S12.8

Automatic Language Recognition Using Acoustic Features

Masahide Sugiyama

ATR Interpreting Telephony Research Laboratories
Sanpeidani, Inuidani, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

ABSTRACT

Language recognition (e.g. Japanese, English, German, etc) using acoustic features is an important yet difficult problem for current speech technology. In this paper, two language recognition algorithms are proposed and some experiment results are described. While many studies have been done concerning the speech recognition problem, very few studies have addressed the language recognition task.

The speech data base used in this paper contains 20 languages: 16 sentences uttered twice by 4 males and 4 females. The duration of each sentence is about 8 seconds. The first algorithm is based on the standard Vector Quantization (VQ) technique. Every language, k, is characterized by its own VQ codebook, V_k. The second algorithm is based on a single universal (common) VQ codebook, $U = \{u_i\}$, for all languages, and its occurrence probability histograms, h_k. Every language, k, is characterized by a histogram, h_k. The experiment results show that the recognition rates for the first and second algorithms were 65% and 80%, respectively, each using just 8 sentences of unknown speech (about 64 seconds).

1 Introduction

Language recognition (e.g. Japanese, English, German, etc) using acoustic features is an important yet difficult problem for current speech technology. In this paper, two language recognition algorithms are proposed and some experiment results are described. While many studies have been done concerning the speech recognition problem, very few studies have addressed the language recognition task [1],[2],[3],[4],[5].

The first paper concerning to automatic language recognition algorithm was reference[1]. In their paper, each language is characterized and represented by one Hidden Markov Model. Their study was just theoretical and produced no experiment results. Recently several studies have been addressed to language recognition in noisy environments. However, in these studies the number of languages is too small, in most cases, two or three languages. This paper reports large number of language recognition techniques based on the VQ algorithm. The first algorithm is based on the standard Vector Quantization (VQ) technique. The second algorithm is based on a single universal (common) VQ codebook, $U = \{u_i\}$, for all languages, and its occurrence probability histogram, h_k.

This paper contains the following sections: in the second section, the multilingual speech database is described. In the third section, two recognition algorithms are proposed. In the fourth section, the results of recognition experiments are described.

2 Speech Database and distance measures

2.1 Multilingual Speech Database

The multilingual speech database contains 20 languages[6]: 16 sentences uttered twice by 4 males and 4 females. The duration of each sentence is about 8 seconds. The list of languages and their abbreviations is shown in Table 1. The speech data was carefully divided into training and test sets, recognition experiments being designed as both speaker-independent and text-independent. The amount of data for each part is approximately the same: 20.9 min for training and 20.5 min for test.

Table 1: 20 languages and their abbreviations

language	abbre-viation	language	abbre-viation
American	AM	Hindi	HI
Arabic	AR	Hungarian	HU
Chinese	CH	Italian	IT
Danish	DA	Japanese	JA
Dutch	DU	Norwegian	NO
English	EN	Polish	PL
Finnish	FI	Portuguese	PR
French	FR	Russian	RU
German	GE	Spanish	SP
Greek	GR	Swedish	SW

Table 2: Acoustic features and specification for speech analysis

acoustic feature	autocorrelation coefficient LPC cepstrum coefficient Δ cep coefficient
autocorrelation analysis order	13
LPC analysis order	10
frame length	128 (16ms)
sampling frequency	8 kHz
VQ relative distortion	0.01

2.2 Acoustic parameters and spectral distance measures

LPC analysis is applied to produce acoustic features[7], each speech waveform generating a sequence of these feature vectors. The acoustic feature parameters and specification for speech analysis is shown in Table 2.

Calculating the VQ distortion, several LPC spectral distortion measures shown in Table 3 were applied: LPC Cepstrum distance (CEP), Weighted Likelihood Ratio (WLR), Frequency Weighted Cepstrum distance (FWCEP), Differential Cepstrum distance (Δ CEP), Quefrency Weight Cepstrum distance (WCEP). Each distance measure has its physical meaning and corresponds to logarithmic spectral distance, spectral peak weighted distance, low frequency axis weighted distance, time axis differential of LPC spectrum and quefrency axis weighted distance, re-

