9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Consistency of base frequency labelling for the F_0 contour generation model using expressive emotional speech corpora

Yoshiko Arimoto[1], *Yasuo Horiuchi*[2], *Sumio Ohno*[3]

[1]Faculty of Science and Engineering, Teikyo University, Japan
[2]Graduate School of Engineering, Chiba University, Japan
[3]School of Computer Science, Tokyo University of Technology, Japan
arimoto@ics.teikyo-u.ac.jp, hory@faculty.chiba-u.jp, ohno@stf.teu.ac.jp

Abstract

To investigate the consistency of base frequency (F_b) labelling of the F_0 contour generation model for expressive and/or authentic emotional speech, a F_b labelling experiment was conducted using three trained labellers employing the parallel corpus of emotional speech, Online-gaming voice chat corpus with emotional labelling (OGVC). Twenty-four utterances from spontaneous dialog speech and emotion-acted speech in the OGVC were labelled with the F_b, phrase command, and accent command by the three labellers. A repeated measure analysis of variance was performed with the factor of the corpus type, gender, speaker, emotion, and labeller, for the F_b value of each utterance. The results show a significant main effect on gender, speaker, and emotion and the significant interaction between speaker and emotion. The results also indicate that the value of F_b varied when the different emotions were expressed, even when uttered by the same speaker. Moreover, the precise inspection for the F_b of each utterance suggests that the F_b also varied when the linguistic content of the utterances differed, even if the same emotion was expressed in those utterances.
Index Terms: base frequency, the F_0 contour generation model, labelling consistency, emotional speech

1. Introduction

The fundamental frequency (F_0) contour generation model (the Fujisaki's model) [1] is one of the effective methods for analyzing and synthesizing the F_0 pattern of speech using relatively small numbers of parameters that correspond to the linguistic content. It has often been adopted to analyze prosodic information of emotional or intentional speech or to examine the appropriate F_0 contour pattern for generating expressive synthesized speech. The Fujisaki's model enables to analyze complicatedly fluctuating F_0 patterns with three types of components: phrase, accent and baseline. The baseline component, i.e., the base frequency (F_b) of utterance, affects the analysis of the phrase and accent components, while its determination with the observed F_0 contour is quite difficult for labellers.

There have been several studies on speech analysis with the Fujisaki's model using read speech material [2, 3, 4, 5, 6, 7, 8, 9, 10]. In those studies, the obtained F_bs for the read speeches were comparatively accurate and consistent within labellers because the read speech was less expressive and the F_bs were less variable for a single speaker. Therefore, the F_bs of those read speech could be fixed based on the speaker's innate F_0 baseline. By contrast, current spoken language research adopts various types of speaking styles such as affective speech or spontaneous dialog speech. Those speeches are quite different from read speech in its realizations of their prosodies. The F_0 contours of expressed speech vary for individual speakers, even if they speak the same linguistic contents. Therefore, it is quite difficult for labellers to consistently annotate F_b for each utterance because the same value of F_b cannot be used for all utterances owing to the F_b variability.

Previous research has reported on these F_b variability as one of the components of the Fujisaki's model. Mixdorff *et al.* analyzed both read speech and spontaneous speech from the same speaker using the Fujisaki's model [6]. They reported that the same F_b value was not used for both the read and spontaneous speech. The F_b value was fixed based on the corpus type (either read speech or spontaneous speech) and was not based on the speaker. Thus, the F_bs of the utterances varied, even for the same speaker. In another research [11], Mixdorff demonstrated the influence of speaking rate on the command of the Fujisaki's model. Using the Bonn Tempo-Corpus, he found that the value of F_b increased with an increased speaking rate. However, no research has focused on the F_b of emotionally-expressed spontaneous and acted speech.

In this study, the consistency of the F_b was demonstrated using the utterances from both emotionally expressed spontaneous dialog speech and emotion-acted speech. If the value of F_b varies among utterances of one speaker according to the expressed emotion, the same value could not be used for utterances expressed with various types of speaking styles, even if those utterances came from the same speaker. This paper reports the results of an experiment to demonstrate which factors influence F_b variability by conducting command labelling with the Fujisaki's model.

