Using Perceptually-related *F0*- and Power-based Parameters to identify Accent types of Accentual Phrases

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

* ATR Human Information Sciences Laboratory, JST/CREST
** Dept. of Frontier Informatics, Grad. School of Frontier Sciences, Univ. of Tokyo
*** Dept. of Info. and Comm. Eng., Grad. School of Info. Science and Technology, Univ. of Tokyo

[c_t_ishi, hirose, mine]@gavo.t.u-tokyo.ac.jp

Abstract

Based on a represention of *F0* in mora units (*F0mora*), several parameters and methods were introduced in order to identify accent types of accentual phrases: new candidates for *F0mora* estimation method; the addition of relative *F0mora* parameters and relative power parameters; and a new accent type identification method. The candidates for *F0mora* were investigated in order to find the best matching with the perceived pitch values. As for the relative *F0mora* parameters, new delta-*F0mora* parameters were proposed, in order to take phrase contextual effects into account, and to supply additional information in segments with missed *F0* data (like in devoiced vowels). Relative power parameters were also investigated, because power also seems to influence in the accent type identification. As for the identification method, neural network models were proposed to find a suitable weighting for each parameter, and a transformation of the input parameters were proposed using Gaussian distributions in order to deal with the parameters with missed data.

1. Introduction

In past researches, several methods were proposed to identify Japanese accent types. (Section 3 gives a brief explanation of Japanese accent types.) In [1] and [2], the use of *HMM* to identify the accent type of words were proposed. *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 of the *F0* peak to the word length, and the declination of the *F0* contour are used as parameters to the identification. In [4], codebooks for *F0* contour shapes of mora units (**"Mora"** is a rhythmic unit of Japanese) plus the difference of *F0* averages between adjacent morae were used to model and recognize accent types of accentual phrases. However, accent type identification was limited to categorize only accent type *0*, type *1* or the others.

In our previous works [5,7], we proposed to represent the pitch movement of an utterance as a sequence of pitch values in mora units (***F0mora***), instead of using directly the *F0* contour in frame units. In [5], average *F0* of *CV* units were used as parameters, and a method was developed to identify the accent type of isolated words based on perceptual thresholds for accent nucleus. The method worked well for isolated words but the performance degraded when applied to accentual phrases. In [7], several candidates were proposed for *F0mora* estimation, and Gaussian models were used to identify accent types of accentual phrases.

In the present research, new candidates for *F0mora* and addition of new parameters were proposed in order to improve the accent type identification of accentual phrases. Section 2 introduces new *F0mora* candidates that better match with the perceived pitch values. Section 3.1 explains the proposed *F0*- and power-based parameters. A method for accent type identification using neural network with transformed input parameters is explained in the section 3.2, and finally, the results of the improvements are summarized in the section 3.3.

2. *F0mora* estimation related to perceived pitch

First of all, all *F0* values (obtained for frame units in 10 ms intervals) were converted to musical scale using the expression (1), before any calculation using *F0* data.

$$F0[semitone] = 12 * \log_2(F0[Hz]) \quad (1)$$

In our previous work [8], *F0mora* estimation methods were investigated from a perceptual viewpoint. Several candidates were proposed taking the segment type (*VC*, *CV* or *V*), the estimation method (average or target) and the weighting of *F0* values by power values (weighted or non-weighted) into account. Then, the matching between these proposed parameters and perceived pitch values by humans (***F0human***) were evaluated. Although VC-avg-w had shown the best matching for global results (as shown in the left portion of Table 1), detailed analysis on each syllable sample showed better matching with target values in some cases where *F0* changes occurred. The relationship between *F0* changes (*F0slope*) and the mismatches between *F0human* and *F0mora* was then investigated. Results showed correlation in both *avg* and *tgt* cases, but in different directions, as shown in the left portion of Figure 1. This indicates problems in both *avg* and *tgt* estimation methods, when pitch change occurs. The reason for *avg* case is that it does not consider *F0* movements within the segment. This was an expected result and that's why target estimation had been proposed. However *tgt* case also showed problems possibly because of excessive extrapolation by using first-order regressive analysis. For our purposes of finding a parameter that matches with the human pitch perception, it's desirable that the distribution of the mismatches concentrates along the abscissa, i.e., a parameter that result small errors even if *F0* changes occur.

