Identification of Accent and Intonation in sentences for CALL systems

Carlos Toshinori Ishi, Nobuaki Minematsu*, Ryuji Nishide*, Keikichi Hirose***

* University of Tokyo, Dept. of Information and Communication, School of Engineering
** University of Tokyo, Dept. of Frontier Informatics, School of Frontier Sciences
[c_t_ishi, mine, nishide, hirose]@gavo.t.u-tokyo.ac.jp

Abstract

In order to construct a CALL (Computer Aided Language Learning) system that can teach learners accent and intonation of Japanese, it's necessary to automatically identify accent types and intonation types in sentence utterances. For this purpose, several acoustic (prosodic) features of speech were investigated taking their effects on human perception into account. For the accent type identification method, the use of average values of F0 in mora and target values of F0 in mora final was evaluated in CV and VC units. Average values of VC units and target values of CV units showed better performance in the identification task. As for the intonation identification, several acoustic features were investigated to represent 6 types of sentence final tones, each conveying different information of intention and perceptual impression. The proposed acoustic features for relative duration and sentence final pitch change showed good correspondence to perceptual features.

1. Introduction

When teaching the pronunciation in sentence level, accent and intonation become important points to be considered. 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, listener's impression, and so on. Therefore, mispronouncing accent or intonation may cause not only unnaturalness, but also misunderstanding between speakers and listeners. Especially in Japanese, subtle changes in intonation may cause the pronunciation sounds impolite, therefore a proper pronunciation of accent and intonation becomes an important topic to foreign learners. Nevertheless, though programs to teach accent and intonation by human teachers have already been developed, researches to apply them in computer systems are not so advanced (especially for intonation).

In our previous works [1], we developed a technique for automatic identification of the accent type of isolated words. However, the performance severely degraded when it was applied to phrases in a sentence. In this work, in order to construct a CALL system to teach the pronunciation in the sentence level, we defined several F0-related parameters and analyzed pitch movement through these parameters. Using the obtained findings, a new method to identify the accent types of words/phrases in a sentence was proposed. Moreover, we also investigated the acoustic features related to sentence final intonation.

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 along the mora sequence of the word. When it is produced in a sentence, however, interaction may occur between adjacent words and a new accent type will be constituted. Generally, a relative pitch height, high or low, is assigned to each mora of the word/phrase to describe its accent pattern.

In this study, we proposed a technique to estimate one representative value of pitch for each mora, called ***F0mora***. Also, in order to characterize the pitch movement along the mora sequence, we defined a variable called ***F0ratio*** as the logarithm of the ratio of the *F0mora* between two adjacent morae. By scaling to the musical scale (in semitone units), *F0ratio* can be represented by:

$$F0ratio(i) = 12\, log_2 \frac{F0_{mora}(i)}{F0_{mora}(i-1)}\,, \tag{1}$$

where *i* represents the mora number.

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

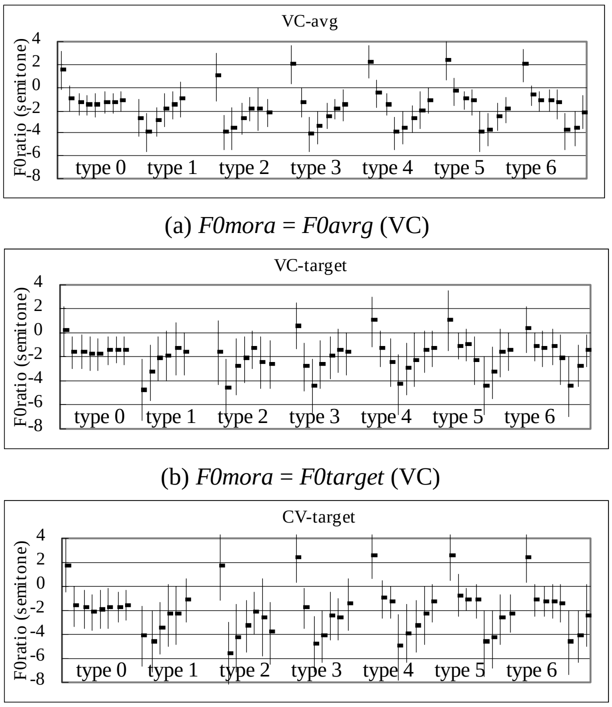
In order to extend the previously proposed technique to identify accent type for isolated words to continuous speech, we used the ATR continuous speech database to analyze the accent types of phrases. The database contains 503 sentences read by a male speaker. For analysis, we used the labeling information and F0 data contained in the database.

