TR－A－0058

11

Prosody and Expression of Emotions in Speech

－韻律と音声における感情表現－

Yoshinori KITAHARA and Yoh'ichi TOHKURA

北原義典　東倉洋一

1989. 8. 21

ATR視聴覚機構研究所

〒619-02 京都府相楽郡精華町乾谷　☎07749-5-1411

ATR Auditory and Visual Perception Research Laboratories

Inuidani, Seika-cho, Soraku-gun, Kyoto 619-02 Japan

Telephone: +81-7749-5-1411
Facsimile: +81-7749-5-1408
Telex: 5452-516 ATR J

Prosody and Expression of Emotions in Speech

Y. KITAHARA and Y. TOHKURA

ATR Auditory and Visual Perception Research Laboratories
Seika-cho, Soraku-gun, Kyoto 619-02, Japan

※ This paper has been submitted to "Computer Speech & Language"

ABSTRACT

For the purpose of application to natural and high quality speech synthesis, the role of prosody in speech perception has been studied. Prosodic components, which contribute to the expression of emotions and their intensity, are clarified by analyzing emotional speech and by performing listening tests of synthetic speech. It has been confirmed that prosodic components, which are composed of pitch structure, temporal structure and amplitude structure, contribute to the expression of emotions more than the spectral structure of speech. The results of listening tests using prosodic substituted speech showed that temporal structure was the most important for the expression of anger, while for the intensity of anger, all the three components were much more important. Pitch structure also played a significant role in the expression of joy and sadness and their intensity. These results made it possible to convert a neutral utterances (i.e., ones with no particular emotion) into utterance expressing various kinds of emotions. The results can also be applied to controlling the emotional characteristics of speech in synthesis by rule.

1. Introduction

We have investigated the role of "prosody" on the spoken language, aiming at improving the quality of synthesized speech. Our goal is to synthesize speech which is rich in naturalness and intelligibility. Among various kinds of media, speech has the advantage of easiness for emotional representation and communication. This advantage greatly depends on the prosody of speech.

Prosody mainly consists of three factors; amplitude structure, temporal structure and pitch structure. Amplitude structure includes stress and prominence. Pause, rhythm and phoneme duration are the parts of temporal structure. Pitch structure participates in accent (pitch accent in Japanese) and intonation. Combined complexly, these factors contribute much to speech perception and are closely related to production of human speech (see Fig. 1). In previous studies of speech synthesis, researchers have mainly focused on improving phoneme intelligibility, but there has been little research on prosody, in particular on the emotional expression in speech. Giving and controlling emotions are indispensable techniques for generating high-quality synthesized speech.

In this paper, prosodic components that contribute to the expression of emotions and their intensity are clarified by analyzing speech with various emotions, and by performing listening tests of synthesized speech. Moreover, we constructed rules for controlling prosody and attempted to give some kinds of emotions to arbitrary neutral speech.

2. Emotion and prosodic information of speech

Before performing the main experiment, we conducted preliminary experiments to confirm the suitability of our speech samples and the contribution of prosody to the expression of emotions, as well as initial listening tests using an excitation source signal given by an LPC analysis/synthesis technique, with the prosody of some emotions.

2.1 Speech samples

Two sentences, neither of which was associated with any particular emotion, were chosen. Each sentence was uttered with four kinds of emotions by a male announcer (a professional broadcaster). These emotions were anger, joy, sadness and neutrality (no particular emotion). The eight speech samples (2 sentences x 4 emotions) were analyzed and synthesized by the PARCOR technique. They were labelled as Va(anger), Vj(joy), Vs(sadness) and Vn(neutrality), respectively. The conditions of PARCOR analysis and synthesis were as follows.

sampling rate	: 10kHz
order of analysis	: 10
time window	: 30msec Hamming Window
time window shift	: 10msec
pitch extraction	: AMDF method

excitation : pulse / noise

synthesizer : two-multiplier lattice filter

2.2 Preliminary experiment I

In this experiment, a set of Va, Vj, Vs and Vn was used as stimuli. First, they were presented to eight subjects over headphones. The order of presentation was random. Their task was to identify which emotion was accompanied with the speech. They had five choices; neutrality, anger, joy, sadness and other. Prior to this listening test, neutral speech Vn was presented to these subjects as a reference speech for the calibration of judgement. The results are shown in Table I. In this table, the numbers indicate the percentage of the responses over the number of subjects. Each emotion in the stimuli was reliably perceived by the subjects in good agreement. Among these stimuli, Vj, for which the speaker intended to express joy, had relatively fewer responses for joy, but the 75th percentile is dominantly higher than the responses for other emotions. Therefore, we consider that these stimuli were distinct from each other and adequate for use in experiments concerned with emotions.

2.3 Preliminary experiment II

For the second step of the study, we used the excitation source signals, En, Ea, Ej and Es (corresponding to Vn, Va, Vj and Vs

respectively). They had no phonemic features but prosodic information, namely amplitude, temporal and pitch structures were left intact. These speech samples, En, Ea, Ej and Es, were presented to seven subjects over headphones. Their task was to identify which emotion was accompanied with the speech. The condition of presentation was the same in as the previous test (Section 2. 2).

Table II shows the results of preliminary experiment II. It suggests that every emotional speech which includes prosodic information only can be perceived with the intended emotion. Consequently, it is assumed that prosody plays an important role in the expression of emotions in speech. Of course, there is a possibility that spectral information may also contribute to the expression of emotions.

3. Prosodic components contributing to the expression of emotions

In this section, we investigate the prosodic components which contribute to the expression of emotions by means of synthesized speech with substituted prosodic components.

3. 1 Prosodic substitution

We synthesized speech Vna1, Vna2 and Vna3 by substituting the amplitude, pitch and temporal structures of Vn with those of Va, respectively, as shown in Fig. 2. Substitutions were performed preserving

the temporal correspondence of the two speech samples with Dynamic Time Warping (DTW). The optimal DTW path was calculated using Euclidean distance of LPC-derived cepstral coefficients (order=10). Here, in particular, 'the substitution of temporal structure' implies the expansion and compression of Vn so that the synthesized speech has the temporal correspondence with Va. Using the same technique, we synthesized the stimuli Vnj1, Vnj2 and Vnj3 by substituting the prosodic components with those of Vj, and Vns1, Vns2 and Vns3 by substituting the prosodic components with those of Vs.

3.2 Listening test

The nine stimuli, Vna1~ Vns3, were presented to nine new subjects who were different from those in the two preliminary experiments using headphones. The experiment task was the same. The subjects were required to identify which emotion was accompanied with the speech.

3.3 Results

The results are shown in Table Ⅲ. Anger was the most frequent response (66.7%) when listeners were presented with the speech Vna2. This indicates that temporal structure is quite important for the expression of "anger". Majority responses of joy for Vnj3(100.0%) and Vna3(88.9%) indicate that pitch structure plays an important role for the expression of "joy". The high response rate for sadness observed in

Vns3 also indicates the importance of pitch structure for the expression of "sadness".

The response of joy was most frequent not only in identifying the stimuli Vnj3 but also the stimuli Vna3. These results suggest that the pitch structure in the expression of anger and joy is similar.

