AUTOMATIC DETECTION OF ACCENT NUCLEI AT THE HEAD OF WORDS FOR SPEECH RECOGNITION

Nobuaki MINEMATSU *Seiichi NAKAGAWA*
mine@tutics.tut.ac.jp *nakagawa@tutics.tut.ac.jp*

Dept. of Information and Computer Sciences, Toyohashi Univ. of Tech.,
1-1 Hibarigaoka, Tempaku-chou, Toyohashi-shi, Aichi-ken, 441 JAPAN

ABSTRACT

A new scheme is proposed to incorporate prosodic processing into speech recognition, where the accent nuclei at the head of words are detected automatically and used to limit the searching space in speech recognition, that is, to preselect candidate words. Especially in this paper, the proposed method for the automatic detection of the accent nuclei and its performance are described. Using this scheme, it is expected that the recognition speed is improved.

This scheme is derived from a finding by perceptual experiments conducted previously by the first author. Results of the experiments indicated that the accent nucleus at the first mora has acceleration effect on perceiving the word. This effect can be explained by the earlier identification of the word accent type as *type 1* by its nucleus at the first mora. In other words, the accent nucleus at the head of a word can limit the searching space effectively in the mental lexicon.

This mechanism was implemented using HMMs and examined for isolated words on a machine, where the vowel detection by broad segmental features and the rejection of words with a devoiced vowel at the first or second mora were introduced at the same time. Evaluation experiments showed 94.7% and 90.0% as recall factor and precision factor of the accent nucleus detection respectively.

1. INTRODUCTION

It is well-known that the current speech recognition technology is almost completely built only on the processing of segmental features in speech, that is, the prosodic processing is not yet sufficiently developed in the technology. In considering the reasons, as well as the difficulties peculiar to the prosodic processing, such as adequate handling of local accents, the difference between the quality of information transmitted by the segmental features and by the prosodic features should be noted. Namely, while the former mainly convey discrete information such as phonemes, the latter are generally considered to transmit continuous information such as emotion. This substantial difference should also let the recognition technology still in the current situation.

However, the roles of the prosodic features in the human processing of spoken language have been reported by a great number of researchers[2]. This can be the case even with the communication between a human and a machine. Namely, humans should use the prosodic features as one of the media to convey some of their thoughts. If the machines cannot properly deal with the features nevertheless and, consequently, they neglect the conveyed information, it is natural that the communication should sometimes end in vain.

Although the main information transmitted by the prosodic features is para-linguistic as told above, they also convey some of the linguistic information, such as word accents, phrase boundaries, syntactic structures, etc. This kind of information is thought to be relatively easy to extract and process compared to the para-linguistic information. After this consideration, the word accents and their facilitation of lexical access are focused upon in this paper.

In the previous studies on the word accents and their use in speech recognition, they were often identified automatically as one of the accent types[3][4], or stressed syllables were automatically detected[5]. These studies were based on the hypothesis that accent nuclei in different positions in a word should play the same roles in speech recognition. Considering the linguistic information conveyed by the segmental features and by the prosodic features, and also their interaction, however, the accent nuclei should play their own roles differently according to their positions in a word. Namely, if an accent nucleus exists in a position of a word and only the segmental features to the position can completely identify the word, the nucleus will have little effect on recognizing the word. On the contrary, if the nucleus exists in an initial portion, it will have great effect. This hypothesis will be proved to be valid at least in word perception in Section **2.**.

2. EFFECT OF WORD ACCENT ON SPOKEN WORD PERCEPTION

The prosodic features of Japanese content words are fully represented by the accent type, denoted by a high-low pattern of F_0 for each constituent mora. And the number of accent types in actual use is strongly limited. In the Tokyo dialect, only $n+1$ types are used for n-mora words. **Figure 1** schematically shows all the accent types for the case of n=4. *Type* i accent has a rapid downfall in F_0 at the end of the i th mora, which is the accent nucleus in Japanese. To examine the difference of the role of word accent among the above types, the following two experiments were carried out[2].

