Text - Independent Speaker Identification using Hidden Markov Models

Mangesh S. Deshpande
SRES College of Engineering, Kopargaon, Maharashtra, India
mangesh8374@yahoo.com

Raghunath S. Holambe
SGGS Institute of Engineering and Technology, Nanded, Maharashtra, India
rsholambe@sggs.ac.in

Abstract

This paper presents a closed-set, text-independent speaker identification using continuous density hidden Markov model (CDHMM). Each registered speaker has a separate HMM which is trained using Baum-Welch algorithm. The system performance has been studied for different system parameters such as the number of states, number of mixture components per state and the amount of data required for training. Identification accuracy of 100% is achieved by conducting the experiments on TIMIT database.

Key Words- Speaker identification, hidden Markov model, admissible wavelet packet tree**.**

1. Introduction

Various classifiers have been proposed for speaker identification. These classifiers include vector quantization (VQ) [1], Gaussian mixture model (GMM) [2] and hidden Markov model (HMM) [3]. In HMM, the nonstationary speech signal is represented as a sequence of states. We can imagine that the vocal configuration is in one of a finite number of articulatory states at any time. In each state a short (in time) signal is produced which has one of a finite number of prototypical spectra depending on the state. Thus the power spectra of short intervals of the speech signal are determined solely by the current state of the model while the variation in the spectral composition of the signal with time is governed predominantly by the probabilistic state transition law of the underlying Markov chain. The probabilistic state transition law reflects the variation in target vocal tract positions for the dynamic aspects of the speech, such as speaking style. This concept motivates to use HMM for speaker identification.

Initially, Poritz proposed the use of a five state ergodic discrete hidden Markov model (DHMM) for text-independent speaker recognition in [4], which is expanded by Tishby to an eight state autoregressive continuous density hidden Markov model (CDHMM) in [5]. In [6] Yuan has applied circular HMM for text-independent speaker identification. Rosenberg has reported a method using left-to-right HMM's for talker verification in [7]. Rose [8] has examined the effects of the number of mixture components in a single state HMM on speaker recognition performance. Matsui and Furui in [9] demonstrated that a concatenated phone HMM system outperforms conventional template matching and vector quantization approaches.

Key issues in this statistical modeling are the amount of data required for training, selection of the topology and the structure of the model. The present work address the issues like amount of data required for training, number of states and number of mixture components per state to achieve the highest identification rate.

The remainder of the paper is organized as follows. Section II provides a brief description of HMM. Section III describes the experiments conducted on TIMIT database and the results. Finally, conclusions are given in section VI.

2. Hidden Markov models

In recent studies, HMM has become the most popular statistical tool for speaker identification. Two main variations of HMM's have been widely used: DHMM and CDHMM. The DHMM uses nonparametric discrete output probability distributions, due to a previous vector quantization (VQ) process. CDHMM uses parametric densities to model the output probabilities. The main problem of DHMM is the loss of information about the input signal during the VQ process. CDHMM avoid this problem by using probability density functions. Thus, CDHMM modeling seems to be a more flexible and complete tool for speaker modeling.

978-0-7695-3267-7/08 $25.00 © 2008 IEEE
DOI 10.1109/ICETET.2008.46

Left-to-right HMMs have one absorbing state at which once the Markov chain arrives, the underlying Markov chain can not leave that. In the structures of left-to-right HMM, the absorbing state governs the fact that the rest of a single, long observation sequence provides no further information about earlier states, once the underlying Markov chain reaches the absorbing state. In text-dependent speaker recognition, a left-to-right HMM can be used because the phoneme sequence of the input speech is predetermined. However in text-independent speaker recognition, it is difficult to know the phoneme sequence beforehand. It is also true that a Markov chain should be able to revisit the earlier states, because the states of an HMM reflect the vocal configuration of a speaker and the variations of vocal configuration may repeat in pronunciation. Therefore an ergodic model which allows transitions to any other states has been assumed to be effective for text-independent speaker identification as it automatically forms broad phonetic classes corresponding to each state.

A parameter set of HMM is given by $\lambda=(A,B,\pi)$, where *A*, *B* and π denote a set of state transition probability, a set of output probability density functions, and a set of initial state probabilities, respectively. For an ergodic HMM, every state can be reached from every other state. The probability density function (pdf) of certain observations o being in state j has the following general form:

$$b_j(o)=c_{jk}\,\aleph(o,\mu_{jk},U_{jk}),\ 1\le j\le N \qquad (1)$$

where $\aleph(x,\mu_{jk},U_{jk})$ and c_{jk} respectively represent the multi-dimensional Gaussian pdf and the weight for the k[th] mixture component of state *j*.

3. Experiments and results

3.1. Speech database

The TIMIT database consists of 630 speakers, 70% male and 30% female from 10 different dialect regions in America. The speech was recorded using a high quality microphone at a sampling frequency of 16 KHz. The speech is designed to have rich phonetic contents which consist of 2 dialect sentences (SA), 450 phonetically compact sentences (SX) and 1890 phonetically diverse sentences (SI).

