Proceedings of the 6th WSEAS Int. Conf. on FUZZY SYSTEMS, Lisbon, Portugal, June 16-18, 2005 (pp7-11)

Fuzzy Estimation of Priors in Speaker Recognition

DAT T. TRAN
School of Information Sciences and Engineering
University of Canberra
Canberra, ACT 2601
AUSTRALIA

TUAN D. PHAM
School of Information Technology
James Cook University
Townsville, QLD 4811
AUSTRALIA

Abstract: This paper proposes a method to estimate the a priori probability for speakers based on the training data set, speaker models and a fuzzy estimation technique. Speaker identification experiments performed on 138 Gaussian mixture speaker models in the YOHO database using the priors estimated by the fuzzy estimation method showed lower error rates than using those estimated by the probabilistic estimation method.

Keywords: Fuzzy Priors**,** Fuzzy Gaussian Mixture Model, Speaker Recognition.

1 Introduction

Speaker recognition is the process of automatically recognizing a speaker by using speaker-specific information included in speech waves [2]-[3]. 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. 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, the identity claim is accepted. 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.

We have applied different fuzzy-set algorithms for speaker recognition [8]-[11] and found that fuzzy Gaussian mixture models (FGMMs) are effective models capable of achieving high identification accuracy [11]. The FGMM algorithm can be referred to as a prototype-based algorithm, that is, a number of prototypes are generated from the training data. Each prototype consists of a set of model parameters including fuzzy mean vector, fuzzy covariance matrix and fuzzy mixture weight. Parameters are trained in an unsupervised classification using the expectation maximisation (EM) algorithm [1].

Given an unknown utterance and a set of speaker models trained by the fuzzy Gaussian mixture modelling (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 [7].

We have proposed a probabilistic method to estimate the *a priori* probabilities for speakers based on the training data set and speaker models [12]. In this paper, we propose a new fuzzy estimation method to estimate those probabilities. The *a priori* probabilities are randomly initialized and then iteratively updated by applying a fuzzy re-estimation equation until a convergence is reached. Speaker identification experiments performed on 138 fuzzy Gaussian mixture speaker models in the YOHO database using the fuzzy estimation showed lower error rates than using the probabilistic estimation rule

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

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

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

$$f_k(X) = P(\lambda_k \mid X) \tag{2}$$

Using the Bayes rule

$$P(\lambda_k \mid X) = \frac{P(\lambda_k)P(X \mid \lambda_k)}{P(X)} \tag{3}$$

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 (2) is equivalent to the following likelihood function [7]

$$f_k(X) = P(X \mid \lambda_k) \tag{4}$$

Since

$$\log P(X \mid \lambda) = \sum_{t=1}^{T} \log P(x_t \mid \lambda) \tag{5}$$

Using the log-likelihood in (5), 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 for any k, $P(x_t \mid \lambda)$ is calculated as follows

$$P(x_t \mid \lambda) = \sum_{i=1}^{C} w_i N(x_t, \alpha_i, \Sigma_i) \tag{12}$$

where λ denotes a prototype consisting of a set of model parameters $\lambda = \{w_i, \alpha_i, \Sigma_i\}$, w_i, i = 1,…, C, are the fuzzy mixture weights and $N(x_t, \alpha_i, \Sigma_i)$, i = 1,…, C, are the n-variate Gaussian component densities with fuzzy mean vectors α_i and fuzzy covariance matrices Σ_i

$$N(x_t, \alpha_i, \Sigma_i) = \frac{\exp\left\{-\frac{1}{2}(x_t - \alpha_i)'\Sigma_i^{-1}(x_t - \alpha_i)\right\}}{(2\pi)^{d/2}\,|\Sigma_i|^{1/2}} \tag{13}$$

$$\overline{w}_i = \sum_{t=1}^{T} u_{it}^m \Big/ \sum_{i=1}^{C}\sum_{t=1}^{T} u_{it}^m \tag{14}$$

$$\overline{\alpha}_i = \sum_{t=1}^{T} u_{it}^m x_t \Big/ \sum_{t=1}^{T} u_{it}^m \tag{15}$$

$$\overline{\Sigma}_i = \frac{\sum_{t=1}^{T} u_{it}^m (x_t - \alpha_i)(x_t - \alpha_i)'}{\sum_{t=1}^{T} u_{it}^m} \tag{16}$$

$$u_{it} = \left[\sum_{k=1}^{C} (d_{it}/d_{kt})^{\frac{2}{m-1}}\right]^{-1} \tag{17}$$

The decision rules using (1) and (11) are called the MAP rule and the maximum likelihood (ML) rule, respectively.

