# A PROPOSED DECISION RULE FOR SPEAKER IDENTIFICATION BASED ON A POSTERIORI PROBABILITY

*Dat Tran, Minh Do, Michael Wagner and T. Van Le*

**Human-Computer Communications Laboratory**
**University of Canberra, PO Box 1, Belconnen, ACT 2601, Australia**
**E-mail: (dat, minhdo, miw, vanl)@hcc1.canberra.edu.au**

## RÉSUMÉ

Dans la reconnaisance du locuteur, la regle du maximum de vraisemblance (ML) est utilise comme critere pour affecter une sequence donnee de vecteurs acoustiques au modele maximum de vraisemblance du locuteur. Parcontre cette regle n'est pas toujours applicable dans certains cas. Une autre regle de decision, moyenne normalisee maximum de vraisemblance (MANL) est proposee dans ce papier. L'analyse theorique et les resultats experimentaux montre que la regle MANL peut etre utilisee pour l'identification du locuteur et est plus efficace que la regle ML dans les approches qui sont basees sur le modele de melange Gaussien et quantification vectorielle.

## ABSTRACT

In speaker recognition, the maximum likelihood (ML) rule is used as a criterion to assign a given sequence of acoustic vectors to the maximum likelihood speaker model. However, this rule is not flexible in some cases. An alternative decision rule, the maximum average normalised likelihood (MANL), is proposed in this paper. The theoretical analysis and the experimental results show that the MANL rule can be used in speaker identification and it is more effective than the ML rule in the approaches based on Gaussian mixture model (GMM) and vector quantisation (VQ).

## 1. INTRODUCTION

Let $S$ be the population of speakers $s_i$ that are represented by speaker models $\lambda_i$, $i = 1,\ldots, n$, and $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ is a sequence of acoustic vectors extracted from a continuous speech signal. This speech is uttered by an unknown speaker in the population $S$. The task of the speaker identification is to use a decision rule to assign the sequence $X$ to one of the models $\lambda_i$. Three approaches to decision rules are geometric, topological, and probabilistic rules where probabilistic decision rule is the most popular [2].

A probabilistic decision rule, which assigns an acoustic vector $\mathbf{x}$ to a model $\lambda_i$ of highest *a posteriori* probability $Pr(\lambda_i|\mathbf{x})$ is called the maximum *a posteriori* (MAP) decision rule [2] or Bayes decision rule for minimum error rate [1].

From the MAP rule, an equivalent rule is derived using Bayes theorem and an assumption that the *a priori* probabilities of speakers are the same. This rule is called *the maximum likelihood* (ML) *decision rule*, in which a vector $\mathbf{x}$ is assigned to a model $\lambda_i$ of highest probability density function (pdf) $f(\mathbf{x}|\lambda_i)$. For a sequence of acoustic vectors $X$, with the assumption the vectors are statistically independent, this rule assigns $X$ to a model $\lambda_i$ of highest likelihood $f(X|\lambda_i)$, which is the product of the above pdfs.

However, for speaker recognition the ML rule could be thought as an incomplete rule since some methods proposed by Matsui and Furui [4, 6, 8] increased the speaker recognition rate. Based on analysing those cases, an alternative decision rule is proposed in this paper. A sequence of acoustic vectors $X$ is assigned to a speaker model $\lambda_i$ if the probability of the selected model is the highest. This probability is computed directly as the average of the *a posteriori* probabilities of those vectors, which are the normalised pdf. Therefore, this rule can be called *the maximum average normalised likelihood* (MANL) *decision rule*.

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

Consider the case of identifying a speaker from a single feature vector $\mathbf{x}$. Let $Pr(\lambda_i|\mathbf{x})$ be the probability that vector $\mathbf{x}$ was produced by speaker $s_i$. The identified speaker is then the speaker with the highest probability. This decision rule is expressed as follows

*assign* $\mathbf{x}$ *to speaker* $s_i$ *if*

$$Pr(\lambda_i|\mathbf{x}) > Pr(\lambda_k|\mathbf{x})$$

*for all* $k \neq i$ (2.1)

This rule is also called Bayes decision rule that minimises the error rate. For a sequence of vectors, using the Bayes theorem, this rule is extended into the maximum likelihood decision rule.

