

## Recognition of accent and intonation types of Japanese using F0 parameters related to human pitch perception

Carlos Toshinori Ishi\*, Nobuaki Minematsu\*\*, Keikichi Hirose\*\*\*

\* Dept. of Info. and Comm. Eng., Grad. School of Eng, Univ. of Tokyo

\*\* Dept. of Info. and Comm. Eng., Grad. School of Info. Science and Tech, Univ. of Tokyo

\*\*\* Dept. of Frontier Informatics, Grad. School of Frontier Sciences, Univ. of Tokyo

[c\_t\_ishi, mine, hirose]@gavo.t.u-tokyo.ac.jp

### Abstract

Based on a new approach of representing F0 in mora units ( $F0_{mora}$ ), several parameters (average and target values of F0 in CV and VC segments) were proposed. Then, a variable ( $F0_{ratio}$ ) was defined representing the F0 movement between two consecutive morae, and their distributions were analyzed and used to create models for each accent type. Evaluation results indicated that average values of VC units and target values of CV units showed the best performances in the accent type identification task. In order to investigate the causes of these results from a perceptual viewpoint, the candidates for  $F0_{mora}$  were checked considering how they were related to perceived mora pitch values. For this purpose, MIDI sounds were used as references to perceived mora pitch ( $F0_{human}$ ). Analysis on the mismatches between  $F0_{human}$  and the proposed  $F0_{mora}$  parameters showed mismatches especially when pitch change occurs within the syllables. As for the intonation type identification, several acoustic features were proposed to represent 6 types of sentence final tones, each conveying different information of subject's intentions and attitudes. The proposed acoustic features for relative duration and sentence final pitch change showed good correspondence to perceptual features.

### 1. Introduction

In Japanese, each word own a unique accent pattern in such a way that words with the same phoneme sequence can have different meanings depending on the accent type. While, intonation takes important roles not only in transmission of linguistic information such as syntax, but also in the transmission of paralinguistic information such as speaker's intention and attitudes. Therefore, a correct identification of accent and intonation is important not only to understand properly, but also to identify speaker's intentions, that would be useful for example in speech dialog systems. However, many researches showed the difficulties to apply prosody to speech recognition.

In past researches, several methods were proposed to identify Japanese accent types. [1] and [2] proposed the use of HMM to identify the accent type of words. F0 and delta F0 parameters are used as the input parameters for the HMM. Another approach for word accent-type identification was proposed in [3]. Global F0 contour is obtained by interpolation and smoothing of F0 by spline functions, and relative position (to the word length) of the F0 peak, and the declination of the F0 pattern are used as parameters to the identification. In [4], codebooks for F0 contour shapes of mora units plus delta F0 averages between adjacent morae

were used to model and recognize accent types of accentual phrases. Since the main topic of his research was the accentual phrase boundary detection, the accent type identification was limited to categorize only accent type 0, type 1 and the others. In our previous work [5], a new approach was proposed to represent F0 patterns in mora units, a method was developed to identify the accent type of isolated words based on perceptual thresholds for accent nucleus, and a CALL system was developed to train the pronunciation of Japanese lexical pitch accent. In this case, average values of F0 in the mora were used as the representative F0 value of the mora. The method worked well for isolated words. However, when applied to accentual phrases in sentence level, the performance degraded. In this study, several ways were investigated to estimate the best representative and quantitative F0 value of the mora, in order to recognize accent types in accentual phrases.

### 2. Japanese Pitch Accent

Accent is a relative positioning of prominence in pitch (or stress) for each word/phrase. In Japanese, there is a unique accent type for each word, which is defined as the relative positioning of pitch (high or low) along the mora sequence of the word. When it is produced in a sentence, a relative pitch height can also be assigned to each mora of the accentual phrase to describe its accent pattern. Figure 1 shows examples of accent types.

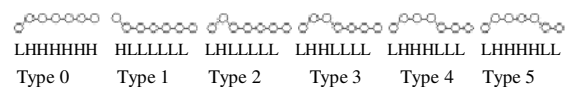


Fig. 1. Binary pitch height representation for Japanese accent.