Although CV is a basic unit (mora) of Japanese pronunciation, from perceptual viewpoint, Japanese rhythm seems to be realized in VC units rather than CV units [3,4]. Therefore, we decided to analyze both forms of the unit, considering that rhythm may be related to pitch perception.

Here, we defined ***F0avrg*** of a segment as the average value of all F0 data in this segment. At first, we used *F0avrg* of VC units as *F0mora* of the equation (1). Figure 1 (a) shows the *F0ratio* distribution for each accent type. A minimum value of *F0ratio* is expected to be found at the nucleus of accent. However, we can note in the figure 1 (a) that the minimum value for type 1 locates between the second and third morae, which surely causes errors in identifying the accent type.

Next, we defined ***F0target*** of a segment unit as the extrapolated value of F0 at the segment final by the linear regression line, and used it as *F0mora*. Figure 1(b) shows the distributions of *F0ratio* calculated by *F0target* of VC units. Here, we can observe that the minimum value for each accent type appears in the correct position (accent nucleus). This result indicates that there is a possibility of improvements in the identification performance between type 1 and 2 by using the *F0target* of VC units.

The distributions of *F0ratio* based on *F0target* of CV units are shown in figure 1(c). The results are similar to the *F0avrg* case: with confusion between type 1 and type 2 accents.



(a) *F0mora* = *F0avrg* (VC)

(b) *F0mora* = *F0target* (VC)

(c) *F0mora* = *F0target* (CV)

Figure 1: Average and standard deviation of the *F0ratio* distributions for each accent type.

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

Based on the analysis results of 2.1, we carried out several experiments on automatic accent type identification for phrases in a sentence. For modeling the accent patterns, we first categorized them according to the length (number of morae *n*) and the accent type (*i*). Then, the modeling was done for each category. As mentioned in the following paragraph, this model can be viewed as a special case of HMM, and it is called *HMM(n,i)* here in after.

The input of the HMM is a (*m-1*) dimensional vector of *F0ratio*(*j*), *j=2,3,..,m*, where *m* is the phrase length in mora number. That is, one phrase is represented by one point in the (*m-1*) dimensional space. Each phrase of length *N* in the training set is also represented as one point in the (*N-1*) dimensional space, and the distribution of these points (phrase set of the same category) is approximated by a multi-dimensional normal distribution. In this way, an (*n-1*) dimensional normal distribution is provided for each accent type *i*, and these distributions are used in the accent type identification. Therefore, the modeling method here can be viewed as a special case of HMM, where length of an input sequence is always 1, and consequently, there will be only 1 state without self-loop transitions. In the experiments below, two types of covariance matrix in HMMs are 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.

In the experiments, several methods were tested for calculating *F0mora*; the average of F0 in VC/CV units (*avrg*), the target value of F0 in the final of VC/CV units (*target*), and the target value in VC unit using only F0 values in the vowel portion (*V(C)-target*). The upper portion of table 1 shows the identification results for each case.

