TEMPORAL ENVELOPE MODULATION USING SYLLABLE SEARCH METHOD FOR ROBUST LANGUAGE IDENTIFICATION

PACS: 43.72.Ar

Aoki, Toshimasa[1]; Komatsu, Masahiko[1, 2]; Arai, Takayuki[1]; Murahara, Yuji[1]
[1]Dept. of Electrical and Electronics Eng., Sophia University
7-1 Kioi-cho, Chiyoda-ku
Tokyo 102-8554
JAPAN
TEL: +81-3-3238-3417
E-mail: aoki-t@splab.ee.sophia.ac.jp
[2]Dept. of Linguistics, University of Alberta
Edmonton T6G 2E7
CANADA

ABSTRACT
Humans are quite capable of identifying languages under noisy conditions while computers still struggle. Robust automatic language identification systems are needed, because there is no place that is totally silent. In this study, multi-layer perceptron is applied using Temporal Envelope Modulation (TEM), a speech signal with reduced spectral information, which is similar to a speech signal in a noisy environment. The experiment used a new method for feature extraction, the Syllable Search Method (SSM), and as a result, the identification rate increased as the number of bands of TEM increased from one to four.

INTRODUCTION

In recent years, the boundaries of countries and spoken languages are becoming transparent because of growing access to the Internet and telecommunications, as well as increased travel. The need to identify which language is spoken is of growing importance. Considering the naturally unpredictable environment in which speech production and perception takes place in the real world, robustness in an automatic language identification (LID) system is imperative.

Previous research has focused on the source information used for automatic LID, that is, segmental and non-segmental information. The majority of LID research has focused on segmental information, using the acoustic property of segments and their alignment ("acoustic phonetics" and "phonotactics" as defined by Muthusamy et al. [1]) [2,3,4].

Much less attention has been directed at non-segmental information ("prosodics" [1]). However, Mori et al. used a speech signal with reduced segmental information as source information for human LID [5], a Temporal Envelope Modulation (TEM) signal, and we used their TEM signal in our experiment.

To construct a noise-robust LID system, it is helpful to use signals with reduced information. Therefore, learning from Mori, et al., we conducted an automatic LID experiment using a TEM signal. Although they did not identify an LID cue from the TEM signal, we focused on syllable structure, which is the main topic of this paper.

The Syllable Search Method (SSM) used in our experiment is a new method that extracts syllable features from the speech signal. This method is very effective when searching for a syllable length feature. This new method is valuable because it captures the significant cues to syllable structure available in a signal with a low sound-to-noise ratio.

We conducted this experiment to test our robust automatic LID system. The experiment will be discussed in three parts. First is the source of information, the Temporal Envelope Modulation signal. Second is the feature extraction of the TEM signal where we used the new Syllable Search Method. Third is the neural network, a multi-layer perceptron described later, which identifies the language being spoken.

To simplify matters, we limit our task to the discrimination of English and Japanese. This will provide us the basic of this approach, and our method could be used for future work by introducing additional languages. In the next section we describe the input signal used in the experiment, the TEM signal. The third section describes the experimental procedure, which includes SSM and the neural network. Finally, our results are discussed in the fourth section.

EXPERIMENTAL DATA

Temporal Envelope Modulation (TEM)

A Temporal Envelope Modulation signal was used as input. The processing of this signal was done by Mori et al. [5]. Starting with a speech signal from the OGI-TS corpus of telephone speech [6], they removed pitch information and reduced spectral information by means of TEM processing. The resulting TEM signal is meant to simulate a noisy environment, which is also characterized by reduced spectral information.

The TEM signal is a white-noise driven signal retaining the intensity information of several frequency bands of the original speech signal but not its pitch information. To create the TEM signal, the temporal envelope of intensity was extracted in each of several broad frequency bands, and these envelopes were used to modulate noises of the same bandwidths. The bands are divided into one to four as shown in Fig. 1 (TEM 1, 2, 3 and 4), following Shannon et al. [7].

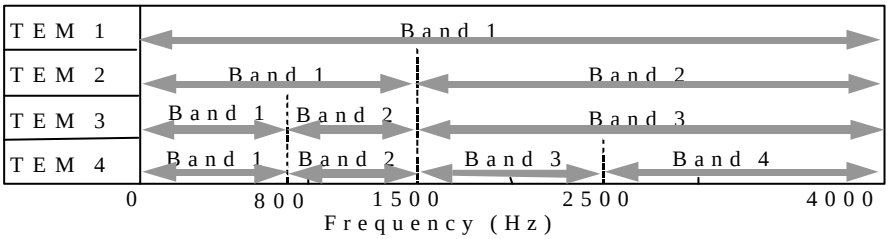

Fig. 1.- Frequency of bands

Figure 2 shows the block diagram of TEM 4. The speech signal was divided into four signals by bandpass filters designed by the Kaiser window (transition region width: 100 Hz; tolerance: 0.001). The outputs of the bandpass filters were converted to Hilbert envelopes, which were further low-pass filtered with the cutoff at 50 Hz. These signals represent the temporal envelopes of the respective bands. Then the white noise, limited by the same bandpass filters used for the speech signal, was modulated by the temporal envelopes and summed up. The amplitude of the signals was then normalized using their peak values.

