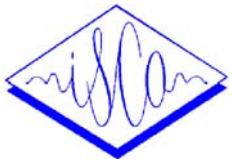

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

DETECTING ACCENT SANDHI IN JAPANESE USING A SUPERPOSITIONAL F0 MODEL

† *A. Sakurai, H. Kawanami, and K. Hirose*
The University of Tokyo, School of Engineering, Dept. of Inf. and Comm. Eng.
Hongo 7-3-1, Bunkyo-ku, Tokyo 113-8656, Japan
{atsuhiro,hirose}@gavo.t.u-tokyo.ac.jp
† also with Texas Instruments, Tsukuba R&D Center

10.21437/Eurospeech.1999-465

ABSTRACT

In this report, we propose a method for automatic prosodic structure recognition of Japanese utterances based on a superpositional F0 model, focusing particularly on the accent sandhi phonemenon in compound nouns. The method enables automatic labeling of F0 contours using the model, which can be useful for creating prosodic databases containing F0 contours in a parametric form. The prosodic structure is identified by comparing the distances between F0 contours generated by hypothetical model configurations and the extracted F0 contour, and choosing the configuration that yields the smallest distance. In this paper, we apply the method to detect the accent sandhi pattern of compound nouns made up of 2 or more words, and show that the method can correctly identify their prosodic structure, except for 1-mora deviations in the position of the accent nucleus.

Keywords: prosodic database, F0 contour model, accent sandhi

1. INTRODUCTION

We are designing a speech database containing prosodic features in a parametric form [1]. A parametric representation of prosodic features has the double advantage that parameters are more mathematically treatable than physical quantities, and the original physical quantities can be obtained from the parameters in a straightforward manner. In this work, we concentrate on the problem of automatic prosodic labeling when word boundaries and POS information are given, particularly the case of compound nouns. We propose a method to automatically characterize the accent sandhi phenomenon in Japanese, by finding the F0 contour model configuration that best fits the F0 contour extracted from the speech signal. Given an F0 contour, the best F0 contour model configuration is found by trying several F0 contour model configurations. The distances between the actual F0 contour and the contours generated by the models are computed and compared, and the configuration that yields the smallest distance is selected. Evaluation experiments are carried out for two cases: 1) two-word compound nous, and 2) compound nouns containing more than 2 words. The evaluation results confirm the validity of the method.

2. DETECTING THE ACCENT SANDHI TYPE OF TWO-WORD COMPOUND NOUNS

2.1 Accent Sandhi Type

For compound nouns made up of two component nouns, the accent type is determined by the second component [2]. The type of accent sandhi that occurs can be classified into the following categories, according to the position of the new accent nucleus of the compound noun, represented by an apostrophe (').

1. The accent nucleus is located at the first mora of the second component. Example: asobia'ite (asobi + aite, "playmate")
2. The accent nucleus is located at the last mora of the first component. Example: seifu'an (seifu + an, "government proposal")
3. The accent nucleus is located one mora before the last mora of the first component. Example: genze'ian (genzei + an, "tax reduction proposal")
4. Flat type, without accent nucleus. Example: akitaken (akita + ken, "akita dog").

In this work, we respectively refer to these types as types A, B, B*, and F. Our objective in this section is to automatically classify compound nouns recorded in a continuous speech database into the types above.

2.2 Partial Analysis-by-Synthesis

The Partial Analysis-by-Synthesis (partial AbS) method has been proposed as a method to use the information contained in the F0 contour for continuous speech recognition [3]. In this method, the F0 contour model is used to synthesize F0 contours for each hypothesized recognition candidate, and then the generated contours are compared with the contour extracted from the actual speech. The candidate corresponding to the closest prosodic structure is the one that yields the smallest distance with respect to the actual F0 contour. This distance is computed as the mean squared difference between the extracted and the model-generated F0 contour, and is called AbS error.

The problem of recognizing the accent sandhi type of compound nouns in a continuous speech database can be interpreted in a similar manner, with the difference that phoneme and word boundaries are given, as well as their POS. Consequently, the Partial AbS Method can also be used for this purpose. Based on this idea, we use a modified Partial AbS Method to automatically determine accent sandhi types. The Partial AbS Method is modified so that candidates (hypothetical model configurations) are created for different possible accent sandhi patterns, and not according to different positions of phrase boundaries as in [3]. *Figure 1* shows an overview of the system.

