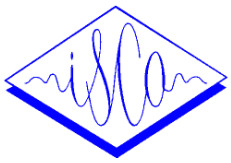

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

Performance Improvement of Speaker Recognition System for Small Training Data

Seong Jin Yun, Yung Hwan Oh

Department of Computer Science
Korea Advanced Institute of Science and Technology
373-1, Kusong-dong, Yusong-gu, Taejon 305-701 Korea

ABSTRACT

Presently many speaker recognition algorithms can provide high accuracy. They need reliable samples of the parameter space, long utterances and large amount of training data for each speaker. In practical applications, data acquisition constraints thus limit their domain of application. This paper proposes a speaker recognition method that creates models to specify speaker information accurately by using only a small amount of training data and very short utterance for each speaker, both for training and recognition. The proposed methods based on a discrete hidden Markov model(HMM) improves modeling of output probability estimation. The basic idea is that the proposed hidden Markov VQ model(HMVQM) uses the state dependent codebook, and each state represents a partition of speaker information. This method can be considered as multisection codebook model with stochastic transitions between section. Speaker identification experiments based on single syllable word tests give a 1.5% error rate for the proposed HMVQM method, whereas the discrete HMM method gives error rate of 24.12%. In this experiment, We use only two training data for each speaker.

1. Introduction

Speaker recognition is an automatic extraction of personal identity information from speech signal. The task of a speaker recognition system is either to identify an unknown speaker among several speakers of known speech characteristics, or to verify whether a speaker is the person he claims to be. It can be directly applied to security checks and also to speaker selection or speaker clustering.

In recent years, many speaker recognition algorithms can provide high accuracy, but these algorithms mostly require a large amount of training data to cover speaker specific features. They need reliable samples of the parameter space, requiring generally about 5 seconds or more of speech and large amount of training data for each speaker. In practical application, data acquisition constraints thus limit their domain of application.

The VQ-based method using speaker-specific codebooks is one of the well-known speaker recognition methods. This method is robust against utterance variations, if sufficient training test data are available[4]. When the amount of available data is small, however, the performance is greatly decreased. HMM-based methods have been successfully used for modeling of speaker recognition. Excellent results have been reported employing continuous mixture density HMM[4]. However, the large number of parameters describing these HMMs requires an excessive number of training data to be able to obtain acceptable performance. The effectiveness of HMM-based speaker recognition methods has not been made clear.

This paper proposes a new robust text dependent speaker recognition method that incorporates the integration of VQ distortion approach and HMM. The proposed method creates a model to specify speaker information accurately by using only a small amount of training data and very short utterance for each speaker, both for training and recognition.

In the next section, we will give a brief overview of speaker recognition system. In section 3, VQ distortion based method and HMM methods are discussed. Then the proposed hybrid method and the application to speaker identification are presented in detail. The experiments and the results are described in section 4 and the conclusion is given in section 5.

2. Overview of Speaker Recognition System

In this section, we will give an overview of speaker recognition systems, mainly the text dependent speaker identification. Figure 1 shows a block diagram of the overall configuration of system. The process of the recognition system consists of three phases - feature extraction, similarity comparison, and decision phase. The functions of each part are as follows.

2.1 Feature Extraction

The speech data are sampled at 16 kHz, and preemphasized with a first order difference filter whose transfer function is $1 - 0.95z^{-1}$. To get feature parameters, a Hamming window with a length of 32 msec is applied every 8 msec and 12 FFT cepstral coefficients are computed. After that, a set of 12 delta cepstral coefficients is computed by fitting a regression coefficient with respective to cepstral trajectories over a time window of five frames long. Each parameter set represents the characteristics of vocal tract and its temporal change respectively.

2.2 Similarity Comparison

Speaker recognition systems typically operate in one of two input modes, text dependent or text independent. In the text dependent mode, speakers must provide utterances of the same text for both training and recognition trials. The general approach to automatic speaker recognition in the text dependent mode is the spectral template matching approach. In this approach, each speaker is represented by a sequence of feature vectors extracted for each word.