Data sets for TEM 1, 2, 3, and 4 each had 80 signals, 40 English and 40 Japanese. In each set there were 20 English males, 20 English females, 20 Japanese males, and 20 Japanese females. Each signal was 10 s long in duration.

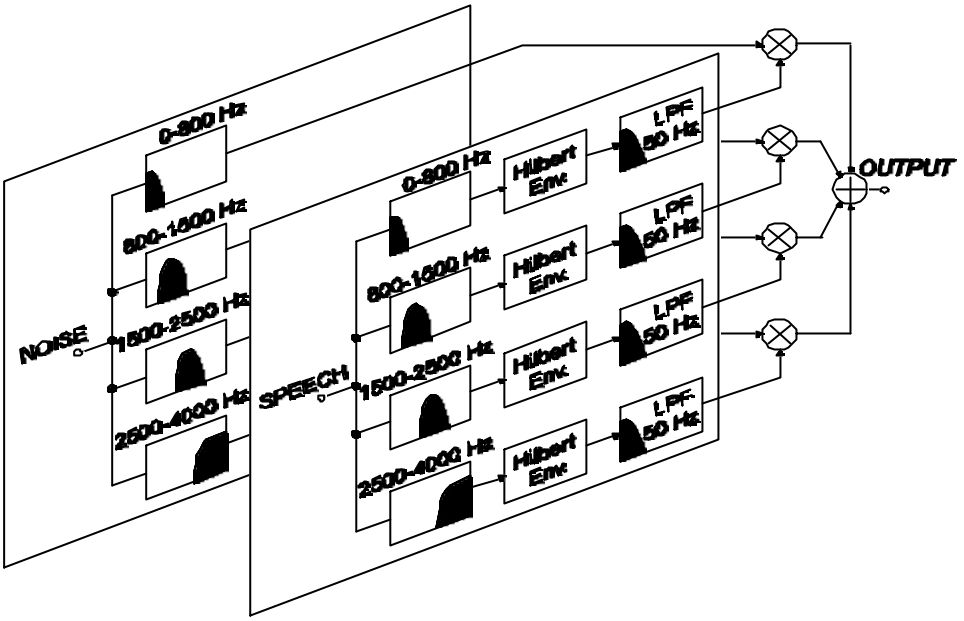

Fig. 2.- Block diagram of TEM 4

EXPERIMENTAL PROCEDURE

Preprocessing

First, each TEM signal was bandpass filtered to each band: TEM 1 was left alone; TEM 2 was divided into two bands; TEM 3, into three bands; and TEM 4, into four bands. The envelope of power for each band was extracted. These envelopes were down-sampled to 32 Hz from 8000 Hz. Then these envelopes were summed up across the bands, creating an envelope of the total power of all of the bands. The total power envelope was used for SSM, and the band power envelopes were input to the neural network.

Feature Extraction: Syllable Search Method (SSM)

To identify which language is spoken, the multi-layer perceptron needs features from the TEM signal. For the process of feature extraction we present the Syllable Search Method.

As shown in Fig. 3, the correlation between the total power envelope of the speech signal (the top panel) and the reference signal (the middle panel) was calculated (the bottom panel). The reference signal used was one cycle of $1-\cos$ at 4 Hz, that is, 250 ms, which is approximately the same length as a syllable [8]. The reference signal slid from the beginning to the end of the total power envelope of the speech signal, and searched the correlation rate with the speech signal. Thus, the ten 250ms syllables in the total power envelope of the speech signal which corresponded with the top ten correlation rates were selected for each TEM signal, and the band power envelopes of those syllables were used as the features. Given the 80 signals in TEM 1, 2, 3, and 4, there were 800 syllables for each TEM, and these became the inputs to the neural network. Since the feature was an envelope, it contained phase information as well as amplitude information.

