Production and Perception of 'Paralinguistic' Information

Kikuo Maekawa

Department of Language Research
National Institute for Japanese Language
kikuo@kokken.go.jp

Abstract

Phonetic manifestation of paralinguistic information (PI) like speaker's attitude and intention is a unique property of speech communication. Production and perception of six PI types were examined using Japanese.

In speech production, acoustic and articulatory analyses revealed that speech signal and the underlying articulatory gesture differed systematically and considerably under the specification of PI. Further it was shown that the planning of PI could be classified into two different processes; one that makes reference to phonological structure of utterance, and the other that does not.

As for perception, identification experiments followed by MDS analysis revealed that native subjects could identify the PI types correctly in three dimensional perceptual space, and, regression analysis revealed high correlation between the acoustic measures and the perceptual space.

Lastly, cross-linguistic perception experiments followed by MDS analyses revealed partly language-dependent nature of PI perception. This finding was in congruence with the finding that production of PI makes partial reference to the phonological structure of utterance.

1. Introduction

An English phrase "Really", for example, can transmit non-textual meanings like "I don't believe it", "I'm surprised to know that", "I'm disappointed to know that", and so forth, depending on the way it is realized phonetically. This kind of information like speaker's mental attitude or intention could be called 'paralinguistic' information, or PI. An important characteristic of PI is that it is a volitional message: PI is manifested under the deliberate will of the speaker like linguistic information [1]. On the other hand, PI is different from 'non-linguistic' information (NI) like speaker's sex, age, and most part of emotions, because non-linguistic information is expressed without speaker's volition (Note the denotation of PI as used in this paper is very different from that of the British usage of 'paralanguage' [2]).

Simultaneous transmission of linguistic, paralinguistic, and non-linguistic information is one of the most important characteristics of spoken language. It is hence expected that study of PI and NI occupy crucial area in phonetics, but it is not the case in reality. Traditional and today's phonetics did not pay enough attention for PI, and concentrated heavily upon the linguistic or intellectual aspect of language. Even the handbook of IPA reads,

Although phonetics as a science is interested in all aspects of speech, the focus of phonetic notation is on the linguistically relevant aspects. (p. 4 of [3])

This statement shows correctly the status of phonetics in the past and the present, but misses the fundamental goal of the discipline. It seems to the present author that phonetics in the coming era should be extended so that it is capable of describing all aspects of information involved in spoken communication. With this in mind, I will summarize the results of recent phonetic experiments about the production and perception of PI in Japanese [4-9].

2. Data

Speakers of Standard (or Tokyo) Japanese were asked to read ten to fifteen semantically neutral sentences aiming at the transmission of specific PI.

The specified PI includes *admiration* (A), *suspicion* (S), *disappointment* (D), *indifference* (I), in addition to *neutral* (N) and *focused* (F). The former four types were chosen because they were listed as the representative "emotions" in literature [10]. N and F were chosen as the reference utterances. N was explained to speakers as the 'utterance without any special implication,' and F was the 'same utterance as N but with your voice raised so that your interlocutor standing on the opposite side of a large room can hear you.' All other PIs were explained by rephrasing the intended paralinguistic message into textual message: A and S were rephrased as 'That's great. I love it' and 'I doubt it and I don't believe it', for example (See the appendix of [4] for details).

All speakers were trained until they could produce intended messages constantly. After this was done, the recording began. Speakers were given a pile of small cards specifying the combination of a sentence and a PI. All combinations were recorded at least ten times and in a randomized order.

The details of data recording differ to some extent from one experiment to another, but what was written above was common to all experiments.

As a whole, four speakers took part in the experiments. When necessity arises, they will be referred to as ST, YS, KM, and JH in the rest of this paper. ST, YS (males), and JH (female) are teachers of Japanese as a foreign language who have knowledge of phonetics but knew nothing about the aims of experiments at the time of experiments. KM (male) is the present author.

Perceptual screening was carried out before starting acoustic analysis. Forced identification test was carried out using three sentences (/so'Hdesuka/, /ana'tadesuka/, and /ya'manosandesuka/) uttered by ST, YS, and JH. The whole 436 utterances were presented to 20 naïve subjects in a randomized order. Table 1 shows that the overall correct perception rate (the diagonal elements of the table) was higher than 80% with the exception of F. Most of the confusion occurred between N and F, and concentrated particularly in the utterance of YS, most of whose F were identified as N. In the analyses reported in section 3.1 below, 33 utterances whose

correct identification rate was lower than 0.5 were excluded from the analysis.

Table 1: *Confusion matrix of PI identification test. Pooled data. Data cited from[8].*

	A	D	F	I	N	S
A	**0.89**	0.01	0.09	0.01	0.00	0.00
D	0.01	**0.99**	0.00	0.00	0.00	0.00
F	0.01	0.00	**0.59**	0.05	0.34	0.00
I	0.01	0.03	0.01	**0.81**	0.14	0.00
N	0.00	0.02	0.05	0.07	**0.86**	0.01
S	0.00	0.01	0.01	0.00	0.00	**0.98**

3. Production of PI

3.1. Acoustic analyses

3.1.1. Duration

Figure 1 shows the averaged duration of the constituent morae of the sentence /ya'mano saN desuka/ (*Is this Mr. Yamano?*) as uttered by ST. In this and other phonological notation, apostrophe is used to refer to lexical accent.

