Estimation of Prior Probabilities in Speaker Recognition

Dat Tran

School of Information Sciences and Engineering
University of Canberra, ACT 2601, Australia

ABSTRACT

According to Bayesian decision theory, the maximum a posteriori (MAP) decision rule is used to minimize the speaker recognition error rate. The a posteriori probability is determined if the a priori probability and the likelihood function are known. However, there was no method to determine the a priori probability, therefore the maximum likelihood (ML) decision rule is used instead. This paper proposes a method to estimate the a priori probability for speakers based on the training data set and speaker models. Speaker identification experiments performed on 138 Gaussian mixture speaker models in the YOHO database using the MAP rule showed lower error rates than using the ML rule.

1. INTRODUCTION

Speaker recognition is the process of automatically recognizing a speaker by using speaker-specific information included in speech waves [5]. This technique can be used to verify the identity claimed by people accessing certain protected systems; that is, it enables access control of various services by voice [2]. Voice dialing, banking over a telephone network, database access services, security control for confidential information, and remote access of computers are important applications of speaker recognition technology.

Speaker recognition can be classified into two specific tasks: identification and verification. Speaker identification is the process of determining which one of the voices known to the system best matches the input voice sample. When an unknown speaker must be identified as one of the set of known speakers, the task is known as *closed-set* speaker identification. If the input voice sample does not have a close enough match to anyone of the known speakers and the system can produce a "no match" decision [8], the task is known as *open-set* speaker identification. Speaker verification is the process of accepting or rejecting the identity claim of a speaker. An identity claim is made by an unknown speaker, and an utterance of this unknown speaker is compared with the model for the speaker whose identity is claimed. If the match is good enough, that is, above a given threshold, the identity claim is accepted. Most of the applications in which voice is used to confirm the identity claim of a speaker are classified as speaker verification.

Speaker recognition methods can also be divided into text-dependent and text-independent. When the same text is used for both training and testing, the system is said to be text-dependent. For text-independent operation, the text used to train and test the system is completely unconstrained.

It has been shown that as long as the training data set covers a sufficient variety of the speaker's speech sound, Gaussian mixture models (GMMs) are effective models capable of achieving high identification accuracy for short utterance lengths from unconstrained conversational speech [7]. In general, the GMM is a statistical clustering method. Its algorithm can be referred to as a prototype-based algorithm, that is, a number of prototypes are generated from the training feature vectors by representing the feature space as a mixture of Gaussian distributions. Each prototype consists of a set of model parameters including mean vector, covariance matrix and mixture weight. Parameters are trained in an unsupervised classification using the expectation maximisation (EM) algorithm [4]. This algorithm provides an iterative maximum likelihood estimation technique.

Given an unknown utterance and a set of speaker models trained by the GMM method, based on Bayesian decision theory, the maximum *a posteriori* (MAP) decision rule is used to minimize the speaker recognition error rate. The *a posteriori* probability is determined if the *a priori* probability and the likelihood function are known. However, there was no existing method to determine the *a priori* probability, therefore an assumption of likely equal speakers is always applied and the maximum likelihood (ML) decision rule is used.

This paper proposes a method to estimate the *a priori* probability for speakers based on the training data set and speaker models. The *a priori* probabilities are randomly initialized and then iteratively updated until a convergence is reached. Speaker identification experiments performed on 138 Gaussian mixture speaker models in the YOHO database using the MAP rule showed lower error rates than using the ML rule.

2. GAUSSIAN MIXTURE MODELS

Let $X = \{x_1, x_2, ..., x_T\}$ be a set of T vectors, each of which is a d-dimensional feature vector extracted by digital speech signal processing. Assuming a statistical independence between these vectors, the probability of the set X given the model λ can be calculated as follows

$$\log P(X \mid \lambda) = \sum_{t=1}^{T} \log P(x_t \mid \lambda) \tag{1}$$

Since the distribution of these vectors is unknown, it is approximately modeled by a mixture of Gaussian densities, which is a weighted sum of c component densities, given by the equation

$$P(x_t \mid \lambda) = \sum_{i=1}^{c} w_i N(x_t, \mu_i, \Sigma_i) \tag{2}$$

where λ denotes a prototype consisting of a set of model parameters $\lambda = \{w_i, \mu_i, \Sigma_i\}$, w_i, $i = 1,..., c$, are the mixture weights and $N(x_t, \mu_i, \Sigma_i)$, $i = 1,..., c$, are the d-variate Gaussian component densities with mean vectors μ_i and covariance matrices Σ_i

$$N(x_t, \mu_i, \Sigma_i) = \frac{\exp\left\{-\frac{1}{2}(x_t - \mu_i)' \Sigma_i^{-1} (x_t - \mu_i)\right\}}{(2\pi)^{d/2} \mid \Sigma_i \mid^{1/2}} \tag{3}$$

