An analysis of individual differences in the F_0 contour and the duration of anger utterances at several degrees

Hiromi Kawatsu[1] *and Sumio Ohno*[1]

[1]Graduate School of Bionics, Computer and Media Sciences, Tokyo University of Technology, Tokyo, JAPAN

kawatsu@so.cs.teu.ac.jp, ohno@cc.teu.ac.jp

10.21437/Interspeech.2007-602

Abstract

Taking up anger emotion expressed by speech, prosodic features were analyzed in order to find out the relationship between the degree of anger and manifestations on the speech signal in terms of individual differences. As a result of analysis, there were some common features among the speakers, although there were some speaker-dependent features. About the baseline frequency and the magnitude of the first phrase command, common tendencies were found in all speakers. The amplitude of the accent command increases as the emotional degree increases on the whole. Some speakers emphasized accent commands at all positions within a sentence, some emphasized only near the end of a sentence. Speaking rate at the 1st and 4th phrases were faster than those at the 2nd and 3rd phrases for the utterance with emotion, although there was an individual difference in the effect of the emotional degree. It is very interesting that two aspects in prosody, *i.e.*, an F_0 contour and a speaking rate, might be complement each other in order to represent a difference of emotional degrees.

Index Terms: emotional speech, anger, prosody, fundamental frequency contour, duration

1. Introduction

Even if the same linguistic contents are uttered, various impressions and information can be conveyed through a spoken utterance. Prosodic features mainly contribute to a rich emotional expression that cannot be conveyed merely by the linguistic representation [1]. The present authors have been engaged in analyzing prosodic features quantitatively for utterances with various emotions, the changes of the prosodic features for the emotional degree are examined in the utterances which are well acted and which are assumed as typical expressions. For this study, we examined "anger" as a kind of emotion and investigated the relationship between the prosodic features and the degree of anger emotion. A neutral utterance and three emotional ones at several anger degrees were recorded and the features of fundamental frequency contour and speech rate were analyzed in terms of individual differences.

2. Speech material

2.1. Recording conditions

In all, 15 sentences were prepared for reading with emotion. All sentences consist of four Japanese phrases (*bunsetu*) and have various syntactic structures. Each sentence was uttered at three emotional degrees: "weak," "medium," and "strong." A neutral utterance (*i.e.,* no emotions are expressed) was also used. A situation scenario was prepared for each sentence to express the emotion naturally. Some examples of sentences are shown in Table 1 and the scenarios are shown in Table 2. Eight adult speakers (6 males and 2 females), who are experienced in theater, uttered each sentence three times. The speakers uttered each sentence at different degrees in order of "neutral," "medium," "weak," and "strong," referring to the text and the situation scenario displayed on a computer monitor.

Table 1: *Examples of a text.*

	Text
1.	*Hitono monoo katteni toruna.* (Do not take my belongings without permission.)
2.	*Anatano taidono warusaga fuyukaida.* (Your bad attitude is unpleasant.)
3.	*Machinakadeno me:wakuna chu:syawa yamenasai.* (Do not park a car in the center of town.)

Table 2: *Example of the scenario.*

Text	*Sono me:re:ni shitagaunowa murida.* (It is impossible to obey such an order.)
Scenario	Although he is my boss at work, he makes unreasonable demands for work. I have had it with him. Said to him with anger.

2.2. Subjective evaluation for the material

For the prepared speech material, subjective evaluation of the emotional degree was conducted by listening tests in order to quantify how much the emotion is perceived. The subjects were six undergraduate students. They listened to each utterance in the material in a random order, evaluated how much specified emotion was perceived, and selected a number from 0–5: "0" means that the emotion is not perceived at all, "1" for slightly perceived to "5" for most strongly perceived. This test was carried out to elucidate the relationship between the emotional degree which a speaker intends and that which a listener perceives. In all, three male speakers and one female speaker, MTI, MTM, MYH, and FCM, were selected for analyses because the emotional degree that they intended was conveyed appropriately to listeners.