In this study, we proposed an idea of estimating one representative value of pitch for each mora, called  $F0_{mora}$ . Also, in order to characterize the pitch movement along the mora sequence, we defined a variable called  $F0_{ratio}$  as the logarithm of the ratio of the  $F0_{mora}$  between two adjacent morae. By scaling to the musical scale (in semitone units),  $F0_{ratio}$  can be represented by:

$$F0_{ratio}(i) = 12 \log_2 \frac{F0_{mora}(i)}{F0_{mora}(i-1)}, \quad (1)$$

where  $i$  represents the mora position inside the word/phrase. Figure 2 shows the  $F0_{ratio}$  sequence for the accent types of the figure 1.

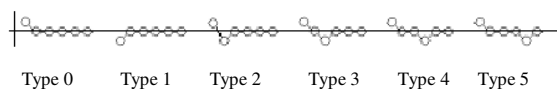


Fig. 2.  $F0_{ratio}$  sequence of the accent types of the figure 1.

## 2.1. Analysis of the accent types of phrases in a sentence

A database of continuous speech containing 503 sentences uttered by a male speaker was used to analyze the accent types of phrases. For analysis, we used the labeling information and F0 data contained in the database.

In order to find out the best definition for  $F0mora$ , investigations were conducted for several candidates, given as combinations of alternatives for F0 calculation method and segment definition; namely, taking average (avg) or target (tgt) values for CV and VC segments.

Average values are the simplest manner to obtain a representative value in a segment, but it doesn't take the movement of the F0 within the segment into account. The target value was introduced from the hypothesis that the most representative pitch value in a segment is the target value to which the pitch moves along the segment. This trend was observed especially in utterance beginning, where average values of F0 was higher in the second mora than in the first mora, being different from the actually perceived pitch. The average values were calculated using all F0's in the segment, while the target values were defined as F0 values obtained by the extrapolation of linear regression line of F0 contour to the segment final position.

Although CV is a basic unit (mora) of Japanese pronunciation from the linguistic viewpoint, Japanese rhythm seems to be realized in VC units rather than CV units from a perceptual viewpoint [6,7]. Thus, we decided to analyze both forms of the unit, considering that rhythm may be related to pitch perception.

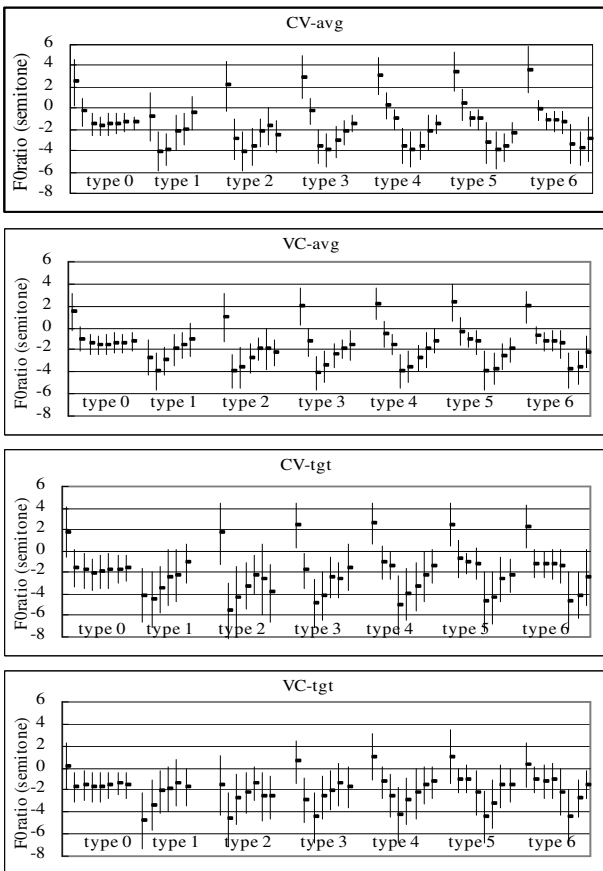


Fig. 3. Distributions (average and standard deviation) of the  $F0ratio$  sequence of each accent type.

Distributions of  $F0ratio$ , calculated using the different definitions of  $F0mora$ , were obtained for each accent type. Figure 3 shows these distributions (average and standard deviation) for each  $F0mora$  candidate (CV/VC; avg/tgt).

