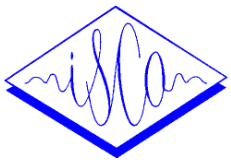

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

EXTRACTION OF PHONEME-DEPENDENT INDIVIDUALITY USING HMM-BASED SEGMENTATION FOR TEXT-INDEPENDENT SPEAKER RECOGNITION

Hideki Noda and Masuzo Yanagida

Kansai Advanced Research Center, Communications Research Laboratory
588-2, Iwaoka, Iwaoka-cho, Nishi-ku, Kobe, 651-24 Japan

ABSTRACT

This paper describes a new method for text-independent speaker recognition which exploits phoneme-dependent voice individuality without direct phoneme recognition. This method uses a segmentation technique based on the Hidden Markov Model (HMM). Appropriate segmentations are expected to be carried out through parameter estimation of models, given enough amount of utterances with their phonetic transcriptions. Segmentations being completed, the difference between feature vectors of reference and input which belong to the same phoneme is used as the dissimilarity measure for speaker recognition. Speaker verification performance has been evaluated by experiments using 20 word utterances of 177 male speakers. In a experiment 95.4% verification rate is achieved using the proposed method, whereas 89.3% by a well-known VQ method.

1. INTRODUCTION

The VQ (Vector Quantization) based approach is known as an excellent method for text-independent speaker recognition[1]. In this method, each speaker is characterized by a VQ codebook, which is constructed from a set of feature vectors from a set of training samples produced by the speaker. The dissimilarity measure between an input (a set of vectors) and a codebook is calculated by accumulating the distance between each input vector and its best matching centroid over all input vectors. However we believe that this method has a drawback; this method does not correctly utilize phoneme-dependent voice individuality because an input vector and its nearest neighbor centroid do not always belong to the same phoneme.

Some methods for text-independent speaker recognition[2,3] have already been proposed which exploit phoneme-dependent individuality through direct phoneme recognition. Since phoneme reference samples are needed for phoneme recognition, segmentations of utterances come to be necessary to obtain them. The most reliable segmentation can be carried out by hand labeling but this is a quite time-consuming work and requires a great deal of labor.

We propose a new method for text-independent speaker recognition which can utilize phoneme-dependent individuality without phoneme reference samples. This method uses the Hidden Markov Model (HMM) framework and is based on expectation that appropriate segmentations can be done by parameter estimation of models, given phonetic transcriptions of utterances. The difference between feature vectors of reference and input which belong to the same phoneme is used as the dissimilarity measure for speaker recognition. Experimental results of speaker verification show that the proposed method provides significant performance improvement over a conventional VQ approach.

2. USE OF HIDDEN MARKOV MODEL

2.1 Model Definition

Less number of states is generally desirable to perform a reliable estimation of model parameters. Considering that a steady portion of each vowel or the syllabic nasal could be described by one state and vowel generally contains a large amount of individuality information, we assign 6 states for steady portions of Japanese 5 vowels (/a/, /i/, /u/, /e/ and /o/) and the syllabic nasal /N/ and one state (/X/) for all other portions. Therefore any utterance is represented by a state sequence composed of 7 independent states. Two examples for Japanese words /deNwa/ and /zero/ (telephone and zero in English respectively) are shown in Fig. 1. The word /deNwa/ is modeled by the state sequence /XeXNXa/ (we refer to this sequence simply as phoneme sequence). The first X represents the

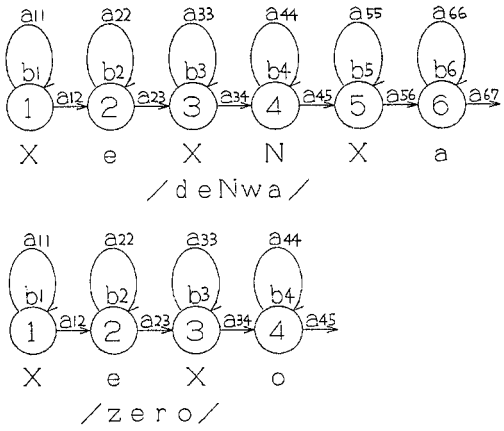

Fig. 1 Left-to-right HMMs for 2 words.