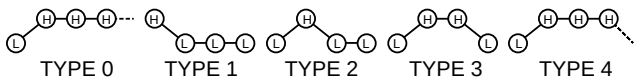

Figure 1: Binary description of F_0 contours of 4-mora words

2.1. Experiment Using Words with Modified Accents

4-mora words were selected as speech material, which went through the analysis re-synthesis process to generate synthetic speech with original and modified accent types. Modification manners are listed below as **CASE**s **1** to **3**. A band-elimination of 0.5 kHz to 3.0 kHz was further performed to all the synthetic speech to produce the speech stimuli. This process was conducted to make it difficult to perceive the stimuli only by the syllable-based matching.

CASE 1 making F_0 constant at 100Hz.

CASE 2 converting the accent type to another type.

CASE 3 no change in F_0 (the original type).

These stimuli were presented to 10 subjects who were asked to reproduce the words orally. **Figure 2** shows the average recognition rates for each case and each original accent type. Among the drops of the rate in **CASE**s **1** and **2**, the largest ones are clearly observed for words with *type 1* accent in both cases. These results indicate that perception of words with *type 1* accent has the greatest dependency on the prosodic features included in the actual utterance. As shown in **Figure 1**, the *type-1* accent has its accent nucleus at the first mora. This early downfall in F_0 should make it possible to identify the accent type before the completion of word recognition process, and therefore, to limit the searching space in the mental lexicon effectively. Namely, words with *type-1* accents are expected to be identified earlier than those with other accent types.

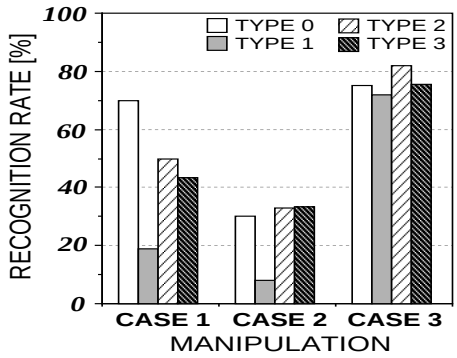

Figure 2: Word recognition rates separately for each case and for each original type

2.2. Experiment Using Words of Gated Duration

In order to prove the above hypothesis of the earlier perception of words with *type 1* accent directly, by using the gating technique, the minimum of initial duration of a word required for correct identification was examined for each accent type. **Figure 3** shows the averages and standard deviations of the minimum durations. According to ANOVA, significant differences were found only between *types 0* and *1* (significant level: $p < 0.01\%$) and between *types 1* and *2* ($p < 0.01\%$). These results clearly show that an F_0 downfall at the first mora definitely makes the earlier identification of the accent type possible and accelerates the word recognition process. However, this effect is not observed in other types. It indicates that the accent nucleus located at or after the second mora has little acceleration effect on word recognition.

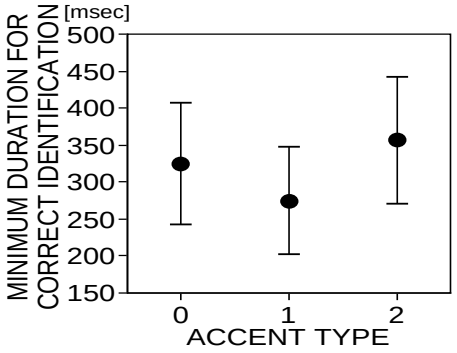

Figure 3: Minimum durations required for the correct identification of words with *types 0* to *2* accents

2.3. Discussion on Use of Accent Nucleus in Speech Recognition

In incorporating the above mechanism into speech recognition, the following two points should be considered. Firstly, the accent nucleus detection must be performed before or at the initial step of the segmental processing. Otherwise, the earlier identification could not be realized. Even if the detection needs some results of the segmental processing, those of the broad processing should be given. Secondly, considering the results of the above experiments, the accent nucleus only at the head of a word should be detected. This is also supported by a previous study[3], where the automatic identification of the word accent type was conducted and it was shown that the accent types with their nuclei at or after the second mora were confusable with each other. It should be also noted that the acceleration effect on word perception by an early F_0 downfall should be peculiar to the Japanese language (and to the Japanese people), which has only the limited patterns of F_0 contour as word accents. These language dependent phenomena have not been sufficiently discussed nor introduced yet into the speech recognition technology. According to [6], the ratio of *type 1* accent words is estimated to be approximately 15%. This small ratio indicates that this mechanism should be implemented on a machine so that the misidentification of *non-type 1* words as *type 1* (false alarm) should occur as rarely as possible.

