PHONETIC AND PHONOLOGICAL CHARACTERISTICS OF PARALINGUISTIC INFORMATION IN SPOKEN JAPANESE

Kikuo Maekawa

Department of Language Behavior

The National Language Research Institute, Tokyo

ABSTRACT

Six paralinguistic information types uttered by three subjects were examined. Considerable changes were observed in duration, pitch, vowel formant, and voice quality. These findings require some reconsideration of the problems in the phrase phonology of Japanese like mora-based timing and phrase-initial pitch movement. On the other hand, lexically specified phonological features were more robust than phrasal features.

1. INTRODUCTION

The study of paralinguistic information is indispensable for the deep understanding of various aspects of speech. It is also indispensable for the implementation of speech synthesis and recognition systems. In this paper, I describe some important phonetic features of paralinguistic information of Japanese observed in an experimental setting.

There is not complete agreement among researchers as to what is meant by 'paralinguistic information'. Here, I follow the definition given by Fujisaki [1,2], according to which, paralinguistic information has two important features. First, paralinguistic information is not inferable from the written counterpart and is deliberately added by the speaker. Second, the strength of a given paralinguistic information type can vary continuously within one and the same category.

One consequence of this definition is that what is generally referred to as 'emotion' is excluded from the paralinguistic information, because emotion can not be controlled deliberately. On the other hand, so-called focus is a part of paralinguistic information, because the placement of focus is a deliberate choice of speaker and the strength of focus can vary continuously from weak focus to very strong focus.

2. DATA COLLECTION

Ten Japanese sentences were used for data collection. Three of these sentences given below are analyzed in this paper.

1) *so'H* *desu* *ka*
so copula particle (*Is that so?*)

2) *Ya'mano-saN desu ka* (*Is that Mr. Yamano?*)

3) *ana'ta desu ka* (*Is that you?*)

Sentences 1)-3) share the same syntactic structure: '*desu*' is the polite form of the copula, and '*ka*' is a sentence particle. These sentences can have different pragmatic meanings, e.g., ordinary question, rhetorical question, proposal, request, blame accompanied by surprise, etc. depending on the interpretation of the particle. In the above English translations, /ka/ is interpreted as an ordinary question.

In sentence 1)-3), the symbol 'H' indicates the second element of a long vowel, and the apostrophes indicate the location of the lexical pitch-accent on the noun before the copula. In 1) and 2) the noun is accented on the first mora, and in 3) on the second. The first syllable of 1) is heavy (consisting of two morae), while the first syllable of 2) and 3) are light (one mora).

Three speakers of Standard Japanese read these sentence ten times in random order trying to realize the following six paralinguistic information types: ADMIRATION (Implication is "*That's great!*"), DISAPPOINTMENT ("*Forget it!*"), SUSPICION ("*I don't believe it*"), INDIFFERENCE ("*I'm not interested*"), FOCUSED, and NEUTRAL. These types will be referred to as A, D, S, I, F, and N respectively below.

The three speakers were all teachers of Japanese as foreign language with knowledge of Japanese phonetics. Speakers ST and YS were male and JH was female.

While it was relatively easy for the speakers to understand what was meant by paralinguistic labels A, D, and S, it was more difficult to explain types F and N. I explained type N as an utterance that has no specific paralinguistic meaning, which sounds rather wooden. In the recording, speakers YS and JH uttered type N as a declarative sentence ending in falling pitch contour, while ST uttered type N as ordinary question ending with rising pitch.

I explained type F as an utterance similar to type N, but pronounced with greater overall vocal strength. The instructions given to the speakers and to the subjects of perception test described below are in an additional file [IMAGE 0997.GIF].

Sentences 1) and 2) were recorded 180 times (i.e., 6 information types with 10 repetitions uttered by 3 speakers), and sentence 3) was recorded only by subject ST. The recording was done in a sound-proof room.

3. PERCEPTION TEST

In order to check the validity of the recorded material, a perception test was conducted. All repetitions of the three sentences by the three speakers were presented in random order to seventeen Japanese-speaking subjects. The subjects were asked to identify the paralinguistic information intended by the speakers in a forced multiple-choice format.

