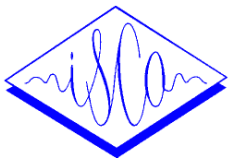

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

STATISTICAL MODELING AND RECOGNITION OF RHYTHM IN SPEECH

Satoru Hayamizu and Kazuyo Tanaka

Electrotechnical Laboratory,
1-1-4 Umezono, Tsukuba, 305 Japan

ABSTRACT

This paper proposes a new framework for processing rhythm in speech where temporal types are recognized using statistical models of mora durations. Temporal patterns, such as rhythm and tempo in speech, contain some basic information about communication through the spoken language. This information has not yet been fully used in speech recognition. This paper proposes that temporal types themselves be modeled and recognized by statistical models. Using the ASJ Continuous Speech Database, experiments for recognizing temporal types of bunsetsu (short phrases) were conducted. Approximately 72% of temporal types were identified correctly using these models, without using information about the length of pauses and fundamental frequencies. The recognized types were very consistent (approximately 94% were of the same types) for closed and open models. These results show the promising potential of the proposed framework.

1. INTRODUCTION

Temporal patterns of speech are thought to play important roles in facilitating communication between two speakers engaged in spoken dialogue [1]. For example, control of dialogue flow, timing cues for turn taking, syntactic structures, relationships between phrases, emotion of speakers, and so on, can be expressed to the other speaker through temporal patterns. Temporal patterns indicate several cues, such as timing for the appropriate flow of dialogue, and segmental information in phrases which corresponds to one unit of semantic information. If machines have the ability to recognize these temporal patterns in human utterances, it will make human-machine interactions more flexible, natural and vivid.

Previous works on prosodic features of speech mainly examined general features of fundamental frequencies, intensities and durations. For temporal aspects of speech, average segment durations and average speech rates (speaking rates, speed of utterances) have been studied using various units of utterances and speaking styles [2, 3, 4]. Constraints on durations of segments have also been utilized to improve the accuracy of speech recognition [5, 6, 7, 8, 9].

Several works have reported recognition of the prosodic events themselves [10, 11, 12, 13]. In these works, features of fundamental frequencies, durations of pauses, and durations of utterances are modeled and used for the recognition.

This paper proposes a new framework of processing temporal patterns. Temporal types themselves are modeled and recognized by statistical models of mora durations. Here, patterns of speaking rates in phrases (bunsetsu) are called "tempo" and patterns of changes in mora units (basically syllables) are called "rhythm". These terms are borrowed from the field of music psychology [14].

2. NORMALIZED DURATIONS

2.1 Mora Durations in Words

Distributions of mora durations and their normalized values were studied for isolated word utterances. Speech data used consisted of a phonetically balanced 1542 word set spoken by 10 males. Segment boundaries of mora units were attached by automatic labeling and manual correction.

Figure 1 shows the distributions of mora durations for three different positions in the words: the *first* mora in the word, the *middle* mora, and the *last* mora. Figure 2 shows the distributions of normalized mora durations for the same three cases. Mora durations were normalized by the average duration of each mora because each mora had a different average duration. For example, each duration of /ka/ was divided by the average duration of /ka/.

Figure 3 shows the distributions of the durations normalized both by the average duration of each mora (same as in Figure 2) and the speech rate of each word. The speech rate was taken as the ratio of the actual length of the word to the sum of the average mora durations in the word. It is shown that by normalizing, the shapes of the distributions become more Gaussian-like and that the distributions of all three cases have better separations.

2.2 Mora Durations in Read Sentences

The temporal patterns of read sentences display both syntactic structures and a relationship between a modifying phrase and a modified one. A pause between phrases in a sentence also displays these features and vowels preceding the pause tend to have longer durations than average. In addition to this pre-pausal