This figure shows relative duration of each mora with the averaged duration of N utterances set to 1.0. With respect to N, F is nearly the same, I is shorter, and, A, D, and S are by far longer. In addition to this overall tendency, there is a remarkable tendency that the above-mentioned durational change becomes prominent at both edges of an utterance, especially in the last mora. This tendency was observed for all sentences and subjects.

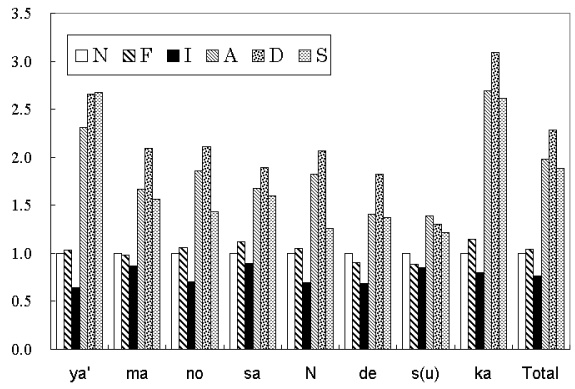

Figure 1: *Mora duration as a function of PI types. Speaker is JH. From [8].*

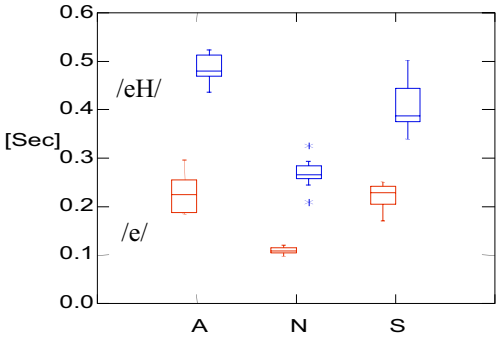

Figure 2: *Contrast of vowel quantity under three PI types. Speaker is ST.*

An interesting problem raised by this finding is the relation between the transmission of PI and the phonological contrast of vowel quantity (Japanese has the phonological contrast between the short and long segments). Is the contrast preserved under the durational change caused by PI?

Figure 2 compares the duration of the first syllables of sentences /e'desuka/ (*Is this a picture?*) and /e'H desuka/ (*Is this (letter) A?*) uttered by the same speaker, where /eH/ denotes phonological long vowel as opposed to the short /e/. This figure shows clearly that the phonological contrast of quantity is maintained even when there is prominent segment elongation caused by PI. Other subjects who read this sentence maintained the contrasts also.

3.1.2. F0 contour

Figure 3 shows typical examples of the F0 (speech fundamental frequency) contours of sentence /so'Hdesuka/ (*Is this so?*) uttered by ST. The utterances N and F show canonical F0 contours of one accentual phrase utterance of Tokyo Japanese beginning with an accented heavy syllable. Since this utterance begins with a heavy syllable (i.e. a syllable consisting of two morae, a long vowel in this case) and the syllable is accented, F0 rise that usually marks the beginning of an accentual phrase cannot be seen clearly in N and F (see [11] for the weakening of AP initial rise). This is also true in utterance I. On the other hand, however, long stretch of low F0 was observed in utterances A and S before the F0 went up for accentual peaks. Figure 4 compares the difference of pitch range caused by phrase-initial rise.

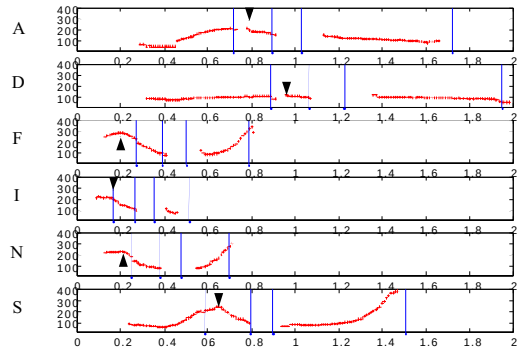

Figure 3: *Typical fundamental frequency contours of /so'Hdesuka/ under six PI types. Vertical lines show syllable boundaries. Speaker is ST. From [8]*

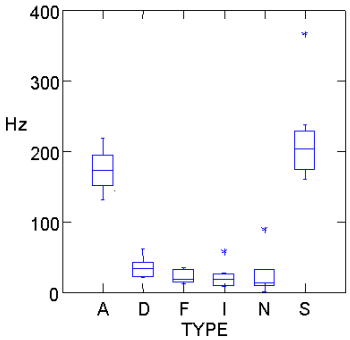

Figure 4: *Comparison of the magnitude of phrase-initial pitch rise in /so'Hdesuka/. Speaker is ST. From [8].*

There seems to be three more F0 characteristics of the manifestation of PI. Comparison of the utterance final F0 rise

in N, F, and S reveals that two different types of rising contour are used. Contour found in S consists of sustained low F0 followed by a rise, while the contour found in N and F is a simple rise. These two rising contours were first described by Kawakami [12] and represented as L%H% and L%LH% respectively in the X-JToBI intonation labeling scheme [13].

The triangles is figure 3 denote the timing of accentual F0 fall, and figure 5 shows that the timing differs systematically according to the type of PI: timing in A, D, and S is systematically later that that in N, F, and I.