As a baseline, we built F0-based HMMs where a pitch curve was treated as a temporal sequence of frames of *log(F0)* and/or Δ*log(F0)*. Here, F0 values were linearly interpolated in the unvoiced segments. As for the HMM configuration, we used left-to-right, duration controlled HMMs with *n* distributions. As input features, we used a 2-dimensional vector of *log(F0)* and Δ*log(F0),* called HMM_{ref1}, and another using Δ*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 (*m*) of the input phrase is known, only *HMM(m,i)* ($0 \leq i \leq m-1$) will be used to the accent type identification. The lower portion of table 1 shows the identification results for both cases.

Table 1: Results of the accent type identification under several conditions.

		Accent type identification (%)					
	Accent type	0	1	2	3	4	all
	total	111	80	55	30	24	306
F0ratio based	*VC-target*	84.7	68.8	45.5	33.3	64.0	**67.3**
	V(C)target	80.2	82.5	65.6	50.0	68.0	**73.8**
	VC-avrg	82.0	70.0	63.6	66.7	76.0	**77.5**
	CV-target	83.8	90.0	63.6	56.7	64.0	**78.4**
	CV-avrg	78.4	73.8	56.4	60.0	72.0	**71.5**
F0 based	HMM_{ref1}	75.7	72.5	45.5	50.0	44.0	**63.7**
	HMM_{ref2}	70.3	57.5	52.7	33.3	24.0	**52.3**

The results in table 1 show that *F0ratio*-based methods have better performance than HMM_{ref1} and HMM_{ref2}. As for the accent type identification in a sentence, the F0 variations due to phrase components can be given as a cause of the degraded performance. Moreover, comparing the results for *CV-avrg* and *VC-avrg*, the former shows better performance, indicating that the rhythm perception influences pitch perception. However, for *VC-avrg* and *VC-target*, the former has better performance. This reason can be found in the estimation method of *F0target*, which is based on first regression analysis. We are presently investigating this point. Finally, we can remark that the use of covariance in the *F0ratio*-based HMM didn't affect significantly in the identification.

3. Japanese intonation

Intonation convey linguistic and paralinguistic information like speaker's intention and listener's impression. In Japanese, the meaning and role of the sentence change mainly due to the change of pitch in the sentence final. So, in this research we focused upon the sentence final intonation.

3.1. The intonation types

In this research, 6 types of intonation [2] are considered, which are defined according to the speaker's intention and listener's impression (see table 2).

Table 2. Six intonation types.

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

3.2. 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 [2]. 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. (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 3.

Table 3. Identification of intonation types by native speakers.

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

According to the results shown in the table 3, 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: "It was difficult to distinguish between *Long* and *Short* patterns, and between *Flat* and *Fall* patterns." A reason of confusion between *Long Flat* and *Long Fall* could be that different intentions can be realized by the same prosodic pattern. As for the mismatching between *Short* and *Long* patterns, there were *Long* pattern samples that all the subjects identified as *Short*. This implies that there could be problems in the speaker's pronunciation in the tape recording.

3.3. 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 3.2. 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 above features were estimated for the sentence final of the 133 utterances and separated according to the results on the intonation type identification by native speakers. Figure 3(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.