CH2977-7/91/0000-0813 $1.00 © 1991 IEEE

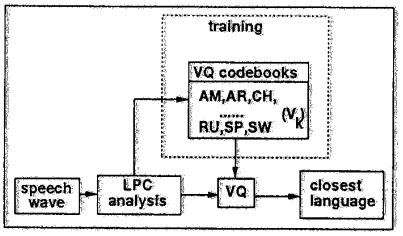

Figure 1: Procedure of the Standard VQ algorithm

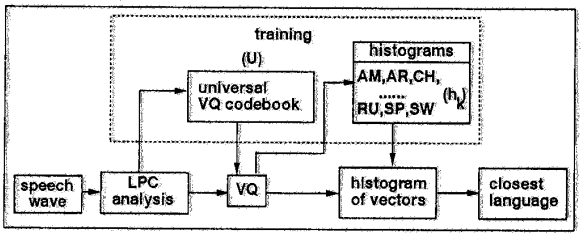

Figure 2: Procedure of the VQ histogram algorithm

spectively.

3 Recognition algorithms

The first algorithm is based on the standard Vector Quantization (VQ) technique. Every language, k, is characterized by its own VQ codebook, V_k. In the recognition stage input speech is quantized by V_k and the accumulated quantization distortion, d_k, is calculated. The language which has the minimal accumulated distortion is recognized. Calculating the VQ distortion, several LPC spectral distortion measures are applied.

The second algorithm is based on a single universal (common) VQ codebook, $U = \{u_i\}$, for all languages, and its occurrence probability histogram, h_k. Every language, k, is characterized by a histogram, h_k. In the recognition stage, input speech is quantized by U and its histogram function, $h(u_i)$, is calculated. The language which minimizes the distance between h and h_k is recognized. The same LPC distortion measures as above are applied in the VQ stage. A Euclidean distortion measure is used to measure histogram separation.

3.1 Standard VQ algorithm

A codebook, $V_k = \{v_{k,j}\}$, for each language is generated using training sentences. The accumulated distance for input vector in sentence, x_i, is defined as;

$$d_k = \sum_{i=1}^{n} \min_j d^2(x_i, v_{k,j}). \quad (1)$$

The distance 'd' can be any distance which corresponds to the acoustic features and it must be the same as the one used for codebook generation. Each language is characterized by its VQ codebook, V_k. The procedure is shown in Fig.1.

3.2 VQ histogram algorithm

A universal codebook, $U = \{u_j\}$, is generated using all training data. Each language is characterized by the occurrence probability histogram, $h_{k,j}$ of each vector, u_j, of the universal codebook. Each testing sentence is also quantized by U and calculated by its occurrence probability histogram, h. The distance between a test language and a reference language, k, is the Euclidean distance between the histograms, h and h_k.

$$d_k = d(h, h_k). \quad (2)$$

The procedure is shown in Fig.2. This similar idea has already been applied to the speaker identification problem[16].

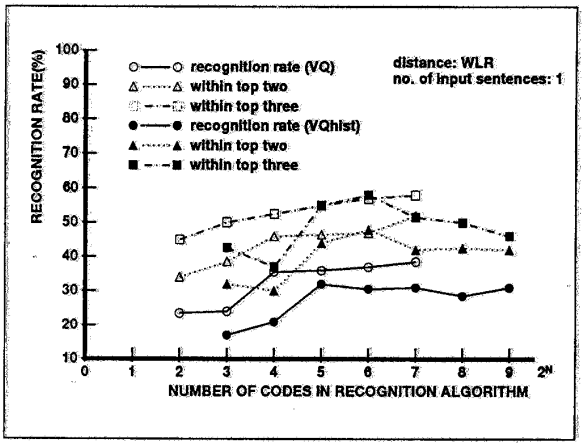

Figure 3: Relationship between codebook size and recognition rates

Table 4: Specifications of recognition experiments

specification	value
frame shift in training	128 (about 3800 training frames)
frame shift in test	64
VQ codebook size (VQ)	4-128
VQ codebook size (VQhist)	8-512

