every phoneme extracted from a manually segmented voice. We modify the duration according to the related acoustic features. In this step, we modify the phoneme boundary for power envelope modification in the target prediction model [18] and modify time constants for the F0 contour modification in the Fujisaki model [2]. Second, we obtain the modified parameter values of Fujisaki model using the estimated displacements of the acoustic features from that of the neutral speech. Then, we apply the Fujisaki model to obtain the modified F0 contour. Next we modify the spectral tilt by shifting formants. STRAIGHT is used to synthesize speech utilizing the modified duration, F0 contour and spectral sequence. Third, by modifying the magnitude of power envelope 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. Lastly, the final synthesized speech can be obtained.

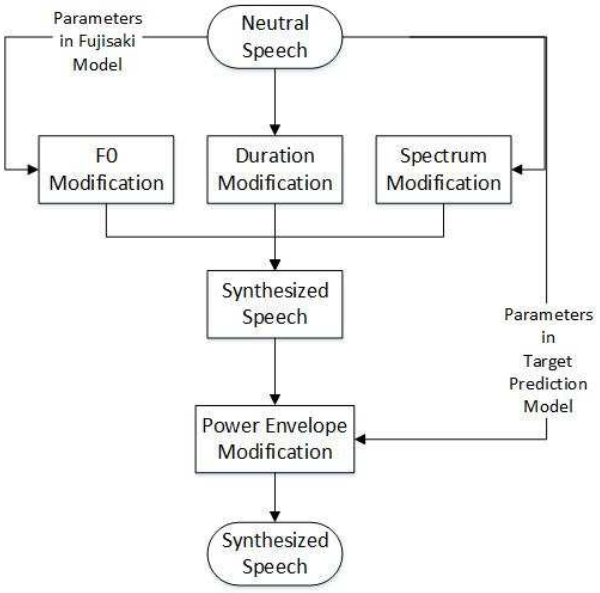


Fig. 2. *Procedure of the modification step.*

IV. Applying Conversion System to Other Languages

As we investigate whether, given the same direction from neutral speech to other emotional states, the converted speech can give the same impression among multiple languages, two different neutral speech as well as Japanese one are given as input to the emotion conversion system.

One utterance is spoken in Chinese by a professional female voice actor which is selected from the Chinese emotional corpus developed by Institute of Automation, Chinese Academy of Sciences (CASIA). The content is in Chinese means "He ends an meaningless love" in English. Another utterance is spoken in English by a female US English speaker from English CMU ARCTIC database. The content is "LORD BUT I'M GLAD TO SEE YOU AGAIN PHIL".

This emotion conversion system built for Japanese shown in Fig.1 is applied to Chinese and English sentences without changing parameter values in the estimation and modification steps. The inputted positions are the same among three languages, that we choose the 8 positions with either large or small value of valence or activation to represent the intended and intensity of emotion; Four in the 1st quadrant that represent joy and four in the 3rd quadrant which represent sad

emotion. The input position is shown as dashed line in Fig.3 and the hollow shape in Fig.4.

A. Listening Test

To verify whether the synthesized voices care well perceived by humans, we carried out subjective listening tests to let subjects evaluate the synthesized speech in the V-A space.

1) Subjects and Stimuli: In the listening test, 9 subjects (two Vietnamese female, one Vietnamese male, three Japanese male, one Chinese female and two Chinese male, mean 26 years old) with normal hearing ability gave evaluation scores on three aspects: activation, valence and naturalness. 27 stimuli are presented to the subjects. The 27 stimuli contains three languages and each language has 9 stimuli. Among the 9 stimuli, four voices of joy, four voices of sad and one voices of neutral. The neutral speech is the original speech given as inputs of the system and the 8 stimuli for joy and sad are prepared with either largest or smallest values of valence or activation.

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 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: In Fig.3, evaluated positions in the valence and activation space are shown with solid line in each quadrant among multiple languages. The oval is calculated using average and standard deviation in each quadrant. Here, the dashed line are the inputs of the system and the solid lines are the evaluated results by subjects in the listening test. The dashed lines are what we want, and the solid lines are what we actually obtained. The blue lines represent the joy voices and the green lines show the sad voices. As among three languages, blue lines are all in the fist quadrant and green lines are all in the third quadrant. It means that subjects can perceive emotions category well among three languages. In Fig. 4, 9 stimuli with either largest or smallest value of valence or activation are used to represent the intended and obtained intensity of emotion in each quadrant. The red dashed line shows directions from the intended positions to obtained positions and the rectangles are the intended intensity and the quadrilaterals are the obtained intensity from the listening test in the V-A space. From Fig. 4, we can see that the quadrilaterals in the first quadrant shows a small area but the tendency of emotion intensity is the same except one valence values of Japanese voices. And the quadrilaterals in the third quadrant shows a large area and the tendency of emotion intensity are similar as intended except for the synthesized voice whose intended position is VA(-1.6,-1.2) which is caused by the estimation part. From Fig.3 and 4, we can confirms that this system can convert neutral speech to emotional ones with the same intended category and similar intensity for multiple languages.

2) Naturalness: The evaluation result for the naturalness of synthesized speech is shown in Fig. 5. Mean opinion score (MOS) of each quadrant is calculated separately. From these results, we can see that all naturalness scores are above or near 2, that means not bad. The excellent synthesized speech in terms of naturalness was Chinese joy voices. MOS of English synthesized voices are all above 3, that means ordinary natural and naturalness of sad voices is low for Chinese and Japanese. The reason why sadness was not good is because the duration control of sad speech sometimes fails in some points so that the synthesized voices was long but the interval in each phrase was not obvious. Therefore, the synthesized speech seemed like machine-like. More precise control of duration ratios between voiced and unvoiced periods is needed to be researched.