The most important problem of a text dependent speaker recognition system is normalizing time and aligning utterances.

Moderate differences in the timing of speech events can be normalized by aligning the feature vectors sequence of a test utterance to the template using a DTW (dynamic time warping) algorithm. In the HMM based methods, the speech is aligned according to the maximum likelihood path or all possible path through word models.

2.3 Decision

In the speaker verification, it only requires comparing the test pattern with one reference pattern and involves a binary decision whether the test speech matches the template of the claimed speaker. In the speaker identification, a speech from an unknown speaker is analyzed and compared with models of known speakers' pattern. The unknown speaker is identified as the speaker whose model best matches the input speech pattern. The fundamental difference between the verification and identification modes is the number of decision alternatives.

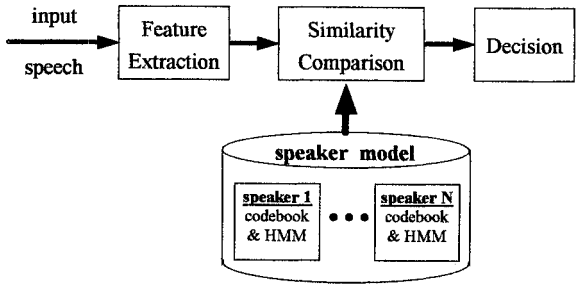

Figure 1. Speaker Recognition Procedure

3. Hidden Markov VQ Model (HMVQM)

Recently, new approaches to extracting and measuring acoustic and phonetic events for application to speaker recognition have met with more success. The bases for these approaches lie in statistical techniques for extracting and modeling reduced sets of optimally representative single feature vectors or feature sequences. These techniques fall under the related categories of vector quantization (VQ) and hidden Markov modeling (HMM). In this section, we discuss the VQ based method and the HMM based method. Then we present a hybrid method that integrates the VQ distortion approach and the HMM.

3.1 VQ Based Method

The VQ based speaker recognition method is the well-known text independent speaker recognition method and its performance has been reported to be high[4]. In the VQ based method, a codebook is designed for each speaker from training data. Input Speech is quantized using each speaker's codebook and the input speaker is recognized as the speaker whose codebook gives the minimum distortion.

It uses the static information of speaker individuality in a shot term spectrum. However, speaker individuality contains not only static, but also dynamic features. One of the approaches that incorporate time sequence information is the multisection codebook VQ by means of sequences of codebooks. A multisection codebook is designed for each speaker's utterance into equal-length section and produce a standard VQ codebook for each section. This method is regarded as the combination of the original VQ distortion approach and a linear time warping method. However, the time warping method is insufficient. The advantage of this method is that it reduces the recognition error and decreases the computation and memory requirements.

3.2 HMM Based Method

There are two types of HMM: the discrete HMM and the continuous HMM. In the discrete HMM, VQ produces the closest codeword from the codebook for each acoustic observation. This mapping from continuous feature parameter space to quantized discrete space may cause serious quantization errors. To reduce VQ errors, various techniques have been proposed. Another disadvantage of the discrete HMM is that the VQ codebook and the discrete HMM are separately modeled, which may not be an optimal combination for pattern recognition. On the other hand, continuous mixture density HMM models the feature parameter directly using estimated continuous probability density functions without VQ, and has been shown to improve the recognition accuracy compared with the discrete HMM. However, mixtures of a large number of probability density function will considerably increase not only the computational complexity, but also the number of free parameters that require to be reliably estimated.

The maximum likelihood estimate of the parameters of a HMM converges to the true values as the number of training data tends to infinity. However, in the practical application, only finite training data are available. If the training data are limited, this will result in some parameters being inadequately trained and the classification based on the poorly trained model will result in fatal error. One solution to this problem is to reduce the size of the model. Although this is possible, there are trades off between the size of model and the modeling power. Another possible solution is to interpolate one set of parameters with another set of parameter estimates from a model for which an adequate amount of training data exists[1].