The intended information was perceived correctly in at least 80% of the cases for all types, with the exception of type F, which was perceived correctly only 59% of the cases, being mostly confused with type N. This confusion occurred mostly in the utterances of speaker YS, whose type F utterances were perceived as N 75% of the cases. A similar tendency was observed in the data of speaker JH, whose type F utterances was perceived as N 22% of the cases. Also, 19% of JH's type A utterances were perceived as N. Utterances with correct perception rates lower than 50% were excluded from the ongoing acoustic analyses. The numbers of utterances excluded were 2 for speaker ST, 11 for JH, and 20 for YS.

4. ACOUSTIC ANALYSES

The utterances used in the acoustic analyses were digitized with 16bit-16kHz sampling condition via a DAT interface connected to a Sun workstation and analyzed using Entropic's *esps* speech-analysis package.

4.1 Duration

Figure 1 shows the mean duration obtained for sentence 2) of speaker ST. In these figures, the duration of constituting morae is shown by the shading in the bar. Type N and F had almost the same duration, type I was significantly shorter than N and F, and types A, D, and S were longer than the others. The same result was obtained in all of the sentences of all the speakers.

An interesting finding was the non-linear relationships between the duration changes of the whole sentences and those of constituting morae. The first and last morae, e.g., /ya'/ and /ka/ were the most elastic, again, in all sentences of all speakers.

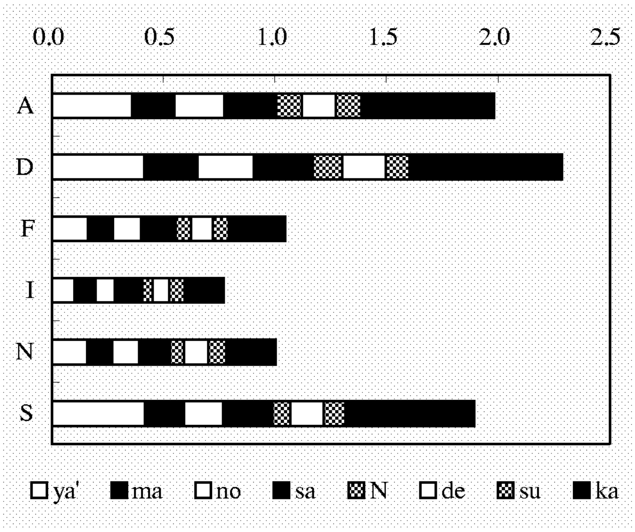

Figure 1: Mean duration of ST's utterances Average of ten repetitions. Abscissa in [sec].

4.2 Pitch

The top six panels of Figure 2 show the typical pitch contours of sentence 1) uttered by speaker ST. In type N, the pitch begins high and falls at the end of the first syllable (/so'H/), because the first mora (/so'/) is lexically accented. When the accent is on the second mora, as in sentence 3), the pitch begins low and rises to the accentual peak and falls at the location of accent, as in the last panel of Figure 2. These panels give the neutral or 'canonical' pitch shapes that determine the linguistic information of the sentences. The same kind of pitch shapes were observed for types F and I.

• **Pitch Range** Pitch range is strongly compressed in type D, and enlarged in A, F and S. I summarize the variation of pitch range in Table 1. I define a pitch range as the difference between the pitch value of the accentual peak and the lowest pitch value observed during the first *n-1* morae of a sentence, where *n* denotes the mora length of the sentence. I excluded the last mora (/ka/) to prevent differences in the final renditions from influencing the pitch range computation.