3 Probablistic Estimation of Priors

We present a method in which the prior probabilities can be estimated directly from the training data set using the Lagrange method [11]. Let X be the whole training data set used to train the model set $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_M\}$ for M speakers, the probability of X given Λ is as follows

$$\begin{aligned}\log P(X \mid \Lambda) &= \sum_{t=1}^{T} \log P(x_t \mid \Lambda) \\ &= \sum_{t=1}^{T} \log \sum_{i=1}^{M} P(x_t, \lambda_i \mid \Lambda) \\ &= \sum_{t=1}^{T} \log \sum_{i=1}^{M} P(\lambda_i \mid \Lambda) P(x_t \mid \lambda_i, \Lambda)\end{aligned} \tag{18}$$

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

$$\sum_{i=1}^{M} P(\lambda_i \mid \Lambda) = 1 \tag{19}$$

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

$$L = \sum_{t=1}^{T} \log \sum_{i=1}^{M} P(.\,_i \mid .\,) P(x_t \mid .\,_i, .\,) - . \left[\sum_{i=1}^{M} P(.\,_i \mid .\,) - 1 \right] \tag{20}$$

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

$$\overline{P(.\,_i \mid .\,)} = \frac{1}{T} \sum_{t=1}^{T} \frac{P(x_t \mid .\,_i, .\,) P(.\,_i \mid .\,)}{\sum_{k=1}^{M} P(x_t \mid .\,_k, .\,) P(.\,_k \mid .\,)} \tag{21}$$

The prior estimation algorithm:

Step 1: Generate the probability $P(.\,_i \mid .\,)$ at random satisfying (19)

Step 2: Compute the probability $P(x_t \mid .\,_i, .\,)$ using (12), (13), (14), (15), (16) and (17)

Step 3: Update the probability $\overline{P(.\,_i \mid .\,)}$ according to (21)

Step 4: Stop if the difference between the probability $P(.\,_i \mid .\,)$ and its update $\overline{P(.\,_i \mid .\,)}$ is below a chosen threshold, otherwise go to step 2.

4 Fuzzy Estimation of Priors

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 $. = \{.\,_1, .\,_2, \ldots, .\,_M\}$ for *M* speakers, the probability of *X* given $.$ is as follows

$$J_m(U, . \,; X) = \sum_{i=1}^{M} \sum_{t=1}^{T} u_{it}^m d_{it}^2 \tag{22}$$

where $U = \{u_{it}\}$ is a fuzzy *M*-partition of *X*, each u_{it} represents the degree of vector x_t belonging to the *i*th speaker and is called the fuzzy membership function. For $1 \le i \le M$ and $1 \le t \le T$, we have

$$0 = u_{it} = 1, \ \sum_{i=1}^{M} u_{it} = 1, \text{ and } 0 < \sum_{t=1}^{T} u_{it} < T \tag{23}$$

$m > 1$ is a weighting exponent on each fuzzy membership u_{it} and d_{it} is the distance from x_t to speaker $.\,_i$, known as a measure of dissimilarity

$$d_{it}^2 = -\log P(x_t, .\,_i \mid .\,) = -\log[P(.\,_i \mid .\,) P(x_t \mid .\,_i, .\,)] \tag{24}$$

Substituting (24) into (22) gives

$$J_m(U, . \,; X) = -\sum_{i=1}^{M} \sum_{t=1}^{T} u_{it}^m \log P(.\,_i \mid .\,) - \sum_{i=1}^{M} \sum_{t=1}^{T} u_{it}^m \log P(x_t \mid .\,_i, .\,) \tag{25}$$

Minimising J_m is performed by minimising each term on the right hand side of (25). To minimise the first term, we apply (19) to maximize the following augmented objective function

$$f(P) = \sum_{i=1}^{M} \sum_{t=1}^{T} u_{it}^m \log P(.\,_i \mid .\,) + . \left(\sum_{i=1}^{M} P(.\,_i \mid .\,) - 1 \right) \tag{26}$$

over $P(.\,_i \mid .\,)$. The updated prior probabilities $\overline{P(.\,_i \mid .\,)}$ is calculated as follows

$$\overline{P(.\,_i \mid .\,)} = \sum_{t=1}^{T} u_{it}^m \Big/ \sum_{i=1}^{M} \sum_{t=1}^{T} u_{it}^m \tag{27}$$

where

$$u_{it} = \left[\sum_{k=1}^{M} (d_{it} / d_{kt})^{\frac{2}{m-1}} \right]^{-1} \tag{28}$$

d_{it} is determined in (24). The second expression on the right-hand size of (24) after minimization gives the parameter estimation equation for speaker models as shown in (14), (15) and (16).

The fuzzy prior estimation algorithm:

Step 1: Generate the probability $P(.\,_i \mid .\,)$ at random satisfying (19)

Step 2: Compute the probability $P(x_t \mid .\,_i, .\,)$ using (12), (13), (14), (15), (16) and (17)

Step 3: Update the probability $\overline{P(.\,_i \mid .\,)}$ according to (27)

Step 4: Stop if the difference between the probability $P(.\,_i \mid .\,)$ and its update $\overline{P(.\,_i \mid .\,)}$ is below a chosen threshold, otherwise go to step 2.

