Study on method to control fundamental frequency contour related to a position on Valence-Activation space

Yasuhiro Hamada[1]*, Reda Elbarougy[2], Yuawn Xue[3] and Masato Akagi[3]
[1]Meiji University, Tokyo, JAPAN
[2]Damietta University, Damietta, EGYPT
[3]Japan Advanced Institute of Science and Technology, Ishikawa, JAPAN

Paper Number: P12000176

1. Abstract

Speech-to-speech translation (S2ST) system is important for human-machine interface. In our previous study, we have proposed a speech conversion system from neutral to emotional ones by considering emotion space spanned by the Valence and Activation axes (V-A space). To build relationships between V-A space and acoustic features, Adapted Network Fuzzy Inference System (ANFIS) was applied. Neutral speech was converted to an emotional speech to control the values of acoustic features that were related to a position of V-A space. However, the proposed conversion system has some problems to control the acoustic features of the neutral speech.

In this paper we propose a new method to control fundamental frequency (F_0) contour. In order to control the F_0 contour, Fujisaki model was used. F_0 contour was modified by controlling the parameters of Fujisaki model. The results showed the F_0 was able to control using Fujisaki model.

***Index Terms*—Speech-to-speech translation, emotion, fundamental frequency contour, Fujisaki model, valence and activation space.**

2. Introduction

A speech-to-speech translation (S2ST) that responds to a speech signal with recognizing and synthesizing speech in different languages is important for human-machine interface.

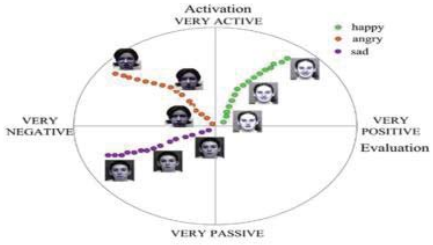

Figure 1. Emotion space spanned by Valence-Activation space [2].

Many studies constructed this kind of systems with linguistic information only. However, speech has not only linguistic but also para- and non-linguistic information [1]. In the previous studies, para-and non-linguistic information such as emotion, individuality and gender have been considered few. To establish human-human like communication, it is needed to consider para- and non-linguistic information. Especially, emotion has a lot of information for human-human communication. In our study, we consider emotion in the S2ST system (Affective S2ST [2]).

So far, emotion is supposed to be categorical such as happy, anger, sad [3] [4]. However, human beings are able to express various and complex emotions. Thus, to add emotion to the S2ST system, it is important to implement various and complex emotions. From psychophysical studies, Valence and Activation are two main axes to span emotion spaces. In the Valence-Activation space (V-A space), emotion can be continuously represented with degrees, such as quiet happy and very happy [5] (Figure 1).

ISBN: 978-981-09-7961-4 — doi: 10.3850/978-981-09-7961-4_P12000176

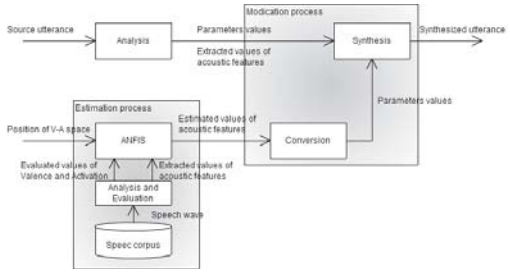

Figure 2. Proposed conversion process.

We have proposed an emotional conversion system on the V-A space [6]. The emotional conversion system was constructed with two processes. One is an estimation process and the other is a modification process of the acoustic features. In the estimation process of the values of the acoustic features, Adapted Network Fuzzy Inference System (ANFIS) [7] was applied to build relationships between V-A space and acoustic features. Using ANFIS, the values of acoustic features can be estimated from a position of V-A space. In the modification process, acoustic features were modified using the values of the estimated acoustic features. However since these estimated parameters were subordination each other, the values of modified acoustic features were far from the values of estimated acoustic features. Therefore, it is needed to convert estimated acoustic features to the other parameters.