2. Speech material

To analyze various types of speaking styles, the speech material was obtained from Japanese speech corpora with different specifications. The adopted corpora were the online gaming voice chat corpus with emotional label (OGVC) [12], which was the parallel corpus of spontaneous dialog speech and acted speech. OGVC was developed specifically for research on emotion recognition and emotional speech synthesis. It includes two types of emotional speech material, spontaneous dialog speech (OGN) and acted speech (OGA) which has the same linguistic content as spontaneous dialog speech. In this study, we used both OGN and OGA.

OGN was recorded while two or three speakers were chatting over the Internet when participating in a Massively Multiplayer Online Role Playing Game (MMORPG). The spontaneous dialog speech corpora of OGVC includes 9,114 utterances of 13 Japanese speakers (4 females and 9 males). The utterances of this corpus were defined based on 400 ms inter-

10.21437/SpeechProsody.2018-81

pausal units (IPUs). Of all the utterances, 6,578 have one of 10 emotion category labels, i.e., joy, acceptance, fear, surprise, sadness, disgust, anger, anticipation, neutral, and others. Eight of the 10 emotion category labels were based on Plutchik's primary emotions [13]. Each utterance was labelled by three labelers so that each utterance has three emotional labels. The 3,845 emotion-labelled utterances for which two of the three labelers agreed with one emotion label were used for the experiment.

OGA contains 2,656 acted utterances spoken by four professional actors (two male and two female). Seventeen short dialogues were selected from the transcribed OGN dialogues. The actors were instructed to perform a specific emotion attached with each utterance in the short dialog under three different levels of emotional intensity (weak, middle, strong) and a neutral state. In this study, only the utterances at the middle level of emotional intensity were used to adjust the number of speech to the other corpus.

The utterances of OGA and OGN were appropriate for our study because the F_b of each utterance was considered to vary between speakers and within a speaker owing to their various types of emotional expression. Sixteen utterances from OGA (2 emotions (joy and anger) $\times$ 2 sentences $\times$ 4 speakers (FOY, FYN, MOY, and MTY)), and eight utterances from OGN (2 emotions (anger and joy) $\times$ 4 speakers (01_MMK, 02_MEM, 06_FTY, 06_FWA)) were selected for this study (24 utterances in total).

3. F_b labelling for the F_0 contour generation model

In the Fujisaki's model, the F_0 contour can be expressed by

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

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

$$G_a(t) = \begin{cases} \min[1 - (1 + \beta t)\exp(-\beta t), \gamma], & t \geq 0, \\ 0, & t < 0, \end{cases} \quad (3)$$

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 indicate

F_b : baseline value of fundamental frequency,
I : number of phrase commands,
J : number of accent commands,
Ap_i : 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, and
γ : relative ceiling level of accent components.

Parameters α, β and γ were assumed to be constant in almost every case and were set equal to 3.0, 20.0 and 0.9, respectively, in this study.

To obtain the F_b for each utterance, the labelling experiment was conducted. The three trained labellers (L1, L2, and L3) were instructed to label the base frequency (F_b), phrase commands, and accent commands for all 24 utterances. They labelled each command through their own decision and without any constraint on labelling.

4. Analysis

A repeated-measure five factorial analysis of variance was performed with the factors of corpus type (OGN or OGA), gender (male or female), speaker (01_MMK, 02_MEM, 06_FTY, 06_FWA, FOY, FYN, MTY, or MOY), emotion (anger or joy), and labeller (L1, L2, or L3) for the logarithmic F_b values of the utterances. As a post-hoc test, a Tukey HSD test was performed.