Finally, there is manipulation of overall F0 range, which is most typically observed in the quite narrow range in D. At this point, it is also to be noted that it might be the case that, in utterances ending in rising contour, the F0 range of the main utterance and that of the last mora have different F0 range values. As seen in utterance S of figure 3, the F0 value of H% (i.e. the peak of final rising) is much higher than the H of the preceding body (i.e. the peak due to lexical accent). Similar relation can be observed in N, and, to lesser extent in F, also. These observations about F0 will be utilized later in the regression analyses in 4.2.

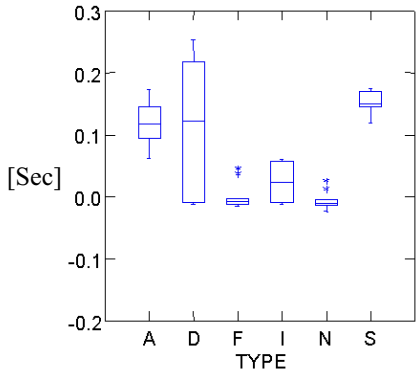

Figure 5: *Comparison of the timing of accentual pitch fall in /ana'tadesuka/.. Ordinate's zero corresponds to the end of accentedsyllable /na'/. Speaker is ST. From [8].*

3.1.3. Vowel formant frequency

So far, we have examined so-called prosodic features, but so-called segmental features also change according to PI.

Figure 6 shows the distribution of the first and second formant frequencies (F1 and F2) of four /a/ vowels contained in /ana'tadesuka/ ('*Is it you?*') uttered by ST. Only the samples of A, N, and S are plotted for the sake of visibility. The female subject (JH) was excluded form analysis, because the formant analysis was unstable due to high F0.

S and A samples are clearly separable on the F2 axis, suggesting relatively forward and backward tongue positions in S and A respectively. This tendency was observed for all male subjects. The Euclidian distance between the centroids of S and A samples differed according to the position of the vowels: it was the first and/or last vowels that the distance became the largest.

3.1.4. Spectral tilt

Perceptual impression suggests clear difference of phonation types in our data, at especially both edges of utterance. Figure 7 shows the difference of spectral tilt of vowels in sentence /sasadaga/ utterd by ST and KM that is a part of articultaroy data described in the next section. Here, spectral tilt was defined as H1-A3, where H1 is the level in dB of the first harmonics, and, A3 is the level of the harmonics whose frequency is the closest to the third formant. Larger and smaller H1-A3 values in D and S suggest larger and smaller spectral tilts respectively, which in turn suggests breathy and pressed phonation in these PI types.

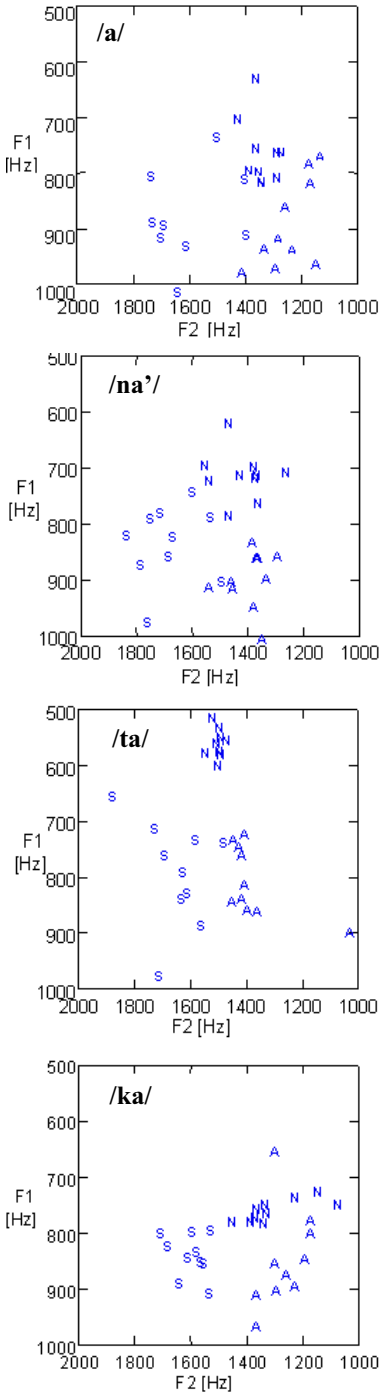

Figure 6: *Distribution of four /a/ vowels on the F1-F2 plane. Plot symbols denote PI types of S, A, and N. Speaker is ST. Data from [8].*

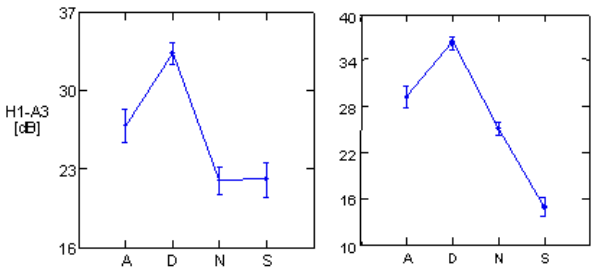

Figure 7: Average spectral tilt (H1-A3) of pooled four /a/ vowels in /sasadaga/. Vertical bar shows S.D. Speakers are ST (left) and KM (right).

3.2. Articulatory analyses

3.2.1. Tongue position

In order to examine if it is really the tongue position that yields the difference in figure 6, articulatory gestures of subjects ST and KM were recorded using the EMA device of the NTT basic research laboratory (courtesy of Dr. Masaaki Honda). Figure 8 shows the distribution of the tongue-dorsum censor of the four /a/ vowels in sentence /sasadaga/ (surname 'Sasada' followed by an particle) measured at the timing of maximum jaw opening. The separation between the A and S samples are literally perfect, showing that tongue shifts forwards and backwards respectively in S and A utterances.