4. Prosodic structure in the expression of emotions

In the previous section, the most important prosodic component in the expression of each emotion was clarified by substitution of prosodic features of speech. In this section, prosodic structures are discussed in detail.

4.1 Temporal structure

Figure 3 shows the temporal duration of an utterance with neutral emotion and three kinds of emotions for two sentences. For the expression of anger, the duration is about 20% shorter, and for sadness it is about 20% longer than that with neutral emotion. Joyful speech has almost the same duration as neutral speech.

What is an appropriate measure for temporal structure? In order to obtain the temporal structure measure, we first investigated the relationship between the spectral change and time compression in comparing pairs of speech; one is neutral speech used as a reference, and the other contains one of the three different emotions in it. Prior to the analysis, we defined the rate of spectral change Δ using the cepstrum regressive coefficient δ_i, where δ_i is a gradient of the cepstrum regressive line of the i-th cepstral coefficient. $\Delta(t)$ at a time t is defined as follows:

$$\Delta(t) = \sqrt{\sum_{i=1}^{10} \{\delta_i(t)\}^2}$$

This regressive line is computed over a range of 70ms.

We calculated the rate of time compression α by using Dynamic Time Warping spectral matching. Here the Euclidean distance of the cepstral coefficient is used as a spectral matching measure. An optimal DTW spectral matching path, thus obtained, is shown in Fig. 4(a). On this optimal path, the local gradient represents a degree of segmental time elasticity. The total time compression rate α is given by averaging the local gradients.

Figure 4(b) shows the degree of time elasticity of Va, Vj and Vs, and the curve lying at the bottom of the figure is the spectral changing rate Δ at every frame on Vn. In Va (solid line), it shows that a small value of Δ corresponds to small value of α. In other words, if the spectrum change is slight, the time tends to be highly compressed. If there is a remarkable spectrum change, there is little elasticity. On the other hand, in Vs, when the spectrum change is small, the time length is expanded. Vj and Vn have nearly the same temporal structure, and the elasticity of Vj is quite small compared with that of Vn.

4.2 Pitch structure

Fundamental frequency (F_0) contours in three kinds of emotions are shown in Fig. 5. It is observed that the pitch pattern of joyful speech is remarkably similar to that of angry speech. Their dynamic range is wide. For another speech sample, also, F_0 of angry speech and joyful

speech are 50 ~ 100 Hz higher than that of neutral speech for the whole sentence. However, in sad speech, F_0 stays around 100Hz and its dynamic range is very narrow.

4.3 Amplitude structure

Figure 6 shows the power (value of 0th auto-correlation) patterns of three emotions. Compared with neutral speech Vn, the dynamic range of the amplitude is wider in angry speech Va, but is smaller in sad speech Vs. On the other hand, joyful speech Vj has a pattern similar to Vn except for the end of the sentence.

5. Intensity of emotions and prosody

In this section, we discuss the analytical and perceptual aspects of the factors dependent of the emotion intensity in speech. Speech samples produced with three emotion intensity levels of anger, joy and sadness each by a male announcer were used. These speech samples were converted to synthesized speech with the same conditions as in section 2.1.

5.1 Intensity of "anger"

Figure 7 shows comparisons of pitch patterns, power patterns, and temporal durations among the speech samples with three levels of anger. So far as the amplitude is concerned, the increase in the intensity of

anger is correlated with the power at the portion of higher pitch in every clause. The pitch pattern shows the rising pattern throughout the sentence until its end where it drops as the intensity level increases. The degree of the temporal duration compression has no dependency upon the intensity of anger.

In order to investigate how these prosodic components are related to the perception of emotion intensity, we synthesized the following speech samples.

Va1 ; increasing the amplitude in the portion of higher pitch (more than a certain threshold) at every clause for 3 times more than that of Va.

Va2 ; compressing the temporal duration for 80% of Va (depending on the spectral change).

Va3 ; increasing F_0 of Va by 30Hz over the whole sentence.

Using these speech samples, we compared the perceptual intensity of anger by performing paired comparison listening tests. The number of subjects was seven. The task was to answer which speech sounded more angry. Table Ⅳ shows the results. It is shown that three prosodic components, which were modified to synthesize the three speech samples Va1, Va2 and Va3 affect the intensity of anger in the table.

5.2 Intensity of "joy"

Components that the intensity of joy depends on are discussed. Figure 8 shows comparisons of pitch patterns, power patterns, and temporal

durations among the speech samples uttered with three intensity levels of joy.

Firstly, it seems that the strength of power and the intensity of joy are not closely correlated with each other. Temporal duration also has no relation with the intensity of joy. On the other hand, F_0 increases throughout the sentence as the intensity increases.

Performing a perceptual experiment on the intensity level of joy, we synthesized the following speech samples.

Vj1 ; increasing the amplitude in the portion of higher pitch (more than a certain threshold) at every clause for 3 times more than that of Vj.

Vj2 ; compressing the temporal duration for 80% of Vj (depending on the spectral change).

Vj3 ; increasing F_0 of Vj by 60Hz over the whole sentence.

Using these samples, we compared the intensity of joy by performing the paired comparison listening tests. The number of subjects was seven. The task was also to answer which speech sample sounded more joyful. Table V shows the results. It is shown from the responses in comparison of Vj and Vj3 that pitch structure contributes mainly to the intensity of joy. We also find that the temporal structure is related to the intensity of joy. Vj2 is perceived to be less joyful than Vj. This is because the more the temporal duration is compressed, the more they respond for anger rather than joy. Conversely, when the speech sample Vj4, whose temporal duration is extended to 20% more than that of Vj was given, the

subjects' responses were more for Vj than for Vj4. In other words, Vj was perceived to be more joyful than Vj4. It indicates that temporal structure, especially time elastisity, has little positive effect on increasing joyful emotion.

5.3 Intensity of "sadness"

Figure 9 shows comparisons of pitch patterns, power patterns, and temporal durations among the speech samples with three levels of sadness. As the intensity of sadness increases, the dynamic range of F_0 and power become narrower. While temporal duration tends to expand.

We synthesized the following speech samples in a perceptual experiment on the intensity of sadness.

Vs1 ; decreasing the amplitude in the portion of originally higher pitch by 1/3 times less than that of Vs.

Vs2 ; expanding the temporal duration by 20% more than that of Vs (depending on the spectral change).

Vs3 ; decreasing F_0 of Vs by 20Hz over the whole sentence.

Using these speech samples, we compared the intensity of sadness by performing paired comparison listening tests. The number of subjects was seven. The task was to answer which speech sounded more sad. Table VI shows that pitch information contributes to the intensity of sadness as well as joy.

6. Giving emotions by prosodic control

Based on the previous analysis and various trial-and-error perceptual experiments, we constructed prosodic control rules to give emotions to neutral speech. By applying the rules, we attempted to convert a new speech sample uttered by the same announcer into three kinds of emotional speech, and evaluated the rule.

6.1 Prosodic control rule

The prosodic control rules to convert neutral speech into each of three kinds of emotional speech are described below.