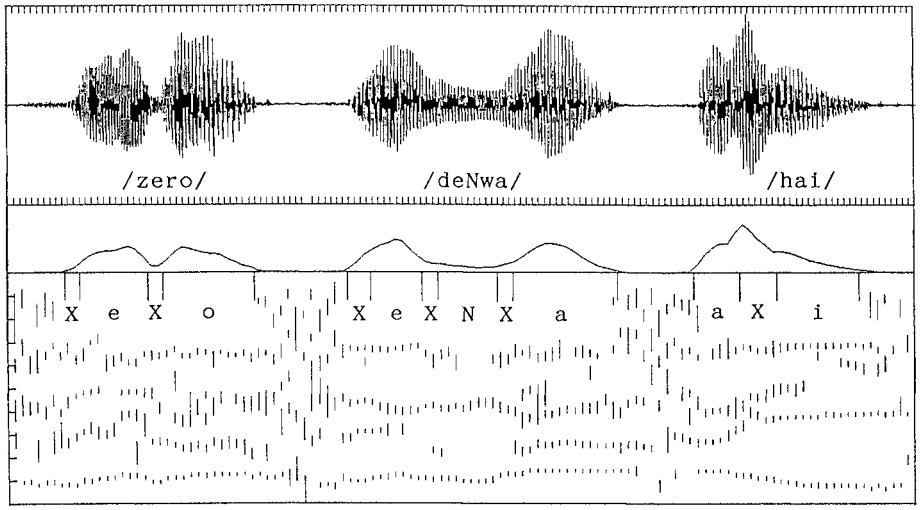

Fig. 2 An example of segmentation results for 3 words.

portion from the beginning and to just before the steady portion of /e/. The second X represents the transition part between the steady portion of /e/ and that of /N/. The third X represents the portion between after the steady /N/ part and before the steady /a/ part including the /w/ part. The word /zero/ is similarly modeled by /XeXo/.

A left-to-right model is used in which output vectors are produced at each state. We use a single Gaussian multivariate probability density function (PDF) with a diagonal covariance matrix for each state. Given phoneme sequences of training utterances, model parameters are estimated with a constraint that the PDFs for all states corresponding to the same phoneme are common. For example, the PDFs represented by b1, b3, b5 for the word /deNwa/ and b1, b3 for the word /zero/ are set to be common (tied to each other). As for state transition probability, $a_{i,i}$ at each state is assumed to be independent, provided that $a_{i,i} + a_{i,i+1} = 1$, since duration of each state varies from sample to sample even for the same phoneme of the same utterance. Appropriate segmentations might be carried out through parameter estimation in the following way.

2.2 Model Parameter Estimation

Parameter estimation of HMM is carried out with a reestimation procedure, which converges a set of model parameters to the optimal one in the sense that the likelihood of the training data given the model reaches a local maximum. The employed reestimation procedure using the Viterbi algorithm is described according to the paper[4].

The sequence of acoustic feature vectors from training samples is written as O = o1o2...oT and a set of parameters of models denoted by L, then the optimal set of parameters is obtained by solving (1). P(O|L) is the probability of the training sequence O, given the model L.

$$\max_{L} P(O|L) \qquad (1)$$

(1) can be rewritten by (2) using state sequence Q = q1q2...qT.

$$\max_{L} \sum_{Q} P(O|Q,L)P(Q|L) \qquad (2)$$

The summation in (2) is carried out for any state sequence Q which can produce the sequence of feature vectors O. Here the sum is approximated by the maximum as shown in (3).

$$\max_{L} \max_{Q} P(O|Q,L)P(Q|L) \qquad (3)$$

This double maximization can be done by the following iterative procedure (4) and (5). Assuming an initial set of parameters Lo given, the state sequence which maximizes (4) is obtained using the Viterbi algorithm.

$$\max_{Q} P(O|Q,Lo)P(Q|Lo) \qquad (4)$$

Then using the derived optimal state sequence Qo, a new set of parameters is reestimated by maximizing (5).

$$\max_{L} P(O|Qo,L)P(Qo|L) \qquad (5)$$

Since the sequence of training vectors O is segmented by Qo; each frame of O is labeled to some state of HMM, a new set of parameters is easily calculated. Given the number of frames N for the state i, the transition probability $a_{i,j}$, $a_{i,i+1}$ for the state i is calculated by N/(N+1), 1/(N+1) respectively. As for PDF, since a Gaussian distribution is assumed for states corresponding to the same phoneme, it is enough to calculate the mean and variance of all frames for the same phoneme. Using the new parameter set as Lo in (4), the maximizations of (4) and (5) are continued until a reestimated parameter set converges to the optimal one.

