SPEAKER AND TEXT INDEPENDENT LANGUAGE IDENTIFICATION USING PREDICTIVE ERROR HISTOGRAM VECTORS

*Qian-Rong Gu and Tadashi Shibata**

Department of Electronic Engineering, The University of Tokyo
tt07142@mail.ecc.u-tokyo.ac.jp
*Department of Frontier Infomatics, School of Frontier Science, The University of Tokyo
shibata@ee.t.u-tokyo.ac.jp

ABSTRACT

A predictive vector quantization[1][2] based speaker and text independent language identification system is proposed, which uses statistical distribution of predictive error vectors to recognize the language spoken by native speakers. According to Stan C. Kwasny[3], most high level features of speech, such as tone of voice, rhythm, style, pace, accent, etc, appear to be related to distributional patterns or statistical aggregates of speech waveforms. We further develop the method used in [4][5] to extract these statistical distributional patterns directly from raw speech waveforms and then use them to identify language. The system has been trained and tested by speeches from English and Japanese native speakers. A best identification ratio of 76.8% can be achieved by our system.

1. INTRODUCTION

Automatic language identification is a problem of identifying the language being spoken by an unknown speaker. Speech plays an increasingly important role in telecommunications as the global economic community expands. In order to translate among languages, inform or instruct speakers and so on, there is a growing need for automatic spoken language identification technologies.

Languages have characteristic sound patterns and significantly different prosodic patterns. They differ in the inventory of phonological units used to produce words, the frequency of occurrence of these units and the order in which they occur in words[6][7]. The key point of solving language identification problem is how to detect and exploit such differences between languages. In the field of speech recognition, the objective is to determine the content of the speech, typically implemented by phoneme recognition coupled with word recognition and sentence recognition. This requires investigating features of small portions of the speech, such as frames, phonemes, syllables, and so on, to determine what the speaker said. In contrast, in text-independent language identification, such small sub-word units alone are not enough, since some phonemes and syllables and even words are common across different languages. The complexity of this phenomenon makes it impossible to capture unique characteristics that can make one language sound distinct from another by any simplistic models such as Hidden Markov Models[6][7].

When humans listen, they constantly make a variety of judgments about features of current speeches, such as tone of voice, rhythm, style, pace, accent, etc. These high-level features are not tied to any particular, conventional set of phonetic or acoustic features of the speech. Instead, they appear to be related to distributional patterns or statistical aggregates of raw speech waveforms[3]. While, Stan C. Kwasny has used a feed-forward network and a recurrent neural network to identify language of English and French directly from raw speech waveforms[3], we further develop the method used in [4][5] to extract statistical distributional patterns of languages from predictive error of the raw speech waveforms, and then use these patterns to identify the language between English and Japanese.

The remainder of this paper is organized as follows. Section 2 describes our method in details, Section 3 provides experiments and experimental results of this system, and discussions are contained in Section 4.

2. METHOD OF ESTABLISHING LANGUAGE MODELS

The current speech waveform can be predicted by observing its past data.

$$\hat{x}_n = \sum_{k=1}^{m} a_k x_{n-k} \qquad (1)$$

Where $\hat{x}_n$ is the prediction of the signal X at time n, $x_{n-m} \ldots x_{n-1}$ are m observations of the signal X before

time n. $a_1 \ldots a_m$ are corresponding coefficients of x_{n-m} … x_{n-1}. m is order of the predictor. If the current value of the waveform is x_n, then the predictive error at time n can be written as:

$$e_n = x_n - \hat{x}_n \qquad (2)$$

This fact implies there are some kinds of general properties hidden in the waveforms. The predictor can "figure out" such common properties and leave specific properties in the predictive error.

If we build a predictor from a combination training data of English speeches and Japanese speeches, then this predictor represents the common properties of both languages, and leaves the individual properties specific to each language in predictive error. The language identification problem becomes to how to detect and exploit these individual properties in the predictive error.