5 The Proposed MAP Decision Rule

Given an unknown utterance *X* and a set of *M* speaker models $. = \{.\,_1, .\,_2, \ldots, .\,_M\}$, the proposed MAP decision rule is stated as follows

Decide speaker k^* if

$$k^* = \arg\max_{1=k=M} P(X \mid .\,_k, .\,)P(.\,_k \mid .\,) \qquad (29)$$

6 Experimental Results

6.1 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 [13] 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

6.2 Algorithmic Issues

Fuzzy GMMs and 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 . j, where σ_k^2, $1 \leq k \leq C$ are variances. A variance limiting constraint was applied to all fuzzy GMMs using diagonal covariance matrices [7]. This constraint places a minimum variance value $\sigma_{\min}^2 = 10^{-2}$ on elements of all variance vectors in the fuzzy GMM and 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. Fuzzy GMMs and GMMs were trained in text-independent mode.

6.3 Experimental Results

Figure 1 shows the speaker identification error rates versus the number of speakers consisting of 16 Gaussian mixture models. The line 16-ML shows the identification error rate for speaker models trained by Gaussian mixture modeling method and tested by the maximum likelihood decision rule in (11). The line 16-MAP shows the identification error rate for speaker models trained by Gaussian mixture modeling method and tested by the maximum a posteriori decision rule in (1). The line 16-FUZZY MAP shows the identification error rate for speaker models trained by fuzzy Gaussian mixture modeling method and tested by the maximum likelihood decision rule in (29).

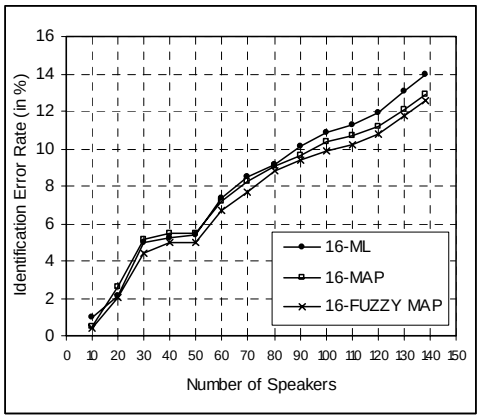

Fig. 1: Speaker identification error rate (in %) versus the number of speakers for speaker models consisting of 16 fuzzy Gaussian mixture models using the maximum likelihood (ML), maximum *a posteriori* (MAP), and fuzzy MAP decision rule

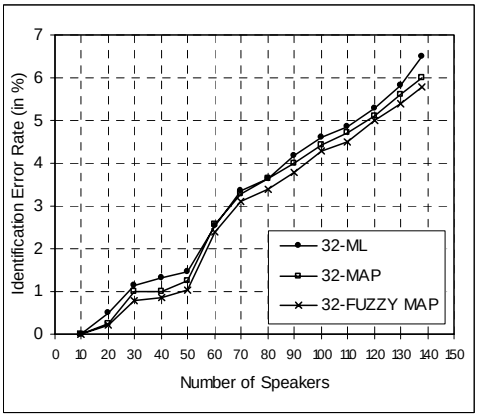

Fig. 2: Speaker identification error rate (in %) versus the number of speakers for speaker models consisting of 32 fuzzy Gaussian mixture models using the maximum likelihood (ML), maximum *a posteriori* (MAP), and fuzzy MAP decision rule

A similar result for speaker models consisting of 32 Gaussian mixture models is presented in Figure 2. In general, the higher the identification error rate is, the larger the number of speakers is. In both the figures, the fuzzy MAP decision rule provides lower identification error rates compared to the ML and MAP decision rules.

7 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.

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, *Pattern Recognition Letters*, 18, 1997, pp. 859-872.

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

[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, May, 1998, pp. 24-48.

[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, 1994, pp. 97-100.

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

[8] T.D. Pham and M. Wagner, Similarity normalization for speaker verification by fuzzy fusion, *Pattern Recognition*, 33:2 (2000) 309-315.

[9] T.D. Pham and M. Wagner, Speaker verification with fuzzy fusion and genetic optimization, *Int. J. Advanced Computational Intelligence*, 3:6 (1999) 451-456.

[10] D. Tran, T. Pham and M. Wagner, Speaker recognition using Gaussian mixture models and relaxation labeling, P*roc. 3rd World MultiConf. Systemics, Cybernetics and Informatics* (SCI'99) and *5th Int. Conf. Information Systems Analysis and Synthesis* (ISAS'99)} (Orlando, USA 1999), Volume 6 (Image, Acoustic, Speech and Signal Processing) pp. 383-389.

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

[12] D. Tran, Estimation of Prior Probabilities in Speaker Recognition, Proceedings of the 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, Hong Kong, October 2004, pp. 141-144

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

[14] 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*, 1994, pp. 55-58.

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