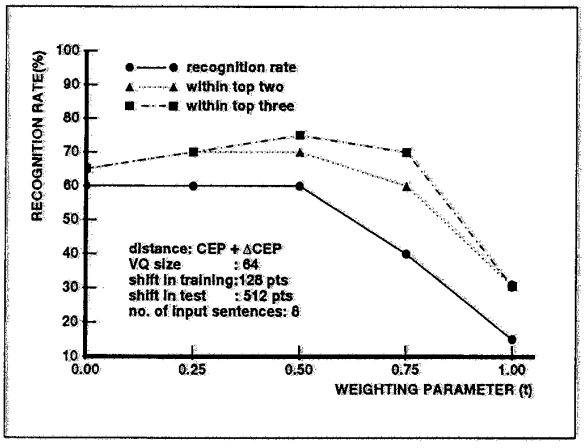

Figure 4: Recognition accuracy for CEP + ΔCEP distance

Table 3: LPC spectral distance measures

distance measure	integrant	parameter	note
CEP[8]	$(\log f(\lambda) - \log g(\lambda))^2$	$\sum_{i=1}^{N}(c_i^{(f)} - c_i^{(g)})^2$	$N = 16$
WLR[11],[13]	$(f(\lambda) - g(\lambda))(\log f(\lambda) - \log g(\lambda))$	$\sum_{i=1}^{N}(r_i^{(f)} - r_i^{(g)})(c_i^{(f)} - c_i^{(g)})$	$N = 16$
ΔCEP	-	$\sum_{i=1}^{N}(\Delta c_i^{(f)} - \Delta c_i^{(g)})^2$	$N = 10$, Δ frame size = 5
CEP + ΔCEP	-	$(1-t)\frac{1}{\sigma_{CEP}}CEP + t\frac{1}{\sigma_{\Delta CEP}}\Delta CEP$	$\sigma_{CEP}, \sigma_{\Delta CEP}$: deviation of CEP, ΔCEP $(0 \le t \le 1)$
WCEP[14]	-	$\sum_{i=1}^{N} w_i(c_i^{(f)} - c_i^{(g)})^2$	$w_i = \begin{cases} i^2 & (1 \le i \le i_0) \\ i_0^2 & (i_0 < i) \end{cases}$ $(i_0 = 6, N = 10)$
FWCEP[12],[13]	$(\log f(\lambda) - \log g(\lambda))^2 w_F(\lambda)$	-	w_F: frequency weighting function

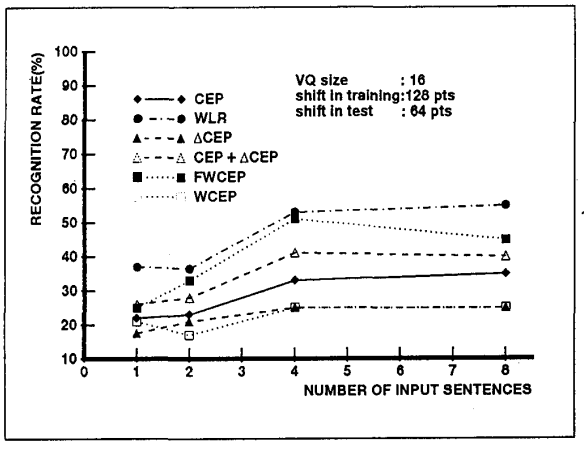

Figure 5: Recognition results using the standard VQ algorithm (VQ size = 16)

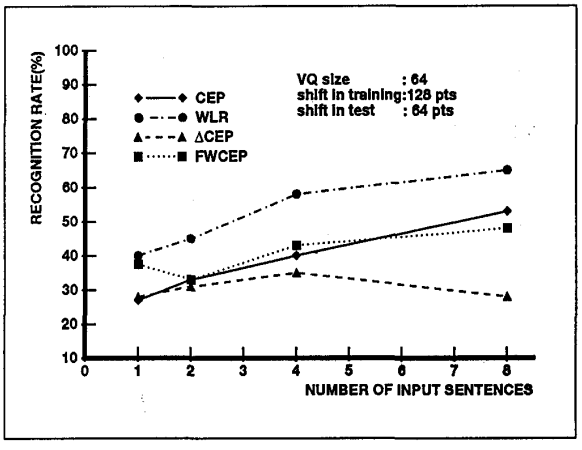

Figure 6: Recognition results using the standard VQ algorithm (VQ size = 64)

4 Experiment Results

4.1 Standard VQ algorithm results

4.1.1 Size of the codebook

The influence of the codebook size has been studied with WLR distance. The specifications of recognition experiments are shown in Table 4. According to the result of the preliminary experiment, the number of shifting points in training and test are set at 128 and 64, respectively. The size of the codebook varies from $4(=2^2)$ to $128(=2^7)$. Fig. 3 shows that the accuracy becomes nearly stable from 16 $(=2^4)$ codes and under 16 it is going down. The size of codebook is set at 16 $(=2^4)$ and 64 $(=2^6)$ for the following experiments.