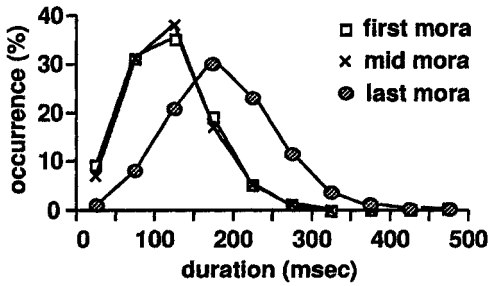

Fig. 1: *Distributions of mora durations.*

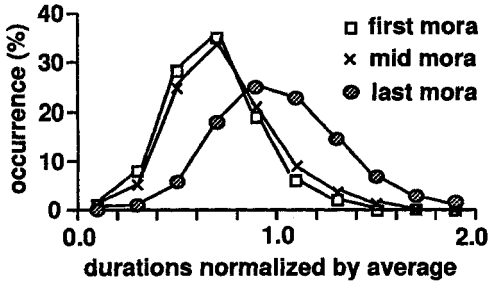

Fig. 2: *Distributions of durations normalized by average durations.*

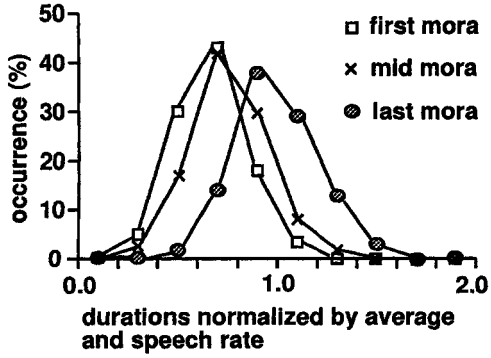

Fig. 3: *Distributions of durations normalized by average durations and speech rate.*

lengthening and the occurrence of pauses, constant durations of preceding mora units are thought to form this type of rhythm.

Distributions of mora durations and their normalized values were studied for read sentences. The speech data used were the half (30 male speakers) of ASJ Continuous Speech Database Vol. 1-3 (total 4518 sentences). Since these speech data were not labeled by phonetic boundaries, they were time-aligned using phonetic models and their transcriptions. As such, these boundaries included some errors due to misalignment. Phonetic models were discrete HMMs [15]. These phonetic models were trained using the above speech corpus.

Each sentence was split into blocks with pauses used as their boundaries. Pauses were identified by a threshold on the length of aligned pauses. The alignment was done using phonetic models (including a

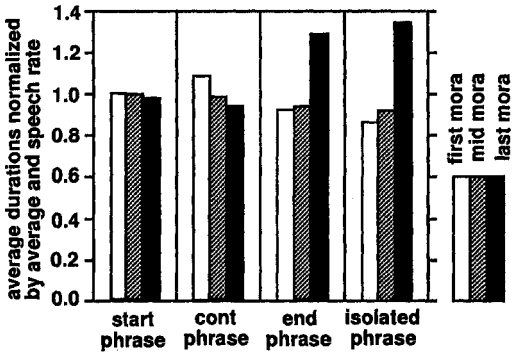

Fig. 4: *Average durations normalized by average durations and speech rate.*

model for pause) and transcriptions where all the phrase boundaries had pauses.

Each phrase (bunsetsu) was classified into one of the following five types according to its position in the block:

```
Type1 - starting phrase in the block,
Type2 - continuous phrase,
Type3 - ending phrase,
Type4 - isolated phrase, and
Type5 - others (one mora phrase).
```

These temporal types represent rhythm patterns in speech. The average durations of the *first* mora, *middle* mora, and the *last* mora for each type of phrases are shown in Figure 4. Only those values normalized by both average mora durations and speech rates are shown. Two levels of normalization were identical to those for isolated word utterances. Speech rates were estimated phrase (bunsetsu) by phrase.

Typical patterns of pre-pausal lengthening were observed and durations of the first mora did not change significantly.

3. RECOGNITION OF RHYTHM

Distributions of normalized mora durations in read sentences suggest that patterns of normalized mora durations can be used to identify the types of temporal rhythm in speech.

Figure 5 shows a block diagram for recognizing rhythm in speech. Input speech was phonetically recognized and time-aligned using phonetic models of acoustic parameters. Mora durations were normalized by their average durations and also by the local speech rate. Estimation of local speech rates can be either for each phrase, for each sentence, or for some fixed period.

The type of temporal rhythm for each phrase is identified (recognized) by comparing the scores of models for all the temporal types. Statistical models are used to improve robustness, since there will be estimation errors in the alignment.

First, speech data were time-aligned using phonetic models and their transcriptions. Time-aligned mora

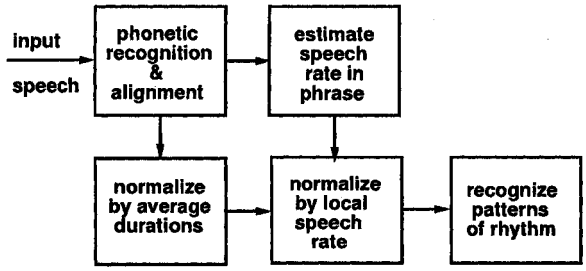

Fig. 5: *Block diagram to recognize rhythm in speech.*

durations were normalized by both each average mora duration and the speech rate estimated using these durations. That is, phonetic recognition was not conducted in the experiments of temporal type recognition experiments.

