ANALYTICAL AND PERCEPTUAL STUDY ON THE ROLE OF ACOUSTIC FEATURES IN REALIZING EMOTIONAL SPEECH

Keikichi HIROSE† *Nobuaki MINEMATSU*‡ *Hiromichi KAWANAMI*‡*
hirose@gavo.t.u-tokyo.ac.jp *mine@gavo.t.u-tokyo.ac.jp* *hkawa@etl.go.jp*

†Dept. of Frontier Informatics, ‡Dept. of Information and Communication Engg.,
University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, JAPAN
∗Currently with Electrotechnical Laboratory

ABSTRACT

Investigation was conducted on how prosodic features of emotional speech changed depending on emotion levels. The analysis results on fundamental frequency (F_0) contours and speech rates implied that humans have several ways to express emotions and use them rather randomly. Investigation was also conducted on what acoustic features were important to express emotions. Perceptual experiments using synthetic speech with copied acoustic features of target speech indicated importance of the segmental features other than the prosodic features. Especially, a high importance was observed in the case of happiness.

1. INTRODUCTION

Recent advancement of multimedia interfaces between man and machine largely increased interests on realizing and recognizing emotions conveyed by speech. Therefore, a rather large number of analyses were conducted on emotional speech mainly from the viewpoint of prosodic features. Although general tendencies of prosodic features of emotional speech as compared to those of neutral utterances were clarified already, such as F_0 and its dynamic range increases in happiness and their decreases in sadness, the prosodic control done for the realization of emotional speech was rather ad hoc. From this point of view, we have been investigating prosodic features of emotional speech of Japanese mainly based on the command response model of F_0 contours[1]. We have pointed out that declination of F_0 contour observable for neutral speech is lessened in emotional speech, which is ascribable to rather large command values of the model in the latter half of the sentence[2]. The major difficulty in the study of emotional speech may reside in a large diversity in the analysis results. From the former analyses, it seems that humans have several ways in realizing an emotion and select them rather independently. With a purpose of obtaining some insight on this issue, in this paper, discussions will be made through the analyses of F_0 contours and segmental duration on how humans control the prosodic features to express degrees in emotional speech.

Although emotions of speech were investigated mostly from the prosodic aspect, importance of segmental features (voice quality) in realizing emotions was also pointed out[3]. Success of realizing high-quality in emotional speech by selection-based concatenative speech using speech database with the emotion to be realized indicated the importance of segmental features[4]. Here, we conduct a perceptual experiment using synthetic speech selectively copying acoustic features of target emotional speech to investigate which acoustic features including segmental ones are important in expressing emotions.

2. LEVELS IN EMOTION

2.1. Speech Samples

In order to collect speech samples with several emotional levels, scenarios were arranged with which simulated dialogues were conducted between speakers A and B. Responding to the speaker A's questions/requests, speaker B repeats utterances with the same content ("ekimademukaeni-ikimasu (I'll meet you at the station.)") but with increased emotional levels as the dialogue proceeds. These utterances by speaker B were used for the analysis. The emotional levels are 5 including a neutral one (levels 0 to 4). Two semi-professional actors of the Tokyo dialect were asked to simulate dialogues by referring to the scenarios and utterances by speaker TI were selected for the analysis. A scenario was arranged for each of anger, happiness and sadness. **Table 1** shows the scenario for anger. Similar dialogues were also collected for several speakers, where the speaker B's responses were fixed to a short phrase sentence "urikiredesu (It was sold out.)." In this case the levels of each emotion is set from 0 to 3.