① anger

amplitude; $a(t) \times 3$ (in the portion of higher pitch)

temporal duration; 30% compression (depending on the spectral change)

F_0 ; $f_0(t)+30$ (except for the end of the sentence)

② joy

amplitude; $a(t) \times 2$ (for stops and fricatives)

temporal duration; unchanged

F_0 ; $(f_0(t)-F_0 min) \times 1.4+F_0 min+30$

③ sadness

amplitude; $a(t) \times 1/10$ (for stops and fricatives)

temporal duration; 15% expansion (depending on the spectral change)

F_0 ; $(F_0(t)-F_0 min) \times 0.6+F_0 min$

Here, a(t) is a time series of linear amplitude values in neutral speech, and $f_0(t)$ implies a time series of fundamental frequency in Hz. F_0min is a minimum value of $f_0(t)$ over the whole sentence.

6.2 Listening test

By means of these rules, we synthesized angry speech Sa, joyful speech Sj and sad speech Ss using a new speech sample S. These speech samples were presented to 11 subjects over headphones. They answered which emotion is accompanied in the speech.

Table Ⅶ shows the results. As far as anger and sadness are concerned, we can conclude that, to a satisfactory extent, emotions of anger and sadness were given to the neutral speech by these rules. For joyful emotion, however, the rules seemed insufficient. This might be due to speech quality degradation caused by spectrum distortion accompanied with excessive increase of pitch in PARCOR synthesis.

7. Conclusion

We investigated the prosodic components that contribute to the expressions of emotions and their intensity, by analyzing the acoustic components of emotional speech and by performing listening tests of synthetic speech whose prosodic components were controlled.

The following conclusions were obtained.

1. Temporal structure is important in the expression of anger. Temporal compression depending on the spectral change works effectively to provide neutral speech with the emotion of anger. The dynamic range of F_0 and its amplitude in angry speech is wider than in neutral speech. The intensity of anger increases as F_0 becomes higher.

2. Joyful speech has nearly the same pitch pattern as that of angry speech. The temporal and amplitude structures of joyful speech are very similar to those of neutral speech. Pitch structure contributes mainly to the intensity of joy.

3. In sad speech, the dynamic range of F_0 and amplitude are narrower than those of neutral speech. The temporal duration of sad speech is comparatively longer than those of neutral speech. The intensity of sadness increases as F_0 decreases.

Additionally, some emotions were given to an arbitrary neutral speech by prosodic control rules based upon the analysis of emotional speech and perceptual experiment.

Further research is necessary to establish more general rules which work well across speech samples and speakers.

REFERENCES

Furui, S. (1986). On the Role of Spectral Transition for Speech Perception, J. Acoust. Soc. Am. 80, (4), 1016-1025

Ichikawa, A., Nakayama, T. & Nakata, K. (1967). Experimental Consideration on

the Naturalness of the Synthetic Speech, 1-3-8, Proc. ASJ Fall Meeting, 95-96 (in Japanese)

Itakura, F. & Saito, S. (1972). On the Optimum Quantization of Feature Parameters in the PARCOR Speech Synthesizer, Proc. 1972 Conf. Speech Commn. Process., 434-437

Ito, K. (1986). A Basic Study on Voice Sound Involving Emotion(Ⅲ), Ergonomics, 22(4), 211-217 (in Japanese)

Kitahara, Y., Takeda, S., Ichikawa, A. & Tohkura, Y. (1987). Role of Prosody in Cognitive Process of Spoken Language, Trans. IEICE, Jpn. J10-D, No. 11, 2095-2101 (in Japanese)

Kitahara, Y. & Tohkura, Y. (1988). Prosodic Components of Speech in the Expression of Emotions, Proc. ASJ-ASA Joint Meeting, FF-13, S98-S99

Komatsu, A., Oohira, E. & Ichikawa, A. (1986). Prosodic Aids to Structural Analysis of Conversational Speech, Proc. ICASSP86, 42. 15, 2283-2286

Nakayama, T., Ichikawa, A. & Miura, T. (1967). Fundamental Consideration on the Naturalness of the Synthetic Speech, 1-3-7, Proc. ASJ Fall Meeting, 93-94 (in Japanese)

Ross, M. J., Shaffer, H. L., Cohen, A., Freudberg, R. & Manley, H. J. (1974). Average Magnitude Difference Function Pitch Extractor, IEEE, Trans. Acoust. Speech and Signal Proc., Vol. 1, ASSP-22, 353-362

Sagayama, S. & Itakura, F. (1979). On Individuality in a Dynamic Measure of Speech, 3-2-7 Proc. ASJ Spring Meeting, 589-590 (in Japanese)

Tohkura, Y. & Kitahara, Y. (1988). Nonlinear Time-scale Modification of Speech Signal with Varied Segmental Duration Characteristics, Proc. ASJ-ASA Joint Meeting, G23, S15

Williams, C. E. & Stevens, K. N. (1972). Emotions and Speech, J. Acoust. Soc. Am. 52, (4), 1238-1250

TABLE I. Results of preliminary experiment I

(%)

	anger	joy	sadness	neutral	other
Va	*100.0*	0.0	0.0	0.0	0.0
Vj	12.5	**75.0**	0.0	12.5	0.0
Vs	0.0	0.0	**87.5**	12.5	0.0

TABLE Ⅱ. Results of preliminary experiment Ⅱ

(%)

	neutral	anger	joy	sadness	other
En	*85.7*	0.0	0.0	14.3	0.0
Ea	0.0	*100.0*	0.0	0.0	0.0
Ej	0.0	14.3	*71.4*	0.0	14.3
Es	0.0	0.0	0.0	*100.0*	0.0

TABLE Ⅲ. Results of prosodic substitution test

(%)

	neutral	anger	joy	sadness	other
Vna1	77.8	22.2	0.0	0.0	0.0
Vna2	22.2	*66.7*	0.0	0.0	11.1
Vna3	0.0	0.0	88.9	0.0	11.1
Vnj1	44.4	44.4	0.0	11.1	0.0
Vnj2	66.7	0.0	22.2	0.0	11.1
Vnj3	0.0	0.0	*100.0*	0.0	0.0
Vns1	77.8	11.1	0.0	11.1	0.0
Vns2	77.8	0.0	0.0	11.1	11.1
Vns3	11.1	0.0	0.0	*88.9*	0.0

TABLE Ⅳ. Results of comparison test

for anger intensity

	JUDGE OF INTENSITY			
	>	=	<	(%)
Va	0.0	14.3	85.7	v a 1
	0.0	0.0	100.0	v a 2
	0.0	0.0	100.0	v a 3

TABLE V. Result of comparison test

for joy intensity

	JUDGE OF INTENSITY			
	>	=	<	(%)
V j	0	100.0	0.0	v j 1
	71.4	28.6	0.0	v j 2
	0.0	0.0	100.0	v j 3

TABLE VI. Result of comparison test

for sadness intensity

	JUDGE OF INTENSITY			
	>	=	<	(%)
Vs	0	57.1	42.9	v s 1
	14.3	71.4	14.3	v s 2
	0.0	0.0	100.0	v s 3

TABLE Ⅶ. Results of listening test of emotional converted synthesized speech

(%)

	anger	joy	sadness	neutral	other
Sa	81.8	9.1	0.0	0.0	9.1
Sj	27.3	54.5	9.1	0.0	9.1
Ss	0.0	0.0	100.0	0.0	0.0