In training the Gaussian mixture model (GMM), these parameters are estimated such that in some sense, they best match the distribution of the training vectors. The most widely used training method is the maximum likelihood (ML) estimation. The following reestimation formulas are used to estimate GMM model parameters [9]

$$\overline{w}_i = \frac{1}{T} \sum_{t=1}^{T} P(i \mid x_t, \lambda) \tag{4}$$

$$\overline{\mu}_i = \frac{\sum_{t=1}^{T} P(i \mid x_t, \lambda) x_t}{\sum_{t=1}^{T} P(i \mid x_t, \lambda)} \tag{5}$$

$$\overline{\Sigma}_i = \frac{\sum_{t=1}^{T} P(i \mid x_t, \lambda)(x_t - \overline{\mu}_i)(x_t - \overline{\mu}_i)'}{P(i \mid x_t, \lambda)} \tag{6}$$

2. MAXIMUM *A POSTERIORI* PROBABILITY (MAP) DECISION RULE

Let λ_k, $k = 1,..., M$, denote speaker models of M speakers. Given a feature vector sequence X, a classifier is designed to classify X into M speaker models by using M discriminant functions $f_k(X)$, computing the similarities between the unknown X and each speaker model λ_k and selecting the model λ_{k^*} if

$$k^* = \arg\max_{1 \le k \le M} f_k(X) \tag{7}$$

In the minimum-error-rate classifier [1], the discriminant function is the *a posteriori* probability

$$f_k(X) = P(\lambda_k \mid X) \tag{8}$$

Using the Bayes rule

$$P(\lambda_k \mid X) = \frac{P(\lambda_k) P(X \mid \lambda_k)}{P(X)} \tag{9}$$

and assuming equally likely speakers, i.e., $P(\lambda_k) = 1/M$, and noting that $P(X)$ is the same for all speaker models, the discriminant function in (8) is equivalent to the following likelihood function [7]

$$f_k(X) = P(X \mid \lambda_k) \tag{10}$$

Finally, using the log-likelihood in (1), the decision rule used for speaker identification is

Decide speaker k^* if

$$k^* = \arg\max_{1 \le k \le M} \sum_{t=1}^{T} \log P(x_t \mid \lambda_k) \tag{11}$$

where $P(x_t \mid \lambda_k)$ is given in (2). The decision rules using (8) and (10) are called the MAP rule and the ML rule, respectively.

3. ESTIMATION OF PRIOR PROBABILITIES

We propose a new method in which the prior probabilities can be estimated directly from the training data set using the Lagrange method. Let X be the whole training data set used to train the model set $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_M\}$ for M speakers, the probability of X given Λ is as follows

$$\begin{aligned} \log P(X \mid \Lambda) &= \sum_{t=1}^{T} \log P(x_t \mid \Lambda) \\ &= \sum_{t=1}^{T} \log \sum_{i=1}^{M} P(x_t, \lambda_i \mid \Lambda) \\ &= \sum_{t=1}^{T} \log \sum_{i=1}^{M} P(\lambda_i \mid \Lambda) P(x_t \mid \lambda_i, \Lambda) \end{aligned} \tag{12}$$

The prior probabilities $P(\lambda_i \mid \Lambda)$ satisfies

$$\sum_{i=1}^{M} P(\lambda_i \mid \Lambda) = 1 \tag{13}$$

The task is to find $P(\lambda_i|\Lambda)$ such that the function $\log P(X|\Lambda)$ is maximized. Maximizing the following Lagrangian with the multiplier k

$$L = \sum_{t=1}^{T} \log \sum_{i=1}^{M} P(\lambda_i|\Lambda)P(x_t|\lambda_i,\Lambda) - k\left[\sum_{i=1}^{M} P(\lambda_i|\Lambda) - 1\right] \quad (14)$$

over $P(\lambda_i|\Lambda)$ is performed by setting its derivative to zero. The updated prior probabilities $\overline{P(\lambda_i|\Lambda)}$ is calculated from $P(\lambda_i|\Lambda)$ as follows

$$\overline{P(\lambda_i|\Lambda)} = \frac{1}{T}\sum_{t=1}^{T} \frac{P(x_t|\lambda_i,\Lambda)P(\lambda_i|\Lambda)}{\sum_{k=1}^{M} P(x_t|\lambda_k,\Lambda)P(\lambda_k|\Lambda)} \quad (15)$$