3. Analysis methods

3.1. Analysis of the F_0 contour

3.1.1. A model for the process of generating an F_0 contour [2]

A model for the process of generating an F_0 contour is used to quantify and parameterize the F_0 contour features. The F_0 contour can be considered as the consequence of motor control for vibration of the vocal folds. Fujisaki and coworkers demonstrated that, in many non-tone languages, $\ln F_0(t)$ can be expressed as the sum of phrase components, accent components, and a baseline component $\ln F_b$. For these languages, the process of F_0 contour generation can be modeled as a block diagram shown in Fig. 1. Although the phrase components and accent components are controlled primarily using linguistic information, our preliminary study indicates that all three types of components might be influenced also by emotion, whereas other model parameters can be considered to remain almost constant.

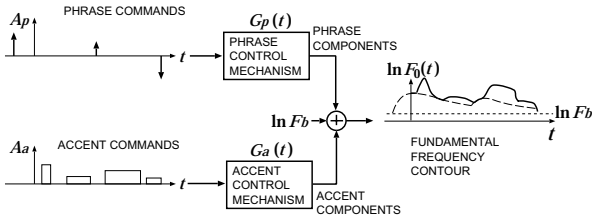

Figure 1: *A model for the process of generating an F_0 contour.*

Using the Analysis-by-Synthesis method, it is possible to determine the number, magnitude, and timing of phrase and accent commands, as well as the baseline value that will generate an F_0 contour, which is the closest approximation to the observed F_0 contour.

In this study, we specifically examine the baseline value of fundamental frequency (F_b), the magnitude of the phrase command (A_p), and the amplitude of the accent command (A_a) among the parameters of the model.

3.1.2. Analysis procedure

The speech signal was digitized at 10 kHz with 16-bit precision. The F_0 was extracted at 10-ms intervals by the modified autocorrelation analysis of the LPC residual. The first approximate values of parameters of a model for the process of generating F_0 contours were determined manually to be consistent with the linguistic information of the text. The model parameters were finally obtained using the Analysis-by-Synthesis method.

3.2. Analysis of duration

3.2.1. Ratio of the duration

To examine the influence of emotional degree on the duration, a ratio of the duration is introduced both for the overall sentence and for each corresponding phrase between a neutral utterance and an emotional one. The ratio of the sentence duration (R_S) represents the global change of the speaking rate in an emotional utterance through comparison with a neutral utterance; the ratio of the phrase duration (R_P) represents a local change of the speaking rate.

3.2.2. Analysis procedure

To obtain each phrase duration, phrase boundaries were determined using Julian [3], which is an automatic speech recognizer. Julian produced every phoneme boundary in a time point by giving a speech waveform and its transcription. The automatically obtained boundaries were then modified by visual inspection of a waveform and a sonagraph.

A pause is an important factor that influences the duration of preceding and/or subsequent phrases. For simplicity, only utterances including no pauses, which constitute the great majority of the prepared material, were analyzed.

After determining the duration for each segment, a ratio of the duration was defined as follows:

$$R_S \quad \text{or} \quad R_P = \frac{d_{\text{target}}}{\bar{d}_{\text{neutral}}}, \tag{1}$$

where $\bar{d}_{\text{neutral}}$ denotes the average of durations over the three neutral utterances for the same sentences and d_{target} denotes the duration of the target segment.

4. Analysis results

4.1. F_0 contour

4.1.1. Baseline value of fundamental frequency (F_b)

Figure 2 shows the ratio of the baseline value of fundamental frequency for an emotional utterance to that for a neutral one. The baseline value tends to become higher as the emotional degree increases. This tendency is remarkable for speaker MYH, although it is only slightly apparent for speaker MTM.

4.1.2. Magnitude of phrase command (A_p)

Figure 3 shows the difference in the magnitude of the first phrase command of emotional utterance from that of a neutral one. The magnitude of the phrase command is generally suppressed for emotional utterances. No great difference exists between the three emotional degrees.