5. Results

Figures 1 – 6 show the mean F_bs (circles in Figs. 1 – 6) and those standard deviations (errorbars in Figs. 1 – 6) of each level of the factors. The number of asterisks shows the significant difference between the levels (*$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$). A result of ANOVA revealed the significant main effects of gender ($F(1, 8) = 68.56, p < 0.001$), speaker ($F(5, 8) = 4.14, p < 0.05$), and emotion ($F(1, 8) = 20.64, p < 0.01$). There are no significant main effects of corpus type and labellers. The ANOVA also revealed a significant interaction between speaker and emotion ($F(5, 8) = 3.79, p < 0.05$).

6. Discussion

There are no significant main effects of labeller (Fig. 3). This result suggested that the labelled F_bs were consistent among the three labellers. In this labelling experiment, all three labellers were trained for the command labelling of the Fujisaki' model. The labellers would be considered to have a certain amount of common knowledge on the F_b labelling. Therefore, the labelling result by all three labellers were rather reliable for this analysis.

There are significant main effects of gender ($F(1, 8) = 68.56, p < 0.001$, Fig. 2) and speaker($F(5, 8) = 4.14, p < 0.05$, Fig. 5). These results are reasonable because the gender difference of F_0 between male and female and the speaker difference of F_0 are well-known facts in acoustic analysis. The individual speakers have innate characteristics on a range of fundamental frequencies. Therefore, the base frequency (F_b) is also different among the individual speakers, and the F_b could be fixed for the utterance of one speaker. Previous research, such as that reported in [3], used the fixed F_b for the utterances of one speaker based on this fact.

There is no significant main effect of corpus type (Fig. 1). In this experiment, the authentic emotional speech derived from OGN and the emotion-acted speech derived from OGA were used for the F_b labelling of the Fujisaki's model. Thus, the results suggested that the F_b variability could not be affected whether the utterance was spontaneous or acted.

The results also revealed that there is a significant main effect of emotion ($F(1, 8) = 20.64, p < 0.01$, Fig. 4). In this experiment, anger and joy were adopted for the analysis. Anger and joy are the opposed emotional states in the valence polarity