Vector Quantization (VQ)[1][2] is a very efficient approach to extract typical features from raw data by mapping raw data vectors X = { x_i | i = 1 ... L} into K (K<< L) clusters such that similar raw data vectors are grouped together and vectors with different features belong to different groups (code vectors)[1][2]. By applying VQ to the predictive error, we can dig out the feature-vectors (codebook) specific to both languages from predictive error. Now the problem is how to use these code vectors to build language model.

The answer is quite straightforward, applying several sets of data containing only English speeches or only Japanese speeches to the predictor, and then VQing the predictive error by the codebook mentioned above to count usage of each code vector, which in all stand for the property of predictive error. The counting results of all the code vectors in the codebook form a usage histogram. From the viewpoint of mathematics, this usage histogram itself can be looked and dealt as a vector, a usage histogram vector of predictive error codebook. It is this histogram that can be used as language model for identification purpose.

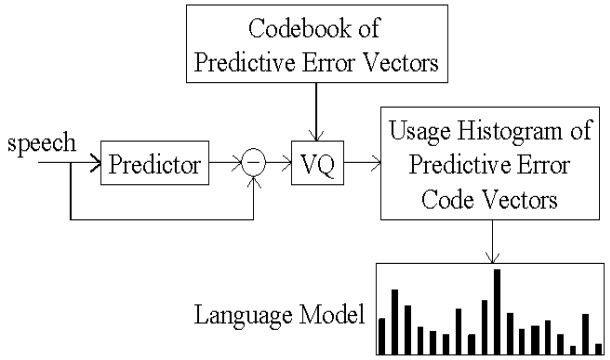

Fig. 1 The procedure of building language models

Figure 1 shows the procedure of building language models. The detailed steps will be described in the following. In training phase:

Stage I:

1. Prepare a training data set containing both English speeches and Japanese speeches.
2. Build a predictor from the above training data.
3. Apply the same training data to the predictor to get predictive error.
4. Use the Generalized Lloyd Algorithm[8] to extract a codebook from the predictive error.
5. Optimize the predictor and the codebook for each other[9].

Stage II:

1. Prepare several new training data sets containing either English speeches or Japanese speeches.
2. Apply these training data to the predictor and the codebook to get usage histogram vectors of each language.
3. Once again use theGeneralized Lloyd Algorithm[8] to extract typical patterns for each language from the above histogram vectors.
4. These two sets of typical patterns are the models of two languages.

In identification phase, the calculating steps are similar to the Stage II of the training phase except that instead of the training data, real-time input speeches are applied to the predictor and the codebook. Then the real-time calculated usage histogram vector is matched against the two sets of language models, the set containing the nearest typical pattern to the real-time histogram vector is considered as identification result.

3. EXPERIMENTS AND EXPERIMENTAL RESULTS

In this section, firstly we explain collection of speeches used to train and test the system; secondly we describe experiments used to test the system; at last we provide experimental results.

3.1. Data Collection

We collected English speeches from 8 native speakers, 4 male and 4 female. Japanese speeches were collected from 10 native speakers, 6 male and 4 female.

The training data set used to calculate the predictor and the codebook is composed of 30 speech segments for each language randomly selected from the above speeches, and each segment is 20 seconds long.

We prepared another 250 20-second-long speech segments not appeared in the above training data for each language to create language models, namely usage histogram of predictive error code vectors.

Additional 50 20-second-long speech segments not used yet for each language were arranged for test purpose.

In order to investigate influence of unknown speakers, whose voice did not appear in the training speeches, on identification ratio, we got further 20 20-second-long speech segments for each language both taken from another 2 native speakers, 1 male and 1 female.

The original speeches are CD quality. All of speech segments used in training and test were re-sampled at a rate of 8.0 kHz with a resolution of 16 bits per sample.

3.2. Experiments

The dimension of raw waveform vector is 8, and the rank of predictor is also 8.

In training phase, we calculated three codebooks with different size, i.e. 64, 128 and 256. It means the usage histogram vectors have three dimensions, 64, 128 and 256. Each language model set was made up of 5 typical patterns extracted from its 250 training usage histogram vectors by the Generalized Lloyd Algorithm[8].