In the present research, we investigated some more candidates for *F0mora* based on the results above. First, a third-order regressive analysis was tested in the target estimation (*tgt3*). However, results were worse than the first-order case (see Figure 1 and Table 1). As a reason for this, we can say that the extrapolation by high-order regressive analysis may not be suitable, since *F0* values only within the segment are used in the estimation. Further, we can say that this kind of estimation method using regressive analysis is more susceptible to small errors in *F0* data than average methods. Another candidate was then proposed taking advantage of the

robustness of average estimation methods, and emphasizing the final portion (target portion) of the segment. The method consists of dividing the segment in two halves and taking the average of the second half (*end*). Right portion of Figure 1 shows the mismatches between *F0human* and the newly proposed *F0mora* candidates. The abscissa represents *F0slope* in semitone/10ms units, while the ordinate represents the *F0mora* mismatches in semitone. Right portion of Table 1 shows the global mismatch values in semitone. The results for *CV-w* case were not included, because no significant difference was found relative to *V-w* case. This is because the weighting of *F0* with respect to power reduces the influence of *F0*s of consonant portions that have smaller power values than in vowel portions. The non-weighted (*nw*) cases were also omitted, because they showed worse results with respect to the weighted counterparts.

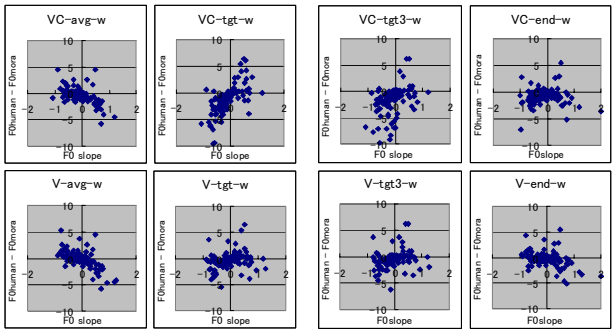

Figure 1: Distributions of the mismatches between F0human and the F0mora candidates, with respect to F0slope.

Table 1. Mean squared errors (in semitones) between F0human and the F0mora candidates.

	avg	*tgt*	*tgt3*	*end*
VC	1.45	2.74	6.1	1.44
V	1.61	1.58	1.71	1.42

According to the results, *V-end-w* showed the best matching to *F0human*. Comparing the distributions of the *F0mora* mismatches in Figure 1, we can note that *V-end-w* parameter globally shows the best concentration of data around the abscissa (mismatches next to 0). Further, slight different tendencies can be observed between falling *F0* (*F0slope*<0) and rising *F0* (*F0slope*>0). These different tendencies for rising and falling *F0* were also reported in [6] for pitch perception experiments using frequency changing sinusoidal stimuli. It possibly indicates that the auditory system works in different manner for rising and falling pitch. Based on Figure 1, we can say that *V-end* shows the best matching for falling *F0*, and maybe a combination of the *V-end* and *V-tgt* could give a better matching for rising *F0*. This alternative was also taken into account in the experiments on accent type identification reported in the Section 3.3.

3. Japanese Accent Type Identification

LHHHHH HLLLLL LHLLLL LHHLLL LHHHLL LHHHHL
Type 0 Type 1 Type 2 Type 3 Type 4 Type 5

Fig. 2. Possible accent types for accentual phrases of 6 morae

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 2 shows examples of accent types for accentual phrases with 6 morae length, in Japanese Tokyo dialect. Hereinafter, accentual phrases will simply be referred as phrases.

The ATR database of continuous speech containing 503 sentences uttered by a male speaker was used to analyze the accent types of phrases. Table 2 shows the distributions of the accent types (*acc_type*) according to the phrase length (number of morae *n_mora*).

Table 2. Distributions of the accent types of the accentual phrases contained in the database

		acc_type							
		0	*1*	*2*	*3*	*4*	*5*	*6*	all
n_mora	2	87	202						289
	3	209	291	249					749
	4	358	244	147	145				894
	5	315	87	80	120	124			726
	6	86	20	40	69	65	73		353
	7	56	5	7	42	51	60	28	249
	all	1111	849	523	376	240	133	28	3260

The phone segmentation, *F0* data, phrase segmentation, and accent type data contained in the database were used. Inspections on the *F0* data indicated some gross errors mainly in sentence finals. The voiced/unvoiced flag of these gross errors were manually corrected to unvoiced to avoid subsequent errors.