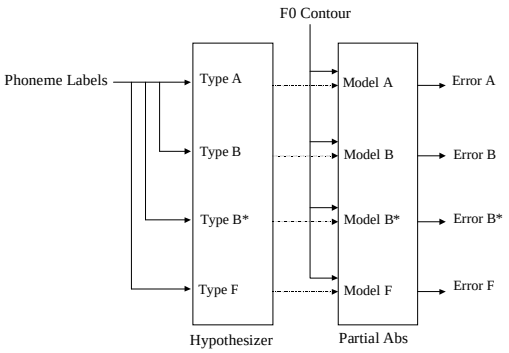

Figure 1. A system for automatic accent sandhi type detection of two-word compound nouns.

2.3 Modeling F0 Contours

A command response model [4] is used to synthesize F0 contours for hypothesized prosodic structures, based on the accent sandhi types of the nouns in question. Even though it is usually necessary to hand-assign initial values for the model parameters (timing and amplitudes/magnitudes of accent and phrase commands), here we automatically find the initial values using the method described below. In principle, this is possible due to the additional knowledge about segmental boundaries and POS tags.

2.3.1 Initial Parameter Values

We introduce an algorithm that automatically assigns suitable initial model parameters for the specific case of compound nouns, based on the hypothesized accent sandhi type.

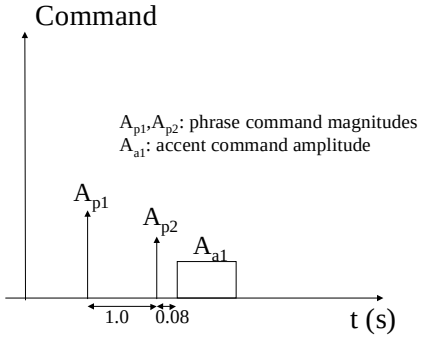

Figure 2. F0 contour model configuration.

First, a model is created containing 2 phrase commands and 1 accent command, as shown in *Figure 2*. The compound noun is represented by the portion corresponding to the accent command. This structure has sufficient degrees of freedom to represent any compound noun formed by two words, uttered in any continuous-speech context.

The accent command is aligned with the utterance so that its onset is placed 70 ms before the beginning of the voiced portion of the second mora of the first noun. The second phrase command is placed 80 ms before the beginning of the voiced portion of the first mora of the first noun, and the first phrase command is placed 1.0 sec before the second phrase.

The offset time of the accent command depends on the hypothesized accent sandhi type. The reference

time is defined as the end of the voiced portion of the mora containing the accent nucleus, or the last mora for type F. The offset time of the accent command is taken 70 ms before this reference time. *Figure 3* illustrates how initial accent timing parameters (t1 and t2) are determined.

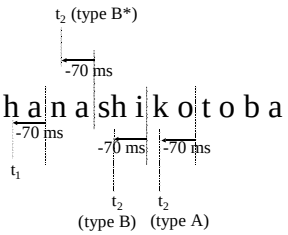

Figure 3. Initial time parameter assignment for compound noun "hanashikotoba" ("hanashi" + "kotoba").

Phrase magnitudes (Ap1,Ap2) and accent amplitude Aa1 are obtained in the following way: all (Ap1,Ap2) pairs in the bidimensional space [0,0.8] x [0,0.8] are taken at discrete intervals of 0.05, and the accent command amplitude Aa1 is calculated for each combination of the other parameters based on the minimization of the minimum-square error between the model-synthesized F0 contour and the extracted contour, as in [5]. The solution (Ap1*,Ap2*) is the one that yields the smallest AbS error between the detected and the model-generated F0 contours.

2.3.2 Fine Parameter Adjustment

Once the initial parameter values (Ap1*,Ap2*) are determined, their values are fine-adjusted using a procedure similar to the one described in [3]. In this case, we allow 20% variation in Ap1, Ap2, Aa1, and ±20 ms in timing parameters (t01,t02,t1,t2), but we fix α (natural angular frequency of the phrase control mechanism) and β (natural angular frequency of the accent control mechanism) as α=3.0 and β=20.0 for simplicity. F0min is calculated from the extracted F0 contour by subtracting 20 Hz from the minimum F0 value of the utterance.

After fine adjustment, the AbS error between the modeled F0 contour and the extracted F0 contour is used to compare accent sandhi type hypotheses.

2.4 Evaluation Experiments

We carried out experiments on 85 samples extracted from ATR's continuous speech database (speakers MAU and MHT). Table 1 shows the obtained results. The phoneme labeling was carried out automatically using HMM-based forced alignment [6].

Table 1 – Experimental results for 2-word compound nouns.

A	19/45	
B	2/10	24/27
B*	10/17	
F	10/13	

The results show that except for type A, the method can detect accent sandhi types with reasonable accuracy. We also noted that most of the incorrect results are due to a deviation of 1 mora in the position of the accent nucleus, and consequently the correct recognition rate improves dramatically if B and B* are counted in the same category.