In the following experiments initial parameters are calculated by dividing the training sequence of feature vectors equally according to the number of states. The logarithm of the probability P is

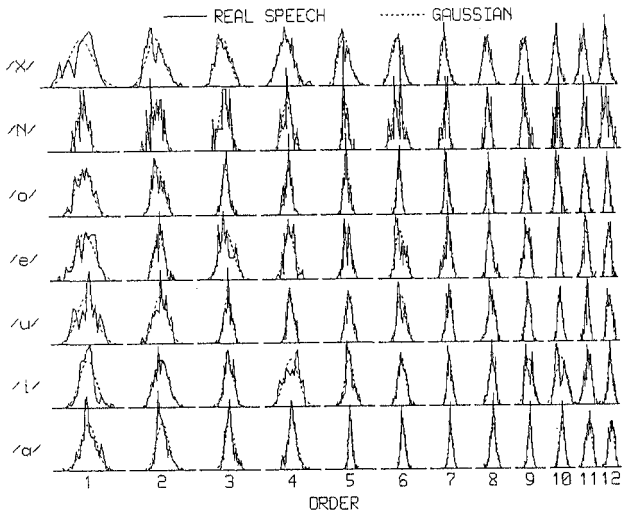

Fig. 3 Distributions of 1st to 12th cepstral coefficients for 7 independent states derived from utterances by one speaker.

here used in the above mentioned procedure to prevent underflow in computation.

2.3 Segmentation Performance

We have confirmed by visual inspection for 20 speakers that the above-mentioned method is able to perform a satisfactory correct segmentation. A numerical precise evaluation of segmentation performance was not carried out because a very precise segmentation might not be necessary for our purpose of extracting phoneme-dependent individuality and a precise evaluation needs time- and labor-consuming hand labeling. An example for one speaker is shown in Fig. 2, where segmentation results for 3 words are shown. These segmentations are carried out using 10 words shown in the right half of Table 1.

A single Gaussian probability density function is assumed for each phoneme in the modeling. We have also visually investigated for the same 20 speakers whether this assumption is allowable or not. Fig. 3 shows distributions of 12 LPC cepstral coefficients for 7 phonemes for the same speaker as in Fig. 2. It seems that in general this assumption might be allowable even though this is obviously inadequate for some distributions.

3. SPEAKER VERIFICATION EXPERIMENTS

3.1 Data Base

The data base used in this study consists of isolated 20 word utterances produced by 177 male speakers. Their ages cover from 20 to 60 years with a good balance. The utterances were recorded over conventional extension telephone lines. Each speaker provided 2 repetitions for each word in 2 sessions spaced three to four months apart. The average value of SNR (signal-to-noise ratio) over all data is approximately 29 dB.

Table 1 Word utterances and their phoneme sequences for reference and input.

Reference		input	
word	model	word	model
/ici/	/iXi/	/zero/	/XeXo/
/saN/	/aXN/	/ni/	/Xi/
/go/	/Xo/	/joN/	/XoXN/
/nana/	/XaXa/	/roku/	/XoXu/
/kjuu/	/Xu/	/haci/	/aXi/
/kuruma/	/uXuXa/	/deNwa/	/XeXNXa/
/keisacu/	/eXiXaXu/	/doku/	/XoXu/
/reNraku/	/XeXNXaXu/	/bakudaN/	/XaXuXaXN/
/giNkoo/	/XiXNXo/	/zikaN/	/XiXaXN/
/mosimosi/	/XoXiXoXi/	/hai/	/aXi/

10 words among 20 words are used as reference samples and the other 10 words as input samples. The phoneme sequences for these words are here determined as shown in Table 1. It is assumed that any word starts with one of vowel phonemes (this means one of steady vowel portions) except words starting with voiced consonants and any word ends with one of phonemes except /X/ (this means a steady portion of one of vowels or /N/). This assumption is based on the following reasons. Unvoiced portions are almost removed because of a pretty high energy threshold applied in the acoustic analysis described below and speech material being telephone speech. Unsteady portions of vowels in the beginning and those portions of vowels and /N/ in the end of utterances are also thought to be removed by a high energy threshold value.

3.2 Acoustic Feature Vectors

Speech wave is low-pass filtered at 4.5 kHz and digitized at a 10 kHz sampling rate. The digitized speech wave is pre-emphasized with a first order adaptive filter and subjected to 12th order LPC analysis with a 25.6 msec Hamming window and a 12.8 msec frame rate. The frames corresponding to silence and unvoiced portions are almost removed because a pretty high threshold on energy is applied and speech material used is telephone speech. In fact, "Selective linear prediction"[5] analysis is here applied to use the spectral information only up to 4 kHz of telephone speech. The first to twelfth cepstral coefficients obtained by this analysis is used as a feature vector for one frame and a sequence of feature vectors is derived from each word utterance.