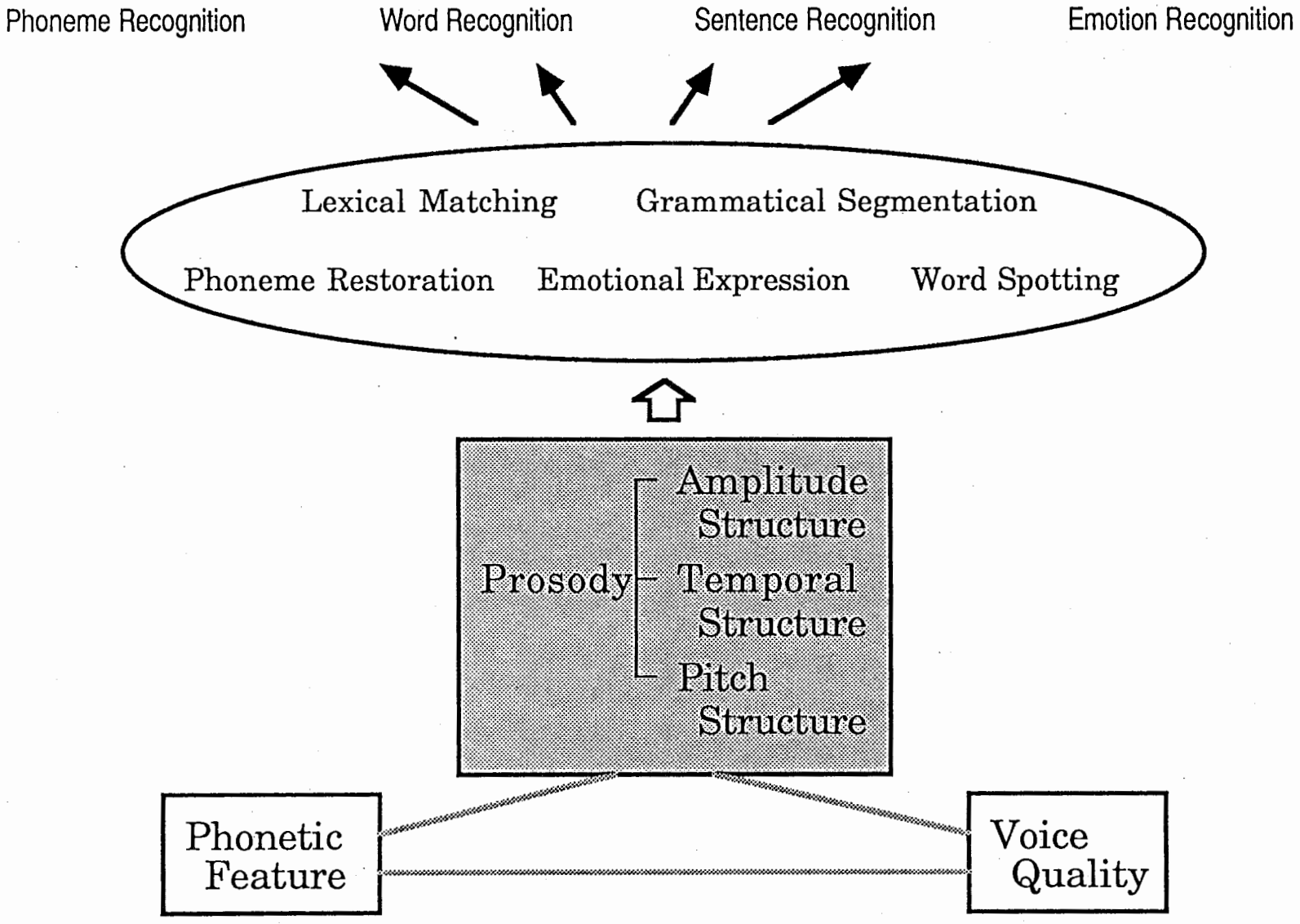
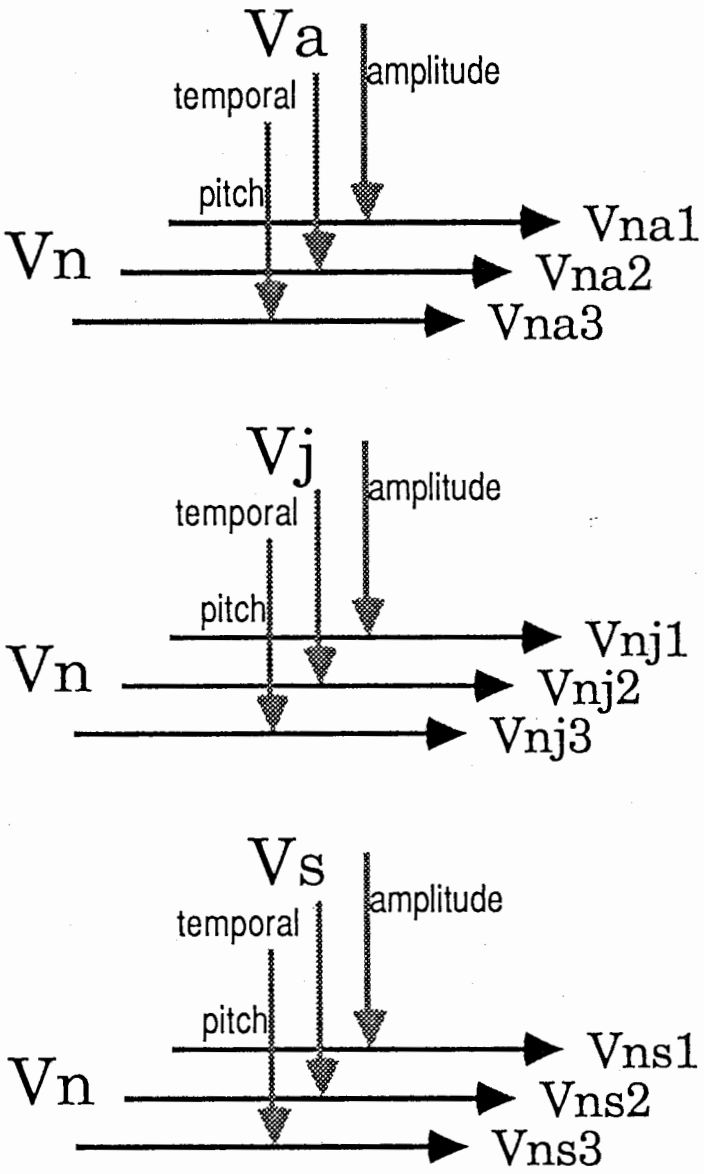
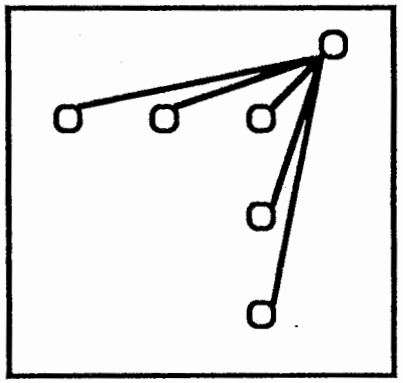