3. AUTOMATIC DETECTION OF ACCENT NUCLEUS

As the first step to realize the above mechanism, we examined the accuracy of detecting the accent nuclei at the head of words, that is, automatic detection of *type-1* words.

3.1. Speech Material

5 sets of 5240 words from **ATR** database were used, each of which were uttered by a different speaker (**SP1~5**). Before the experiment, the human identification of the accent type was conducted for every word. A group of *type 1* words shall be called **A** and the other **B** in each set henceforth. After that, each of the groups was divided into two subgroups, even-numbered words (**A-0**&**B-0**) and odd-numbered words (**A-1**&**B-1**). The actual ratio of the *type 1* words here was about 20%, 5% larger than the value in [6]. For each word, it was also investigated using the label information whether a vowel at the first or second mora was devoiced and whether a geminate obstruent existed at the second mora. This is because an F_0 downfall cannot be detected in such words. In **Table 1**, the second and third columns show the number of *type 1* and *non-type 1* words respectively. In the forth and fifth columns, the number of words with a devoiced vowel and that with a geminate obstruent were shown separately for *type 1* and *non-type 1* words. As for the words with a devoiced vowel, the number was represented in the form of addition of two numbers, one for the words having a devoiced vowel at the first mora, the other at the second mora.

Table 1: Number of words with a devoiced vowel and with a geminate obstruent ($\overline{type\ 1}$=*non-type 1*).

	type 1	$\overline{type\ 1}$	# words with a devoiced vowel		# words with a geminate obstruent	
			type 1	$\overline{type\ 1}$	*type 1*	$\overline{type\ 1}$
SP1	1094	4146	27+25	352+111	25	202
SP2	1096	4144	24+27	321+118	26	201
SP3	1082	4158	15+51	328+111	25	202
SP4	1094	4146	24+101	332+115	25	202
SP5	1091	4149	20+52	318+106	25	202

3.2. Acoustic Analysis

F_0 extraction was carried out to all the material every 10.0 msec. Median filtering was then conducted for smoothing. In this experiment, an F_0 curve of a given segment was obtained after the polynomial approximation of the median filtered F_0 values. By this procedure, even for a segment including unvoiced parts, an F_0 curve covering the entire segment could be obtained. Then, the F_0 curve at the head of a word, namely, initial two morae, was modeled separately for *type 1* and *non-type 1* words, using 4-state single Gaussian continuous HMMs with the parameters of $\log(F_0)$ and its first order differential coefficient[7]. The accent nucleus detection at the head of a word was conducted based upon the spotting procedure with an adequate threshold, where the subtraction of the likelihood of being *non-type 1* from that of being *type 1* was compared to the threshold.

In modeling an F_0 downfall at the head of a word by HMMs, the voiced consonant at the first mora should be carefully handled. In other words, it should be examined whether the F_0 curve should be modeled including the voiced consonant at the first mora. This is due to the following two reasons.

- When the first mora consists of a voiced consonant and a vowel, a rising pattern of F_0 is sometimes observed in the consonant, even if the accent type of the word is *type 1*.
- It is reported that speech perception by Japanese should be conducted with a unit of mora, namely, synchronously with the occurrence of vowels[8].

Based upon these considerations, two schemes of handling an F_0 curve at the head of a word were compared. One is to model the F_0 curve including the consonant part at the first mora (scheme-1). The other is for the F_0 curve excluding the consonant part (scheme-2).