Figure 9 shows the mean position of the same censor for all segments of the sentence. The same forward-backward difference can be seen between S and A samples in all segments covering both vowels and consonants.

3.2.2. Phonation

Imaging of laryngeal articulation was conducted to obtain direct evidence of phonation difference suggested by the spectral tilt analysis mentioned above. Vibration of the vocal fold was recorded digitally with the rate of 4500 frame/sec at the (late) Research Institute of Logopedics and Phoniatrics, University of Tokyo. One subject (KM) uttered one word sentences like /e'ki/ ('*station*') under three PI types, i.e., N, D, and S.

Figure 10 compares the time courses of glottal area among the three PI types during the last 200 frames of the /e'/ in /e'ki/. Glottal area does not reach the bottom line in D utterance, because closure of the vocal folds is incomplete due mainly to so-called 'glottal chink'. In N and S, the closure is complete, but S has smaller open quotient (longer closure) than N. These observations agree completely with the results of spectral tilt analysis. These characteristics were observed in both vowels of /e'ki/.

There is one more interesting finding about the glottal control. Usually, glottal area during the articulation of a voiceless consonant like /k/ is much wider than in vowel articulation, which is exactly the case in utterance N. But in utterances D and S, glottal area during the consonant is identical to or narrower than that of vowels. It is likely that, in these utterances, glottis is not controlled just for a segment but for utterances as a whole [9].

Acoustic consequences of these differences in glottal control are discussed in literatures [5] and [6].

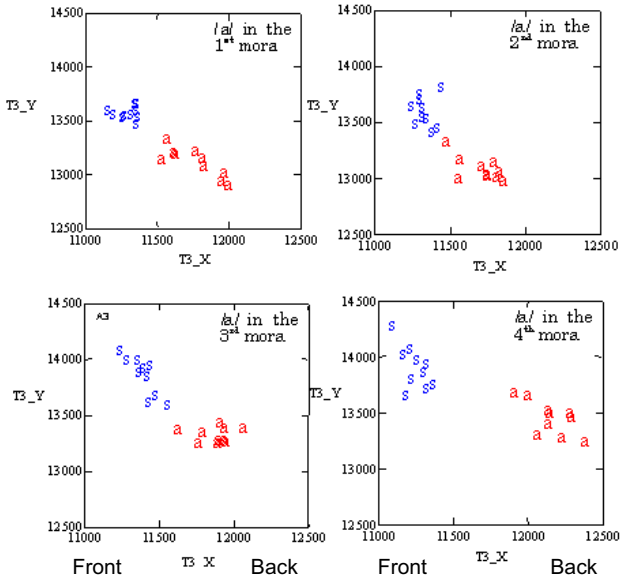

Figure 8: *Distribution of the tongue dorsum sensor T3 in articulatory space. Plot symbols stand for PI types. Unit of both axes is 10^{-2}mm. Speaker is KM. From [7].*

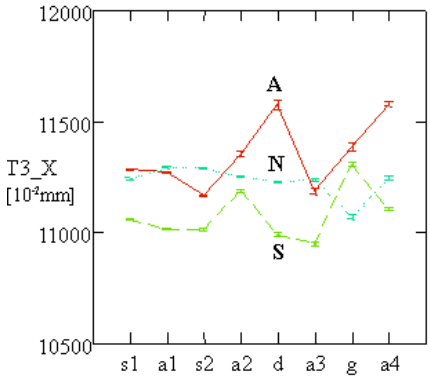

Figure 9: *Tongue-dorsum horizontal position averaged for each segment of /sasadaga/ as a function of PI type. Vertical bars show S.D. Speaker is KM. From [7].*

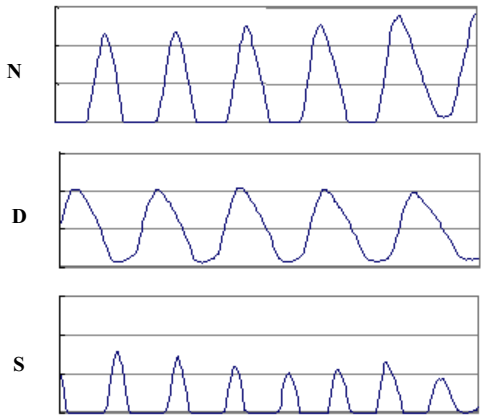

Figure 10: *Glottal area near the end of vowel /e'/ under three PI types. Abscissa corresponds to about 45ms. Ordinate is in arbitrary scale. Speaker is KM. Data reported in [9].*

4. Perception of PI

4.1. Perceptual space of PI

In order to construct perceptual space of PI, multidimensional scaling (MDS) techniques was used. Similarity matrix for the MDS analysis was computed using the identification data reported in section 2 above.

Similarity between any given two stimuli was defined as the probability that they were identified as belonging to the same PI type. In the example shown in table 2, the similarity is 0.6 because 6 subjects out of the total of 10 identified the two stimuli, *I* and *J*, as being the same PI type. Note, here, that similarity thus defined has nothing to do with the correctness of identification. This computation was conducted for all pairs and resulted in a triangular matrix of 436 rows and columns, which is the number of stimuli used in the identification experiment. This matrix was analyzed using the MDS procedure of the SAS system.