These findings, in terms of both the baseline value and the magnitude of the first phrase command, showed the same tendency as that identified in the authors' previous study of other speech materials [4].

4.1.3. Amplitude of accent command (A_a)

Figure 4 shows the difference in the amplitude of accent command for emotional utterance from that for a neutral one by each phrase position within a sentence. The change of the amplitude of the accent command seems to be different among speakers.

MTI There seems to be no change for 1st and 2nd phrases as the emotional degree increases. The amplitude of the accent command for the 3rd and 4th phrases increases remarkably as the emotional degree becomes strong, especially at the end of the sentence.

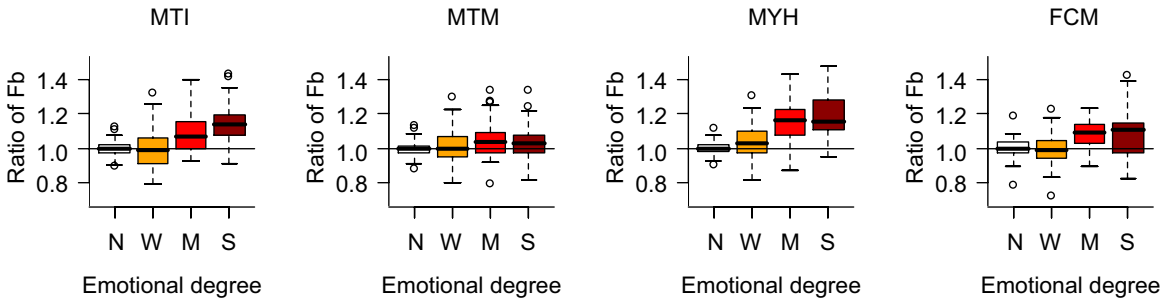

Figure 2: *Ratio of baseline value of fundamental frequency versus emotional degree: (***N,** *neutral;* **W,** *weak;* **M,** *medium;* **S,** *strong).*

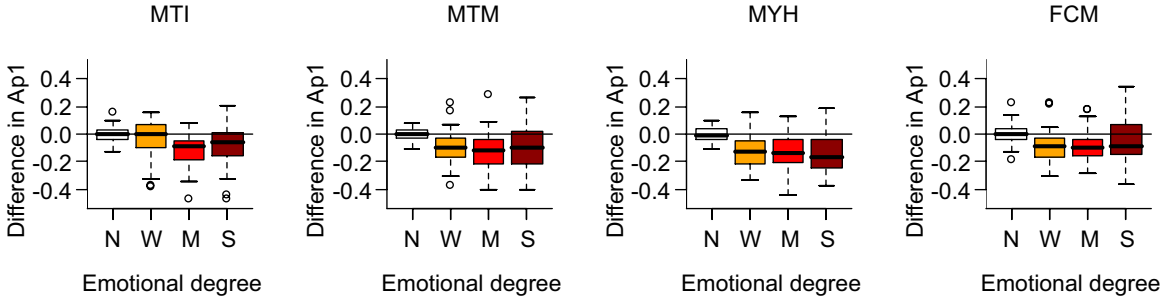

Figure 3: *Difference in magnitude of the phrase command versus emotional degree.*

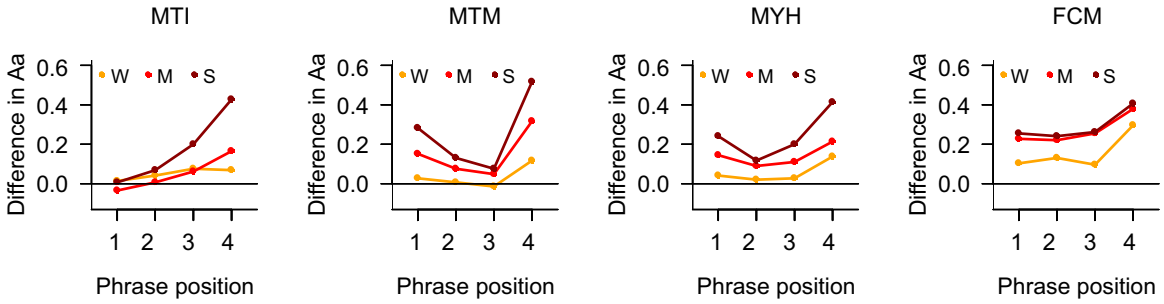

Figure 4: *Average of difference in amplitude of the accent command as a function of the phrase position within a sentence by emotional degrees.*