## 3. MAXIMUM LIKELIHOOD (ML) DECISION RULE

Using likelihood functions, Bayes theorem is written

$$Pr(\lambda_i \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid \lambda_i) Pr(\lambda_i)}{\sum_{h=1}^{n} f(\mathbf{x} \mid \lambda_h) Pr(\lambda_h)} \qquad (3.1)$$

where $f(\mathbf{x}|\lambda_i)$ is the probability density function or likelihood that a vector $\mathbf{x}$ is observed in model $\lambda_i$, $f(\mathbf{x})$ is the unconditional probability density function for all speaker models, and $Pr(\lambda_i)$ is the *a priori* probability of speaker $s_i$ being the unknown speaker. Using (2.1), we have

*assign* $\mathbf{x}$ *to* $s_i$ *if*

$$\frac{f(\mathbf{x} \mid \lambda_i) Pr(\lambda_i)}{\sum_{h=1}^{n} f(\mathbf{x} \mid \lambda_h) Pr(\lambda_h)} > \frac{f(\mathbf{x} \mid \lambda_k) Pr(\lambda_k)}{\sum_{h=1}^{n} f(\mathbf{x} \mid \lambda_h) Pr(\lambda_h)}$$

*for all* $k \neq i$ (3.2)

Assume that all speakers are equally-likely to be the unknown speaker, i.e. $Pr(\lambda_i) = 1/n$, $i = 1,\ldots, n$, and statistical independence between each vector for a sequence of vectors $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ [1, 3, 5]

$$f(X \mid S) = \prod_{t=1}^{T} f(\mathbf{x}_t \mid S) \qquad (3.3)$$

the rule in (3.2) is extended for a sequence $X$ as follows

*assign* $X$ *to* $s_i$ *if*

$$\prod_{t=1}^{T} f(\mathbf{x}_t \mid \lambda_i) > \prod_{t=1}^{T} f(\mathbf{x}_t \mid \lambda_k)$$

for all $k \neq i$ (3.4)

This rule is also known as the maximum likelihood decision rule and is usually presented as *the maximum log-likelihood decision rule*

*assign* $X$ *to* $s_i$ *if*

$$\sum_{t=1}^{T} log\ f(\mathbf{x}_t \mid \lambda_i) > \sum_{t=1}^{T} log\ f(\mathbf{x}_t \mid \lambda_k)$$

*for all* $k \neq i$ (3.5)

Using a mixture of multivariate Gaussian pdfs and the "winner-take-all" assumption [11] we can relate mixture models and hard clustering algorithms. According to this asumption, a VQ clustering with Euclidean distance is equivalent to maximum likelihood parameter estimation when the component Gaussians are spherically symmetric and the mixing proportions are equal. So from (3.5) a rule for VQ method is as follows

*assign* $X$ *to* $s_i$ *if*

$$\sum_{t=1}^{T} d_{ti} < \sum_{t=1}^{T} d_{tk}$$

*for all* $k \neq i$ (3.6)

where $d_{ti}$ is the Euclidean distance between vector $\mathbf{x}_t$ and the *nearest* codevector in the codebook of speaker $s_i$ . (3.6) is known as the *minimum overall average distortion rule* in VQ method.

A remark could be made about the above rules from their expressions. The product of probabilities in (3.4) will be very small if a very small probability is included. In (3.6) this means the sum of distances is very large if it includes a long distance. This problem may happen in text-independent speaker recognition using a short utterance with intrinsically wide variability, and where the test vector distribution deviates from the training vector distribution. The recognition will be poor in such a case. To overcome this, Matsui and Furui [4] defined a distortion-intersection measure (DIM). If a test vector is out of the scope of the VQ codebook vectors (a long distance will appear), the corresponding distance will be set to the boundary of the scope. In one-state continuous HMMs (equivalent to Gaussian mixture models GMMs) the idea of DIM is that if a test vector corresponds to the tail of the Gaussian distribution, the output probability is set to the floor value.

A second remark is the variations that arise from the speaker him/herself due to noise or differences in recording. Tokens of the same utterance recorded in one session are much more highly correlated than tokens recorded in separate sessions [8]. Matsui and Furui proposed normalisation methods [6, 8] based on the *a posteriori* probability in speaker verification to reduce these variations.

## 4. MAXIMUM AVERAGE NORMALISED LIKELIHOOD (MANL) DECISION RULE

From the above second remark, the proposed decision rule to assign a sequence of vectors $X$ to a speaker model $\lambda_i$ should be based on the *a posteriori* probabilities $Pr(\lambda_i|\mathbf{x}_j)$ instead of the pdfs as in (3.4). To avoid the influence of the very small probabilities, the probability for a sequence of vectors should not be represented as the product of probabilities of vectors as in (3.4).