Figure 1. Structure of speech acceptance





local path constraint

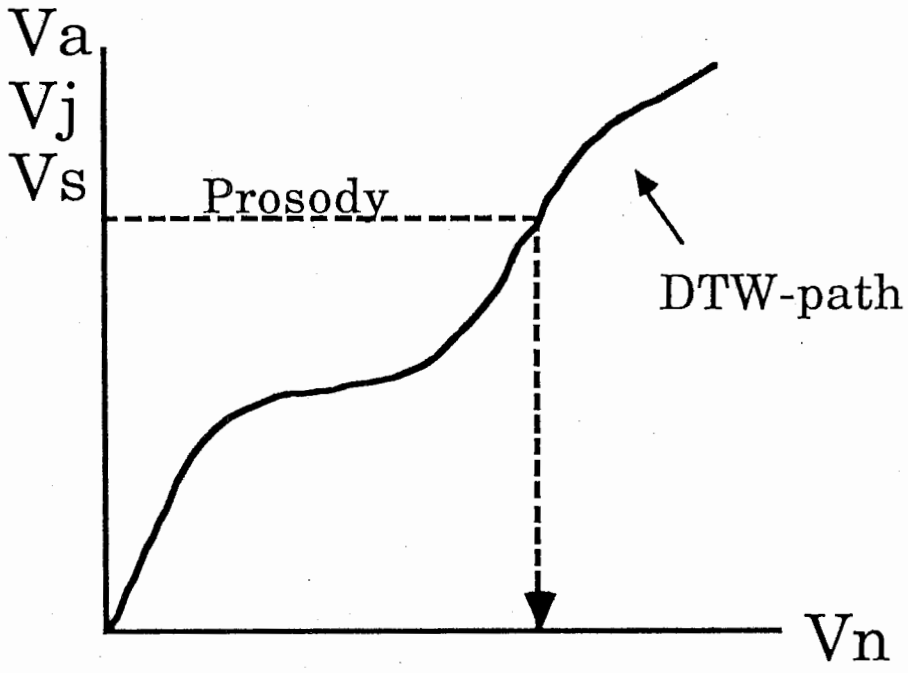


Figure 2. Prosodic substitution

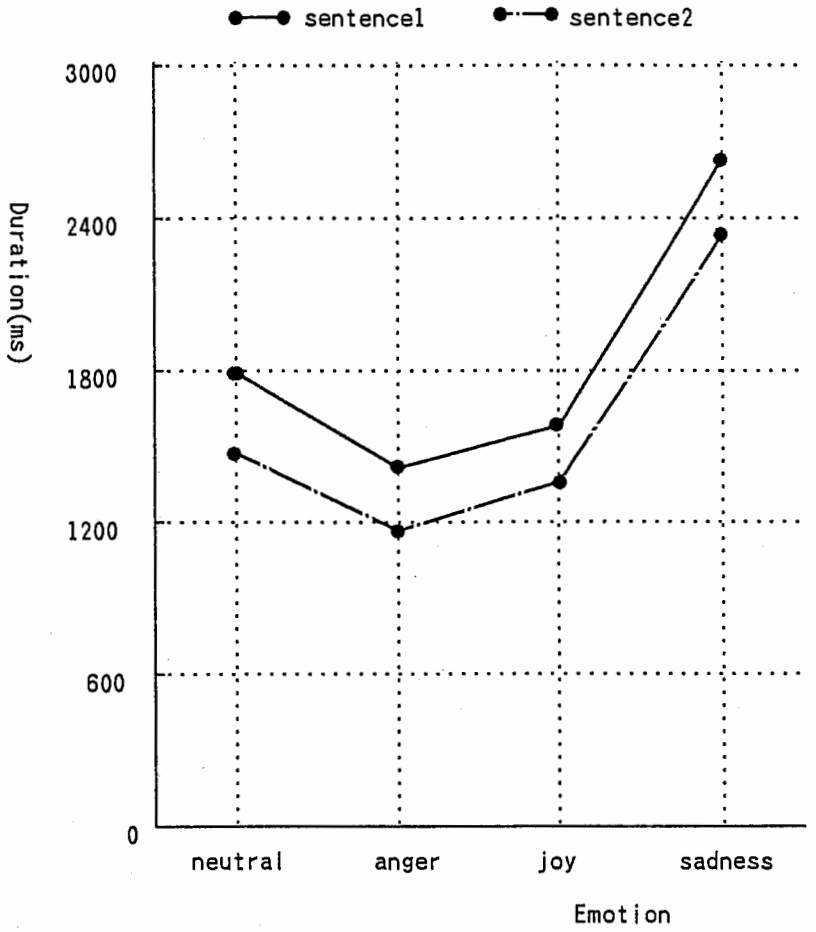


Figure 3. Temporal duration of the sentences with each emotion

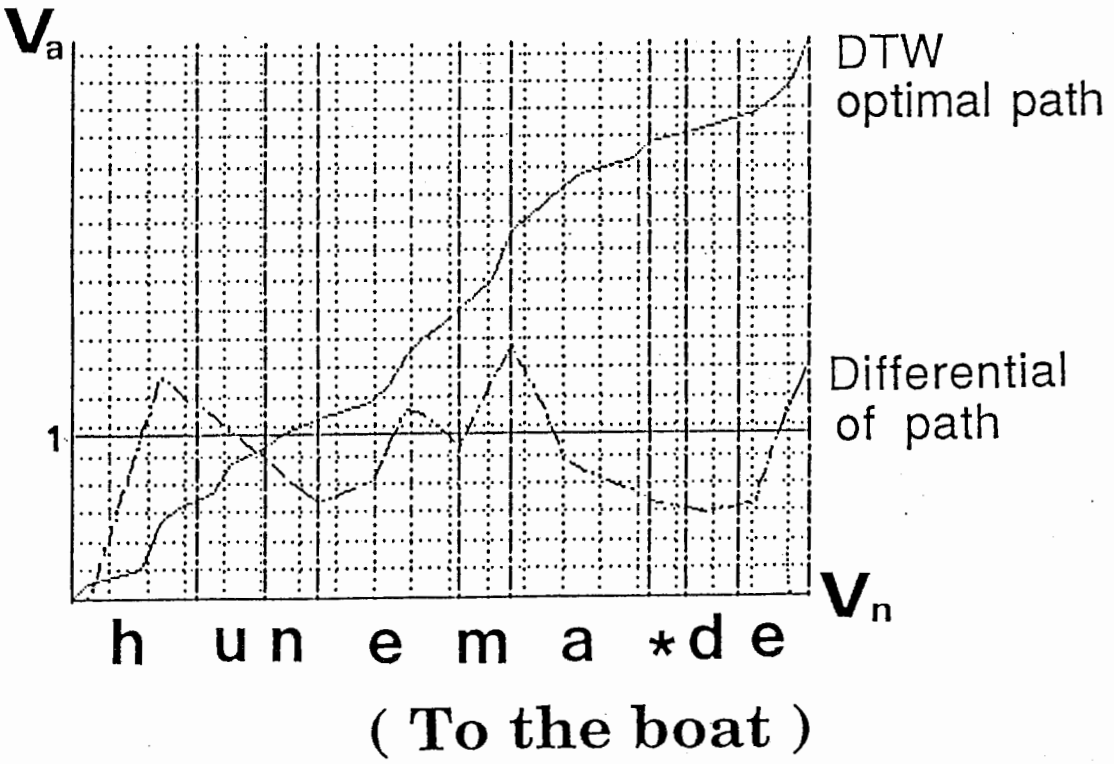