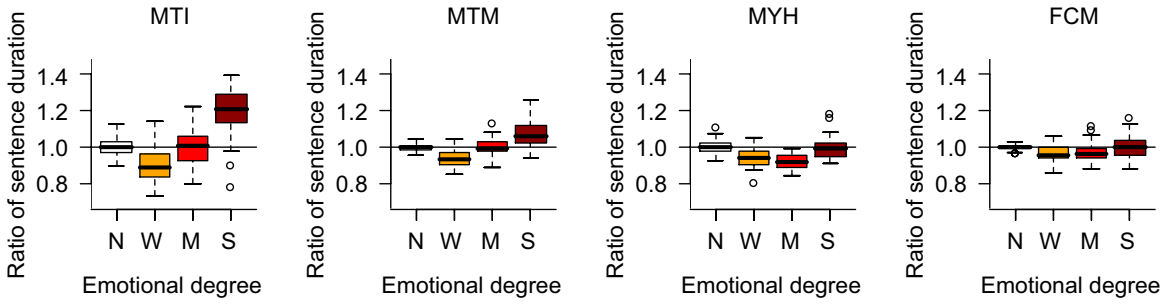

Figure 5: *Ratio of sentence duration versus emotional degree.*

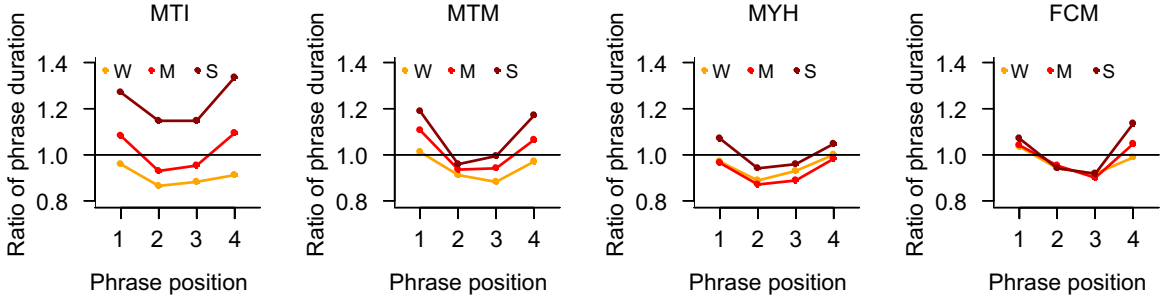

Figure 6: *Average of ratio of phrase duration as a function of the phrase position within a sentence by emotional degrees.*

Table 3: Factors that might influence a change of the phrase duration.

	Factor (Range of variables)
1.	Emotional degree (weak, medium, strong)
2.	Phrase position within a sentence (1, 2, 3, 4)
3.	Boundary depth of a subsequent phrase (1, 2, 3)
4.	Number of morae (2∼7)
5.	Accent type (flat, head-high, mid-high)
6.	Number of geminate consonants (0∼1)
7.	Number of syllabic nasals (0∼2)
8.	Number of long vowels (0∼2)

MTM, MYH The amplitude of the accent command increases as the emotional degree increases for all positions within the sentence. The most remarkable increase is at the 4th phrase, whereas the next marked increase is found at the 1st phrase.

FCM The increased amplitude of the accent command is found for all positions within the sentence. The difference between a neutral utterance and an emotional one is large, whereas the influence of the emotional degree is slight.