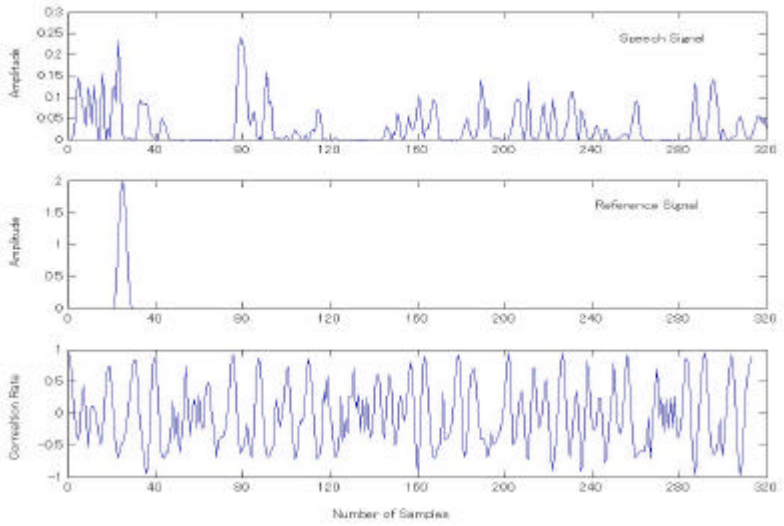

Fig. 3.- Process of SSM

Neural Network

Our neural network was a three-layer perceptron. The number of inputs depended on the TEM signal: 8 for TEM 1, 16 for TEM 2, 24 for TEM 3, and 32 for TEM 4, because each band power envelope of a syllable had eight samples. There were 500 hidden layer neurons, and two outputs, "English" and "Japanese".

In the training session we used a back-propagation algorithm. The band power envelopes of a syllable were put into the neural network. If the syllable was from an English utterance, 1 was given to the "English" node, and 0 was given to the "Japanese" node, and vice versa for the Japanese utterances.

In the test session, the values of the two output nodes were compared for each syllable. The values were between 0 and 1. When the English output value was greater than the Japanese output value, the answer was judged as English; and vice versa for Japanese.

To divide the TEM signals into training and test sets, a five-fold cross-validation was applied. Eighty utterances were divided into five blocks so that each block was balanced with respect to language and gender (one of each: Japanese, male/female; English, male/female). Out of the five blocks, four blocks were used for training, and the other was further divided into half for the development test and the final test. The development test was used to check for over-training of the neural network. Because one block had 16 utterances and 10 syllables were selected from each utterance, 640 syllables in total (4 blocks * 16 utterances * 10 syllables) were used for training, 80 syllables (0.5 block * 16 utterances * 10 syllables) were used for the development test, and 80 syllables were used for the final test. By rotating the blocks, five trials were conducted.

EVALUATION

Result

The identification rates in the final tests are shown in Table 1. These rates are the averages of the five trials of the cross-validation test. The overall identification rates are shown in Fig. 4. Figure 5 shows the identification rates for each of the four types of TEM signals in the test set: English male, English female, Japanese male, and Japanese female. Figure 6 shows the identification rates in two categories, language and gender.

Table 1.- Identification rates [%]

	Overall	English	Japanese	Male	Female	English Male	English Female	Japanese Male	Japanese Female
TEM 1	54.0	54.0	54.0	54.3	53.8	53.0	55.0	55.5	52.5
TEM 2	56.4	55.3	57.5	61.5	51.3	60.0	50.5	63.0	52.0
TEM 3	60.1	59.3	61.0	63.5	56.8	63.0	55.5	64.0	58.0
TEM 4	61.4	61.0	61.8	64.8	58.0	68.0	54.0	61.5	62.0