Table 2: *Example of similarity computation.*

	SUBJECTS									
STIMULI	1	2	3	4	5	6	7	8	9	10
I	**N**	S	**D**	F	**N**	**D**	D	**A**	S	**D**
J	**N**	N	**D**	D	**N**	**D**	A	**A**	D	**D**

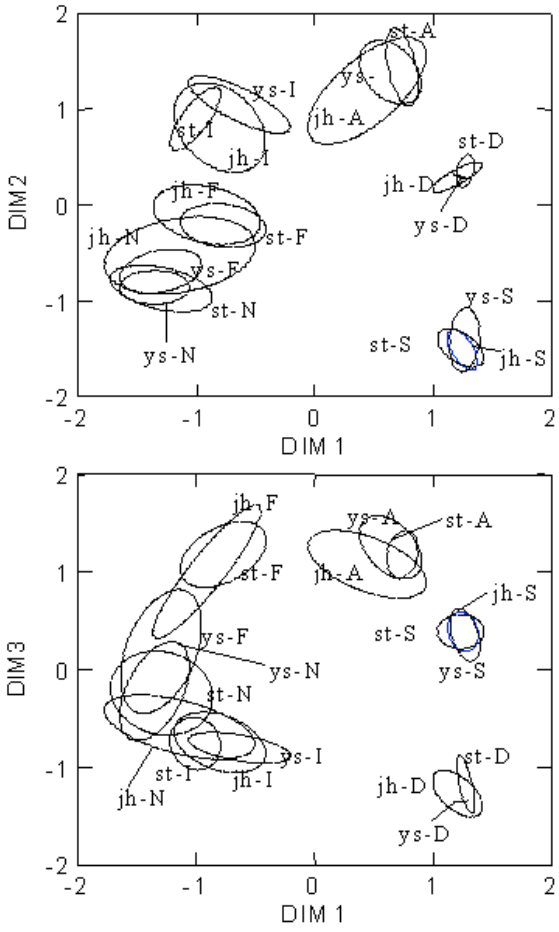

Figure 11: *Perceptual space derived by MDS. Japanese-speaking subjects. Cited from [8].*

Figure 11 shows the distribution of stimuli in the resulting 3-dimensional perceptual space (STRESS value was 0.04 when convergence measure was 0.1). In this figure, 68% probability ellipses were used to denote the distribution of six PI types for each speaker. Plot symbols like 'st-N' stands for the distribution of the stimuli of type 'N' uttered by speaker 'ST'. All PI types were separated clearly with the exception of F and N uttered by YS and JH. Also, stimuli of the same type uttered by different speakers are all close to each other.

DIM1 could be interpreted as the axis of the 'salience'. A, D, and S are all acoustically salient because they are marked with phonetic features like longer duration, higher intensity, characteristic pitch movement at the end, and characteristic phonation. DIM 2 could be interpreted as the axis of 'query'. This axis separates distribution of S, which requires the answer on the part of interlocutor, and those of A and I, which do not require interlocutor's response. Lastly, DIM3 could be interpreted as the axis of 'loudness'. While A, F and S are characterized by their higher loudness; N, I, and D are marked by lower loudness.

Table 3: *Result of regression analyses using all dependent variables. Each cell shows standardized partial regression coefficient. Bold and underlined cells were significant at the 0.01 and 0.001 levels respectively (both sides).*

INDEPENDENT VARIABLES	DEPENDENT VARIABLES		
	DIM1	DIM2	DIM3
F0	0.117	0.037	**0.299**
F0_I	**-0.210**	0.049	0.149
F0_F	0.035	-0.139	-0.137
PR	-0.034	**0.470**	0.208
PR_I	-0.013	-0.010	**0.302**
PR_F	-0.044	**-1.047**	-0.046
RMS	-0.098	0.051	**0.497**
RMS_I	0.045	0.153	-0.211
RMS_F	0.081	0.080	0.069
DUR	**0.616**	0.047	**0.173**
DUR_I	0.041	-0.083	0.168
DUR_F	**0.408**	**0.371**	**-0.208**
Peak Timing	**0.204**	**0.180**	**-0.180**
F1_Dev	0.046	0.011	0.061
F2_Dev	**0.103**	**-0.242**	**-0.286**
F3_Dev	-0.009	0.036	0.144
R^2	0.892	0.601	0.642

Table 4: *Result of regression analyses without using independent variables of formant.*

INDEPENDENT VARIABLES	DEPENDENT VARIABLES		
	DIM1	DIM2	DIM3
F0	0.217	0.032	**0.671**
F0_I	**-0.254**	0.003	-0.152
F0_F	0.032	-0.147	**-0.359**
PR	0.027	**0.415**	**0.293**
PR_I	0.035	0.119	-0.066
PR_F	-0.071	**-1.050**	0.054
RMS	-0.059	0.009	**0.625**
RMS_I	0.058	0.106	-0.186
RMS_F	0.057	0.127	0.109
DUR	**0.598**	0.020	**0.198**
DUR_I	**0.116**	-0.075	0.139
DUR_F	**0.425**	**0.374**	-0.137
Peak Timing	**0.264**	**0.169**	**-0.212**
R^2	0.876	0.645	0.519

4.2. Regression analysis using acoustic measures

In order to clarify the relationship between the structure of perceptual space and the acoustic characteristics of speech signal, regression analyses were conducted using various acoustic measures as independent variables. Three separate analyses were conducted for DIM1-3 using the value of each dimension as the dependent variables. Tables 3 and 4 show the results.