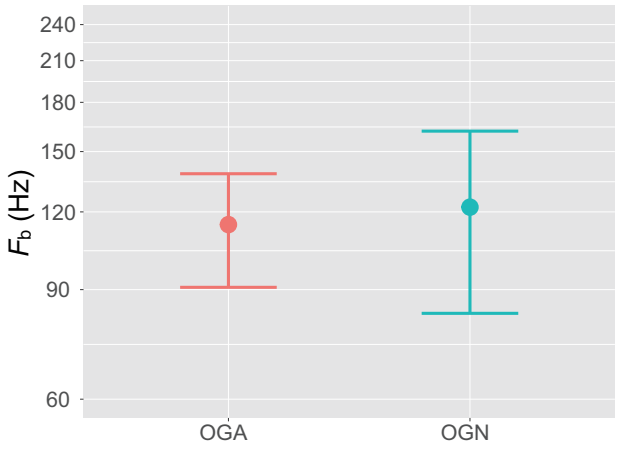

Figure 1: *Mean and standard deviation of F_b for each corpus.*

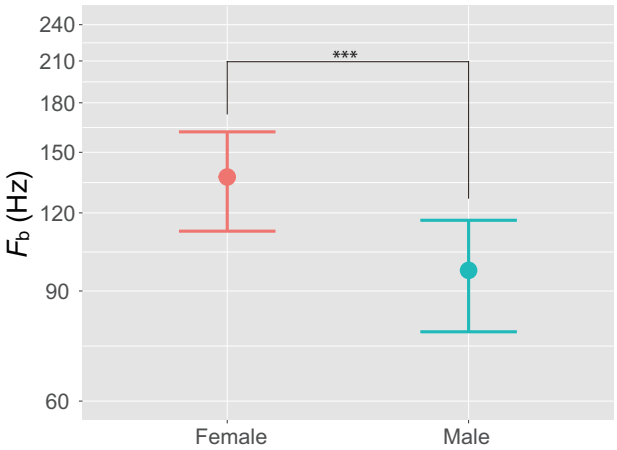

Figure 2: *Mean and standard deviation of F_b for each gender.*

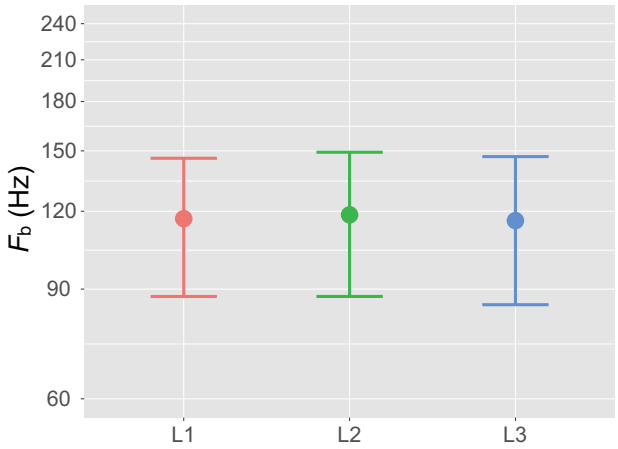

Figure 3: *Mean and standard deviation of F_b for each labeller.*

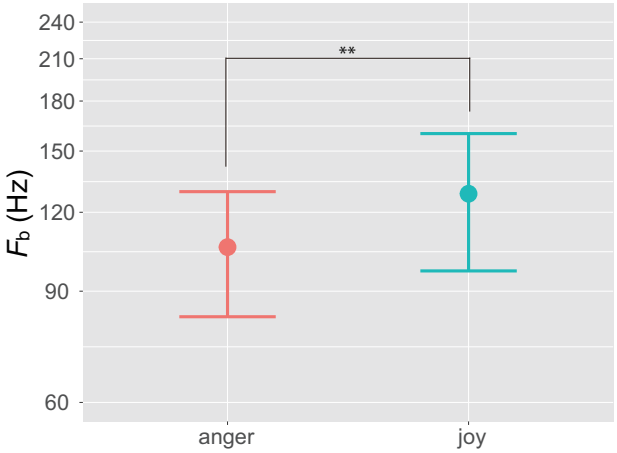

Figure 4: *Mean and standard deviation of F_b for each emotion.*

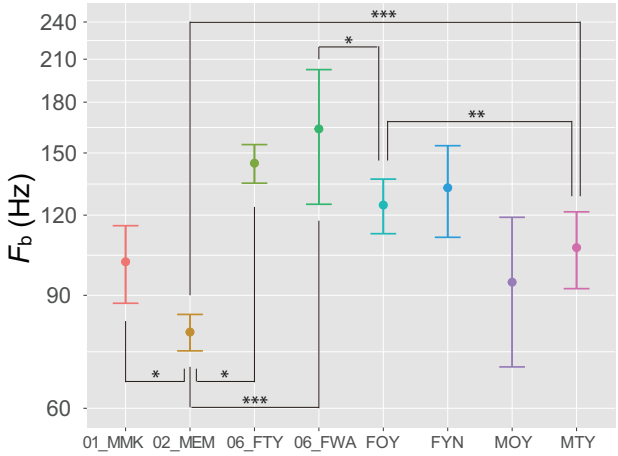

Figure 5: *Mean and standard deviation of F_b for each speaker.*

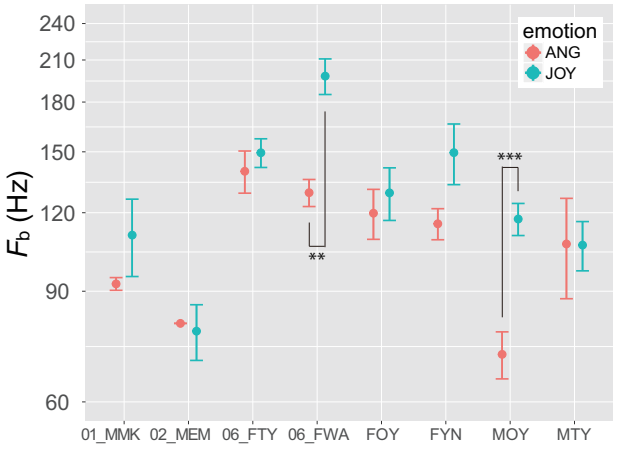

Figure 6: *Mean and standard deviation of F_b between speaker and emotion.*

according to Russell [14]. Anger is a more negative emotion, while joy is a more positive emotion. This result suggested that the F_b of the emotional speech varies based on where its emotion locates in the valence polarity.

Moreover, the result revealed that there is a significant interaction between speaker and emotion ($F(5,8) = 3.79, p < 0.05$, Fig. 6). The 06_FWA from OGN and MOY from OGA exhibited the same tendency. The F_b of the joyful speech is significantly higher than that of the angry speech ($p < 0.01$) as a result of the post-hoc test. This tendency is the same as that shown in Fig. 4 by the significant main effect of emotion. Interestingly, this result also revealed that there are speakers whose F_b did not vary between anger and joy. It suggests that the F_b variability for emotion is a speaker specific phenomenon.

For more precise inspection of the labelled F_b, Fig. 7 shows the three F_b values for each utterance annotated by the three labellers. The shapes and colors represent the three labelers, and the utterance names were lined in the y-axis. Each utterance name consisted of the speaker's name and the emotion separated by a hyphen ([speaker name]-[emotion]). For the utterance from OGA, there are two utterances of the different linguistic contents with the same emotion. To distinguish those utterances, the numbers are attached to the utterance name after emotion ([speaker name]-[emotion][number]). Figure 7 exhibits the tendencies on the F_b variability in one emotion and the labellers' fluctuations for F_b labelling.