In test phase, we test the system by two different strategies called Top 1 and Top 3 scoring method with speeches from the speakers same as those of training speeches, and with speeches from the speakers whose voice did not appear in the training speeches.

The Top 1 scoring strategy means only the nearest distance of 5 typical patterns in each language model set was compared to give out identification result. While, the Top 3 implies the mean of the first 3 nearest distances of 5 in each model set was compared.

3.2. Experimental Results

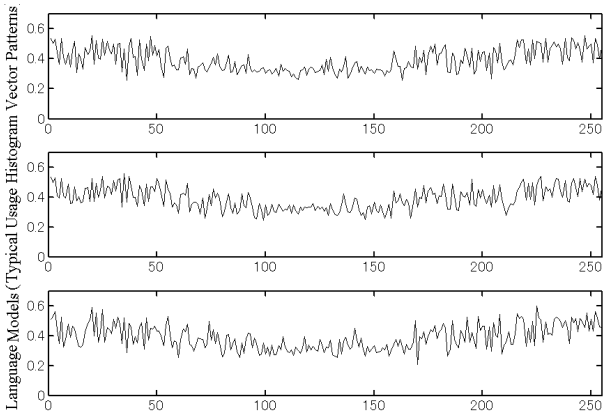

Fig.2 Three of five typical patterns in the English language model-set. (predictive error codebook size: 256)

Figures 2 and 3 show 3 of 5 typical patterns in the language model-set of English and Japanese with codebook size of 256, respectively. From these two figures we can see two apparently different “tendency” of usage histogram vectors of two languages. It is this different “tendency” that is employed for identification purpose.

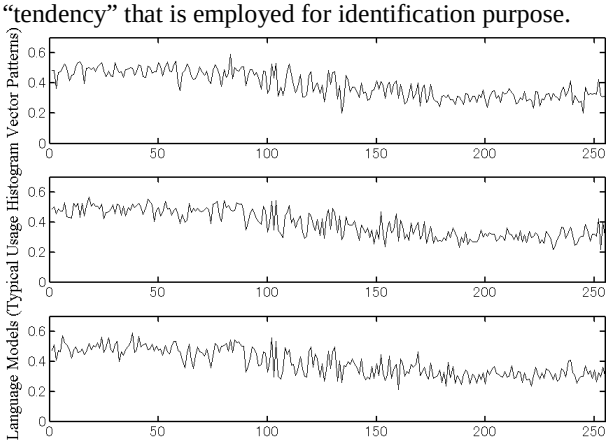

Fig.3 Three of five typical patterns in the Japanese language model-set. (predictive error codebook size: 256)

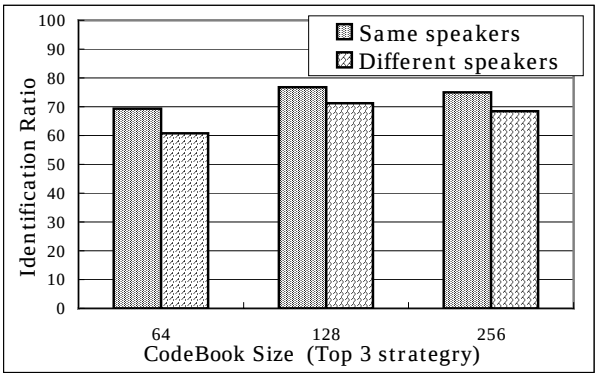

Fig.4 Identification ratio of 3 different codebook sizes and different speaker-groups with the Top 3 strategy.

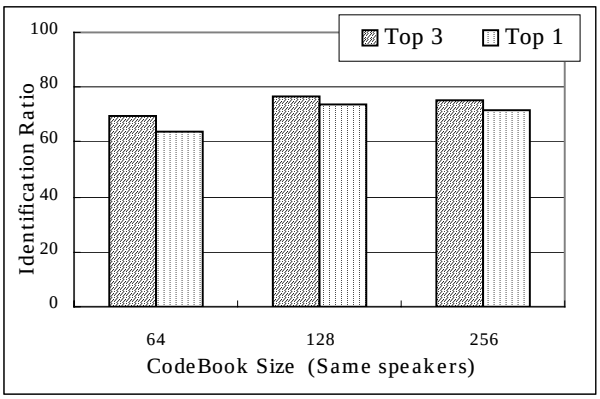

Fig. 5 Identification ratio of 3 different codebook sizes and 2 different scoring strategies with the same speaker-groups