In statistics, the most important estimate of the sample values is the sample mean that indicates where the center of these values is located. So we propose to use the mean of the *a posteriori* probabilities to specify assigning a sequence of vectors $X$ to a model $\lambda_i$

$$\overline{Pr(\lambda_i \mid \mathbf{x})} = \frac{1}{T}\sum_{t=1}^{T} Pr(\lambda_i \mid \mathbf{x}_t) \qquad (4.1)$$

Assuming that all speakers are equally likely to be the unknown speaker, i.e. $Pr(\lambda_i) = 1/n$, $i = 1, \ldots, n$, the Bayes theorem (3.1) is rewritten as

$$Pr(\lambda_i \mid \mathbf{x}_t) = \frac{f(\mathbf{x}_t \mid \lambda_i)}{\sum_{h=1}^{n} f(\mathbf{x}_t \mid \lambda_h)} \qquad (4.2)$$

Using the MAP rule in (2.1) together with (4.1), a decision rule is proposed as follows

*assign* $X$ *to* $s_i$ *if*

$$\sum_{t=1}^{T} Pr(\lambda_i \mid \mathbf{x}_t) > \sum_{t=1}^{T} Pr(\lambda_k \mid \mathbf{x}_t)$$

*for all* $k \neq i$ (4.3)

Since $Pr(\lambda_i \mid \mathbf{x}_t)$ can be expressed as the normalised likelihood function in (4.2), this rule can be named *the maximum average normalised likelihood* (MANL) *decision rule*. With the pdfs in (4.2), we have

*assign* $X$ *to* $s_i$ *if*

$$\sum_{t=1}^{T}\left( \frac{f(\mathbf{x}_t \mid \lambda_i)}{\sum_{h=1}^{n} f(\mathbf{x}_t \mid \lambda_h)} \right) > \sum_{t=1}^{T}\left( \frac{f(\mathbf{x}_t \mid \lambda_k)}{\sum_{h=1}^{n} f(\mathbf{x}_t \mid \lambda_h)} \right)$$

*for all* $k \neq i$ (4.4)

Similarly, for the VQ method, we have

*assign* $X$ *to* $s_i$ *if*

$$\sum_{t=1}^{T}\left( e^{-u_{ti}} \Big/ \sum_{h=1}^{n} e^{-u_{th}} \right) > \sum_{t=1}^{T}\left( e^{-u_{tk}} \Big/ \sum_{h=1}^{n} e^{-u_{th}} \right)$$

*for all* $k \neq i$ (4.5)

where $u_{ti} = d_{ti}/2\sigma^2$, $d_{ti}$ as in (3.7) and $\sigma^2$ is referred to as the average covariance of the training set of all speakers.

With (4.5) we need not define a distortion intersection measure (DIM) as in section 3. If a test vector is out of the scope of VQ codebook vectors, corresponding to a long distance, its contribution in (4.5) is very small due to the exponential function. In the one-state HMMs, flattening the tail of each Gaussian distribution is implemented in (4.4) because the influence of the very small probability in the sum is less than that in the product.

To compare the ML rule and MANL rule, after some simple calculations, we can transform the expressions in (3.5) following the *a posteriori* probabilities. The ML rule is equivalent to

assign $X$ to $s_i$ if

$$\sum_{t=1}^{m} log\, Pr(\lambda_i \mid \mathbf{x}_t) > \sum_{j=1}^{m} log\, Pr(\lambda_k \mid \mathbf{x}_t)$$

for all $k \neq i$ (4.6)

The ML rule in (4.6) and the MANL rule in (4.3) are different because of the logarithm function essentially. Thus we can see that the variations of small probabilities in the ML rule are more strongly influential than the ones of large probabilities. In the MANL rule, the role of these probabilities is the same.

## 5. EXPERIMENTAL RESULTS

According to the above remarks, in this paper we present the results of GMM-based and VQ-based speaker identification experiments. The commercially available TI46 speech data corpus is used to compare these decision rules. There are 16 speakers, 8 female and 8 male, labelled f1-f8 and m1-m8, respectively. The vocabulary contains a set of ten single-word computer commands which are: *enter, erase, go, help, no, rubout, repeat, stop, start,* and *yes*. Each speaker repeated the words 10 times in a single training session, and then again twice in each of 8 later testing sessions. The corpus is sampled at 12500 samples per second and 12 bits per sample. The data were processed in 20.48 ms frames (256 samples) at a frame rate of 125 frames per second (100 sample shift). Frames were Hamming windowed and preemphasised with $\mu$ = 0.9. For each frame, 46 mel-spectral bands of a width of 110 mel and 20 mel-frequency cepstral coefficients (MFCC) were determined [10].