For the training, all the phrases were labeled by pauses as one of 5 temporal types (Type1, Type2, ... , and Type5 in the previous section). Using the normalized mora durations and assigned temporal types, statistical models of temporal types were trained. They were represented as continuous hidden Markov models (HMM) using average values and variances of normalized mora durations as their parameters. There were 5 temporal type models (Type1, Type2, ... , and Type5). The structure of each model had 3 states and each state roughly corresponds to the first mora, the middle mora, and the last mora in the phrase.

Recognition of rhythm types was conducted sentence by sentence. Constraints on successiveness between temporal types were utilized for recognition. For example, "Type1 (starting phrase)" was not allowed to follow another "Type1". This is similar to the grammar for continuous speech recognition.

In this experiment, speech rates were estimated for each sentence. This was done because speech rates estimated for each phrase produced little difference in recognition rates for this speech corpus.

Experiments for recognizing temporal types of bunsetu (short phrase) in sentences were conducted. The speech data used were the 9636 sentences from the ASJ Continuous Speech Database, Volume 1-3. The 30 male speakers' 4518 sentences were split into two parts. The former part was tested using the models that were trained using the latter part, and vice versa. The 34 female speakers' 5118 sentences were used in the same way.

Figure 6 shows the relationship between speech corpus and models for recognizing rhythm types for the case involving male speech corpus.

Table 1 shows the recognition results of temporal types of rhythm. Results are shown for male set 1 (sentence numbers 1 - 2259), male set 2 (sentence numbers 2260 - 4518), female set 1 (sentence numbers 1 - 2559), and female set 2 (sentence numbers 2560 - 5518).

From the total 62664 bunsetu phrases, the temporal types of 44962 bunsetu were identified correctly

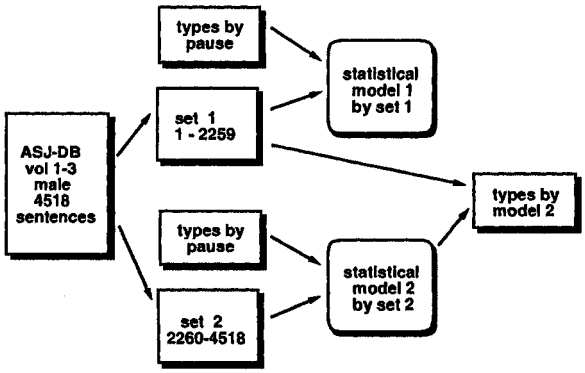

Fig. 6: *Speech corpus and models to recognize temporal types of rhythm.*

	Bunsetsu	Correct	Error	Rate
male set1	14706	10436	4270	71.0%
male set2	14591	10517	4074	72.1%
female set1	16864	12421	4443	73.7%
female set2	16503	11588	4915	70.2%
Total	62664	44962	17702	71.8%

Table 1: *Recognition results of temporal types of rhythm.*

(i.e.; were identical to the temporal types assigned by pauses) using these models. The average recognition rate of temporal types was around 72%. In these experiments, only the temporal information was used and neither the length of pauses nor the fundamental frequencies were used for recognition.

4. CONSISTENCY OF RECOGNIZED TYPES

In the above experiments, the recognized temporal types of rhythm were compared with the types assigned by pauses. But the types assigned by pauses are not necessarily "correct" types because the pauses were automatically decided using only the length of the aligned pauses. Intrinsically, no one knows what the "correct" types are for all the speech corpus.

In order to test the robustness of the trained models of temporal types, the consistency of recognized types was studied. This was done by comparing two sets of recognized temporal types. One set contains the types generated by closed models and the other contains those generated by open models. Figure 7 shows the relationship between the speech corpus and models for testing consistency in recognized types for the case of male speech corpus. In Figure 7, statistical models from set 1 are the closed models for testing the speech corpus of set 1 (sentence numbers 1 - 2259) and statistical models from set 2 are the open models for testing the same speech corpus of set 1.

Table 2 shows the consistency in temporal types recognized by the closed and open models. Results

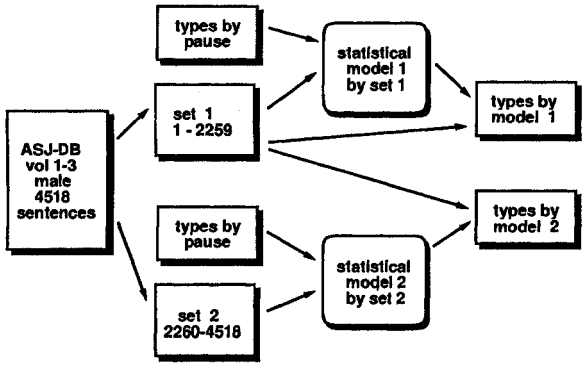

Fig. 7: *Test of consistency in recognized types.*

	Bunsetsu	Same	Different	Rate
male set1	14706	13859	847	94.2%
male set2	14591	13852	739	94.9%
female set1	16864	15933	931	94.5%
female set2	16503	15460	1043	93.7%
Total	62664	59104	3560	94.3%

Table 2: *Consistency in recognized types.*

are shown for male set 1, male set 2, female set 1, and female set 2.