In the case of modeling an F_0 contour excluding the voiced consonant at the first mora, it is necessary to detect a vowel following the consonant. Although the detection is realized by the segmental processing, the results of the detail processing should not be utilized here as discussed in Section **2.3.**. Then, the introduction of the broad segmental processing was examined in each of the following two manners: 1) additional inclusion of the broad segmental features in a parameter vector (method-A), 2) pre-processing (vowel detection) by the broadly built HMMs before the accent nucleus detection by F_0 parameters only (method-B). While, in method-A, LPC mel cepstrum coefficients of 0 to 4 dimensions were additionally included in a vector[5], the vowel detection in method-B was carried out using the coefficients of 0 to 16 dimensions. In the latter case, each model for vowels/non-vowels was obtained as a one-state HMM with mixture densities (Gaussian mixture model), each density corresponding to each of the 5 vowels in case of the vowel modeling, or to each of the 7 categories derived from the consonants in case of the non-vowel modeling. Mean vectors and covariance matrices for the HMMs were simply calculated from the frames located around the middle of the vowel/non-vowel segments, that is, the re-estimation process was not conducted here. In accordance with the discussion in Section **2.3.**, these HMMs can merely represent the vowels/non-vowels broadly. The vowel detection by these HMMs were performed based upon the spotting procedure with an adequate threshold as well as the accent nucleus detection told above.

3.3. Introduction of Rejection Function

As discussed in Section **3.1.**, words with a devoiced vowel at the first or second mora and those with a geminate obstruent at the second mora should be rejected out of the detection process. According to [9], most of the devoiced vowels in Japanese occur between two unvoiced consonants or between an unvoiced consonant and a silent segment. This means that an unvoiced vowel is inclined to be found in a relatively larger unvoiced segment. Then, in scheme-1, a threshold for the duration from the onset of input word to the onset of the initial *voiced* part was prepared for the rejection. In scheme-2, it was prepared as the duration from the onset of input word to the onset of the first *vowel* in the word. As for the words with a geminate obstruent at the second mora, an adequate threshold for the ratio of silent frames in a segment is expected to reject these words.

3.4. Results and Discussions

Accuracy of the accent nucleus detection was investigated for each speaker using **A-0**&**B-0** as training data and **A-1**&**B-1** as testing data, and vice versa.

Two schemes were compared to examine whether a (voiced) consonant at the first mora should be considered in the detection process. **Figure 4** shows the average detection rates over all the speakers for two cases. One is for testing **A-0**&**B-0** using the HMMs trained by **A-1**&**B-1** and the other is vice versa. These results were obtained after carrying out the rejection procedure as post-processing only to the words judged as *type 1* through the preceding detection process. The rates in the figure were calculated including the results of the words with undesirable events as shown in **Table 1**. In scheme-2, method-B was used as the vowel detection method. Namely, in both schemes, a parameter vector in the detection process consisted of F_0 related parameters only. The figure clearly indicates that the evaluation including an F_0 curve at the voiced consonant results in the severe degradation of detection performance. As for the rejection performance in scheme 2, 50.5% of incorrectly identified words were rejected on the average, while it was only 1.4% for correctly identified words. Re-evaluation of these results except the words with undesirable events listed in **Table 1** shows that about 94.7% of the *type 1* words are correctly identified by scheme-2.

Since the above results indicated that it was quite necessary to introduce the segmental processing to the accent nucleus detection, methods A and B, described in Section **3.2.**, were compared. The difference between the two methods is whether the explicit detection of vowels should be conducted. Results for each method, shown in **Figure 5**, indicate that method-B outperforms method-A, which implies that the segmental processing in the accent nucleus detection is required only to detect the onset of the first vowel in a word.

Preliminary experiments on the normalization among the speakers were also carried out, where $\log(F_0)$ was simply replaced with $\log(F_0)-\log(F_{0\mathrm{avg}})$. Although the identification rates were about 4% lower than those in the above experiments, it is expected that the performance may be improved by introducing other normalization techniques.