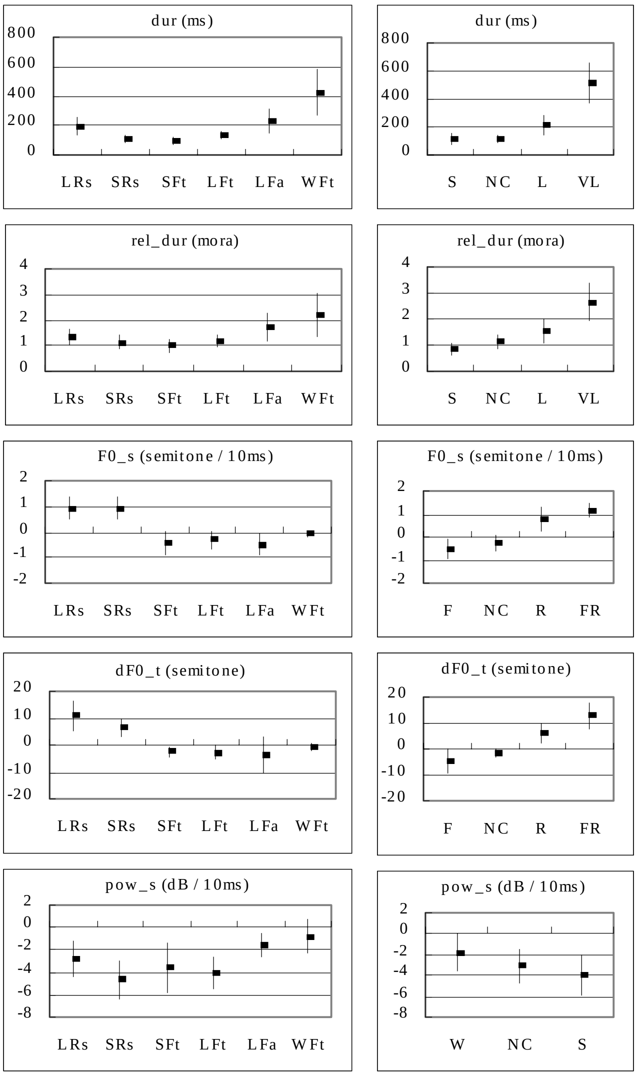
F0_s represents the slope of F0 in the sentence final. As expected, *Rise* patterns have positive values of *F0_s*, *Flat* and *Fall* patterns have negative values. However, it's difficult to separate *Long Rise* and *Short Rise* only from this parameter. Further, it's also difficult to separate *Flat* 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.

3.4. Correlation between acoustic features (related to the intonation types) and perceptive features

In section 3.2, we examined how well native speakers are able to identify the intonation types, and in section 3.3, 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 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.



(a) (b)

Figure 3: Correlation of acoustic features to identified intonation types by native speakers (a), and to perceptive features by non-native speakers (b).

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 3(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. 3a to those in fig. 3b, we can associate each intonation type (in 3a) with one of the perceptual categories (in 3b) for each acoustic feature. Also, we can say that the results obtained in the section 3.3 can be used to automatically identify the intonation type, and the results in section 3.4 can be used to give an instructive feedback to learners.

4. Discussion

In section 2, we presented the results for analysis and a method to automatically identify accent types of phrases in a sentece; however, for application to CALL systems, learners may utter with an accent pattern that doesn't exist in standard Japanese. For this purpose, the features of *F0ratio* distribution in the beginning of the phrase and around the accent nucleus may be used to construct a model that includes non-existing accent types.

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.

Finally, for application to CALL systems, it's necessary to elaborate automatic scoring methods both for accent and intonation.

5. Conclusion

With the goal of constructing a CALL system to train the Japanese accent and intonation, we investigated several acoustic features from the perceptual viewpoint. With respect to the accent type, though *F0target* of VC units have shown better results from the visual feedback viewpoint, *F0target* of CV units and *F0avrg* of VC units showed better performance in actual automatic identification task. For future works, we intend to investigate more about the estimation method of the target value. As for the intonation, we investigated the correlation between sentence final acoustic features and the subjects' perceptual impression for 6 intonation types. We observed that the proposed features rel_dur (relative duration) and *dF0_t* (change of F0target in sentence final) showed good correlation. For intensity, it's necessary to take the relative power into account.

6. References

[1] Kawai, G and Ishi, C.T. "A system for learning the pronunciation of Japanese Pitch Accent," *Proceedings of Eurospeech 99'*, Vol. 1, pp. 177-181, sep.1999

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

[3] Ishi, C.T., Hirose, K., and Minematsu, N. "A study on isochronal mora timing of Japanese," *Proceedings of Acoustic Society of Japan*, Vol.1, pp. 199-200, sep. 2000

[4] Sato, H. "Temporal characteristics of spoken words in Japanese," J. Acoust. Soc. Am., Vol. 64, Sup. No. 1, S113, 1978.