Independent variables include means of the F0, pitch range (PR), RMS amplitude (RMS), and duration (DUR). Means were computed not only for the whole utterance but also for the initial and last syllables. For example, 'F0_I' and 'F0_F' in the tables denote respectively mean F0 of the initial and last syllable. The timing of accentual pitch fall (PT) was measured as the distance between the accentual peak and the end of accented syllable. Lastly, deviations of the first three formant frequencies are used as the measures of formant characteristics using the vowel of the last syllable, which was fixed to /a/.

Table 3 shows the result of analysis that used all independent variables. Since formant measures were obtained only from male speakers (cf. 3.1.3), stimuli uttered by female speakers were excluded from dependent variables. The table shows that DIM1 is deeply correlated with duration. As for DIM2, pitch range of the last syllable is the predominant factor, and, DIM3 is mainly concerned with the overall RMS.

Table 4 shows the case where formant measures were excluded from independent variables (hence all three subjects' stimuli were used as dependent variables). The result is essentially the same as in table 3.

4.3. Case of non-Japanese subjects

Whether perception of PI is language-dependent like linguistic information or largely language-independent like emotion [14] is an interesting research question. In this section, perception of Japanese PI by English speaking subjects is examined.

The same identification experiment described earlier was conducted with two groups of subjects. The first group consists of 11 paid American English-speaking participants who stayed in Japan at the time of the experiment. They were language teachers (of either English or Japanese) and researchers of information sciences. Two of them were advanced level learners (i.e. passed level 1 proficiency examination or higher), but all the rest were intermediate level learners or beginners. This group is called 'learners'. The second group consists of 15 paid participants who have not learned Japanese at all. They were undergraduate students of the Black Hill State University in South Dakota. The latter group is called 'non-learners'.

The perceptual spaces of these two groups were derived using exactly the same analysis technique as in the case of Japanese subjects. STRESS values were 0.07 and 0.16 respectively for learners and non-learners. Figures 12 and 13 are the perceptual spaces of learners and non-learners. Unlike figure 11, speaker difference was not shown in these figures because separation of speakers was not observed clearly.

Learners' space has close resemblance to that of Japanese subjects. DIM1 could be interpreted as the axis of 'salience' as in the case of Japanese subjects. Learners' DIM2 and DIM3 seem to correspond to DIM3 and DIM2 of Japanese subjects respectively.

Non-learner's space is different, however. The most striking difference consists in that the perceptual space of non-learners is virtually two-dimensional.

Non-learner's DIM1 separates A, D, S and F, I, N as in figures 11 and 12, but the separation is far from being complete. Non-learner's DIM2 separates D from all other meanings. And on DIM3, which was not shown in the figure, all PI types overlap considerably to the extent that it was impossible to find any meaningful interpretation.

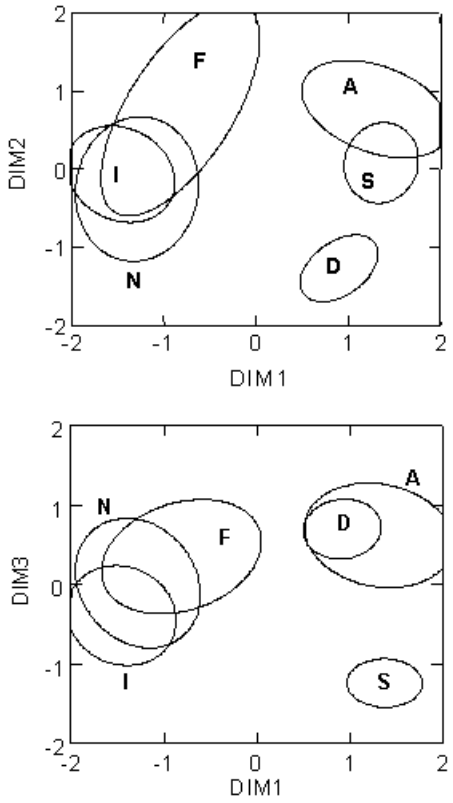

Figure 12 *Perceptual space of 'learners' derived by MDS. Only PI types are shown.*

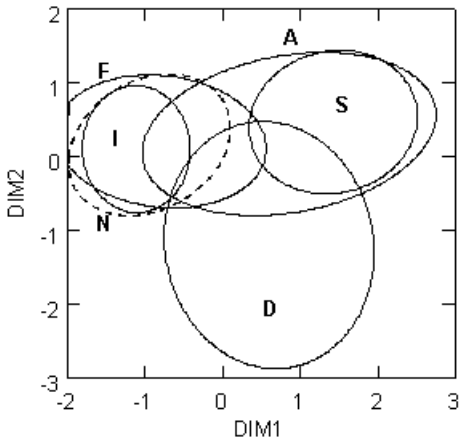

Figure 13 *Perceptual space of 'non-learners' derived by MDS. Only PI types are shown.*

5. Discussions

Because the study of PI has not been involved in the mainstream of phonetic/phonological studies, there are many

basic issues that are to be discussed. Among them, two issues that seem to be the most central of all will be discussed below.

5.1. Status of PI in speech production

One important objective of the study of PI is to fix the place of PI in the process of speech production, thereby obtaining wider view of speech production. Figure 14 is a conceptual model of speech production encompassing not only linguistic information but paralinguistic and non-linguistic information as well. Concerning the planning and implementation of PI, this model has two parallel paths: one that makes reference to LI and the other that does not.