3. DETECTING THE ACCENT SANDHI PATTERN OF COMPOUND NOUNS CONTAINING MORE THAN 2 WORDS

3.1 Long Compound Nouns

For compound nouns containing just 2 words, the accent sandhi phenomenon can be relatively well predicted using linguistic knowledge [7]. If the compound noun contains more than 2 words, however, there are no rules capable of predicting how accent sandhi occurs, i.e., which component words concatenate to form new accentual phrases. Here, we refer to such clustering process as accent sandhi pattern. This is, then, a typical case where the present method can be applied. In this section, we use the partial AbS method to detect how accent sandhi occurs in the case of compound nouns formed by more than 2 nouns. We selected 2 sentences containing long sequences of nouns (S1 and S2) from a continuous speech database [8]. For each sentence, we selected 2 speech samples spoken by different individuals (I1 and I1' for S1, and I2 and I2' for S2) in different manners (different accent sandhi patterns). We denominate these accent sandhi patterns as H1 and H1' (for I1 and I1'), and H2 and

H2’ (for I2 and I2’). Then, we used the partial AbS method to cross-compare accent sandhi structures. The selected sentences are shown below, where the compound nouns are indicated by slashes.

S1: SaikiNno fukyo:o riyu:ni so:ru /goriN ko:ho seNshu/kara hazusu kotoo happyo:shita (“His exclusion from the team of candidate athletes for the Olympic Games in Seoul is due to the recent economic recession, said the announcement.”)

S2: Watashitachi fu:fuwa guNno iraide /chu:gokujiN uNtenshu/no tsu:yakuo shinagara shanhaimade to:hiko:no do:haNo shita. (“By request of the army, we accompanied the group up to Shanghai, and helped the Chinese driver as interpreters.”)

Accent sandhi structures H1, H1’, H2, and H2’ are shown below, where the position of the accent nucleus is indicated by an apostrophe (’), and the space indicates the separation between two accent commands.

H1: So’oru goriNkoohose’Nshu

H1’: SoorugoriN koohose’Nshu

H2: ChuugokujiNuNte’Nshu

H2’: ChuugokujiN uNte’Nshu

The model parameters are found as in the previous section.

3.2 Experimental Results

Figure 4 shows the obtained results. It can be seen that the system can correctly identify the actual prosodic structure of the utterances.

4. CONCLUSION

A method for automatic detection of prosodic structure based on the F0 contour model was presented. The method provides a partial solution for the automatic labeling of a prosodic database where F0 contours are represented in a parametric form. For now on, we intend to generalize the method to other prosodic phenomena.

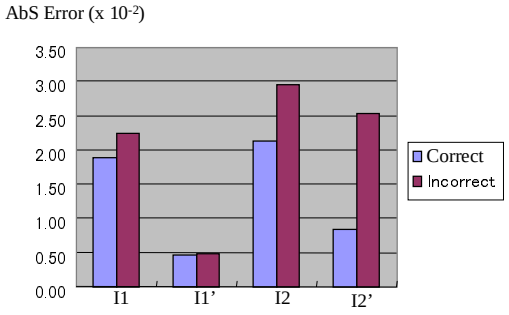

Figure 4. Experimental results on detecting accent sandhi patterns of long compound nouns.

5. REFERENCES

[1] A. Sakurai, T. Natsume and K. Hirose (1998), A Linguistic and Prosodic Database for Data-Driven Japanese TTS Systems. *Proceedings of ICSLP-98*, pp. 2843–2846.

[2] Nippon Hoso Kyokai (NHK) (1985), Japanese Pronunciation and Accent Dictionary.

[3] K. Hirose and A. Sakurai (1996), Detection of Phrase Boundaries by Partial Analysis-by-Synthesis of Fundamental Frequency Contours. *Proc. of ICASSP-96*, pp. 809-812.

[4] H. Fujisaki and K. Hirose (1984), Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese. *J.Ac.Soc.Jpn(E)*, Vol. 5, No. 4, pp.233-242 (1984-10).

[5] T. Hirai, N. Iwahashi, Y. Sagisaka (1996), Automatic Extraction of F0 Control Rules Using Statistical Analysis. In: Springer, *Progress in Speech Synthesis*.

[6] Young, S. et. al., “The HTK Book, version 2.1”, Cambridge University, 1996.

[7] Y. Sagisaka and H. Sato, Accentuation rules for Japanese word concatenation. *IEICE Trans.*, Vol. J66-D, No. 7 (in Japanese).

[8] T. Staples, J. Picone and N. Arai (1994), The Voice Across Japan Database – The Japanese Language Contribution to Polyphone. *Proc. of ICASSP-94*, Vol. 1, pp.89-92, Australia.