V. Conclusion

This paper shows that, given the same direction from neutral speech to emotional states in V-A space, the conversion system for emotion can convert neutral speech to emotional ones among multiple languages. The conversion system for emotion trained in Japanese is used to convert a tone language, Chinese and a stress language, English. The results from the listening tests by subjects confirmed that the synthesized voices can convey the same category and similar intensity of emotion among different languages which means that the conversion system for emotion built for Japanese is compatible to other languages.

Acknowledgements

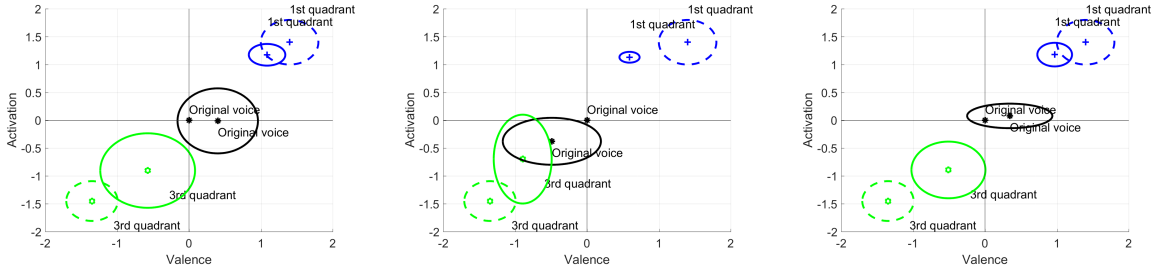
This study was supported by a Grant-in-Aid for Scientific Research (A) (No. 25240026) and the JSPS A3 Foresight program.

References

[1] M. 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.

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

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

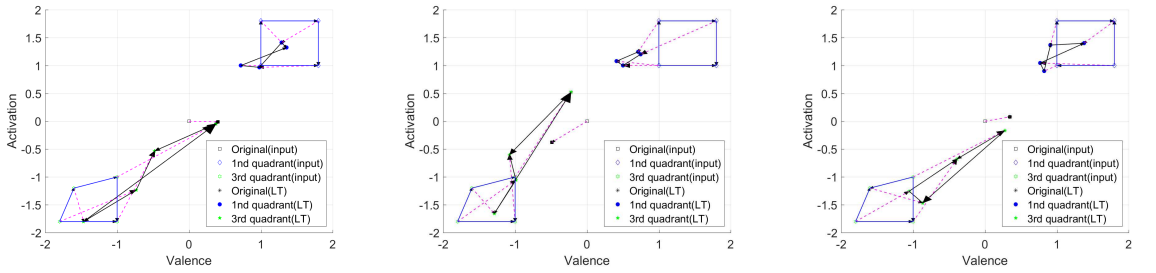


(a) Evaluation results of synthesized voices in Chinese.

(b) Evaluation results of synthesized voices in English.

(c) Evaluation results of synthesized voices in Japanese

Fig. 3. The evaluation results (emotion category) for multiple languages in V-A space.

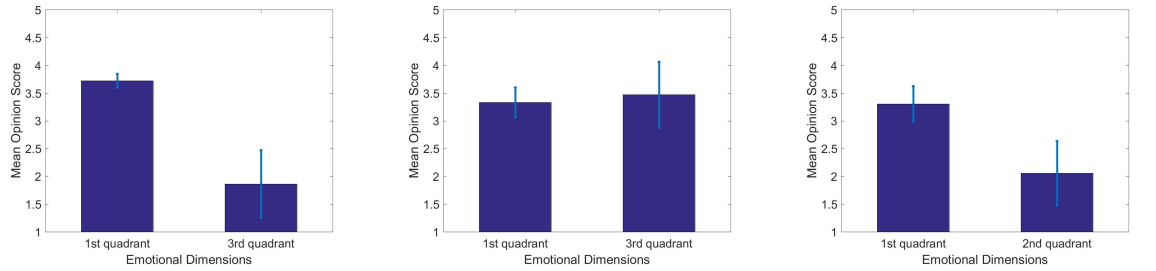


(a) Evaluation results of synthesized voices in Chinese.

(b) Evaluation results of synthesized voices in English.

(c) Evaluation results of synthesized voices in Japanese

Fig. 4. The evaluation results (emotioni intensity) for multiple languages in V-A space.



(a) Evaluation results of Chinese synthesized voices.

(b) Evaluation results of English synthesized voices.

(c) Evaluation results of Japanese synthesized voices..

Fig. 5. The mean opinion score for each quadrant.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

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

[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] X. Han, R. Elbarougy, M. Akagi, J. Li and T. Ngo, "A study on perception of emotional states in multiple languages on Valence-Activation approach ". *NCSP* 2015 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing, (NCSP'15), pp 86-89,2015.

[14] R. Elbarougy, X. Han, M. Akagi, and J. Li, "Toward relaying an affective speech-to-speech translator: Cross-language perception of emotional state represented by emotion dimensions," *O-COCOSDA* International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques, Proceedings, pp 48-53,2014.

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

[16] 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.

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

[18] 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.

[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.