A minimum value in  $F0ratio$  sequence is expected to lie in the accent nucleus position. However, we can note in the figure 1 that the minimum value for type 1 locates between the second and third morae for CV-avg, CV-tgt and VC-avg cases. This is characteristics of accent type 2, meaning that the use of these parameters may cause confusions to identify these accent types.

## 2.2. Identification of the accent type of phrases in a sentence

### 2.2.1. $F0ratio$ based models

In order to create models to recognize accent types, we first categorized them according to the word/phrase length (number of morae  $n$ ) and the accent type ( $i$ ). Then, the  $F0ratio$  (Gaussian) distribution parameters (mean vector and covariance matrix) were used to build a multi-variate Gaussian Model for each category.

Thus, the model is constituted by the  $F0ratio$  distributions, and the input of the model is a  $(n-1)$  dimensional vector of  $F0ratio(j)$ ,  $j=2,3,\dots,n$ , where  $n$  is the word/phrase length in mora number. That is, one word/phrase is represented by one point in the  $(n-1)$  dimensional space.

Each phrase of length  $n$  in the training set was also represented as one point in the  $(n-1)$  dimensional space, and the distribution of these points (phrase set of the same category) was approximated by a multi-dimensional normal distribution. In this way, an  $(n-1)$  dimensional normal distribution was provided for each accent type  $i$ , and these distributions were used in the accent type identification.

Further, two types of covariance matrix were examined, diagonal ones and full-covariance ones. In the former, the correlation between pitch movements separated by more than 1 mora is ignored, and in the latter, it is explicitly modeled.

### 2.2.2. $F0$ based models

As a baseline, we also built F0-based HMMs where a pitch curve was treated as a temporal sequence of frames of  $\log(F0)$  and/or  $\Delta\log(F0)$ . As for the HMM configuration, left-to-right duration controlled HMMs with  $n$  distributions were used. As input features, we used a 2-dimensional vector of  $\log(F0)$  and  $\Delta\log(F0)$ , called  $HMM_{ref1}$ , and another using  $\Delta\log(F0)$  only, called  $HMM_{ref2}$ . The last one was intended to make correspondence to the proposed  $F0ratio$ -based model, where only the differences between adjacent  $F0mora$  are used. Since the number of morae ( $n$ ) of the input phrase is known, only  $HMM(n,i)$  ( $0 \leq i \leq n-1$ ) will be used to the accent type identification.

## 2.3. Evaluation of the $F0mora$ candidates

As described in the previous section,  $F0ratio$ -based models for each  $F0mora$  candidate and the F0-based models were prepared to evaluate their performances. For training the model parameters, 90% of the speech database was used as the training set, and 10% as the test set. Since the test set is small, we rotated these 10% along all database, in the way that the experiments were repeated 10 times using different test sets. Table 1 shows the global results for these 10 experiments.

The results in table 1 shows that  $F0ratio$ -based methods had a better performance than the baseline method (F0-based HMMs), indicating that the use of representative values for mora units is more effective to the recognition task than using

straightly F0 values of each frame. Moreover, comparing the results for *CV-avg* and *VC-avg*, the latter showed a better performance, indicating that the pitch perception is possibly related to rhythm perception. However, for *VC-avg* and *VC-tgt*, the former had a better performance. This reason may reside in the method of the target value estimation, which is based on first regression analysis. Maybe a higher order regression analysis is more suitable. Among the *F0ratio*-based parameters, *CV-tgt* and *VC-avg* showed the best overall performances.

Table 1: Results of the accent type identification (in %) for several parameters.

<i>F0ratio</i> based	<i>VC-tgt</i>	69.5
	<i>VC-avg</i>	75.1
	<i>CV-tgt</i>	75.5
	<i>CV-avg</i>	68.0
F0 based	<i>HMM<sub>ref1</sub></i>	65.3
	<i>HMM<sub>ref2</sub></i>	57.4

## 2.4. Pitch perception experiment

Although some hypotheses with respect to several characteristics of speech perception had been taken into account in the definition of *F0mora* candidates in the previous section, their actual relationship to the perceived features was not verified quantitatively. In this section, we investigated how humans perceive the pitch of speech, i.e., how they obtain a representative F0 value for each mora, and verified its relationship to the proposed *F0mora* candidates.