When investigating emotions, a first problem will be the content of target utterances. In several researches, utterances of one or few syllables, mostly interjections, were selected. Although this selection is reasonable to avoid emotions being affected by the linguistic contents of utterances, no knowledge will be available on the prosodic control for speech synthesis from the analysis results. Since our final goal is to construct prosodic rules for the synthesis

Table 1: Text used for the control of anger. (Originally in Japanese)

A: How can I reach from the station?
B: I'll meet you at the station.(neutral, Level 0)
A: No, thanks. Let me know the way. That is enough.
B: I'll meet you at the station.(slightly displeased, Level 1)
A: I'll come by myself. Let me know the way only.
B: I'll meet you at the station.(irritated a bit, Level 2)
A: I'll go by myself.
B: I'll meet you at the station.(clear anger, Level 3)
A: Are you really come to meet me?
B: I'll meet you at the station.(a burst of anger, Level 4)

of emotional speech, we select a sentence with a syntactic boundary, which is possibly accompanied by a prosodic phrase boundary in actual utterances.

2.2. Methods of Analysis

Analyses were conducted on F_0 contours and speech rates. As for F_0 contours, they were first extracted from speech waveform through a pitch extraction method based on LPC residual autocorrelation, and were then analyzed by the Analysis-by-Synthesis (AbS) method using the command response model[1]. Pitch extraction errors were manually corrected before the AbS analysis. The model represents an F_0 contour in logarithmic frequency scale as a superposition of phrase and accent components on a flat baseline. These components are represented as responses of second-order critically damped linear systems to the corresponding commands; impulse-like commands for the phrase components and step-wise commands for the accent components.

As for the speech rates, utterances were first segmented into phoneme units by the forced alignment using tri-phone HMM's[5]. All phoneme boundaries were manually checked and corrected if necessary. Then, the following reduction rate was calculated for each mora (Japanese utterance unit similar to a syllable) as the index representing speech rate increase in emotional speech from neutral speech[6]:

$$RD = (dur_r - dur_e)/dur_r, \quad (1)$$

where dur_r and dur_e are the duration of neutral (level 0) speech and emotional speech, respectively. When emotional speech is uttered faster than neutral speech, RD takes a positive value. The mora duration is subject to change due to various factors (such as the phone constitution and so on). By calculating the reduction rate, these factors can be suppressed, and the features directly related to emotional speech are revealed.

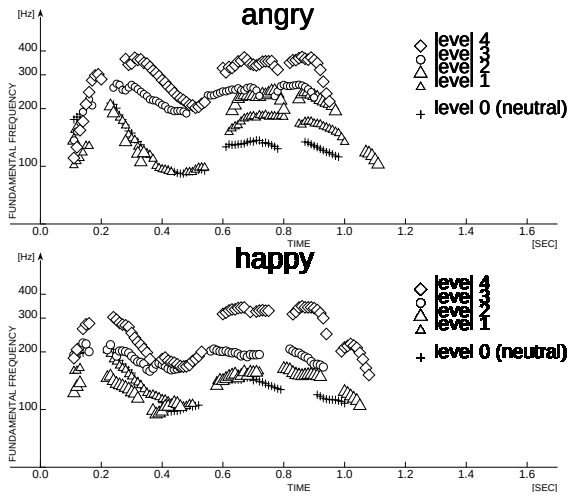

Figure 1: Fundamental frequency contours of "ekimade-mukaeni-ikimasu" uttered by speaker TI in 4 levels of anger and happiness.

2.3. Analysis Results

Figure 1 shows F_0 contours of various levels of anger and happiness. For anger and happiness, F_0 contours shift upwards and were deformed as the level increases. The deformation is due to the command level and timing changes. Through the deformation, the contour loses its declination feature. For sadness, the contour also loses its declination feature as the level increases.

As for the average speech rates of utterances, they are increased in the case of anger, but decreased in the case of sadness, as the level increases. The speech rate reduction for sadness is typical for sadness of level 4; accompanied by a pause at the prosodic phrase boundary.