3.1. *F0*- and power-based parameters to model accent types

3.1.1. F0mora-based parameters

In our previous research on accent type identification [7], we proposed the use of the sequence of relative values of *F0mora* between adjacent morae (expression (2)) to represent the pitch movement along the utterance. Here, we remind that all *F0mora* values are converted to musical scale (in semitone).

$$dF0mora(i,i\text{-}1) = F0mora(i) - F0mora(i\text{-}1), \quad 2 \le i \le n\_mora \qquad (2)$$

In the present research, we added some more parameters in order to improve the accent type identification. First, we proposed the use of the difference between *F0mora*s of the first mora of the current phrase and the last mora of the previous phrase:

$$dF0mora(1,L) = F0mora(1) - F0mora(L) \quad , \qquad (3)$$

where *L* refers to the last mora of the previous phrase. This parameter is intended to take the influence of the pitch of the previous phrase into account. Thus, this parameter will only be used when there are not pauses separating the phrases. Pause information was also extracted from the database labeling.

Another problem in the *F0mora* representation is how to deal with non-voiced segments that does not contain *F0* values, like plosive and fricative obstruents, and devoiced vowels. A suitable handling of these segments is important, because their occurrences are very frequent in Japanese. However, there are no clear evidences about how humans perceive the pitch of these segments. One solution could be the interpolation (extrapolation) of the *F0* contour by smoothing functions, considering that the pitch of such segments are strongly influenced by the adjacent segments. Another way to deal with this problem is simply not to define *F0* values in these non-voiced segments, considering that there's no pitch in such segments. In the present research, we opted not to use interpolated values because even if *F0* extraction is robust, microprosody effects could difficult a suitable interpolation.

The lack of *F0* values in these non-voiced morae will result in lack of information in the *dF0mora(i,i-1)* representation. So, we proposed additional use of the difference between *F0mora*s of morae separated by 2 morae:

$$dF0mora(i,i-2) = F0mora(i) - F0mora(i-2), \quad 3 \le i \le n\_mora \qquad (4)$$

3.1.2. Power-based parameters

Although *F0* is the most important parameter related to pitch perception, power also may influence in the accent type decision. Preliminary analysis on *F0* contour of the phrases in the database showed similar *F0* contours for different accent types. Also, it was observed tendency of lowering the power of the morae after an accent nucleus.

Here, we proposed the representation of power values in mora units (*RMSmora*), and the use of their relative values such as the relative *F0mora* values described in the previous section. As for the *RMSmora* estimation, we decided to use the average of the *RMS* values (scaled in dB) of the central portion of the vowels. As well as in *F0mora*, we decided not to use *RMSmora* values of non-voiced segments.

The expressions (5) and (6) shows the relative power for adjacent morae and for morae separated by 2 morae, respectively.

$$dRMSmora(i,i-1) = RMSmora(i) - RMSmora(i-1), \quad 2 \le i \le n\_mora \qquad (5)$$

$$dRMSmora(i,i-2) = RMSmora(i) - RMSmora(i-2), \quad 3 \le i \le n\_mora \qquad (6)$$

3.2. Accent type identification methods

3.2.1. Gaussian models

In order to create models to identify accent types, we first categorized them according to the accent type (*acc_type*) and the phrase length (*n_mora*), since the number of input parameters (n_p) varies depending on the phrase length. Gaussian distributions of each parameter described in Section 3.1 were obtained for each category, and used to build a multi-dimensional Gaussian model for each category ($GM[\mu_{n\_mora,acc\_type}, \sigma_{n\_mora,acc\_type}]$).

The identified accent type is the category whose Gaussian model parameters give the smallest error (expression (7)) for the input parameters.

$$Err_{n\_mora,acc\_type} = \frac{1}{n_p}\sum_{p=1}^{n_p}\left(\frac{param(p) - \mu_{n\_mora,acc\_type}(p)}{\sigma_{n\_mora,acc\_type}(p)}\right)^2 \qquad (7)$$

3.2.2. Neural network models

In the Gaussian models described above, equal weights are applied on all (*F0*- and power-based) parameters in the accent type decision. However, there are no clear evidences on how power information contributes to the accent type decision jointly with pitch information, i.e., how should be the weights of power parameters with respect to *F0* parameters. Also, we are not sure if the contributions of each parameter can be linearly treated. Here, we proposed the use of neural network models, because of their non-linearity property, in order to find suitable weights for *F0*- and power-based parameters.