In past researches, many efforts had been realized to investigate and model the pitch perception. However, most of them [8,9] used artificial signals such as pure tones as stimuli. Since we cannot guarantee if these results may be applied directly to speech sounds, in the present research, we investigated the pitch perception in actual speech utterances.

### 2.4.1. Experiment procedure

Although we have in mind that the pitch perception may differ according to the global context, we initially decided to investigate the pitch perception of isolated syllables.

We decided to use MIDI synthesized sounds as references of pitch values of the speech utterances. Advantages of using MIDI sounds instead of using synthesized speech are the context-free properties, such that subjects can pay attention only to the pitch feature.

For this purpose, a tool was developed to allow subjects to adjust (in musical scale) the pitch of the instrument so that it matches with the pitch of a syllable segmented from a natural utterance. The adjustable pitch range was set between 32 and 523 Hz, covering 4 octaves. The tool allows subjects to hear the speech sample and the synthesized tone separately, or simultaneously. Further, an option to select two values of a pitch was given, if pitch change is perceived inside the syllable.

100 syllables were selected from 9 sentences and 2 isolated words uttered by male and female speakers. The utterances were manually segmented in “CV(N)C” configuration, where N is geminated nasals and V includes short and long vowels. Hereinafter, we’ll refer syllables as segments with “CV(N)C” configuration.

Six subjects were used in the experiment. The subjects were asked to hear a syllable sample and adjust the pitch of the instrument in order to match it with the perceived pitch of the syllable. The syllable samples could be heard as many times as desired.

### 2.4.2. Results and discussion

The data of each subject were grouped for each syllable sample, and the averages and standard deviations of the perceived pitch were calculated. The standard deviation of most of the samples concentrated around 0.5 semitones, indicating a good accordance of pitch perception among subjects. Average values were used as the human perceived pitch, *F0human* hereinafter.

#### 2.4.2.1. Mismatch between *F0human* and *F0mora*: global analysis

Then, we evaluated the mismatches between *F0human* and the *F0mora* candidates (calculated from the observable F0 values of the speech). Mean square errors were used to quantify the mismatches between *F0human* and *F0mora*. Table 2 shows the global results of these mismatches.

Table 2. Mean squared errors (in semitones) between *F0human* and *F0mora*(CV/VC/V; avg/tgt; w/nw).

	nw			W		
	CV	VC	V	CV	VC	V
Avg	1.81	1.61	1.77	1.61	1.45	1.61
Tgt	1.68	3.11	1.66	1.58	2.74	1.58

Here, we introduced the pair w/nw (weighted/non-weighted). The weighting of the F0 is with respect to the power of the signal, and it’s expected to be more effective than non-weighted estimation, since perception of segments with lower power seem to be masked by adjacent segments with higher power. Results showed that the weighted case had slightly better performances in all (VC/CV/V; avg/tgt) cases, showing the effectiveness of the weighting.

Global results showed that *VC-avg-w* had the best matching to human perceived pitch, followed by *V-tgt-w* and *CV-tgt-w*. These results show similar trends to results obtained in the pitch accent recognition (section 2) that was based on the perceived accent types by humans. From this viewpoint, we can say that if we can find a parameter that produces similar values to the human perceived pitch, this parameter will probably produce good performance when applied to the recognition task.

#### 2.4.2.2. Local analysis

Local analysis was also realized verifying each syllable. We observed that *VC-tgt-w* was better than *VC-avg-w* in some syllables, especially when pitch change occurs. Figure 4 shows the correlation between *F0slope* and the mismatch between *F0human* and *F0mora* for weighted case. The correlations for non-weighted case were not included in the figure, but they showed similar tendencies as for the weighted case.

Results showed negative correlation for all average cases, and positive correlation for all target cases. It means that when pitch rises (*F0slope*>0), the estimated *F0mora*(avg) is lower than the actually perceived pitch (*F0mora* < *F0human*), and *F0mora*(tgt) is higher than the perceived pitch (*F0mora* > *F0human*). And when pitch falls (*F0slope*<0), the inverse events occur. As a reason for the negative correlation characteristic of the average cases, we can say that average values don’t consider the pitch movement along the segments, while human perception process does. However, the target cases showed reverse correlation. As a reason for this reversion, we can say that the proposed target methods (first linear regression analysis) resulted in an excessive

extrapolation. Maybe a higher order regression analysis is more suitable.