Approximately 94% of bunsetsu phrases were recognized as identical types both for the closed and open models. It is shown the recognized temporal types were very consistent.

5. CONCLUSION

A new framework was proposed for automatically recognizing rhythm in speech where temporal types themselves are modeled and recognized by statistical models. Two levels of normalizations on mora durations in isolated word utterances and read sentences were studied and then used in the recognition of temporal types.

Using the ASJ Continuous Speech Database, experiments for recognizing temporal types of (bunsetsu) phrases were conducted. From a total of 62664 phrases, approximately 72% of the temporal types were identified correctly (i.e.; were identical to the assigned temporal types) using these models. The recognized types were very consistent (approximately 94% had same types) for closed and open models.

These results show the promising potential of the proposed framework to model and recognize temporal types using statistical models. This framework can be extended from read sentences to other styles of more casual speech.

In these experiments, only the temporal information was used and neither the length of pauses nor the fundamental frequencies were used for recognition. Future work will include the integration of temporal and other prosodic information [13].

ACKNOWLEDGEMENT

The authors would like to thank Dr. Nobuyuki Otsu for his support, and members in the Speech Processing Section and the Real World Computing Section of ETL for their technical assistance and discussion. Part of this work was done under the Real World Computing Program.

REFERENCES

[1] Hayamizu, S. : "Role of rhythm and tempo in spoken dialogue", Proc. Int. Sym. on Spoken Dialogue, ISSD-93 pp.177-180 (1993).

[2] Miyatake, M., Sagisaka, Y. : "Prosodic characteristics and their control in Japanese speech with various speaking styles", IEICE Trans. Vol.J73-D-II, No.12, pp.1929-1935 (1990) (in Japanese).

[3] Campbell, N. : "Speech timing in English and Japanese", Preprints of International Symposium on Japanese Prosody, pp.207-216 (1992-11).

[4] Takagi, K., Houra, N., Itahashi, S. : "Characteristics of various segment durations in simulated dialogue speech", Tech. Rep. of IEICE, SP92-111 (1992) (in Japanese).

[5] Rabiner, L.R. : "A tutorial on hidden Markov models and selected applications in speech recognition", Proc. of the IEEE, Vol.77, No.2, pp.257-286 (1989).

[6] Matsuo, H., Makino, S., Kido, K.: "Spoken word recognition using a verification method based on phoneme duration model", IEICE Trans. D-II, Vol. J73-D-II, No.12, pp.1936-1944 (1990) (in Japanese).

[7] Takizawa, Y., Tsuboka, E. : "Syllable duration prediction for speaker independent continuous speech recognition", IEICE Trans. A, Vol.J77-A, No.2, pp.173-181 (1994) (in Japanese).

[8] Mitchell, C.D., Jamieson, L.H. : "Modeling duration in a hidden Markov model with the exponential family", ICASSP, II-331-334 (1993).

[9] Singer, H., Takami, J., Matsunaga, S.: "Mora duration models for SSS-LR continuous speech recognition", Preprint of ASJ Meeting 2-7-3 (93-10).

[10] Osaka, N. : "Conversational turn-taking model using Petri net", ICSLP, pp.1297-1300 (1990).

[11] O'Shaughnessy, D. : "Recognition of hesitations in spontaneous speech", ICASSP, I-521-524 (1992).

[12] Geoffrois, E. : "Prosodic event detection from F0 contours using the Fujisaki model", Preprint of ASJ Meeting 1-8-22 pp.187-188 (93-03).

[13] Yoshimura, T., Hayamizu, S., Tanaka, K. : "Word accent patterns modelling by concatenation of mora hidden Markov models", ICASSP, I-69-72 (1994).

[14] Deutsch, D. : "The psychology of music", Academic Press 1982.

[15] Itou, K., Hayamizu, S., Tanaka, H. : "Continuous speech recognition by context-dependent phonetic HMM and an efficient algorithm for finding N-best sentence hypotheses", ICASSP, S10.6 (1992).