Eight sentences, five SX and three SI (approximately 24 seconds) from the TIMIT database are used for training called as training dataset, whereas two SA sentences per speaker are used separately (800 tests of 3 s each) for testing and average identification results are noted. The experimental results are evaluated for 400 speakers.

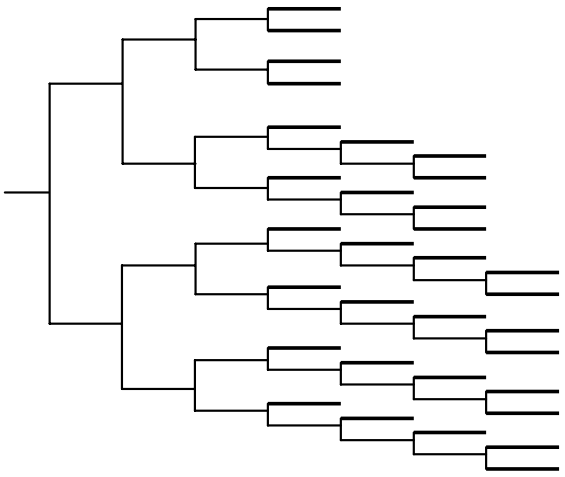

Figure 1. Filter bank structure achieved by admissible wavelet packet decomposition.

3.2. Feature extraction

The Mel frequency cepstral coefficients **(**MFCC) have been the most widely used features for speaker identification. The Mel scale is a mapping between the real frequency scale (Hz) and the perceived frequency scale (Mel). This mapping is virtually linear below 1 kHz and logarithmic above. The drawback of MFCC is that it uses short time Fourier transform (STFT), which has fixed time-frequency resolution. In addition to this it assumes that the signal is stationary during the window period, which may not be strictly stationary. In order to overcome these limitations, we used multiresolution approach based filter structure obtained by admissible wavelet packet (AWP) tree [10]. We found after some experimentation that the tree given in figure 1, gives the best overall result among a reasonable set of AWP trees. The tree structure results in 32 frequency bands.

After performing this decomposition of a 32 ms speech frame, features are derived by taking 24 point DCT of log energies obtained from each frequency band. The energy is normalized by the number of samples in the corresponding band, thereby giving an average energy per frame in each band. The normalization is essential because each band will have a different number of samples.

3.3. HMM parameter estimation

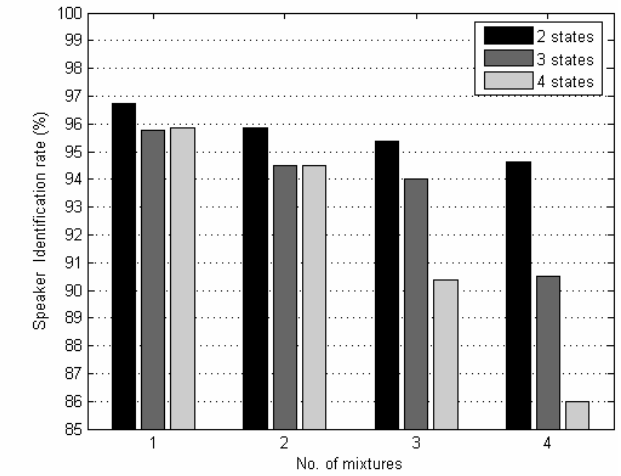

Figure 2. Speaker identification performance as a function of number of mixture components for 2, 3 and 4 states CDHMM.

The most difficult problem of HMM is to estimate the model parameters, λ in order to maximize the probability of the observation sequence given the model. There is no optimal way of estimating the model parameters [11]. However, it is still possible to choose model parameters $\lambda=(A,B,\pi)$ for each model independently such that $P(O/\lambda)$ is locally maximized using an iterative re-estimation procedure such as Baum-Welch algorithm.

The procedure for estimating model parameters has been shown to be very sensitive to initial estimates [12]. An adequate choice for π and A is uniform distribution. Whereas the output probability density functions, B needs good initial estimation, to get rapid and proper convergence. This is done by using uniform segmentation of the training observations into N segments, where N is the number of states. After segmentation, all observations

from the state *j* are collected from all training samples. Then a well known k-means clustering algorithm is used to compute mixture components of state *j* and this procedure is repeated for every state. This simple parameter estimate of each state is followed by a Viterbi segmentation algorithm, where each observation in the training sequence is aligned. Recomputation of the mixture components occur during realignment of the observations. Then the parameters are re-estimation using Baum-Welch algorithm until convergence. Diagonal covariance matrices are used in the estimation of the output distributions.

3.3. Performance evaluation

The speech signal is pre-emphasized with a coefficient equal to 0.97and windowed using 32ms Hamming window with a 16 ms frame shift. AWP based 24 dimension feature vectors are obtained using 6^{th} order Daubechies' orthogonal filter. Baum-Welch algorithm is used to estimate the CDHMM parameters and then the likelihood of the CDHMM for testing speaker frames is used for identification decision. The speaker models are trained by using 5 sentences from the training dataset. Speaker identification experiments using 2, 3 and 4 states ergodic CDHMM are carried out.