4.1.2 CEP + ΔCEP distance

The codebooks of Cepstrum and Δcepstrum are generated separately. Fig.4 shows that CEP + ΔCEP is better than CEP alone or ΔCEP alone. However, ΔCEP alone has very poor result. This means that the instantaneous and transitional spectral information is complementary and that more information is contained in the instantaneous domain than in transitional domain for language recognition using this algorithm. The best value for t is 0.5, which is used in the following experiments.

4.1.3 Distance comparison

Fig.5 shows the recognition accuracy for 6 distances; CEP, WLR, ΔCEP, CEP + ΔCEP, FWCEP and WCEP in function of the amount of input data. The size of VQ codebook is set at 16. Here the shifting in training is 128 and the shifting in test is 64. The same test was done on four distances (CEP, WLR, ΔCEP and FWCEP) with 64 codes shown in 6 and the results shown to be 5 to 10% better. Figs.5 and 6 show that the best distance is always WLR. CEP has behaves in a similar manner.

ΔCEP did not perform well and this confirms the previous test. The good results for WLR measure seems to indicate that the most important information is contained in the peaks in instantaneous spectral domain. FWCEP results seems to show that frequency weighting can improve accuracy.

4.2 VQ histogram algorithm results

4.2.1 Size of the universal codebook

Using the WLR measure, the influence of the universal codebook size is evaluated. The universal codebook is generated from about 10000 frames. According to the result of the preliminary experiment, the number of frame shifting points is set at 128. The shifting in test is 512 and

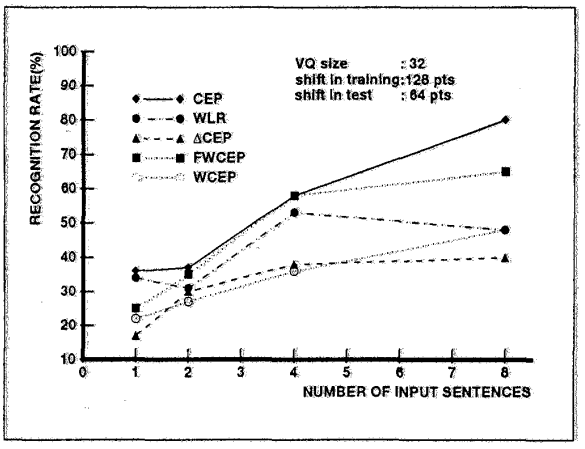

Figure 7: Recognition results using the VQ Histogram algorithm (VQ size = 32)

1 sentence is used. The shape of the result (the filled symbols in Fig.3) is close to the one obtained with the standard VQ algorithm. In this case a peak appeared for 32 codes with the first choice but this peak is located at 64 codes for the top second and top third case. In a further experiment we will use 32 codes but the 64 case will be tried for distance comparison.

4.2.2 Distance comparison

The shifting size in training and test are the same as the standard VQ algorithm. The two universal codebook sizes are evaluated: 32 vectors in Fig.7 and 64 vectors in Fig.8. These results show that the larger codebook did not produce better results.

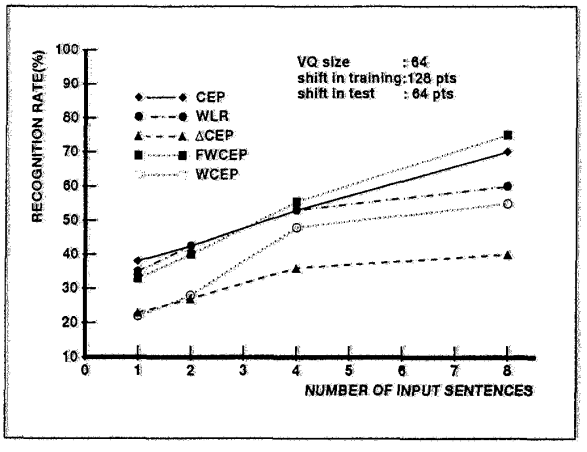

Figure 8: Recognition results using VQ Histogram algorithm (VQ size = 64)