Figure 4 (a). Optimal DTW spectral matching path and gradient of the path

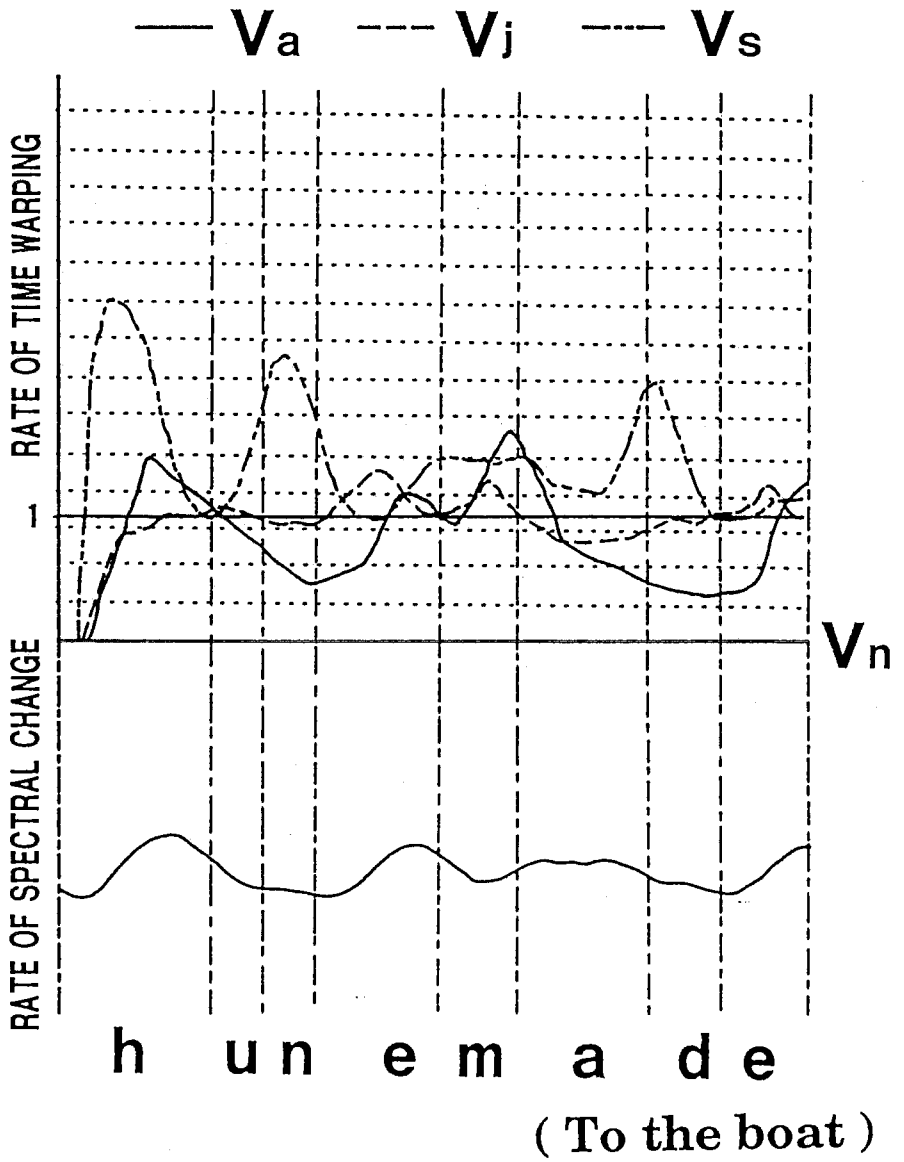


Figure 4 (b). Relationship between rate of time warping and rate of spectral change

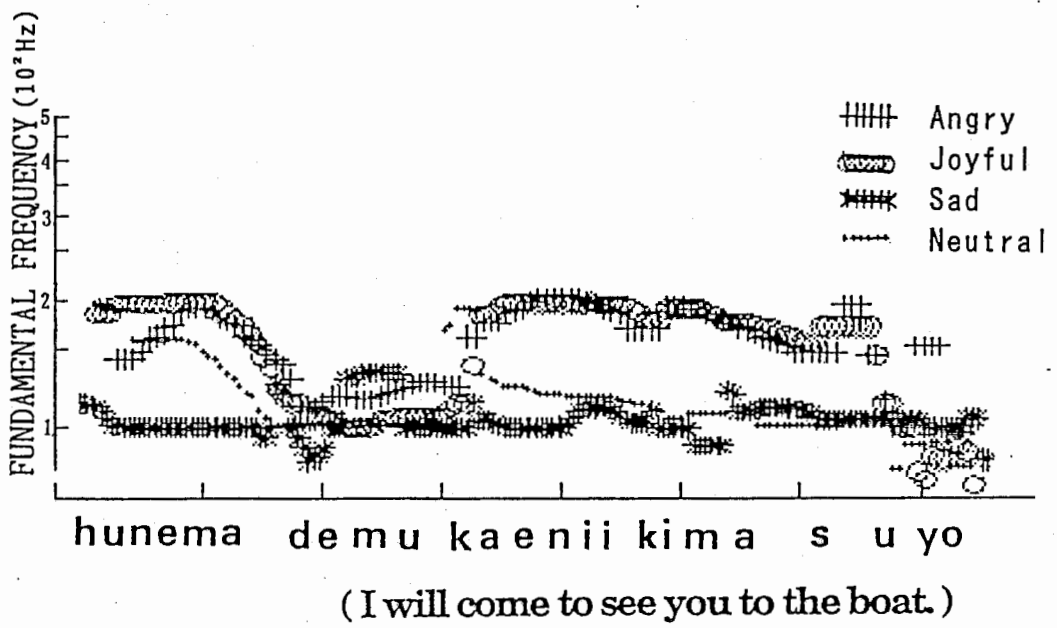


Figure 5. Fundamental frequency contours in three kinds of emotions and neutral speech