In this paper, we propose a new method to control fundamental frequency (F_0) contour. In the modification process, the values of acoustic features are converted to the other incompatible parameters for controlling acoustic features. Since F_0 contour is important for emotions, we focus on fundamental frequency. Four acoustic features related to F_0 that is mean values, highest value, rising slope of the first accentual phrase and mean values of rising slopes of the other accentual phrase were estimated in the estimation process. To modify parameter values of the F_0 contour of the neutral speech to fit the estimated acoustic features, Fujisaki model [8] was adopted. Fujisaki model is a mathematical model that represented by the sum of phrase component, accentual component and base line. With controlling F_0, neutral speech was converted to an emotional speech related to a position on the V-A space.

Figure 2 shows the process of the proposed emotional conversion system. The estimation process was shown in section 3, and modification process was shown in section 4. In section 5, we evaluated this system with calculating the values of F_0 related acoustic features.

3. Estimation of values of acoustic features

To estimate the values of the acoustic features, Adapted Network Fuzzy Inference System (ANFIS) was applied to build relationships between V-A space and acoustic features.

3.1. Speech materials and acoustic features

Used speech materials and acoustic features were the same as that in our previous study [6].

The speech database was the multi-emotion single speaker Fujitu database produced and recorded by Fujitu laboratory. It contains five emotional states: neutral, joy, cold anger, sad and hot anger. Total number of utterances in this database was 179 utterances.

In the emotion recognition system proposed by Elbarougy *et al*. [9], twenty-one acoustic features were used. Following this system, we used the same 21 acoustic features. These 21 acoustic features were extracted using STRAIGHT [10]. Table 1 shows 21 acoustic features.

Table 1. Acoustic features.

F0:	Average F0, highest F0, rising slope of the F0 contour,
Power:	Power range, average power, rising slope of the power
Spectrum:	Formant frequency, spectral tilt, spectral balance
Duration:	Total length, consonant length ratio between consonant and vowel length
Voice quality:	The mean values of first harmonics minus second harmonics of vowel

3.2. Estimation process

In order to model the relationships between the values of acoustic features and the positions of the V-A space, ANFIS was used.

Valence and Activation were evaluated in listening tests while all 21 acoustic features were extracted from the Fujitu database. Using ANFIS, 21 acoustic features and positions in the V-A space were mapped non-linearly.

4. Parameter conversion

The 21 estimated parameters are subordination each other, the values of modified acoustic features were needed to convert to the other parameters. In this section, we focus on the F_0 contour and convert related parameters using Fujisaki model.

4.1. Fujisaki model

Fujisaki model is a mathematical model based on the mechanism of the speech production. This model is represented by the phrase components, accent components and base line. F_0 contour is expressed by the following equation:

$$\ln F_0 = \ln F_b + \sum_{i=1}^{I} A_{P_i} G_{P_i}(t - T_{0_i})$$

$$+ \sum_{j=1}^{J} A_{a_j} \{G_{a_j}(t - T_{1_j}) - G_{a_j}(t - T_{2_j})\}$$

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.

$$G_{p_i}(t) = \begin{cases} \alpha_i^2 t \exp(-\alpha_i t), t \geq 0 \\ 0, t < 0 \end{cases}$$

$$G_{a_j}(t) = \begin{cases} \min[1 - (1 + \beta_j t)\exp(-\beta_j t), \gamma], t \geq 0 \\ 0, t < 0 \end{cases}$$

Here, α is natural angular frequency of the phrase control mechanism and β is the that of accent control mechanism.

4.2. Parameter estimation

To estimate parameter values of Fujisaki model, we adapted a method proposed by Mixdorff [11]. Parameter α is set equal to 1.0/s and β is set equal to 20/s.

4.3. Duration control

The gradient of F_0 and that of envelope are related to the duration. The related acoustic features of the duration are total length, consonant length, and ratio between consonant length and vowel length. The related parameters of Fujisaki model are T_0, T_1 and T_2. According to the ratio between the extracted values of the acoustic features and the estimated values of the acoustic features, T_0, T_1 and T_2 were modified.