Figure 2 shows the command values (magnitudes/amplitudes) of the model obtained by the AbS process for the three

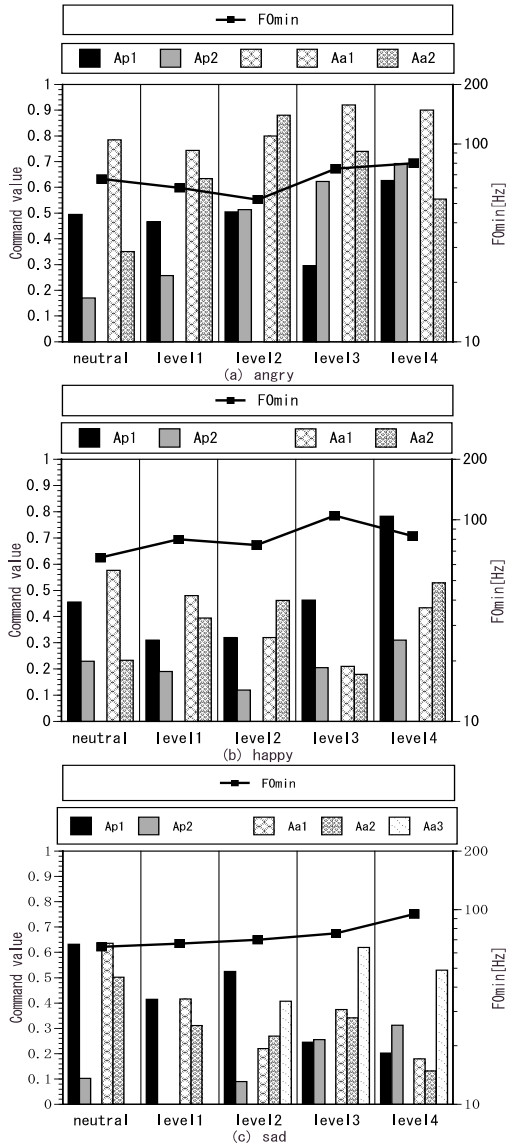

Figure 2: F_0 model command values of "ekimade-mukaeni-ikimasu" uttered by speaker TI in 4 levels of 3 types of emotions.

emotions. The commands are corresponding to the sentence parts as follows:

(Phrase Commands)
A_{p1}: for the first prosodic phrase "ekimade (to station)."
A_{p2}: for the second prosodic phrase "mukaeni-ikimasu (go and meet you)."

(Accent Commands)
A_{a1}: for the first prosodic word "ekimade."
A_{a2}: for the second prosodic word "mukaeni-ikimasu" in the cases of anger and happiness. In the case of sadness, for "mukaeni."
A_{a3}: for the third prosodic word "ikimasu" in the case of sadness.

In the case of anger, command values, especially those with small values in level 0 (neutral), increase as the level increases. It should be noted that the command value increase saturates for higher levels (levels 2 to 4). Increases of F_0 for higher levels are due to the $F_{0\min}$ (flat baseline) rises. On the contrary, for happiness, F_0 rises in the lower levels are mainly due to the rises of $F_{0\min}$, and those for the higher levels are due to the increases of command values. For sadness, as the level increases, values of sentence initial commands decrease, while those of other commands increase, resulting in the level F_0 contours. It should be noted that, in levels 2 to 4, the third prosodic word appears as the division of the second prosodic word.

Mora reduction rates are shown in **Figure 3** for the three emotions. The results for the sentence final mora "su" are not reliable because of occasional devoicing of vowel /u/. In the case of anger, mora reduction rates take smaller values (slower speech rates than those of neutral speech) at the beginning of utterance and larger values at the latter half of the utterance. Enhanced features are observable for higher levels. As for the happiness, mora reduction rates takes positive values (higher speech rate than neutral speech) on the average. No relation with level increase is observable. The mora reduction rates takes negative values for sadness: larger absolute values for higher levels on the average. No systematic result is observed on which mora is lengthened/shortened depending on the level.

Similar analysis was also conducted for "urikiredesu" uttered by three speakers (two males and one female), including speaker TI. Although, for speakers TI and KM, the command values increase corresponding to level increase as in the case of "kurumade-mukaeni-ikimasu," the phrase command value decreases for speaker YY. The decrease is over compensated by the increases of accent command value and $F_{0\min}$, resulting in the upward shifts of F_0 contours.