However, we should take care with the problem that some of the parameters may be missed (because of *F0* absence in devoiced vowels, for example). In this case, entering zero values in the neural network input might cause errors in the output decision. As a solution for this problem, we decided to map all input parameters to zero value vicinity using their Gaussian distributions (expression (8)), before enter them in the neural network input.

$$new\_param(p) = \frac{param(p) - \mu_{n\_mora,acc\_type}(p)}{\sigma_{n\_mora,acc\_type}(p)} \qquad (8)$$

With this transformation, all parameters of phrases with a certain accent type will be mapped to zero value vicinity by using the distributions for this accent type, while some of the parameters of phrases with other accent types are intended to deviate from zero values when mapped using the same distributions. In this way, parameters with missed values could be entered with zero values in the neural network input.

As for the topology of the neural network models, we decided to use a fully connected multi-layer perceptron with one hidden layer, and the *tanh* function as the activation function of the hidden and output layer neurons. Since there are different Gaussian distribution parameters for each accent type, different models were also prepared for each accent type. In this way, each model will have only one neuron in the output layer.

The algorithm of back-propagation was used to train the model parameters. For the training set, 300 phrase samples were prepared such that half of them are samples of the target accent type to be identified, and the other half are samples of all other (non-target) accent types. The samples were arranged in {target, non-target} sequence. Samples were repeated to complete the 300 samples of the training set in the case of lack of training data. Since the *tanh* function is used as the activation function of the output neuron, the target values of the network output was set to {1, -1}. Thus, the identified accent type will be the model whose output gives a value most next to 1.

Experiments showed that the convergence of the network weights was achieved after hundreds of epochs. The training process was programmed to stop after 800 epochs, or when a mean square error smaller than 0.01 is achieved.

As for the number of neurons in the hidden layer, two criteria were taken into account: a) a function of the number of accent types to be identified (n_{acc_type}); b) a function of the number of neurons in the input layer (n_{input}). At first n_{acc_type} was tested, and the convergence was not obtained for phrases larger than 6 morae. Then, $n_{input}/2$ was tested, and the convergence was not obtained for phrases larger than 5 morae. It possibly implies that the convergence becomes more difficult to occur as the number of neurons in the hidden layer increases. However, a possible explanation for these non-convergence could simply be the lack of training data for phrases larger than 5 morae (see Table 2). Anyway, the following heuristic function was proposed such that the number of neurons in the hidden layer becomes not so large for long phrases: $n_{hid} = k + log_2(n_{acc\_type})$, where k is a constant. Reasonable convergences were achieved when k=2. All experiments hereinafter were conducted using this heuristic function to define the number of neurons in the hidden layer.

3.3. Evaluation of the *F0*- and power-based parameters, and the accent type identification methods

As for the *F0mora* estimation, 4 methods were investigated: *VC-avg (f1)*, *V-tgt (f2)*, *V-end (f3)*, *(V-end + V-tgt)/2* for rising *F0* within the segment or *V-end* otherwise (*f4*). The reason for the proposition of method *f4* was explained in the end of Section 2.

4 types were also investigated with respect to the input parameter set: *dF0mora(i,i-1)* only (*df1*), *df2 = {df1, dF0mora(1,L)}*, *df3 = {df2, dF0mora(i,i-2)*, and *df3dp = {df3, dRMSmora(i,i-1), dRMSmora(i,i-2)}*.

As for the accent type identification method, the models proposed in the Section 3.2 were evaluated: Gaussian models (*GM*), and neural network models with transformation of the input parameters using Gaussian distributions. (*NNGM*).

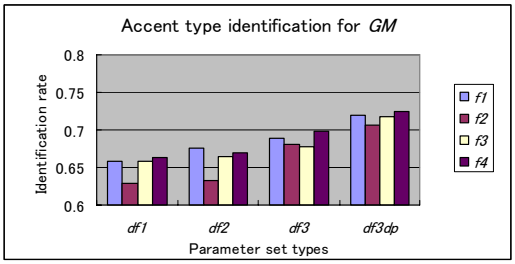

Fig. 3. Comparison of the F0mora estimation methods and parameter set types for the GM.

According to the identification results in Figure 3, we can observe that the addition of each of the proposed parameters in the parameter set improves the accent type identification for all *F0mora* estimation methods. A significant contribution of power-related parameters can be observed comparing the results for *df3* and *df3dp* in the figure.