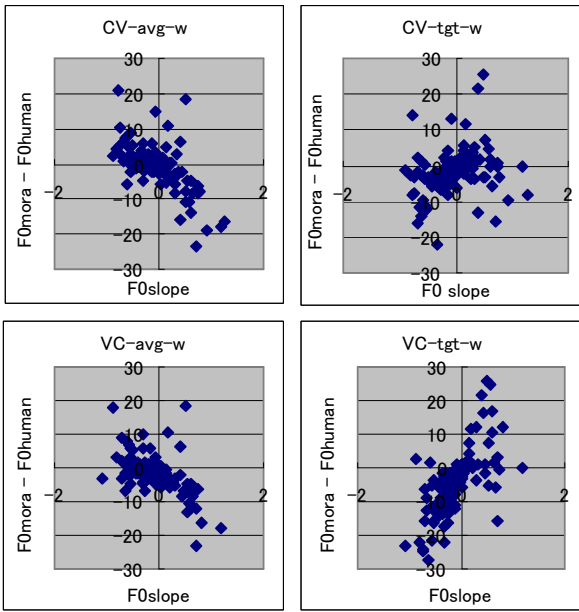


Figure 4: Correlation between  $F0slope$  and mismatch between  $F0human$  and  $F0mora(CV/VC;avg/tgt;w)$

For our purposes of finding a parameter that matches with the human pitch perception, it's desirable that the distribution of the points concentrate along the abscissa axis, that is, a parameter that doesn't depend on pitch changes. Other methods to estimate target values should be investigated to satisfy the above conditions.

### 3. Japanese intonation

Intonation conveys linguistic and paralinguistic information such as speaker's intention and attitudes. In Japanese, the meaning and role of the sentence change mainly due to change of pitch in the sentence final. So, in this research we focused upon the sentence final intonation. 6 types of sentence final intonation [11] are considered, which are defined according to the speaker's intention and attitudes (see table 3).

Table 3. Six intonation types.

Type	Intention	Attitude
<i>Long Rise</i> ( $LRs$ ) ↗	Question, confirmation, offer, invitation	Gentle
<i>Short Rise</i> ( $SRs$ ) ↗	Question, confirmation, agreement	Carefree, cheerful
<i>Long Flat</i> ( $LFt$ ) →	General answer sentences	Calm
<i>Short Flat</i> ( $SFt$ ) ⇨	General answer sentences	Carefree, cheerful
<i>Weak Flat</i> ( $WFt$ ) ⇨	Reserved question, reserved decline	As talking to oneself
<i>Long Fall</i> ( $LFa$ ) ↘	Understanding, discovering, confirmation, doubt, offer	Consent, dissatisfied, disappointed

#### 3.1. Identification of the intonation types by humans

At first, we verified if Japanese native speakers are able to identify the intonation types hearing the utterances.

For this purpose, we used the cassette tape attached to a textbook of Japanese pronunciation learning [11]. This tape includes examples of the above 6 intonation types uttered by one male and one female speaker. Before the listening test, we allowed the subjects hear 40 utterances to become familiar with the 6 intonation types, since native Japanese are generally not so familiar with the 6 types. After that, new 133 utterances were presented and the subjects were asked to identify the intonation types. The experiments were carried out using 6 native speakers. The results are shown in table 4.

Table 4. Identification of intonation types by native speakers.