In order to obtain a better parametric model, a greater number of Gaussian mixtures per state are often utilized. In the following experiments, we investigate the identification performance with a varying number of mixtures from 1 to 4. Figure 2 shows the identification results as a function of number of Gaussian mixtures per state for 2, 3 and 4 state CDHMM. Diagonal covariance matrices are used to represent each output probability density function. . It is observed that the identification rate decreases as the model states are increased from 2 to 4. It is also noted that the identification performance decreases with increase in the mixtures assigned per state from 1 to 4. The best result of speaker identification is 96.75% for 2 state single mixtures CDHMM. In the following experiments a 2 state CDHMM is used.

Separate experiments are conducted for investigating the effect of amount of data required for training the models. We have trained the speaker models by using 5, 7 and 8 sentences from the training dataset. Figure 3 shows the results for varying the number of training sentences (i.e. tokens) from 5 to 8 and number of mixture components from 1 to 4. Significant improvement can be seen in the performance as the amount of training data increases. 100 % identification rate is achieved with 2 state single mixtures CDHMM trained by 8 sentences. It is also noted that for the models trained using 5 sentences, as we increase the number of mixtures per state, the identification performance decreases where as for models trained using 7 sentences the performance increase up to 3 state CDHMM. Both these experiments show that for small amount of training speech, number of mixtures should be less. As the number of sentences used for training the models increases, we can increases the number of mixture components, which gives better identification performance.

4. Conclusion

We conclude that the speaker identification rates using an ergodic CDHMM are strongly correlated with the number of states, number of mixtures per state and the amount of data used for training. The performance of HMM lies in having better description of vocal tract and instantaneous characteristic. But the corresponding model needs more speech samples and longer training time. This approach is based on the theory of statistic. The maximum likelihood estimation algorithm used to estimate the parameters, needs a large number of training samples, otherwise the result does not have statistic characteristics.

5. References

[1] F. K. Soong, A. Rosenberg, L. Rabiner and B. Juang, "A vector quantization approach to speaker recognition," in *Proc. ICASSP*, Apr.1985, pp. 387-390.

[2] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[3] T. Matsui and S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/ continuous HMMs," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 3, pp. 456–459, July 1994.

[4] A. B. Poritz, "Linear predictive hidden Markov models and the speech signal," in *Proc. ICASSP*, vol. 7, May 1982, pp. 1291-1294.

[5] N. Z. Tishby, "On the application of mixture AR hidden Markov models to text independent speaker recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 563–570, Mar. 1991.

[6] Yaun-Cheng Zheng and Bao-Zong Yaun, "Text dependent speaker identification using circular hidden Markov models," *in Proc. IEEE ICASSP*, 1988, pp. 580-582.

[7] A. E. Rosenberg, C. H. Lee, S. Gokcen, "Connected word talker verification using whole word hidden Markov models," in *Proc. ICASSP*, vol. 1, Apr. 1991, pp. 381-384.

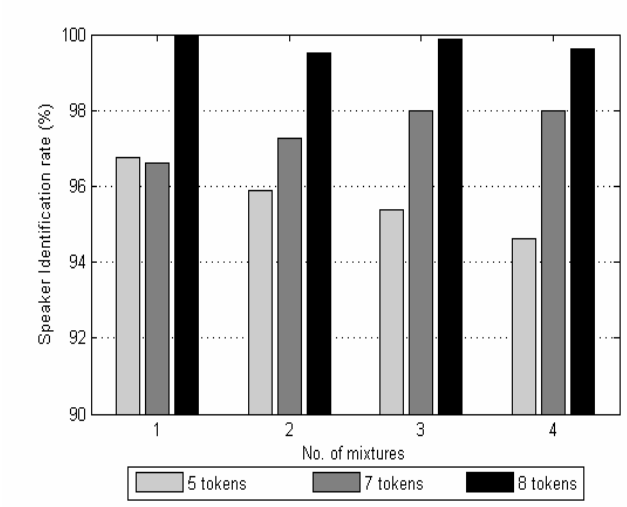

Figure 3. Speaker identification performance as a function of number of training tokens

[8] R. C. Rose and D. A. Reynolds, "Text independent speaker identification using automatic acoustic segmentation," in *Proc. ICASSP*, vol. 1, Apr. 1990, pp. 293-296.

[9] T. Matsui and S. Furui, "Concatenated phoneme models for text-variable speaker recognition," in *Proc. ICASSP*, vol. 2, Apr. 1993, pp. 391-394.

[10] O. Farooq and S. Datta, "Mel filter like admissible wavelet packet structure for speech recognition," *IEEE Signal Process. Lett.*, vol. 8, no. 7, pp. 196-199, July 2001.

[11] L. R. Rabiner, "A Tutorial on Hidden Markov Models and selected applications in speech recognition," *IEEE,* vol. 77, no. 2, Feb. 1989, pp. 257-286.

[12] B-H. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models for speech recognition," *IEEE Trans. Acoustic Speech and Signal Processing*, vol. 33, no. 6, Dec. 1985, pp. 1404-1413.