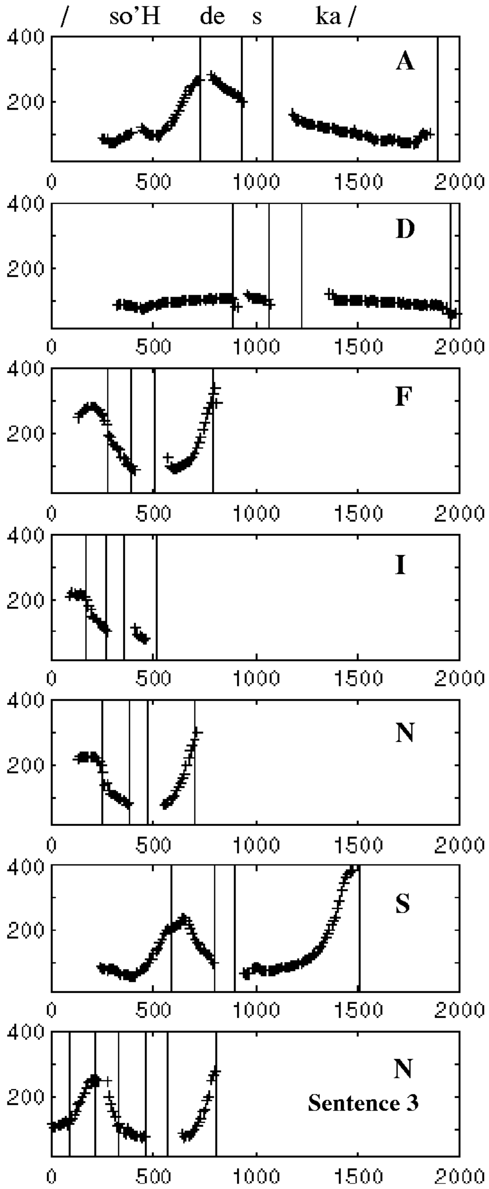

Figure 2: Top six panels are typical pitch contours of sentence 1). The last panel is type N of sentence 3. All uttered by ST. Time axis is in millisecond, and zero point corresponds to the beginning of /s/ for the top six panels. Vertical lines denote mora / syllable boundaries. [SOUND 0997.WAV]

Table 1 shows that the manipulation of pitch range, when uttering sentence 1), depends, to a certain extent, on the speakers. Speakers ST and JH enlarged their pitch ranges for type A, while speaker YS did not. Also, for types A, F, and S, all speakers used wider pitch range in the production of sentence 2) than in sentence 1).

• **Phrase Initial Rise** In Figure 2, the pitch contours of A and S begin very low and maintain this low pitch for a while, clearly deviating from the 'canonical' pitch shape. The same characteristic was observed for sentence 3), which is not initially accented. For sentence 3), the initial pitch is lower in types A and F than in N, F, and I; and this low pitch is maintained for a while before beginning to rise to the accentual peak.

Figure 3 summarizes the differences in pitch between the accentual peak and the lowest pitch preceding the accent. The magnitude of phrase initial rise is by far greater in types A and S than in N. ANOVA showed high significance for the difference (F=37.51, d.f.=2, p<0.001).

Table 1: Pitch range variation across information types. Mean±S.D. in Hz. All utterances of sentence 1) with Type F of speaker YS are excluded as a result of perception test.

Speaker	Type	STC1	STC2	STC3
ST	A	175±29	253±32	244±26
	D	45±12	46±14	64±39
	F	188±12	219±9	204±12
	I	136±16	189±26	188±36
	N	138±13	182±38	156±20
	S	213±57	340±29	267±43
YS	A	111±35	147±40	---
	D	41±12	94±27	---
	F	---	165±4	---
	I	97±17	176±19	---
	N	142±13	104±12	---
	S	162±91	259±45	---
JH	A	305±68	428±37	---
	D	89±17	136±20	---
	F	238±14	291±46	---
	I	298±30	357±39	---
	N	188±26	169±35	---
	S	264±67	360±96	---

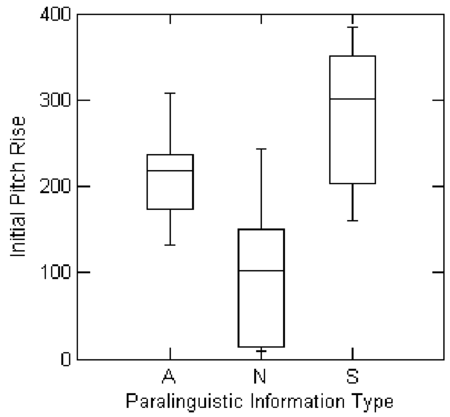

Figure 3: Influence of paralinguistic information on phrase initial pitch rise given in Hz. Pooled data over all sentences of speaker ST.

- **Peak Timing** Presence of non-canonical phrase initial rise pushes the location of pitch accent rightward on the time-axis. In Figure 2, the accentual peaks of type A and S were located in the time domain of the following mora.