3.3 Dissimilarity Measures

Model parameters are estimated and segmentations are carried out for reference and input samples independently. Then the difference between feature vectors of reference and input which belong to the same phoneme is accumulated over all phonemes and used as the dissimilarity measure for speaker verification. The Euclidean distance D1 is here used as the dissimilarity measure between a reference R and an input I.

$$D1(R,I) = \sum_{s=1}^{7} \sum_{k=1}^{12} (m_s^R(k) - m_s^I(k))^2 \qquad (6)$$

$m_s^R(k)$ and $m_s^I(k)$ are the means of k-th LPC cepstral coefficient over all frames labeled as phoneme s (s=1-7 for a, i, u, e, o, N, X) for reference and input respectively.

Experiments using a well-known VQ approach[1] are also carried out for performance comparison with the proposed method. The method described in the paper[6] is used to generate codebooks. Given a codebook for a reference speaker, the Euclidean distance $d(x_i^I, c_*^R)$ between each input vector x_i^I and its best matching centroid c_*^R is computed. Then, the distance accumulated over all input vectors and normalized by the number of input vectors N is used as the dissimilarity measure D2 between a reference and an input.

$$D2(R,I) = \left(\sum_{i=1}^{N} d(x_i^I, c_*^R) \right)/N \qquad (7)$$

In the following experiments the codebook size used is 64, which is determined based on separate experimental results; the size 64 among 16, 32 and 64 provided the best result.

3.4 Experimental Results

Two speaker verification experiments are carried out. In Experiment 1, two repetitions of 10 words (see Table 1) uttered in the former session are used as a reference and two repetitions of the other 10 words (see Table 1) in the latter session as a input. In Experiment 2, the first sample in the former session and the second sample in the latter session for 10 words are used as a reference and the second sample in the former session and the first sample in the latter session for the other 10 words as an input. The number of combinations that a reference and an input are taken from the same speaker is 177 and that from different speakers is 177*176. Speaker verification performance is evaluated with the verification equal-error rate (in fact equal-correct rate here).

The results are shown in Table 2. Verification rates using the proposed method are found to be much better than those by a VQ method. We believe that this performance difference comes from the difference on how effectively the phoneme-dependent individuality is exploited in the two methods. The results also show that in the case that all available samples are those in the same session (Experiment 1) the speaker verification performance is poor because of time variation of speech.

There is a different approach using HMM for text-independent speaker recognition[8]. In this approach, model parameters for each phoneme for a reference speaker are prepared in advance. Given the phonetic transcriptions of input utterances, the models for them are constructed by concatenating models for phonemes. Then the probability that input utterances are produced by the models is computed and used as the similarity measure between reference and input. This method is similar to the method denoted as PLU-based system in [7]. In our evaluation[8], speaker verification performance using this method was worse than that using the proposed method.

Table 2 Speaker verification results (%).

	Proposed	VQ
Experiment 1	89.3	84.7
Experiment 2	95.4	89.3

4. CONCLUSION

We have proposed a new method for text-independent speaker recognition which exploits phoneme-dependent individuality using HMM-based segmentation. It can be concluded that the proposed method is quite promising according to the experimental results that speaker verification rates using the proposed method are much better than those by a conventional VQ method.

The merit of the proposed method should be emphasized that it does not need time- and labor-consuming hand labeling to make use of phoneme-dependent individuality. However it is assumed that the phonetic transcriptions of utterances are available. We think that there are many applications like forensic applications where this assumption is realistic.

ACKNOWLEDGEMENT

The authors wish to thank Horacio Franco at the Laboratory of Sensory Research in Argentina for the use of his HMM program.

REFERENCES

[1] A.E.Rosenberg and F.K.Soong, "Evaluation of a vector quantization talker recognition in text independent and text dependent modes," Proc. ICASSP, pp.873-876, 1986.

[2] H.Matsumoto and T.Nimura, "Text-independent speaker identification on piecewise canonical discriminant analysis," Proc. ICASSP, pp.291-294, 1978.

[3] S.Furui, "Research on individual information in speech waves," Ph.D.Thesis, Tokyo University, 1978.

[4] H.Franco, "Recognition of intervocalic stops in continuous speech using context-dependent HMMs," J. Acoust. Soc. Jpn.(E), vol. 11, pp.131-143, 1990.

[5] J.Makhoul, "Spectral linear prediction: Properties and applications," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-23, pp. 283-296, 1975.

[6] Y.Linde, A.Buzo and R.M.Gray, "An algorithm for vector quantizer design," IEEE Trans. Communications, vol. COM-28, pp.84-95, 1980.

[7] A.E.Rosenberg, C.-H.Lee and F.K.Soong, "Sub-word unit talker verification using Hidden Markov Model," Proc. ICASSP, pp.269-272, 1990.

[8] H.Noda and M.Yanagida, "Investigations on speaker recognition using Hidden Markov Model," Proc. Spring Meet. Acoust. Soc. Jpn., pp.59-60, 1990(in Japanese).