In Fig. 7, there are a few utterances whose F_bs differed from each other, even if those emotional expressions were the same and the speakers were the same, i.e., FOY-ANG1 and FOY-ANG2 and MTY-ANG1 and MTY-ANG2. This suggests that even if one speaker expressed one emotion in the utterances of the different linguistic contents, the F_b of those utterance could vary and could not be fixed at the same value.

From the ANOVA result, there is no significant main effect of labeller. However, the F_b fluctuations among labellers were found in some utterances such as 01_MMK-JOY, 02_MEM-JOY, FOY-JOY1, FYN-JOY2, and MTY-JOY1. The differences between the minimum F_b and maximum F_b among the three labellers ranges from 3.31 semitones (FYN-JOY2) to 4.91 semitones (FOY-JOY1) for those utterances, whereas they range from 0.00 (01_MMK-ANG) semitones to 2.79 semitones (MTY-ANG1) for other utterances. Those huge differences in F_b values among the labellers might be caused by the different interpretations of the F_0 contours of those utterances.

7. Conclusion

To investigate the consistency of base frequency (F_b) labelling of the F_0 contour generation model, F_b labelling experiments were conducted by three trained labellers using the parallel corpus of the Online-gaming voice chat corpus with emotional labelling (OGVC). A repeated measure variance of analysis was performed with the factors of corpus type, gender, speaker, emotion, and labeller for the F_b values of each utterance of expressive and/or authentic emotional speech. The results elucidated that emotional expression in speech strongly affects the value of F_b. Moreover, the precise inspection of the F_b of each utterance suggested that the F_b also varied when the linguistic content of utterance was different, even if the same emotion was expressed in those utterances. Our findings suggest that the F_b values cannot be fixed to one value for one speaker.

In future research, the F_b determination method should be demonstrated to estimate the appropriate and accurate F_b for the utterances of the various types of speaking styles such as spontaneous dialog speech or emotion-acted speech.