3.3 Hybrid Method

The basic idea of proposed HMVQM uses the state dependent codebook, and each state represents a partition of speaker information. This method can be considered as a multisection codebook model with stochastic transitions between the sections. The HMVQM consists of a global codebook, state transition probability and codebooks for every state. The global codebook is used only in the state codebook training procedure.

The proposed method based on a discrete HMM improves modeling of output probability estimation. Instead of output probability at a state for the standard HMM, the VQ distortion measure using the state dependent codebook is calculated in frame by frame. The modified forward-backward algorithm is given by equations (1) ~ (3).

$$\alpha_t(j) = \left[\sum_i \alpha_t(j)\, a_{ij}\right] b_j(\mathbf{X}_t) \qquad (1)$$

$$\beta_t(j) = \left[\sum_i a_{ij}\, b_j(\mathbf{X}_t)\, \beta_{t+1}(i)\right] \qquad (2)$$

$$b_i(\mathbf{X}_t) = \exp\left(\max_k \left[-d(\mathbf{X}_t, \mathbf{C}_k^i)\right] \right) \qquad (3)$$

Where, $d(\mathbf{X}_t, \mathbf{C}_k^i)$ means distance between an input feature vector $\mathbf{X}_t$ and a codevector $\mathbf{C}_k$ of i-th state codebook. Therefore, the output probability is the function of quantization distortion.

The overall flow diagram of a training procedure is shown in Figure 2. The reestimation of model parameters consists of two steps: the reestimation of the state transition probability is equal

to the procedure of the HMM, and the reestimation of the state dependent codebook is performed by the iteration procedure like the k-means clustering procedure for the standard VQ. In the reestimation of state codebook, we interpolate the state codebook with N-nearest codevetor of the global codebook. It is useful for the estimation of the model parameters when the training data are limited.

Step 1. Reestimation of the state transition probability

$$\bar{a}_{ij} = \sum_{t=1}^{T-1}\gamma_t(i,j) \Big/ \sum_{t=1}^{T-1}\gamma_t(i) \qquad (4)$$

where, $\gamma_t(i,j) = \dfrac{\alpha_t(i)\, a_{ij}\, b_j(X_{t+1})\, \beta_{t+1}(j)}{\sum_k \alpha_t(k)\, \beta_t(k)}$

$$\gamma_t(i) = \sum_j \gamma_t(i,j)$$

Step 2. Reestimation of the state codebook

Given a state codebook, $C^i = \{C_k^i\}$, find the optimal partition into quantization cells:

$$R_k = \{X_t : d(X_t, C_k^i) < d(X_t, C_j^i);\ \text{all } k \neq j\}$$

Using the centroid condition, update the centroid:

$$\bar{C}_k^i = \frac{\sum_t \left[X_t + \eta \sum_n V_t^n\, w_t^n \right] \gamma_t(i)}{\sum_t \left[1 + \eta \sum_n w_t^n \right] \gamma_t(i)} \qquad (5)$$

where, η is the global factor $(0 \le \eta \le 1)$

V^n is the n-th nearest codevector of global codebook

w^n is the weight factor of the n-th nearest codevector

$$w^n = 1 - \frac{d(X, V^n) / (N+1)}{\sum_n^N d(X, V^n)} \qquad (\sum_n w^n = 1)$$

4. Experimental Evaluation and Results

4.1 Speech Data

Two data sets are used for training and testing the text dependent speaker identification. The first data set has one to three syllable words uttered by 20 female and 20 male speakers, four times [Table 1]. Two of these are used in training and others are used in testing. The average duration of one to three syllables words are 447 msec, 502 msec, 706 msec respectively. We use this data set for experiment of small training data environment. The second data set has 48 Korean allophone segments uttered by 6 male speakers. 4474 segments are manually extracted from 445 phonetically balanced words. This data set has much more training data, but shorter utterance than the first data set.

4.2 Experimental conditions

Each word unit is represented by a left-to-right HMM, containing three transitions. The modified K-means algorithm is used for creating the global and state dependent VQ codebooks. For HMVQM, we use the global codebook of size 256, and three nearest codewords are used for the state dependent codebook reestimation. In the experiment with each data set, we use the global factors 0.5 and 0.1 respectively.