As for the speech rates, the speaker variations are also rather high as shown in **Figure 4**. For instance, mora reduction rates take positive values for speaker TI, while they take negative values for speakers YY and KM. When looking at each mora, speakers YY and KM shows features quite different to each other; absolute values of reduction rates are larger for morae at the latter part of sentence for speaker YY, while they are lager at the beginning part for

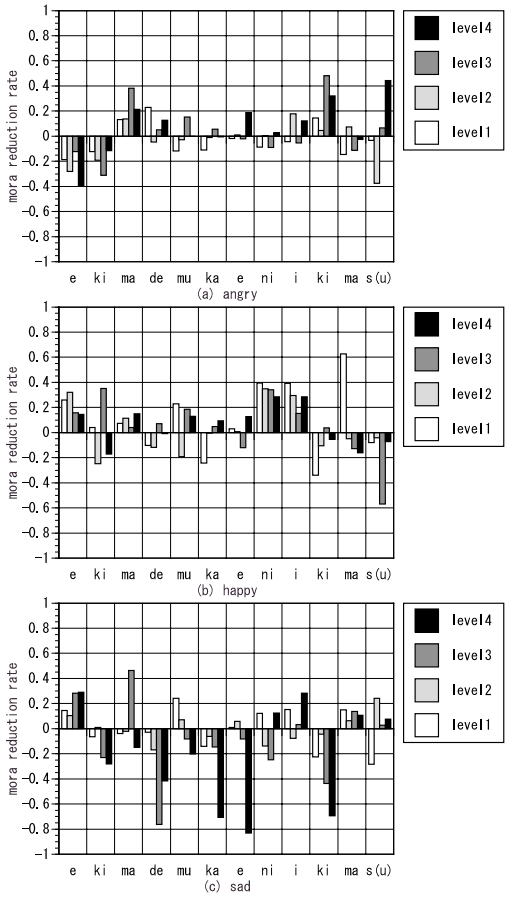

Figure 3: Mora reduction rates of "ekimade-mukaeni-ikimasu" uttered by speaker TI in 4 levels of 3 types of emotions.

speaker KM.

2.4. Discussion

Although the general tendencies (such as F_0 rise in anger) of emotional speech are observable in all the levels, it seemed all the tendencies do not become apparent evenly as the level increased. Humans may have several ways to express emotion and they may select one other than use them all. Selection will be done rather differently depending on each individual.

3. ACOUSTIC FEATURES IN REALIZING EMOTIONAL SPEECH

Perceptual experiments were conducted to find out what acoustic features were more important in realizing emotions. Through the analysis synthesis process of neutral speech, one or several acoustic features of speech with target emotion are copied. The speech thus synthesized is used for the perceptual experiment. The copied acoustic features are the F_0 contours in the model commands, duration of each mora, power of each vowel, and Cepstral coefficients. Content of the sentence is "kurumade-mukaeni-ikimasu (I will meet you with a car.)," and neutral and

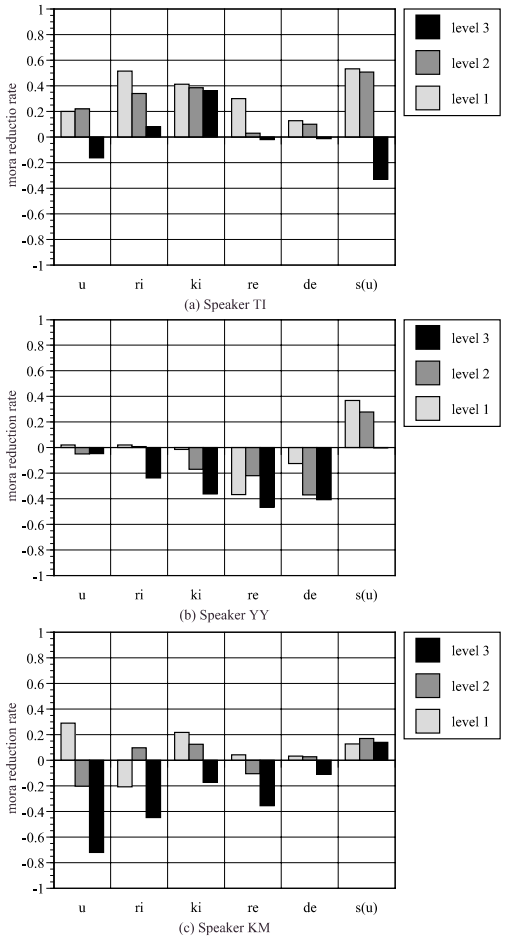

Figure 4: Mora reduction rates of "urikiredesu" uttered by three speakers in 3 levels of anger.