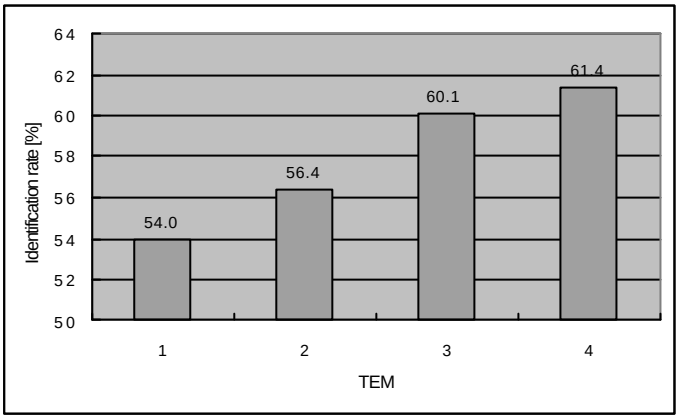

Fig. 4.- Identification rate overall

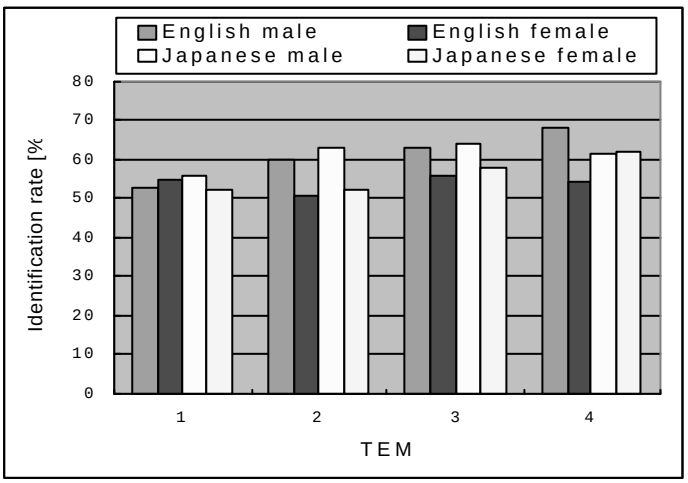

Fig. 5.- Identification rate in types

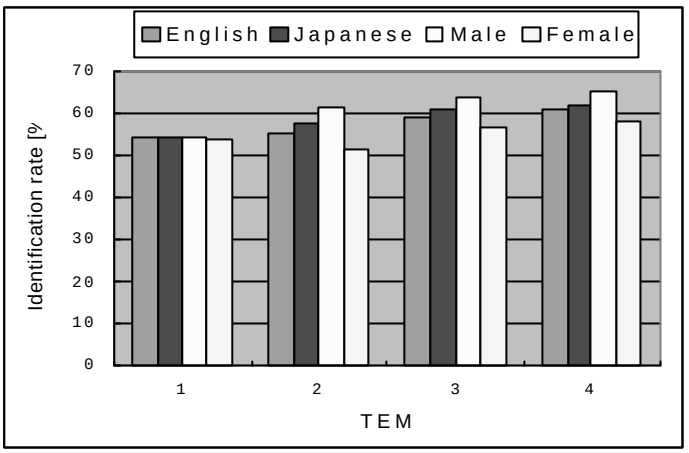

Fig. 6.- Identification rate in categories

Discussion

Figure 4 shows that the identification rate increases as the number of bands of TEM increases from one to four. From Fig. 5 and Fig. 6, the identification rate for English females was low compared to the other utterance types, while male utterances had high identification rates when categorized. Note that the identification rates presented in this study are based on the number of correctly identified syllables and not the number of utterances. Therefore, our results show how effective our method of feature extraction is for LID; they do not show the LID results of utterances. Specifically, our results show that syllable structure is a significant cue for automatic LID.

CONCLUSION

In this paper, a new method of feature extraction is introduced, the Syllable Search Method. This method is effective when combined with other algorithms, when syllable structure is the feature for automatic LID. Our method adds robustness to the system since its features are strong in noisy environments.

I thank Terri Lander for the useful comments on writing this paper.

BIBLIOGRAPHICAL REFERENCES

[1] Y. K. Muthusamy and R. Cole, "Reviewing automatic language identification," *IEEE Signal Processing Mag.*, Vol. 11, No. 4, pp. 33-41, 1994.

[2] Y. K. Muthusamy, K. Berkling, T. Arai, R. Cole and E. Barnard, "A comparison of approaches to automatic language identification using telephone speech," *Proc. of Eurospeech*, Vol. 2, pp. 1307-1310, 1993.

[3] T. Arai, "Automatic language identification using sequential information of phonemes," *Trans. IEICE Japan*, Vol. E78-D, No. 6, pp. 705-711, 1995.

[4] J. Hieronymus and S. Kadambe, "Robust spoken language identification using large vocabulary speech recognition," *Proc. of ICASSP*, Vol. 2, pp. 1111-1114, 1997.

[5] K. Mori, N. Toba, T. Harada, T. Arai, M. Komatsu, M. Aoyagi, and Y. Murahara, "Human language identification with reduced spectral information," *Proc. of Eurospeech*, Vol. 1, pp. 391-394, 1999.

[6] Y. K. Muthusamy, R. A. Cole and B. T. Oshika, "The OGI multi-language telephone speech corpus," *Proc. of ICSLP*, Vol. 2, pp. 895-898, 1992.

[7] R. V. Shannon, F. G. Zeng, V. Kamath, J. Wygonski and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, Vol. 270, pp. 303-304, 1995.

[8] T. Arai and S. Greenberg, "The temporal properties of spoken Japanese are similar to those of English," *Proc. of Eurospeech*, Vol. 2, pp. 1011-1014, 1997.