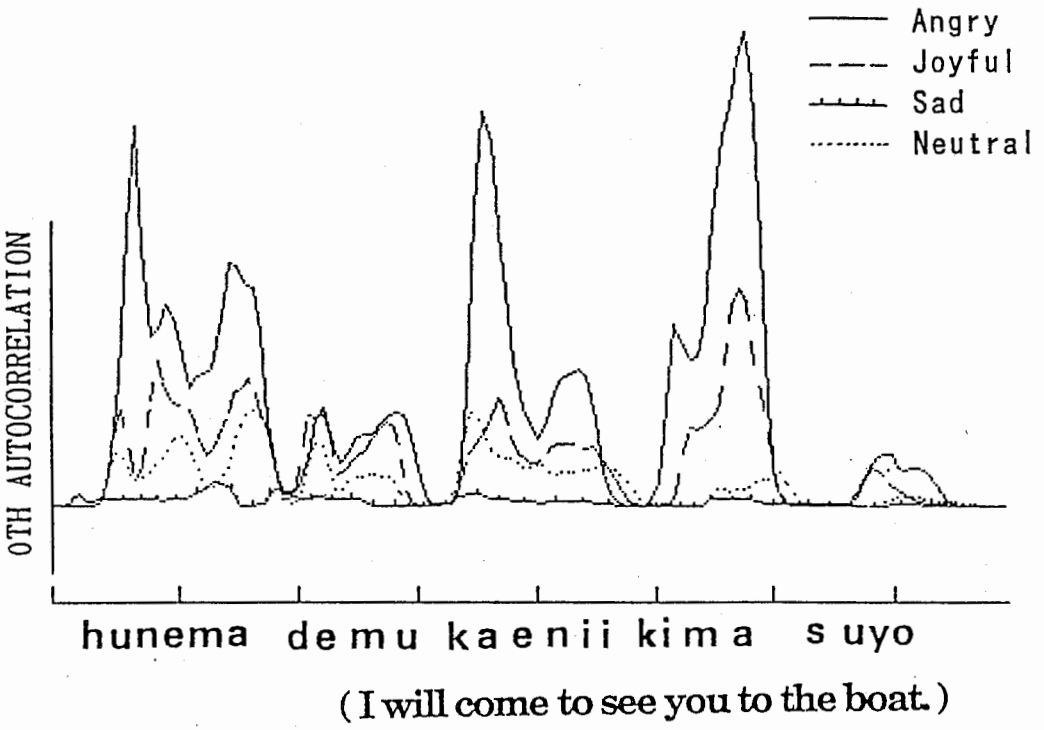


Figure 6. Power patterns in three kind of emotions and neutral speech

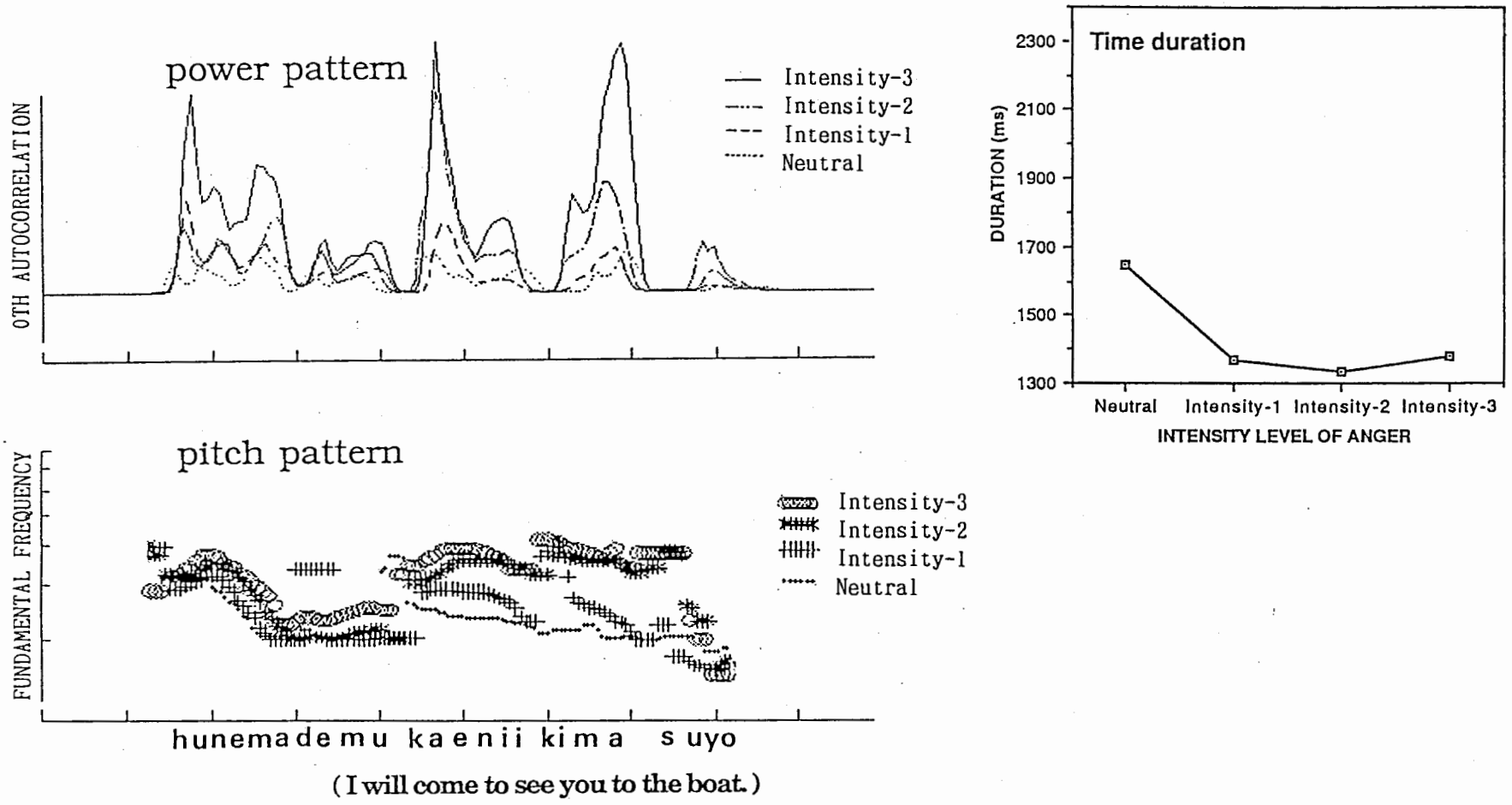


Figure 7. Intensity of "anger"