5 Conclusions

In this paper, two algorithms for automatic language recognition have been proposed. The first algorithm is based on the standard VQ algorithm and the second one is based on VQ and histogram algorithm. The 20 language recognition experiments show that the recognition rates for the first and second algorithms were 65% and 80%, respectively, using 8 sentences of unknown speech (about 64 seconds). On the other hand, using only 1 sentence of input speech, the recognition rate for the first and second algorithms dropped to 40% and 36%, respectively. At 4 sentences, two methods provide almost the same recognition accuracy. Thus, the second algorithm with sufficient input speech is better than the first. In the first algorithm, the WLR measure is the most effective, while in the second algorithm, the CEP measure provides the highest recognition rate. In this study, no speaker normalization technique has been applied, so the evaluation of language recognition methods with speaker normalization is one of the feature studies.

Acknowledgments

This study was done by Paul Hiriart(INSA, Lyon) under the author's supervising at NTT Basic Research Labs. I would like to thank Dr.M.Honda, Group Leader of NTT Basic Research Labs. for his encouragement, and Mr. Y.Shiraki, Senior Research Scientist, NTT Basic Research Labs., for his useful discussions, Dr. K.Mano and Miss T.Matsui, Research Engineers, Human Interface Lab., for their many suggestions.

References

[1] A.E.House and E.P.Neuberg, Toward automatic identification of the language of an utterance, I. Preliminary methodological considerations, Journal of the Acoustical Society of America 62(3), pp.708-713 (1977).

[2] R.A.Cole, J.W.T.Inoue, Y.K.Muthusamy and M.Gopalakrishnan, Language identification with neural networks: a feasibility study, IEEE Pacific Rim Conference on Commu., Comp. and Signal Processing, (June 1989).

[3] J.T.Foil, Language identification Using Noisy Speech, ICASSP86, 17.1 (1986-03).

[4] F.Goodman, A.Martin, Improved Language Identification in Noisy Speech, ICASSP89, 35.S10b.4 (1989-05).

[5] M.Sugiyama, Language Identification Using Acoustic Features, Proc. of ASJ, 3-3-6, pp.81-82 (1989-03).

[6] H.Irii, K.Itoh, N.Kitawaki, Multilingual Speech Data Base for Evaluating Quality of Digitaized Speech, Proc. of ICSLP90, 23.15, pp.1025-1028 (1990-10).

[7] J. D. Markel and A. H. Gray, Linear Prediction of Speech, Springer, 1976.

[8] B.S. Atal, Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification, J. Acoust. Soc. Am., 55, pp.1304-1312 (1974).

[9] F. Itakura, Minimum Prediction Residual Principle Applied to Speech Recognition, Trans. IEEE, ASSP-23, pp.67-72 (Feb. 1975).

[10] A.H. Gray and J.D. Markel, Distance Measures for Speech Processing, IEEE Trans., ASSP-24, 5, pp.380-391 (Oct. 1976).

[11] M. Sugiyama and K. Shikano, LPC Peak Weighted Spectral Matching Measures, Trans. of IECE, Vol.J64-A, No.5, pp.409-416 (May 1981) (in Japanese).

[12] M. Sugiyama and K. Shikano, Frequency Weighted LPC Spectral Matching Measures, Trans. IECE, Vol.J65-A, No.9, pp.965-972 (Sep. 1982) (in Japanese).

[13] M. Sugiyama and K. Shikano, WLR Measure Applied to Word Recognition, Trans. of IECE, Vol. J66-D, No.4, pp.385-391 (Apr. 1983) (in Japanese).

[14] Y. Tohkura, A Weighted Cepstral Distance Measure for Speech Recognition, ICASSP86, 14.17, pp.761-764 (Apr. 1986).

[15] F.K.Soong, A.E.Rosenberg, On the Use of Instantaneous and Transitional spectral Information in Speaker Recognition, ICASSP86, 17.5 (1986-04).

[16] K.Shirai, K.Mano and S.Ishige, Speaker Identification based on Frequency Distribution of Vector-Quantized Spectra, Trans. of IECE, Vol.J70-D, No.6, pp.1181-1188 (Jun. 1987) (in Japanese).