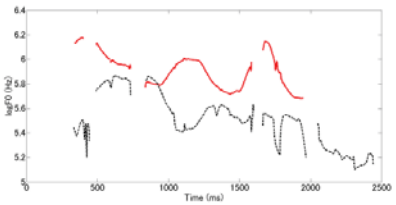

Figure 3. F_0 trajectory of a neutral speech (dashed) and synthesized speech (solid).

4.4. F_0 control

After the modification of the duration, F_0 related features were modified. F_0 related acoustic features are average of F_0 (F0_AP), highest F_0 (F0_HP), rising slope of the F_0 contour (F0_RS) and that of the first accentual phrase (F0_RS1). The extracted acoustic features related to the F_0 of the source voice were modified to be the estimated values as the following steps.

1. The parameter values of Fujisaki model of the source voice was estimated as mentioned 4.2.
2. Since F_b, A_p and A_a are related to the values of the estimated F_0 related acoustic features, F_b, A_p and A_a were modified.
3. By calculating the values of F_0 related acoustic features, the optimized values of F_b, A_p and A_a were estimated.

Using modified F_0 contour modeled Fujisaki model, target voice was resynthesized.

5. Experiments

In order to confirm that the modified F_0 contour expresses the intended F_0 contour, the values of estimated acoustic features that are one of the target voice and calculated values of modified acoustic features were compared.

5.1. Experimental conditions

The source voice was a neutral voice in Fujitu database. The sentence was /Atarashii meil ga todoite imasu/ (in English: You've got a new mail). The value of the Valence was 0 and Activation was 0 that was evaluated by subject shown in 3.1.

This voice was modified to be a voice at which Valence was 1 and Activation was 1.

The values of F_0 related acoustic features of source voice, target voice and modified voice were compared. Figure 3 shows the F0 contour of the source voice and modified voice.

5.2. Experimental results

The values of acoustic features of source voice, target voice and modified voice were shown in Table 2.

Table 2. Experimental results.

	Source voice	Target voice	Modified voice
F0_AP	2.39	2.55	2.55
F0_HP	2.55	2.70	2.71
F0_RS	0.6	1.3	1.5
F0_RS1st	0.8	0.4	0.5

These results showed the values of modified voice became closer to the target ones. From these results, F_0 contour is suggested to express by controlling parameters of Fujisaki model.

6. Conclusion

This paper proposed a method of controlling F_0 contour in order to represent intended F_0 contour to control F_0 related acoustic features independently.

F_0 contour is represented by the Fujisaki model and the parameters of Fujisaki model are modified.

To confirm whether the values of modified acoustic features are closer to the target ones, the values of three groups of acoustic features are compared. The results show that the modified F_0 contour is suggested to express emotions that have intended F_0 contour by controlling parameters of neutral voices with Fujisaki model.

Acknowledgements

This study was supported by the Grant-in-Aid for Scientific Research (A) (No. 25240026).

References

[1] H. Fujisaki, *Proc. Speech Prosody 2004*, pp. 22-26.
[2] M. Akagi *et al*., *Proc. APSIPA 2014*.
[3] O. Pierre- Yves., *Int. J. Hum. Compur. Interact.,* 35 (2), (2003), pp. 157-183.
[4] C. M. Lee and S. Narayanan., *IEEE Trans. Speech Audio Process.,* 13 (2), (2005), pp. 293-303.
[5] J.A Russell and P. Gerldine, *J.Pers. Soc. Psychol,* 38 (2), (1980), pp. 311-322.
[6] Y. Hamada, R. Elbarougy and M. Akagi, *Proc. APSIPA 2014*.
[7] J. -S. Jang, *J.Pers., IEEE Trans. Syst.,* 23 (3), (1993), pp. 665-685. -207.
[8] H. Fujisaki, *Tech. Rep. of IEICE,* 37 (9), (1994), pp. 1-8.
[9] R. Elbarougy and M. Akagi, *Acoust. Sci. Tech..,* 35 (2), (2014), pp. 86-98.
[10] H. Kawahara *et al*., *Speech Commum.,* 27, (1999), pp. 187.
[11] H. Mixdorff, *Proc. ICA 2000,* 3, pp. 1281-1284