4.2. Duration

As for the duration, the ratio of sentence duration for emotional utterance to that of a neutral one was calculated as a function of the emotional degree. Figure 5 shows the ratio for each speaker. When the emotional degree is weak, the ratio of the sentence duration is smaller than 1.0, *i.e.*, the speaking rate of weak anger utterance is faster than that of neutral one. For speakers MYH and FCM, the ratio of the sentence duration does not show a large change when the emotional degree became strong. On the other hand, the ratio tends to become larger as the emotional degree increases for speakers MTI and MTM. For these speakers, the speaking rate of strong anger utterance is slower than that of the neutral one.

Regarding the phrase duration, various factors listed in Table 3 might influence the change duration locally. A forward stepwise regression procedure in a multiple regression model was used to examine the influence of these factors. The emotional degree and the phrase position within a sentence, then, were statistically significant irrespective of the speaker. Figure 6 shows the ratio of the phrase duration as a function of the phrase position within a sentence by the emotional degree. For all speakers, the ratios of 1st and 4th phrases are larger than those of 2nd and 3rd phrases. Especially in the 4th phrase, the change of the ratio is most significantly influenced by the emotional degree. In terms of individual differences, MTI changes the duration both globally and locally to control the emotional degree, while FCM changes only the last phrase duration.

5. Discussion and Conclusions

In this paper, the F_0 contours and the duration were analyzed between utterances at several anger degrees. Those without any emotions used the same sentence, which consists of four Japanese phrases uttered by four speakers. Results of analyses showed common features among the speakers, although some speaker-dependent features are also visible. Characteristics of each prosodic parameter are summarized in Table 4. Regarding the baseline frequency (F_b) and the magnitude of the first phrase command (A_{p1}), common tendencies are apparent for all speakers, although the impact of the change differs among speakers. On the whole, the amplitude of the accent command (A_a) increases as the emotional degree increases. Some speakers emphasize accent commands at all positions within a sentence; some emphasize only those near the end of a sentence. Speaking rates at the 1st and the 4th phrases are faster than those at the 2nd and the 3rd phrases for the utterance with emotion, although there are individual differences in the effect of the emotional degree. When changes between different emotional degrees in an F_0 contour domain are very small, a change in the duration domain is rather large, and vice versa. It is very interesting that two aspects in prosody, *i.e.*, an F_0 contour and a speaking rate, might complement each other to represent a difference of emotional degrees. Further studies of other prosodic features such as intensity and voice quality are needed to examine this complementation.

Table 4: Comparison of characteristics of each prosodic parameter among speakers.

	MTI	MTM	MYH	FCM
F_b	becomes *higher* as the emotional degree increases			
A_{p1}	becomes *smaller* as the emotional degree increases			
A_a	becomes *larger* as the emotional degree increases			
1st		***	***	**
2nd		*	**	**
3rd	**	*	***	**
4th	***	***	***	
R_S	*fast* when the emotional degree is weak			
	becomes *remarkably slow* as emotional degree increases		becomes *slightly slow* as emotional degree increases	
R_P	*faster* at 1st & 4th phrases than at 2nd & 3rd phrases			
1st	***	***	**	*
2nd	***		**	
3rd	**	**		
4th	***	***	**	*

*** : $p < 0.01$ between W and M **and** between M and S.
** : $p < 0.01$ between W and M **or** between M and S.
* : $p < 0.10$ between W and M **or** between M and S.
none : non significant.

6. References

[1] Banse, R. and Scherer, K. R., "Acoustic Profiles in Vocal Emotion Expression", Journal of Personality and Social Psychology, Vol. 70 (3), pp. 614–636, 1996.

[2] Fujisaki, H. and Nagashima, S., "A model for the synthesis of pitch contours of connected speech", Annual Report of the Engineering Research Institute, University of Tokyo, vol. 28, pp. 53–60, 1969.

[3] http://julius.sourceforge.jp/

[4] Kawatsu, H. and Ohno, S., "F_0 contour generation with four kinds and several degrees of emotion for speech synthesis", Proc. Spring Meet. Acoust. Soc. Jpn., Vol. 1, pp. 177–178, 2006 (in Japanese).