	Total units	Identification rate (%)					
		$LRs$	$LFt$	$SRs$	$SFt$	$LFa$	$WFt$
$LRs$	180	<b>75.5</b>	0.0	<b>24.4</b>	0.0	0.0	0.0
$LFt$	108	0.0	<b>62.0</b>	0.0	<b>36.1</b>	1.9	0.0
$SRs$	138	3.6	0.0	<b>96.4</b>	0.0	0.0	0.0
$SFt$	150	0.0	8.6	0.0	<b>88.7</b>	2.7	0.0
$LFa$	138	0.0	<b>26.0</b>	0.0	6.5	<b>60.8</b>	6.5
$WFt$	84	0.0	2.4	0.0	0.0	7.1	<b>90.5</b>

According to the results shown in the table 4, a large number of *Long Rise* and *Long Flat* samples were identified as *Short Rise* and *Short Flat* respectively. Further, about a quarter of *Long Fall* were identified as *Long Flat*. These results accord with comments from the subjects after the test. A reason of confusion between *Long Flat* and *Long Fall* could be that different intentions can be realized by similar prosodic patterns. As for the mismatch between *Short* and *Long* patterns, some sentences of *Long* pattern were identified as *Short* by all the subjects. This implies that there could be problems in the speaker's pronunciation in the tape recording.

#### 3.2. Analysis of the acoustic features related to the intonation types

In this section, we examined the correlation between the observable acoustic features and the intonation types identified in 3.1. For analysis, we used the test set of 133 utterances, and focused on the sentence final. F0 and RMS were estimated for each 10 ms interval, and the following features were obtained for each sentence.

- Sentence final vowel duration ( $dur$ ).
- Average mora duration of the sentence final phrase ( $mora\_dur$ ): average mora duration estimated from the last phrase of the sentence, excluding the sentence final vowel.
- Sentence final relative duration ( $rel\_dur$ ): ratio between  $dur$  and  $mora\_dur$ ; corresponds to the number of morae of the sentence final relative to the overall utterance.
- Sentence final power slope ( $pow\_s$ ): slope obtained by first-order linear regression analysis of RMS.
- Sentence final F0 slope ( $F0\_s$ ): slope obtained by first-order linear regression analysis of F0.
- $F0target$  variation in the sentence final ( $dF0\_t$ ): difference between the  $F0target$  of the last two segments, when the sentence final is segmented according to  $mora\_dur$ .

The features above were estimated for the sentence final of the 133 utterances and arranged according to the results of the intonation type identification test. Figure 5(a) shows the analysis results.

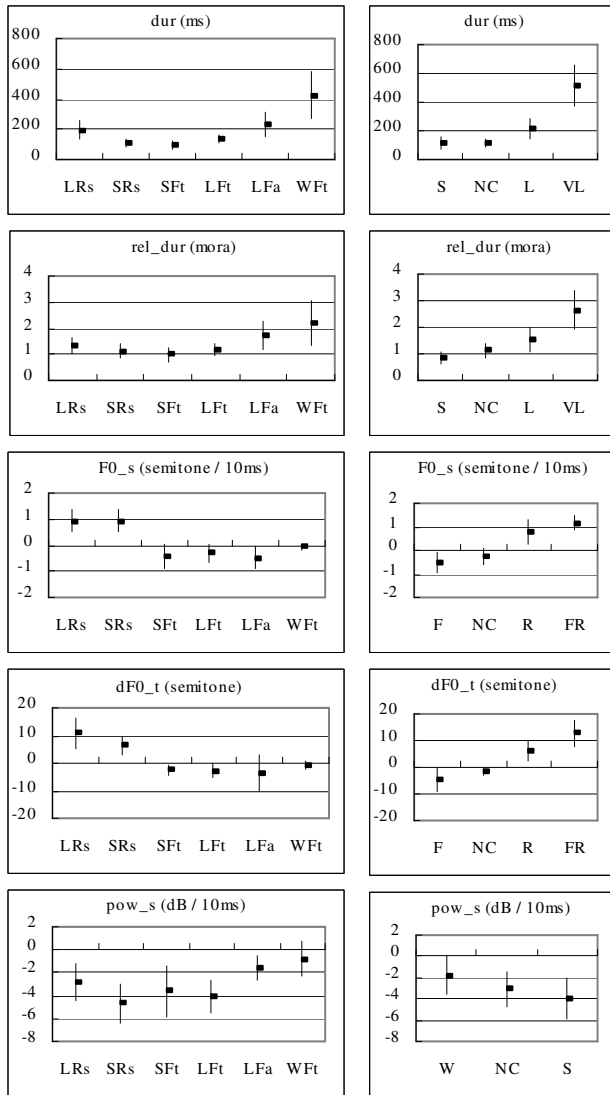
With respect to the parameter  $rel\_dur$ , values upper than 1 mean that the sentence final is lengthened relative to the overall sentence, and values lower than 1 mean that the sentence final is shortened. In *Long Fall* and *Weak Flat*, we can observe a tendency to lengthen the final vowel. In the