Two methods used in this experiment are Gaussian mixture model (equivalent to an one-state continuous HMM) [9] with the assumption of diagonal covariance matrices [5] and variance-weighted VQ, in which distances were weighted on a per-codebook basis according to observed variances in the training data.

In the training phase, 100 training tokens (10 utterances x 1 training session x 10 repetitions) of each speaker were used to train GMMs of 32, 64 , 128 mixtures and codebooks of 32, 64, 128 codevectors using the LBG algorithm. Speaker identification was carried out by

testing all 2560 test tokens (16 speakers x 10 utterances x 8 testing sessions x 2 repetitions) against the GMMs and the codebooks of all 16 speakers in the database using the decision rules in (3.5), (4.3) for GMM method and (3.6), (4.5) for variance-weighted VQ method.

In both methods, the MANL rule gives the better recognition rate than the ML rule.

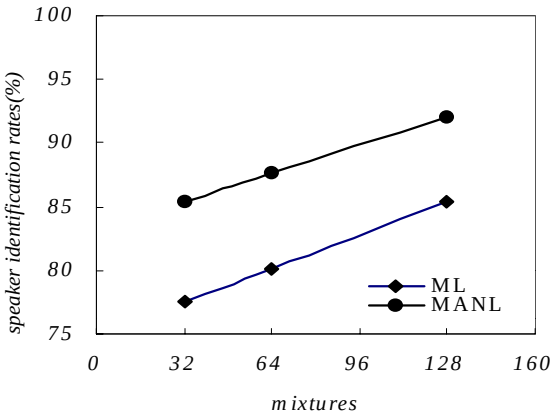


**Figure 1**: Total average speaker identification rates (%) of ML rule and MANL rule using GMM-based method

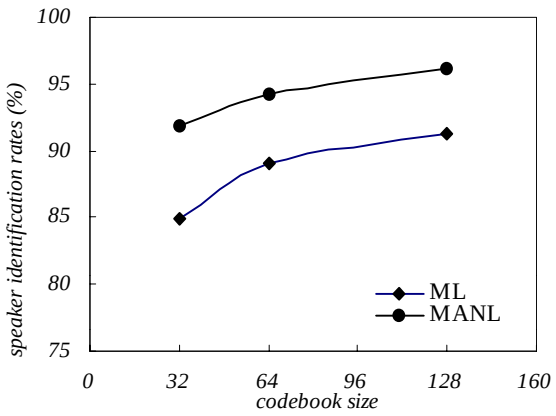


**Figure 2**: Total average speaker identification rates (%) of ML rule and MANL rule using VQ-based method

## 6. CONCLUSION

In this paper, the maximum average normalised likelihood decision rule has been proposed for speaker recognition. This decision rule has been compared with the well-known maximum likelihood rule in GMM-based and VQ-based speaker identification systems. Results show a significant error reduction for the new decision rule.

## 7. REFERENCES


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

[2] Michael Allerhand (1987), "*Knowledge-based speech pattern recognition*", Kogan Page Ltd.

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

[4] Tomoko Matsui and Sadaoki Furui (1992), "*Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs*", Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, San Francisco, pp. II-157-160.

[5] Reynolds, Douglas Alan. (1993), "*A Gaussian mixture modeling approach to text-independent speaker identification*", PhD thesis, pp. 41-43.

[6] Tomoko Matsui and Sadaoki Furui (1994), "*A new similarity normalisation method for speaker verification based on a posteriori probability*", ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 59-62.

[7] Tomoko Matsui and Sadaoki Furui (1994), "*Speaker adaptation of tied-mixture-based phoneme models for text-prompted speaker recognition*", Proc. IEEE, ICASSP, Adelaide, pp. I-125-128.

[8] Sadaoki Furui (1994), "*An overview of speaker recognition technology*", ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 1-9.

[9] Xiaoyuan Zhu, Yuqing Gao, Shuping Ran, Fangxin Chen, Iain Macleod, Bruce Millar and Michael Wagner (1994), "*Text-independent speaker recognition using VQ, Mixture Gaussian VQ and Ergodic HMMs*", ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 55-58.

[10] Michael Wagner (1996), "*Combined speech-recognition/speaker-verification system with modest training requirements*", Proceedings of the Sixth Australian International Conference on Speech Science and Technology, Adelaide, Australia, 1996, pp. 139-143.

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