Table 2 shows the configuration of the speaker models. The comparison with respective to the size of speaker model, the HMVQM has a smaller size of models about half the size of HMM

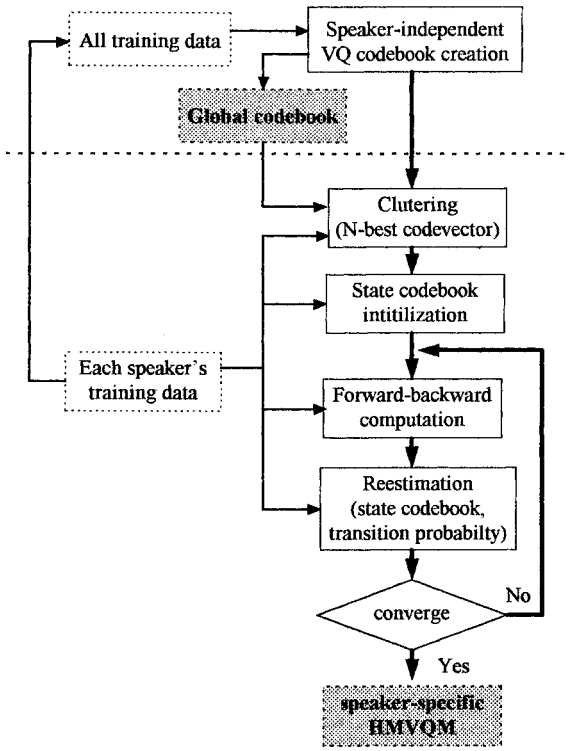

Figure 2. Training procedure of HMVQM

Table 1. Data set

Set 1	1 syllable word (10)	/jʌŋ/, /il/, /i:/, /sam/, /sa:/, /o:/, /juk/, /cʿil/, /pʰal/, /gu/
	2 syllable word (5)	/hana/, /ɖasʌɗʿ/, /jʌsʌɗʿ/, /ilgobʿ/, /jʌɗʌbʿ/
	3 syllable word (3)	/ɖʌhagi/, /gopʰagi/, /nanugi/
Set 2	allophone (48)	/a/, /ʌ/, /o/, /ɯ/, /ɯ/, /i/, /e/, /e/, /ja/, /jʌ/, /jo/, /ju/, /je/, /wa/, /wʌ/, /ɨi/, /wi/, /we/, /we/, /ɥ/, /b/, /bʿ/, /p/, /pʰ/, /ɖ/, /d/, /dʿ/, /l/, /tʰ/, /g/, /g/, /gʿ/, /k/, /kʰ/, /ʒ/, /ʃ/, /c/, /cʰ/, /s/, /z/, /h/, /r/, /l/, /m/, /nʿ/, /m/, /mʿ/, /ŋ/

Table 2. Model configuration

Model \ Data		Set 1			Set 2
		1 syll.	2 syll.	3 syll.	allophone
HMM	state	5	7	10	3
	codebook	64/128	64/128	64/128	64
	size	1881/3737	2033/4017	2276/4452	1737
HMVQM	state	5	7	10	3
	codebook	5	5	5	3
	size	635	889	1300	225

4.3 Results

The method is compared to discrete HMM based speaker recognition method through text dependent speaker identification experiments on the first data set, especially from the viewpoint of robustness against relatively small amount of training data. The speaker identification experiments based on 800 single syllable tests give a 1.5% error rate for the proposed HMVQM method, whereas the discrete HMM method gives error rate of 24.12% [Figure 3]. In this experiment, we use only two training data for each speaker. The evaluation experiments are also carried out with the 48 Korean allophone units using the second data set. The HMVQM gives the average identification error rate of 20.11%. This is 8.15% higher than the error rate of the discrete HMM method [Figure 4].

Table 3 shows the experimental results. It indicate that the HMVQM method is more robust than the discrete HMM method against the small amount of training data and very short utterances.