*Short Rise* and *Short Flat*, the data concentrated on the interval of 0.5 to 1.5 morae, and in *Long Rise*, on the interval of 1 to 2 morae. However, in *Long Flat* the data concentrated on the interval of 1 to 1.3 morae, indicating that some of the *Long* patterns are not actually so long, which makes it difficult to discriminate *Short* and *Long* patterns only using this parameter.

As for the power slope ( $pow\_s$ ), negative values indicate the power decrease. The more negative the slope is, the more rapid the power decreases. In the *Short (S)* patterns this tendency is observed. In the *Weak Flat (WP)*  $pow\_s$  is small, showing that the power decreasing is smooth.

$F0\_s$  represents the slope of F0 in the sentence final. As expected, *Rise* patterns have positive values of  $F0\_s$ , *Flat* and *Fall* patterns have negative values. However, it's difficult to separate *Long Rise* and *Short Rise* only from this parameter. Further, it's also difficult to separate *Flat* and *Fall* patterns.

For the parameter  $dF0\_t$  that takes the change in F0 and the relative duration into account, it's possible to discriminate *Long Rise*, *Short Rise*, and *Fall* patterns.



(a) (b)  
 Figure 5: Correlation of acoustic features to identified intonation types by native speakers (a), and to perceptive features by non-native speakers (b).

### 3.3. Correlation between acoustic features (related to the intonation types) and perceptual features

In section 3.1, we examined how well native speakers are able to identify the intonation types, and in section 3.2, we analyzed the correlation between the acoustic features and the intonation types identified by the native speakers. However, in CALL systems, the objective is to teach to learners how to pronounce based on the speaker's intention and the listener's impression. So it's necessary for the system to instruct through prosodic features that the non-native learners can perceive.

In this experiment we investigated the correlation between the acoustic and the perceptual features. In order to examine the perceptual features as context-free situation as possible, we decided to carry out this experiment using non-native speakers.

5 non-native speakers (2 Chinese, 1 Korean, 1 French, and 1 Brazilian) were asked to hear the 133 utterances and decide a value for each prosodic attribute as follows:

- Relative duration: perception of sentence final length relative to the overall phrase (*Short, Long, Very Long, No Change: S,L,VL,NC*)
- Intensity: perception of sentence final intensity relative to the overall phrase (*Weak, Strong, No Change: W,S,NC*)
- Sentence final tone: perception of sentence final pitch movement (*Fall, Rise, Flat+Rise, No Change: F,R,FR,NC*)

According to the results in figure 5(b),  $rel\_dur$  showed good correspondence to the perception of relative duration by humans. As for the pitch perception,  $dF0\_t$  showed good correlation. However, for the intensity perception, a small correlation was found in the  $pow\_s$  parameter.

Comparing the panels in fig. 5a to those in fig. 5b, we can associate each intonation type (in 5a) with one of the perceptual categories (in 5b) for each acoustic feature.

### 3.4. Preliminary evaluation for intonation type recognition using the proposed parameters

The distributions of the proposed acoustic parameters were used to build Gaussian Models, in order to recognize the intonation types. Since the database is small, all data was used to train and evaluate the parameters.

The recognition task showed a performance of about 87% of recognition rate, when all parameters were used. Most of the errors were caused by confusion between Long and Short patterns. This result was already expected since similar distributions were observed in these two categories. Including power parameters should reduce these errors, but a careful consideration is necessary on how to represent their relative values, since different vowels sound to have different intensities even though they have acoustically the same power value.

### 3.5. Discussion

The mutual effect between accent and intonation must also be considered. As well as the final intonation may influence the accent type identification, the sentence final tone may be influenced by accent, especially when the accent nucleus immediately precedes the sentence final. This causes confusion between *Rise* and *Flat-Rise* tones, for example. So, the effects of accent and intonation must be separated, or the models must include both effects.