Figure 4 showed the identification ratio tested by 100 20-second-long speech segments from the speakers same as those of training speeches and by 40 20-second-long speeches from speakers whose voice did not appear in the training data with the Top 3 scoring strategy. We can see the best identification ratio 76.8% was achieved under the condition of 128-codebook-size with the speeches from the same speakers. Though for all 3 different codebook sizes the identification ratio degraded when test speeches from new speakers were used, the degradations were not very serious, the system showed some kind of speaker-independent property.

Figure 5 showed the identification ration tested by 100 20-second-long speech segments from the speakers same as those of training speeches with two different scoring strategies. For all of three cases, the Top 3 strategy got better identification ratios than the Top 1 strategy.

The identification ratios of another two combinations of speakers and scoring strategies show in Tables 1 and 2.

Table 1: Test with the Top 1 scoring strategy

Top 1 Strategy	Codebook Size		
	64	128	256
Same speakers	63.7%	73.8%	71.4%
Different speakers	58.6%	68.5%	65.7%

Table 2: Test with different speakers

Different speakers	Codebook Size		
	64	128	256
Top 3	60.8%	71.2%	68.4%
Top1	58.6%	68.5%	65.7%

5. DISCUSSIONS

Although our results are preliminary, it is really surprising that such a straightforward approach can achieve a reasonable identification ratio. Comparing to methods based on explicit phonetic identification as well as a variety of other intermediate level structuring typically found in speech recognition system, our approach uses the statistical distributional patterns extracted directly from raw speech waveforms to identify languages. Our system could not get a better performance than Stan C. Kwasny's[3]. Nevertheless, our work has verified his assumption that some high-level features specific to different languages appear to be related to statistical aggregates of the speech waveforms.

6. REFERENCES

[1] M. Gray, "Vector Quantization," *IEEE ASSP Magazine*, pp. 4-29, April 1984.

[2] Anil K. Jain, Robert P.W. Duin, and Jian-chang Mao, "Statistical Pattern Recognition: A Review." *IEEE Trans. on Pattern Analysis and Machin Intelligence*, pp.4-37, Jan 2000.

[3] Stan C. Kwasny, Barry L. Kalam, A.Maynard Engebretson, and Weilan Wu, "Identifying Language from Raw Speech: An Application of Recurrent Neural Networks", In *Proceedings of the 5th Midwest Artification Intelligence and Cognitive Science Society Conference*, pp. 53-57, 1993,

[4] Qian-Rong Gu and Tadashi Shibata, "A Computationally Efficient Modeling Method For Text Dependent VQ-Based Speaker Identification System", *Sixth World Multiconference on Systemics, Cybernetics and Informatics (SCI2002)*, Orlando Florida, July 2002

[5] Qian-Rong Gu and Tadashi Shibata, "Modeling Time Series Signal Patterns By Statistical Distribution Of Prediction Errors And Its Application To Speaker Identification", *6th World Multiconference on Systemics, Cybernetics and Informatics (SCI2002)*, Orlando Florida, July 2002

[6] Y. K. Muthusamy, "A Segmental Approach to Automactic Language Identification", *PhD thesis*, Oregon Graduate Instiute, July 1993

[7] K. Berkling, T, Arai, E. Barnard, and R.A. Cole, "analysis of phoneme-based features for language identification", In *Proceedings 1994 IEEE International Conference on Acoutics, Speech, and Signal Processing*, page I-289-I-292, Adelaide Australia, april 1994

[8] Linde Y., Buzo A., and Gray R.M. "An algorithm for vector quantizer design", *IEEE Trans. on Communications*, 28(1): 84-95, January 1980.

[9] Gersho A., and Gray R.M, "Vector Quantization and Signal Compression", Chapter 11, Dordrecht: Kluwer Academic Publishers, 1992.