Figure 4 summarizes the variation in accentual peak location. It shows the time difference between the location of the accentual peak and the right edge of the vowel to which the accent is linked at the level of phonology. Accordingly, minus values indicate that the peak is within the acoustically determined time domain of the accented vowel, and plus values, that the accentual peak is realized in the domain of the following segment(s). The figure reveals that in types A, D, and S, the location of the accent peak was significantly late in comparison to types N, F, and I (F=13.04; d.f.=5; p<.001.). This tendency was observed for each speaker in all of the sentences.

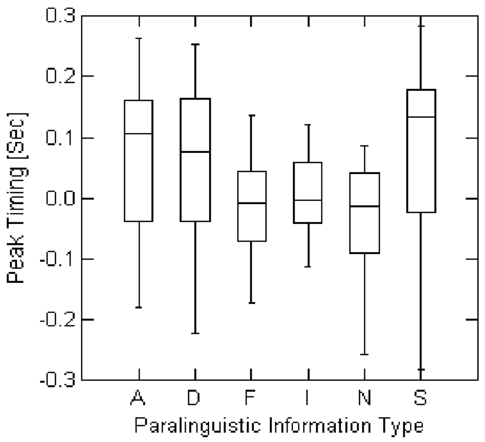

Figure 4: Timing [sec] of accentual peak relative to the end of the phonologically accented vowel. The data was pooled over all speakers and all sentences.

4.3 Vowel Spectrum

The influence of paralinguistic information was observed in segmental domain as well. Figure 5 is a F1 (first formant frequency) versus F2 scatter plot of the vowel in the particle /ka/ of sentence 1) by speaker ST. The mid-point of each vowel was selected for analysis by the covariance-based LPC method (order=20). Obtained formant values were manually checked by comparison with the DFT spectra.

Vowels of type A and D utterances had lower F1 and F2 values than types S and I. MANOVA showed high significance for these differences among the six types (F=34.84; d.f.=6,66; p<.001). It also showed high significance for speakers YS (F=9.90; df=8,84; p<.001) and JH (F=9.90; df=10,76; p<.001).

According to articulatory phonetic introspection vowels are more front in type S and more back in type A utterances, and the spectral analysis shown above supports this introspection.

To rule out the possible criticism that the observed differences in formant frequencies are a mere reflection of differences in vowel pitch, I summarize the mean pitch values at the formant estimation points in Table 2. Given that the mean differences between types S and A is less than 10Hz, the above criticism can hardly be supported. The same conclusion can be drawn for any combinations of A, D and I, S.

Table 2: Mean pitch values at the formant estimation points in Figure 5.

Type	Mean ± SD [Hz]
A	102 ± 14
D	99 ± 7
F	117 ± 21
I	84 ± 10
N	109 ± 13
S	96 ± 8

4.4 Voice Quality

The last phonetic characteristic to be noted is the change in voice quality. In some utterances of some speakers, vowels underwent remarkable change in voice quality, which I will tentatively refer to as 'laryngealization'. This phenomenon was observed, typically, in the first syllable of types S, and sometimes in types A and D in speakers ST and JH. Figure 6 compares two vowel waveforms of the initial mora of sentence 2), /ya'/, with and without 'laryngealization'. These examples were both uttered by ST, and taken from the mid-portion of the vowel segments. Waveform of the 'laryngealized' vowel, excerpted from a type S utterance, is irregular and has smaller amplitude than its non-laryngealized counterpart excerpted from a type N utterance. This result suggests that quantitative analysis of voice quality difference should be the theme of further study.