4. CONCLUSION

In this paper, the automatic detection of accent nuclei at the head of words was investigated to facilitate the lexical access by using the detected nuclei. This mechanism, which should be peculiar to the Japanese language, was found in the perceptual experiments. The proposed detection method includes the vowel detection function and the rejection function. Evaluation experiments were conducted for each of 5 speakers and 94.7% and 90.0% were obtained as recall factor and precision factor respectively. However, the misidentification rate of *non-type 1* words as *type 1* was about 2.8%, which is not sufficiently low. Further reduction of the rate with the detection rate retained is required as the future work.

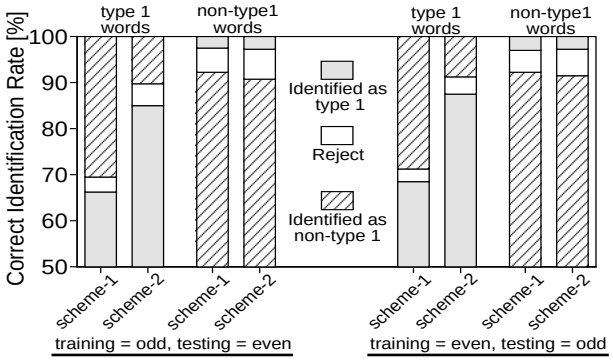

Figure 4: Comparison between scheme-1 and scheme-2 in correct detection rate

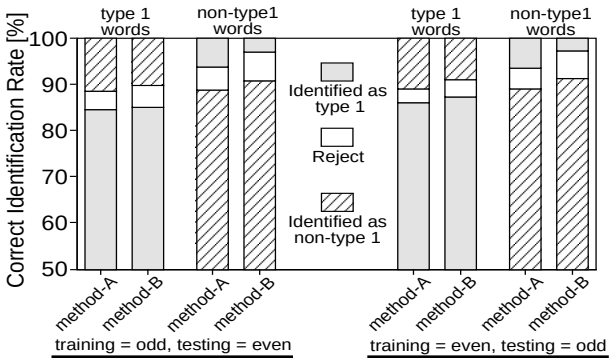

Figure 5: Comparison between method-A and method-B in correct detection rate

REFERENCES

1. N.Minematsu *et al.*, "Use of Accent Nucleus in Speech Recognition," *Report of Spring Meet. Acoust. Soc. Jpn.*, 1-P-22, pp.205-206, (1996).
2. M.Minematsu *et al.*, "Role of Prosodic Features in the Human Process of Perceiving Spoken Words and Sentences," *J.Acoust. Soc. Jpn. (E)*, 16, pp.311-320 (1995).
3. S.Takahashi *et al.*, "Isolated Word Recognition Using Pitch Pattern Information," *Technical Report of IEICE*, SP90-17, pp.65-72 (1990, **J**).
4. T.Yoshimura *et al.*, "Evaluation and Study of Vocabulary-Independent Mora Hidden Markov Models on Word Accent Pattern Identification," *Technical Report of IEICE*, SP93-143, pp.37-44 (1994, **J**).
5. G.J.Freij *et al.*, "Lexical Stress Estimation and Phonological Knowledge," *Computer Speech and Language*, 4, pp.1-15 (1990).
6. S.Hashimoto, "Several Features of Japanese Word Accent," *Trans. IEICE*, 56-D, 11, pp.654-661 (1973, **J**).
7. X.Hu *et al.*, "Recognition of Chinese Tones in Monosyllabic and Disyllabic Speech Using HMM," *Proc. ICSLP'94*, pp.203-205 (1994).
8. Otake *et al.*, "Mora or Syllable ? Speech Segmentation in Japanese," *Journal of Memory and Language*, 32, pp.258-278 (1993).
9. H.Kawai *et al.*, "Devoicing Rules For Text-to-Speech Synthesis of Japanese," *J.Acoust. Soc. Jpn. (J)*, 9, pp.698-705 (1995, **J**).

J = "in Japanese"