The prior reestimation algorithm:

Step 1: Generate the probability $P(\lambda_i|\Lambda)$ at random satisfying (13)
Step 2: Compute the probability $P(x_t|\lambda_i,\Lambda)$ using (2), (3), (4), (5) and (6)
Step 3: Update the probability $\overline{P(\lambda_i|\Lambda)}$ according to (15)
Step 4: Stop if the difference between the probability $P(\lambda_i|\Lambda)$ and its update $\overline{P(\lambda_i|\Lambda)}$ is below a chosen threshold, otherwise go to step 2.

The proposed MAP decision rule:

Given an unknown utterance X and a set of M speaker models $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_M\}$, the proposed MAP decision rule is stated as follows

Decide speaker k^* if

$$k^* = \arg\max_{1\le k\le M} P(X|\lambda_k,\Lambda)P(\lambda_k|\Lambda) \quad (16)$$

4. EXPERIMENTAL RESULTS

Database description

The YOHO corpus was designed for speaker verification systems in office environments with limited vocabulary. There are 138 speakers, 106 males and 32 females. The vocabulary consists of 56 two-digit numbers ranging from 21 to 97 pronounced as "twenty-one", "ninety-seven", and spoken continuously in sets of three, for example "36-45-89", in each utterance. There are four enrolment sessions per speaker, numbered 1 through 4, and each session contains 24 utterances. There are also ten verification sessions, numbered 1 through 10, and each session contains 4 utterances. All waveforms are low-pass filtered at 3.8 kHz and sampled at 8 kHz. Speech processing was performed using HTK V2.0, a toolkit [10] for building hidden Markov models (HMMs). The data were processed in 32 ms frames at a frame rate of 10 ms. Frames were Hamming windowed and pre-emphasized. The basic feature set consisted of 12th-order mel-frequency cepstral coefficients (MFCCs) and the normalized short-time energy, augmented by the corresponding delta MFCCs to form a final set of feature vector with a dimension of 26 for individual frames

Algorithmic Issues

GMMs are initialized as follows. Mixture weights, mean vectors, and covariance matrices were initialized with essentially random choices. Covariance matrices are diagonal, i.e. $[\sigma_k]_{ii} = \sigma_k^2$ and $[\sigma_k]_{ij} = 0$ if $i \neq j$, where σ_k^2, $1 < k < c$ are variances. A variance limiting constraint was applied to all GMMs using diagonal covariance matrices [7]. This constraint places a minimum variance value $\sigma^2_{min} = 10^{-2}$ on elements of all variance vectors in the GMM in our experiments. Each speaker was modelled by using 96 training utterances in four enrolment sessions without end-point detection. Error rates therefore were not too low to allow meaningful comparisons between the current and proposed methods. GMMs were trained in text-independent mode.

Experimental Results

Figure 1 shows the speaker identification error rates averaged on the YOHO 138 speakers. Speaker models consist of 16, 32 and 64 Gaussian mixtures, respectively. The identification error rate obtained by using the MAP decision rule is lower than that obtained by using the ML decision rule in all of the three different model sizes.

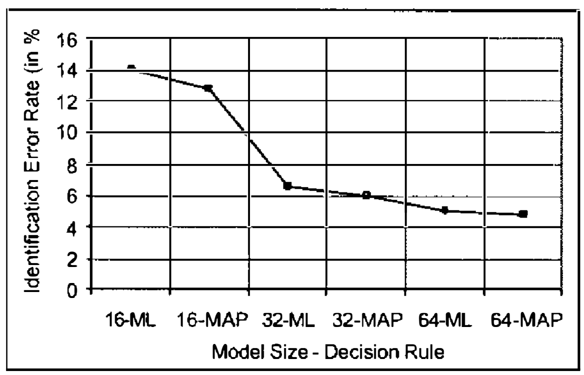

Figure 1: Speaker identification error rate (in %) averaged on 138 speakers for speaker models consisting of 16, 32 and 64 Gaussian distributions using the maximum likelihood (ML) and maximum *a posteriori* (MAP) decision rule

Figures 2 and 3 show the speaker identification error rates versus the number of speakers. In general, the higher the identification error rate is, the larger the number of speakers is. In both the figures, the MAP decision rule provides lower identification error rates compared to the ML decision rule. A similar result is also obtained for experiments using 64 Gaussian mixtures.