Among the *F0mora* estimation methods, the combination of *V-end* and *V-tgt* values in rising pitch segments (*f4*) showed a small improvement relative to the use of *V-end* only (*f3*). *f4* also showed a slight better performance than *VC-avg* (*f1*). As expected, the target method (based on regressive analysis) (*f2*) showed worse results than the others.

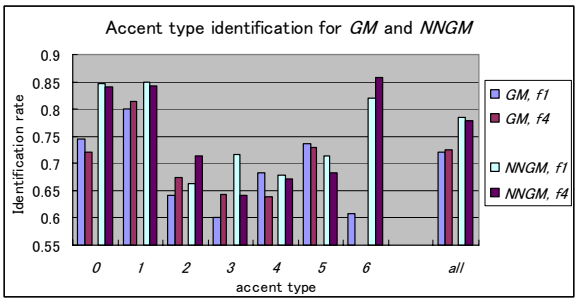

Fig. 4. Comparison between the accent type identification methods (GM and NNGM) for all accent types, for the parameter set df3dp.

Figure 4 shows identification rates of all accent types for *F0mora* estimation methods *f1* and *f4*, and for accent type identification methods *GM* and *NNGM*. We can observe a better performance of *NNGM* in almost all accent types for both *f1* and *f4*, especially in the accent type *0*. This indicates that the neural network could find suitable weights for each input parameter, also improving the identification of accent types with ambiguities in the *F0* contour (observed mainly in the accent type *0*). The best improvement observed in accent type *6* doesn't necessarily mean that the trained model is robust, because the database for accent type *6* is very small (see Table 2), and in this case the training set and test set were the same.

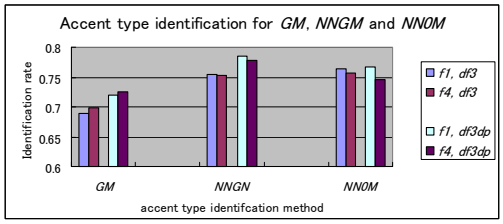

Fig. 5. Comparison between the accent type identification methods GM, NNGM and NN0M.

Additional experiments were also conducted training neural networks **without** transforming the input parameters using the Gaussian distributions (*NN0M*). Figure 5 shows the comparison of identification results with the previous methods. These results indicate that even *NN0M* and *NNGM* have similar performances when only *F0*-based parameters are used (*df3*), *NNGM* shows a better performance when both *F0*- and power-based parameters are used (*df3dp*). It indicates the effectiveness of the transformation of the input parameters using Gaussian distributions in *NNGM*.

4. Conclusion

New parameters related to mora representative *F0* (*F0mora*) were investigated, in order to find a better matching with the perceived pitch values by humans (*F0human*). *V-end* showed the best global matching with *F0human*. In the identification task, *V-end* and a combination of *V-end* and *V-tgt* for rising *F0* showed a slight better performance than the other tested parameters.

As for the input parameters to be used in the accent type identification, the newly proposed *dF0mora* parameters showed a better performance than simply use relative *F0mora* of adjacent morae. The addition of relative power parameters also showed improving in the identification, indicating that power information is also important in the accent type identification.

The proposed neural network models with transformation of the input parameters using Gaussian distributions showed the best performance among the investigated accent type identification methods, indicating that a more suitable weighting for each input parameter were found.

As the next step, we intend to analyze the phrases where identification errors occurred. Also, we intend to take advantage of tendencies observed in the parameter distributions for phrases smaller than 5 morae to create generalized models for any phrase larger than 6 morae.

5. References

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

[2] Minematsu, N.; Nakagawa, S., 1996. 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, 69-74.

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

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

[5] Kawai, G.; Ishi, C.T., 1999. A system for learning the pronunciation of Japanese Pitch Accent. Proc. of *Eurospeech 99'*, Vol.1, 177-181.

[6] Nabelek, I.; Nabelek, A.; Hirsh, I., 2001. Pitch of Tone Bursts of Changing Frequency. *JASA* Vol 48, N.2, 536-553.

[7] Ishi, C.; Minematsu, N.; Hirose, K.; Nishide, R., 2001. Identification of Accent and Intonation in sentences for CALL systems. Proc. of *Eurospeech 2001*, 2455-2458.

[8] Ishi, C.; Minematsu, N.; Hirose, K., 2001. Recognition of accent and intonation types of Japanese using F0 parameters related to human pitch perception. Proc. of *ISCA Workshop on Prosody in Speech Rec. and Und.*, 71-76.