In the analysis of acoustic features for intonation identification, we found that the proposed parameters for relative duration and pitch change presented good correlation with the perceptual features, but the vowel internal power

change (*pow\_s*) did not. So, relative power to the overall utterance must be taken into account, but there is the problem that its representation is difficult because intensity is differently perceived for different vowels.

#### 4. Conclusion

In order to recognize Japanese accent and intonation, several acoustic parameters were proposed and investigated from the perceptual viewpoint.

As for the accent type identification, the proposed *FOratio*-based models showed a better performance than the F0-based models, indicating that the use of representative values for mora units is more effective to the accent type identification task than the direct use of frame F0 values. Among the *FOratio*-based parameters, CV-tgt and VC-avg showed the best performances, indicating possibly that the final portion of the vowel is more important to the representative F0 value of the mora.

Then, we investigated the relationship between the several candidates of *F0mora* and the human perceived pitch (*F0human*). Analysis of the mismatches between *F0human* and *F0mora* calculated by the proposed methods showed that average methods resulted in larger mismatching when F0 change occurs in the syllable. On the other hand, although target values also showed larger mismatch according to the F0 change, it was in reverse direction when compared to the average case. For the next step, we are planning to investigate other estimation methods of the target values that don't depend on the F0 changes. And then, after finding the best parameters that most closely represent the human perception, we intend to apply them to the recognition task, and verify their performance.

As for the intonation, we investigated the correlation between sentence final acoustic features and the subjects' perceptual impression for 6 intonation types. With respect to the identification ability of the intonation types, overall results showed about 80% of agreement among the subject's decision, indicating the possibility of the same intonation type can be realized by different prosodic patterns. The correlation between the acoustical features and the subjects' perceptual impression for 6 intonation types showed that the proposed features *rel\_dur* (relative duration) and *dF0\_t* (change of *F0target* in sentence final) showed good correlation, but improvements are still necessary in the target value estimation. For intensity, it's necessary to take the relative power into account. Preliminary results for recognition task showed a performance of about 87% of recognition rate, indicating that some more improvements in acoustic parameterization are still necessary.

#### 5. References

[1] Yoshimura, T., Hayamizu, S., Tanaka, K. "Identification of word accent patterns by HMM using fundamental frequency features," Proc. of Acoust. Soc. Japan, vol. 1, pp. 173-174. (Oct. 1992)

[2] Minematsu, N., Nakagawa, S. "Automatic identification of words with Type 1 accent based upon the accent nucleus detection at the head of words using HMMs," Technical Report of IEICE, SP96-29, pp. 69-74. (1996)

[3] Sasaki, H., Miwa, J. "Discrimination of Japanese Word Accent Type using Cepstrum Method of Moving Average and Band-Limitation", Proc. of Acoustic Society of Japan, pp.255-256. (Mar 2000)

[4] Iwano, K., Hirose, K. "Representing prosodic words using statistical models of moraic transition of fundamental frequency contours of Japanese," Proc. of ICSLP98, vol. 3, pp. 599-602. (1998)

[5] Kawai, G & Ishi, C.T. "A system for learning the pronunciation of Japanese Pitch Accent," Proceedings of *Eurospeech 99*, Vol.1, pp. 177-181, (Sep.1999)

[6] Sato, H. "Rule-based speech synthesis," PhD thesis (in Japanese), pp. 55-92. (1987)

[7] Ishi, C.T., Hirose, K., Minematsu, N. "A study on isochronal mora timing of Japanese," Proc. of Acoust. Soc. Japan, vol. 1, pp. 199-200, (Sep. 2000)

[8] Aikawa, K. "New approach of speech analysis for speech recognition," Proc. of Acoust. Soc. Japan, vol. 1, pp. 21-24. (Mar 1996)

[9] Nabelek, I., et al. "Pitch of Tone Bursts of Changing Frequency," JASA Vol 48, N.2, pp. 536-553. (1970)

[10] Hart, J. "Differential sensitivity to pitch distance, particularly in speech," JASA Vol. 69, N.3, pp. 811-821. (1981)

[11] Toki, T., Murata, M. *Pronunciation & Task Listening - Innovative Workbooks in Japanese*, Aratake Publishers, pp. 37-55., 1989