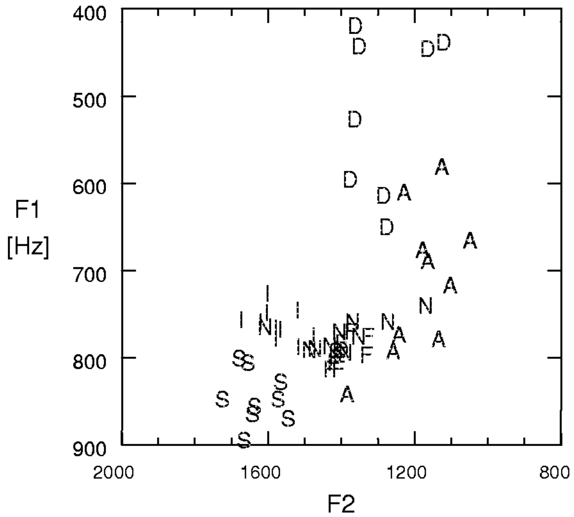

Figure 5: Influence of paralinguistic information on vowel formant frequencies in the case of speaker ST. Information types are given by the letters used as plot symbols. The distributions of types F and N are difficult to see due to considerable overlapping.

5. DISCUSSIONS AND CONCLUSION

I have demonstrated that paralinguistic information influenced various aspects of speech, i.e., duration, pitch, frequency-spectrum, and voice quality. These influences are not previously reported, with the exception of the phrase-initial pitch movement, which was previously reported by Kawakami as early as in 1956 [3].

These findings indicate the need to reconsider some problems of the Japanese phrase phonology. For example, so-called mora-based isochrony was violated in the case of paralinguistic conditions for types A, D, and S in the current data set.

In addition, the stretch of low flat pitch at the beginnings of types A and S suggests the need to modify the current treatment of phrase-initial pitch movement in the analysis of Japanese intonation. Distinction between the weak and strong phrasal L tones can not be described adequately considering only linguistic information such as syllable weight [4]. So-called weak L appears only when paralinguistic information allows it to appear.

On the other hand, it is to be noted that even the strongest paralinguistic modifications (i.e., those triggered by types S, A and D) hardly altered the presence and location of accent at all. This is probably due to the fact that accent is a property of lexical phonology in Japanese. Additional evidence showing the robustness of lexically specified phonological contrast was found in the formant data. The wide dispersion of the /a/ vowel on the formant plane shown in Figure 5 did not overlap considerably with other vowels uttered in the same sentence.

In conclusion, paralinguistic information is a very important variable of the study of speech phenomena. We can not expect to understand the real nature of speech, whether phonetic or phonological, unless we pay more attention to the relevance of paralinguistic information. This point may be all the more important in the analysis of so-called spontaneous speech and dialogue speech. This preliminary report suggests the need to enlarge the scale of investigation, and to develop new ways of analysis, especially for voice quality.

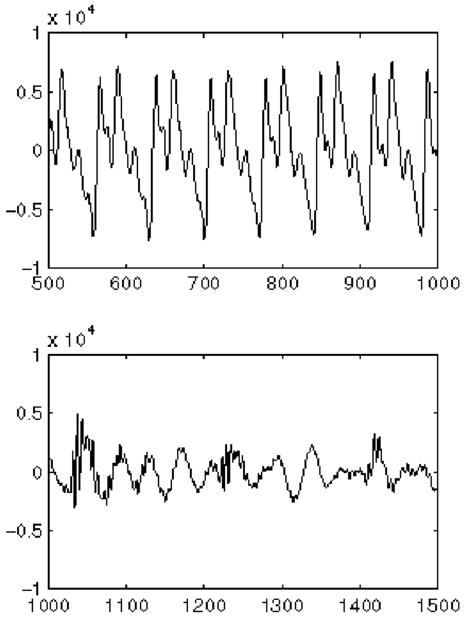

Figure 6: Waveform of normal (top) and 'laryngealized' (bottom) vowel. Speaker was ST. Each panel corresponds to about 30 ms.

6. REFERENCES

1. Fujisaki, H. "Aspects of Prosody Research and Topics for Further Investigation." *Proc. 1994 Autumn Meeting of ASJ*, 1: 287-290, 1994.
2. Fujisaki, H. "Prosody, Models, and Spontaneous Speech." In Sagisaka et al ed. *Computing Prosody*. Springer, 1997.
3. Kawakami, S. "Buntou no intoneeshon" *Kokugogaku*, 25: 21-30, 1956.
4. Pierrehumbert, J. & Beckman., M. *Japanese Tone Structure*, MIT Press, 1988.

Acknowledgment: Part of this work is supported by the grant-in-aid for scientific research from the Ministry of Education to the present author (Grant # 10410110).