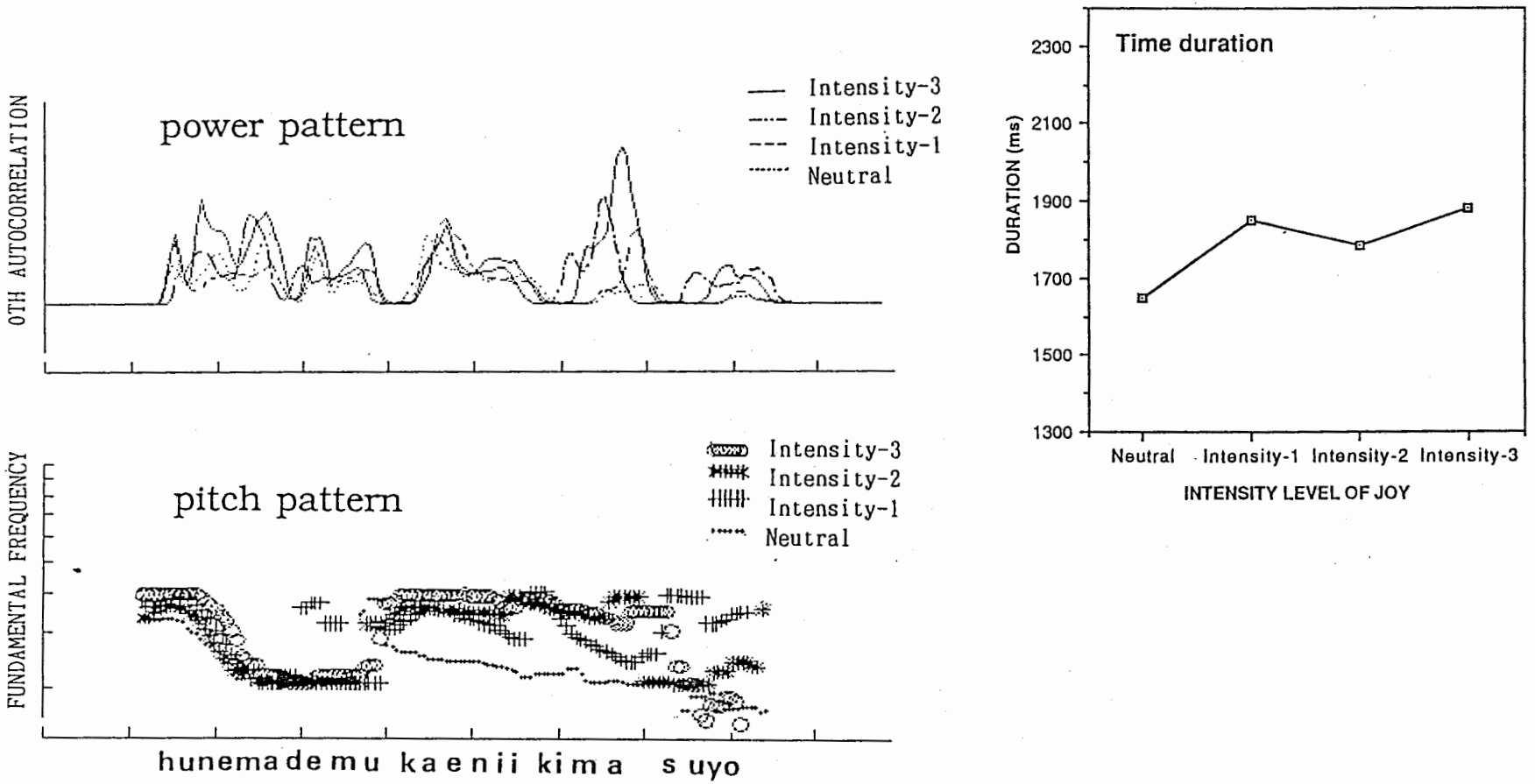
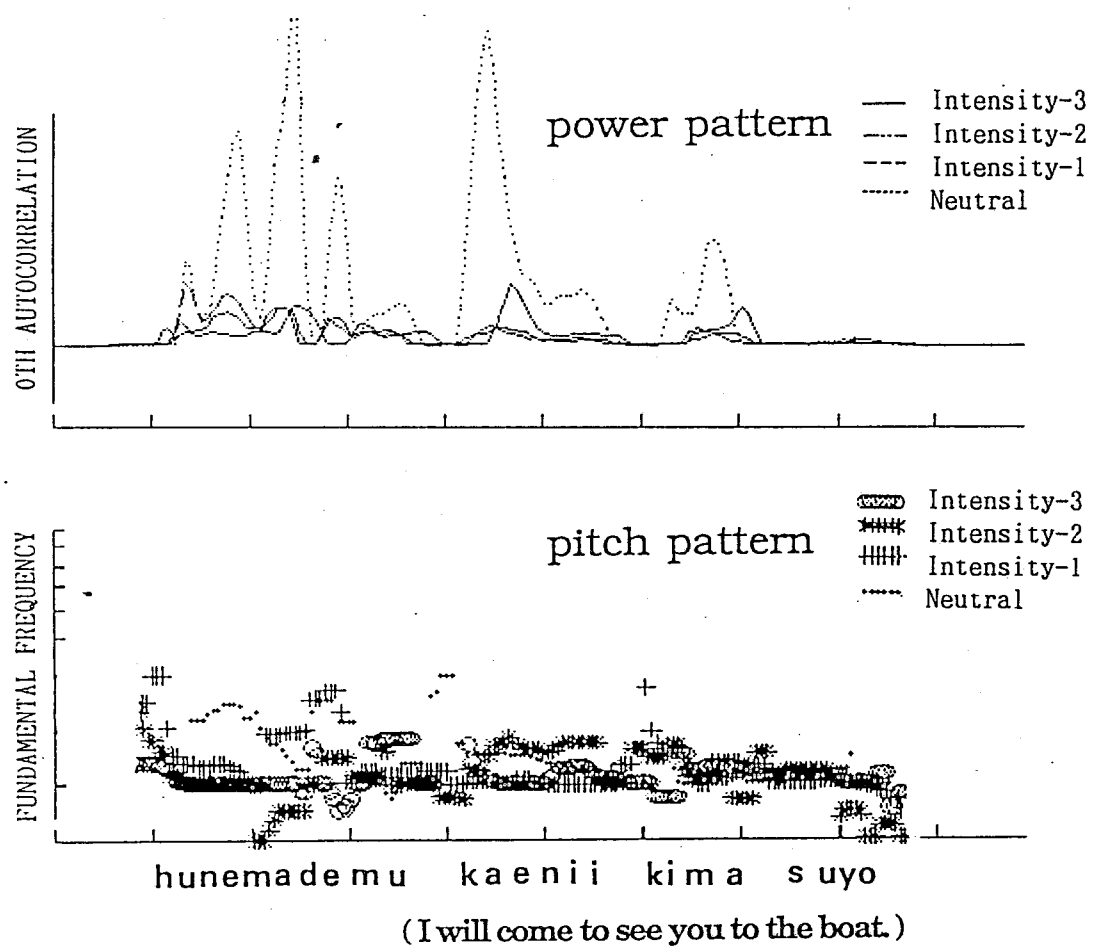


Figure 8. Intensity of "joy"



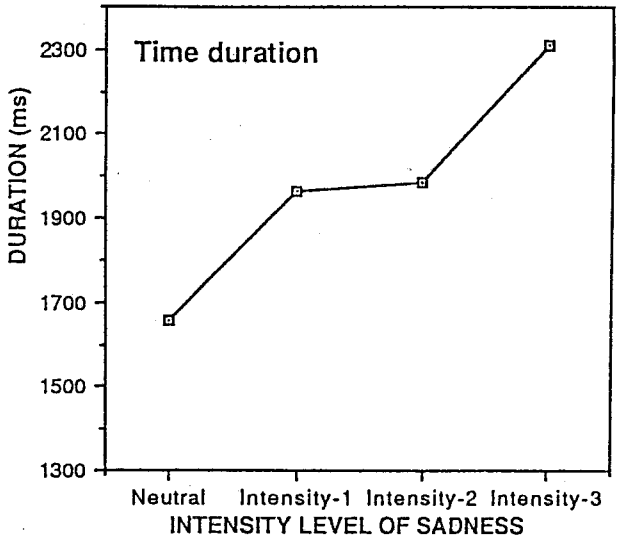


Figure 9. Intensity of "sadness"