Table 3. Average identification rate (%)

Model \ Data		Set 1			Set 2
		1 syll.	2 syll.	3 syll.	allophone
HMM	64 CB	74.25	82.08	89	71.74
	128 CB	75.88	83.75	89.25	-
HMVQM		98.5	98.5	100	79.89

5. Conclusion

We have proposed HMVQM based speaker recognition method. The proposed HMVQM is a hybrid method of VQ distortion and HMM based method. This paper proposes speaker recognition methods that create models to specify speaker information accurately by using only a small amount of training data and very short utterances for each speaker, both for training and recognition. This method reduces the size of model and uses interpolation technique. It is shown that the proposed methods based on a discrete HMM improves modeling of output probability estimation. Through some experiments, the proposed methods are superior to a discrete HMM when sufficient amounts of data are not available. Furthermore, additional advantages of the proposed HMVQM based speaker recognition system is that it requires the small memory requirements and gives fast response.

For further studies and modeling of speaker recognition system, we should evaluate this method for a large database.

References

[1] S. J. Yun; "Performance Improvement of Speaker Recognition System for Small Training Data", M.S. Thesis, Department of Computer Science, KAIST, 1993

[2] L. R. Rabiner; "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proc. IEEE, Vol. 77, pp. 257 - 285, Feb. 1989

[3] David K. Burton, John E. Shore, Joseph T. Buck; "Isolated-Word Speech Recognition Using Multisection Vector Quantization Codebooks", IEEE Trans. on ASSP, Vol. 33, No. 4, pp. 837 - 849, Feb. 1985

[4] Tomoko Matsui, Sadaoki Furui; "Comparison of Text-independent Speaker Recognition Methods Using VQ-distortion and Discrete/Continuous HMMs", Proc. ICASSP, Vol 2, pp. 157 - 160, 1992

[5] Seiichi Nakagawa, Hideyuki Suzuki; "A New Speech Recognition Method based on VQ-Distortion Measure and HMM", Proc. ICASSP, Vol. 2, pp.676 - 679, 1993

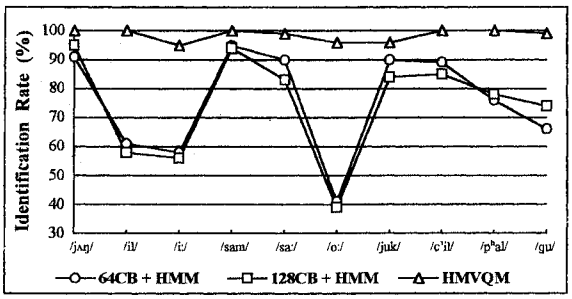

Figure 3. Speaker identification rate (1 syllable word)

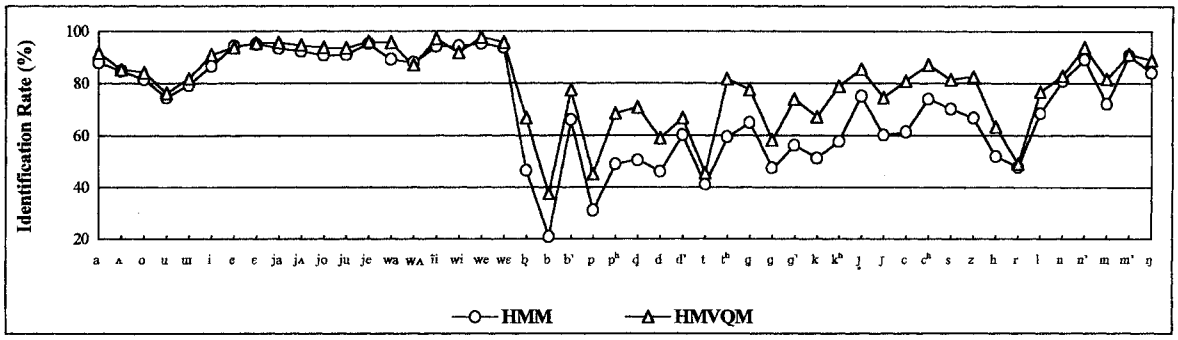

Figure 4. Speaker identification rate (allophone)