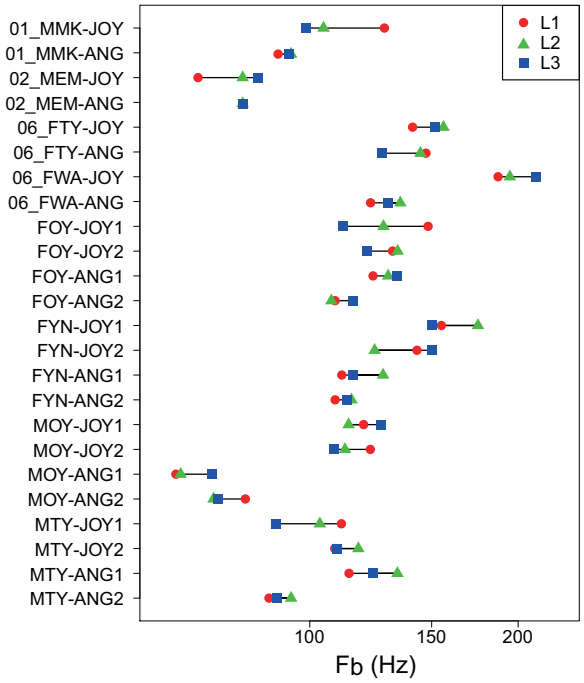

Figure 7: *Three F_b values for each utterance annotated by the three labellers.*

8. References

[1] 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.

[2] E. Geoffrois, "A pitch contour analysis guided by prosodic event detection," in *Speech Communication*, no. September, 1993, pp. 2081–2084.

[3] A. Sakurai, K. Hirose, and N. Minematsu, "Data-driven generation of F0 contours using a superpositional model," *Speech Communication*, vol. 40, no. 4, pp. 535–549, 2003.

[4] H. Kruschke and A. Koch, "Parameter extraction of a quantitative intonation model with wavelet analysis and evolutionary optimization," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, vol. 1. IEEE, 2003, pp. I–524–I–527.

[5] H. Fujisaki, C. Wang, S. Ohno, and W. Gu, "Analysis and synthesis of fundamental frequency contours of Standard Chinese using the command-response model," *Speech Communication*, vol. 47, no. 1-2, pp. 59–70, 2005.

[6] H. Mixdorff and H. R. Pfitzinger, "Analysing fundamental frequency contours and local speech rate in map task dialogs," *Speech Communication*, vol. 46, no. 3-4, pp. 310–325, 2005.

[7] K. Hirose, K. Sato, Y. Asano, and N. Minematsu, "Synthesis of F0 contours using generation process model parameters predicted from unlabeled corpora: Application to emotional speech synthesis," *Speech Communication*, vol. 46, no. 3-4, pp. 385–404, 2005.

[8] H. Kawatsu, D. Nagashima, and S. Ohno, "Rules and Evaluation for Controlling the Fundamental Frequency Contours with Various Degrees of Emotion Based on a Model for the Process of Generation," *The IEICE transactions on information and systems (Japanese edition)*, vol. 89, no. 8, pp. 1811–1819, 2006, (*in Japanese*).

[9] H. Kawatsu and S. Ohno, “An analysis of individual differences in the f0 contour and the duration of anger utterances at several degrees,” *Proceedings INTERSPEECH2007*, pp. 2213–2216, 2007.

[10] Q. Sun, K. Hirose, and N. Minematsu, “A method for generation of Mandarin F 0 contours based on tone nucleus model and superpositional model,” *Speech Communication*, vol. 54, no. 8, pp. 932–945, 2012.

[11] H. Mixdorff, A. Leemann, and V. Dellwo, “The influence of speech rate on Fujisaki model parameters,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, p. 33, 2014.

[12] Y. Arimoto, H. Kawatsu, S. Ohno, and H. Iida, “Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment,” *Acoustical Science and Technology*, vol. 33, no. 6, pp. 359–369, 2012.

[13] R. Plutchik, *Emotions: A psychoevolutionary synthesis*. New York: Harper & Row, 1980.

[14] J. A. Russell, “A circumplex model of affect,” *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.