Figure 14 shows that there is a constraint of word phonology in the planning of PI. The example of vowel quantity shown in 3.1.1 is a clear example of the constraint. Similarly, the delay of the timing of accentual pitch fall shown in 3.1.2 seems to be constrained by the accent location, which is a property of lexical item in Tokyo Japanese. Although accentual peaks of A and S located in the temporal domain of the post-accented syllable, it does not mean phonological shift of accent. Listening to these utterances, native speakers perceive accent in exactly the same location as in utterances N, F, and I. Moreover, informal perception test using F0 re-synthesis revealed that, if the location of peaks in A and S were delayed further so that they located near the end of post-accented syllable, re-synthesized stimuli were perceived as having incorrect accent. There seems to be a general constraint that planning of PI should not alter the linguistic contrasts specified at the level of word-phonology.

On the other hand, it is difficult to find evidence that the features of phrase-phonology work as the constraint on PI planning. On the contrary, it seems to be the case that manifestation of PI plays central role in the implementation of some phrase-level phonological features like phrase-initial pitch rise and phrase-final pitch movements. As noted in literature [12], speaker's mental attitude, as well as the syllable structure of the final syllable, plays central role in the selection of the shape of phrase-final pitch rise. Similarly, the enhanced pitch range of the phrase-initial pitch rise shown in A and S seem to express the paralinguistic meaning of speaker's 'involvement'.

At this point, it is important to note that, as long as the present author knows, the manifestation of emotion does involves manipulation of phonological, hence temporally localized, features; on the contrary, literatures suggest that emotion is realized globally throughout an utterance (See [15] and [16], among many others, for the global nature of the manifestation of emotion). This is the central reason why the present author wants to make distinction between emotions and PI.

However, it is not the intention of the present author to say that PI should be manifested locally. In fact, experiments reported in this paper showed clearly that PI could be manifested globally; the large contributions of independent variables DUR and PR in tables 3 and 4, and the overall articulatory shift and phonation difference shown in figure 9 and 10 are the examples. In figure 14, these phonetic features are referred to as the features of 'voice-quality'.

Lastly, two issues about figure 14 are to be noted. Firstly, since PI involves global manifestation of phonetic features, it is not always possible to make distinction between emotion and PI by means of phonetic analyses. In case of difficulty, the final cue is to be found in semantic analyses.

Secondly, there was at least one case where individual difference was found regarding the reference to LI in the planning of PI. In figure 9, subject KM showed forward versus backward articulatory shirt for all segments of /sadadaga/; and this was also true with the second subject ST. In utterances like /a'ki/ ('autumun') and /sa'ke/ ('salmon'), however, ST shifted only /a/ but not /e/ and /i/, while KM shifted all vowels in the same manner.

The simplest interpretation is that ST makes reference to the phonological specification of the words, while KM did not. Since /a/ is the only open vowel in Japanese, it is highly probable that the vowel is not phonologically specified with respect to its horizontal position; and the plausibility is that ST shifted only those vowels that were not specified in terms of its horizontal position, hence having high degree of freedom in horizontal direction.

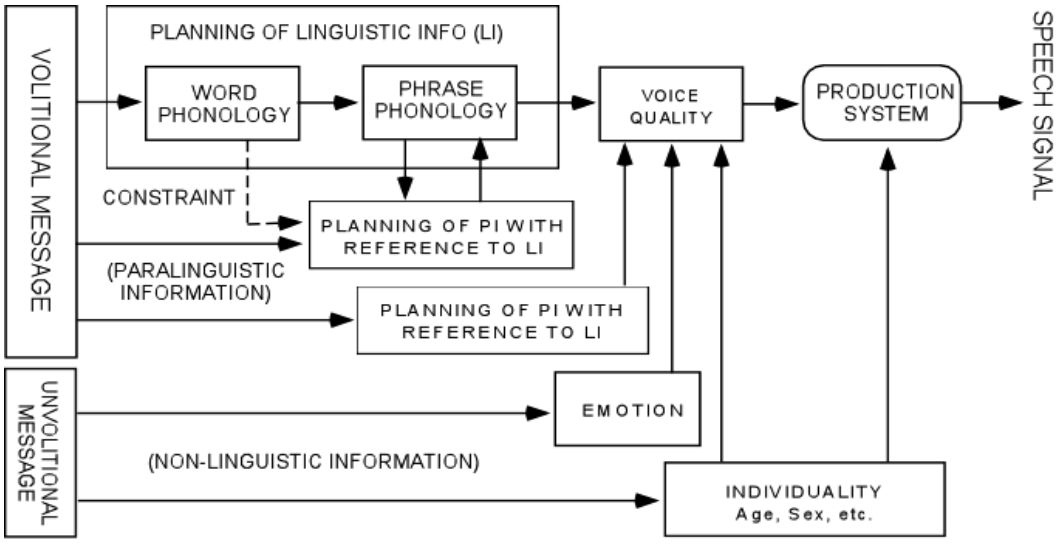

Figure 14: *A conceptual model of speech production covering linguistic, paralinguistic, and non-linguistic information.*

5.2 Language dependency

The model of speech production shown as figure 14 implies that PI production is language-dependent, at least partially, because it makes reference to LI (i.e. specifications of phrase-phonology). If this is true, perception of PI might be also influenced by LI. Results reported in section 4.3 seem to support this prediction.