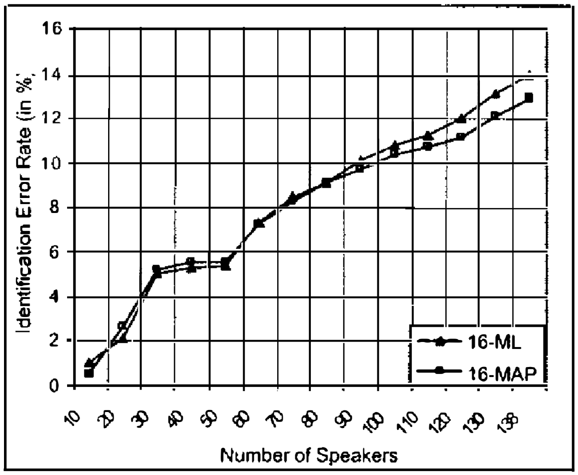

Figure 2: Speaker identification error rate (in %) versus the number of speakers for speaker models consisting of 16 Gaussian distributions using the maximum likelihood (ML) andmaximum *a posteriori* (MAP) decision rule

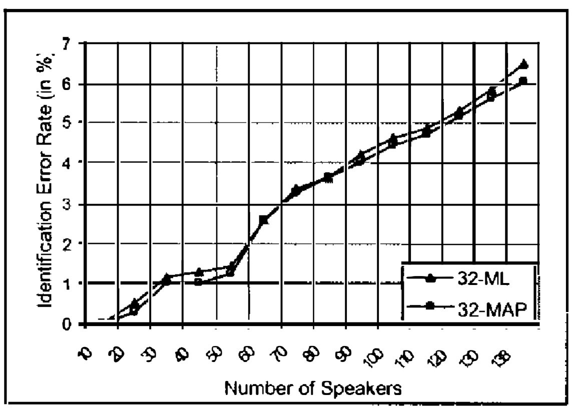

Figure 3: Speaker identification error rate (in %) versus the number of speakers for speaker models consisting of 32 Gaussian distributions using the maximum likelihood (ML) and maximum *a posteriori* (MAP) decision rule

5. CONCLUSION

An estimation method has been proposed to estimate the a priori probability for each speaker. The a priori probabilities are estimated directly from the training data set and speaker models trained by using this data set. Experimental results on 138 speakers showed that using the estimated a priori probability in speaker identification has provided a better performance.

6. ACKNOWLEDEMENT

The author would like to acknowledge the support of the Divisional Research Institute Grant, University of Canberra, Australia.

7. REFERENCES

[1] R.O. Duda and P.E. Hart, "*Pattern classification and scene analysis*", John Wiley & Sons, 1973.

[2] S. Furui, "*Recent advances in speaker recognition*", Patter Recognition Letters, 18, pp. 859-872, 1997.

[3] S. Furui, "*An overview of speaker recognition technology*", in Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 1-9, 1994.

[4] X.D. Huang, Y. Ariki, and M.A. Jack, "*Hidden Markov models for speech recognition*", Edinburgh University Press, 1990.

[5] B. H. Juang, "*The Past, Present, and Future of Speech Processing*", IEEE Signal Processing Magazine, 15:3, pp. 24-48, May, 1998.

[6] J. B. Millar, J. P. Vonwiller, J. M. Harrington, and P. J. Dermody, " The Australian National Database of Spoken Language", in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, pp. 97-100, 1994.

[7] D.A. Reynolds, "*Robust text-independent speaker identification using Gaussian mixture models*", IEEE Transactions on Speech and Audio Processing, vol. 3, no. 1, January 1995.

[8] D.A. Reynolds, "*A Gaussian mixture modeling approach to text-independent Speaker Identification*", PhD thesis, 1993.

[9] D. Tran and M. Wagner, "Fuzzy Gaussian Mixture Models for Speaker Recognition", Australian Journal of Intelligent Information Processing Systems (AJIIPS), vol. 5, no. 4, pp. 293-300, 1998

[10] P. C. Woodland et. al., "*Broadcast news transcription using HTK*", in Proceedings of ICASSP, USA, 1997.

[11] X. Zhu, Y. Gao, S. Ran, F. Chen, I. Macleod, B. Millar and M. Wagner, "*Text-independent speaker recognition using VQ, Mixture Gaussian VQ and Ergodic HMMs*", ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 55-58, 1994.

[12] N. Kambhatla, "*Local models and Gaussian mixture models for statistical data processing*", PhD thesis, Oregon Graduate Institute of Science & Technology, pp. 175-177, 1996.