target utterances utilized are those from speaker YY. Utterances of speaker TI were not utilized because of serious speech quality degradation when Cepstrum coefficients were copied. The following 5 types of coping were conducted:

A) F_0 contours in command level,
B) Mora duration,
C) Vowel power,
D) All the three prosodic features A) to C),
E) Cepstral coefficients.

Here, we should note that the Cepstral coefficients include 0th order coefficient, which corresponds to speech power.

Eleven Japanese native speakers were asked to listen to the synthetic speech, and to evaluate it on five-rank score: 1 for neutral speech and 5 for target speech. Synthetic speech is arranged in our web page together with the neutral and target speech[7]. Since the listening test was conducted through this web page, informants could listen to the speech as much as they wanted. They were asked to select score 1, when they could not feel the target emotion. As clearly shown in **Table 2**, role of segmental features (Cepstral coefficients) for realizing emotion is especially high for happiness, though their contribution is comparative for that of prosodic features in the cases of anger and sadness. These tendencies coincide with those obtained in elsewhere for other languages[8].

Table 2: Results of perceptual experiments averaged for 11 subjects.

Copied acoustic feature	Anger	Happiness	Sadness
F_0 contour	1.3	1.6	2.2
Duration (speech rate)	1.1	1.2	1.3
Power	1.6	1.2	1.6
Three prosodic features	3.6	2.4	3.0
Cepstral coefficients	3.5	4.0	2.9

4. CONCLUSION

Through the analysis of prosodic features on how they changes depending on the levels of emotions, it was implied that humans may have several acoustic cues to express emotions and they use it rather randomly instead of changing them evenly to a target point when realizing higher levels. Perceptual experiments using synthetic speech with copied acoustic features of target speech revealed that segmental features also play an important role in transmitting emotions. Although an analysis was also conducted for formant frequencies, no systematic changes were observed with emotions.

Further studies both on prosodic and segmental features for increased number of utterances are required before our final goal of finding intrinsic features of emotions.

Finally, we would like to acknowledge Prof. Kazuya Takeda for providing us with Japanese tri-phone HMM's.

REFERENCES

1. H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," J. Acoust. Soc. Japan (E), Vol.5, pp.233–242 (1984-10).
2. K. Hirose, H. Kawanami and N. Ihara, "Analysis of intonation in emotional speech," ESCA Workshop on Intonation, pp. pp.185–188 (1997-9).
3. I. Murray and J. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," J. Acoust. Soc. Am., Vol.93, No.2, pp.1097–1108 (1993-2).
4. A. Iida et. al., "Designing and testing a corpus of emotional speech," Report for the Spring Meeting, Acoustical Society of Japan, pp.311–312 (1998-3). (in Japanese)
5. K. Takeda et. al., "Common platform of Japanese large vocabulary continuous speech recognition research: constriction of acoustic model," IPSJ SIG Notes, 97-SLP-18-3 (1997). (in Japanese)
6. M. Sakata and K. Hirose, "Analysis and Synthesis of prosodic features in spoken dialogue of Japanese," Proc. EUROSPEECH, Madrid, Vol.2, pp.1007–1010 (1995-9).
7. `http://www.gavo.t.u-tokyo.ac.jp/~ohura/exfin01c.html`
(Contains several Japanese characters.)
8. J. M. Montero et. al., "Analysis and modelling of emotional speech in Spanish," Proc. ICPhS, San Francisco, Vol.2, pp.957–960 (1999-7).