As noted there, DIM1 of the native speakers was the dimension on which A, D, S were separated from others. And it was the overall duration (DUR) that made the greatest contribution to this dimension. Accordingly, it is natural to suppose that this dimension is concerned mostly with the PI as the voice-quality in figure 14, which should be perceived language-independently because its production makes no reference to linguistic information. As a matter of fact, this dimension was found for all subject groups (i.e., Japanese natives, learners, and non-learners).

DIM2 of native speakers (that corresponds to DIM3 of learners), on the other hand, should be regarded as language-dependent. The acoustic measure of PR_F that showed the largest contribution in regression analysis can be interpreted as the reflection of the choice of phrase-final boundary tones (either H% or LH%), which is a manipulation in phrase-level phonology. This dimension was less clear in the case of learners and not found at all in the case of non-learners who knew nothing about Japanese.

Lastly, DIM 3 of native speakers (that corresponds to DIM2 of learners) could be language-independent because 'loudness' is nothing but feature of voice-quality. It is curious that this dimension per se was not found in the perceptual space of non-learners, but DIM2 of non-learners, which distinguishes D from all others, may have the similar function.

It seems to be the case that non-learners who lack the knowledge of Japanese perceive only the PI that is not linked to LI (i.e. PI as voice-quality), while learners who learned something about the language structure can perceive the PI linked to LI as well. To sum up, the perception of PI as voice-quality is language-independent, or universal, like perception of emotion, while the perception of PI as manifested by the manipulation of the features of phrase-phonology is language-dependent.

6. Conclusion

Study of PI is an important but largely uninvestigated field of speech science. In this paper, I tried to show that production of PI is not a completely independent process from that of linguistic information.

Phonetic realization of PI is planned under the constraint of linguistic contrasts of word-phonology, and, the planning involves manipulation of phrase-level phonological features, most typically the tonal structure of utterance.

As a consequence of this, part of PI is language-dependent and is difficult to perceive correctly for those who lack the knowledge of the language in which PI is manifested.

This paper as a whole is a preliminary trial of systematic investigation of PI. There are many issues of PI that await further, and more systematic, investigations, among which, it seems to the present author, development of the semantics of PI is the most important of all [17].

7. References

[1] Fujisaki, H., 1997. Prosody, Models, and Spontaneous Speech, In Y. Sagisaka, et al., (Eds.), *Computing Prosody: Computational Models for Processing Spontaneous Speech*, New York: Springer, 27-42.

[2] Crystal, D., 1968. *Prosodic systems and intonation of English.* Cambridge Univ. Press.

[3] International Phonetic Association, 1999. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet.* Cambridge Univ. Press.

[4] Maekawa, K. 1998. Phonetic and phonological characteristics of paralinguistic information in spoken Japanese. *Proc. Int. 5th Conf. Spoken Lang. Processing*, Sydney, 2, 635-638.

[5] Kasuya, H.; Maekawa, K.; Kiritani, S., 1999. Joint estimation of voice source and vocal tract parameters as applied to the study of voice source dynamics. *Proc. 14th Int. Cong. Phonetic Sciences*, 2505-2512.

[6] Kasuya, H.; Yoshizawa, M.; Maekawa, K., 2000. Roles of voice source dynamics as a conveyer of paralinguistic features. *Proc. 6th Int. Conf. Spoken Language Processing*, Beijing, 2, 345-348.

[7] Maekawa, K.; Kagomiya, T., 2000. Influence of paralinguistic information on segmental articulation. *Proc. 6th Int. Conf. Spoken Language Processing*, Beijing, 2, 349-352.

[8] Maekawa, K.; Kitagawa, N., 2002. How does speech transmit paralinguistic information? *Cognitive Studies*, 9(1), 46-66 [In Japanese].

[9] Fujimoto, M.; Maekawa, K., 2003. Variation of phonation types due to paralinguistic information: An analysis of high-speed video images. *Proc. 15th Int. Cong. Phonetic Sciences*, Barcelona, 2401-2404.

[10] Mizutani, O.; Mizutani, N., 1979. *Aural Comprehension Practice in Japanese.* The Japan Times.

[11] Pierrehumbert, J; Beckman, M., 1988. *Japanese Tone Structure.* MIT Press.

[12] Kawakami, S., 1963. Bunmatsu nadono jooshoochooni tsuite (On final rise). *Kokugokenkyuu*, 16, 25-46.

[13] Maekawa, K.; Kikuchi, H.; Igarashi, Y.; Venditti, J., 2002. X-JToBI: An extended J_ToBI for spontaneous speech. *Proc. 5th Int. Conf. Spoken Language Processing*, Denver, 3, 1545-1548.

[14] Scherer, K. R., 2000. Cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology, *Proc. 6th Int. Conf. Spoken Language Processing*, Beijing, 2, 379-382.

[15] Williams, C.; Stevens, K., 1982. Emotion and speech: Some acoustical correlates. *Journal of the Acoustical Society of America*, 52(4), 1238-1250.

[16] Murry, I.; Arnott, J., 1993. Toward the simulation of emotion in synthetic speech. *Journal of the Acoustical Society of America*, 93(2), 1097-1107.

[17] Maekawa, K., 2002. Issues in the study of paralinguistic information. *Proc. Autumn Meeting of the Acoustical Society of Japan*, 247-250. [In Japanese]