A PROPOSED DECISION RULE FOR SPEAKER RECOGNITION BASED ON FUZZY C-MEANS CLUSTERING

Dat Tran, Michael Wagner and T. Van Le

Human-Computer Communication Laboratory
School of Computing, Faculty of Information Sciences and Engineering
University of Canberra, ACT 2601, Australia
E-mail: (dat, miw, vanl)@hcc1.canberra.edu.au

ABSTRACT

In vector quantisation (VQ) based speaker recognition, the minimum overall average distortion rule is used as a criterion to assign a given sequence of acoustic vectors to a speaker model known as a codebook. An alternative decision rule based on fuzzy c-means clustering is proposed in this paper. A set of membership functions associated with vectors for codebooks are defined as discriminant functions and the maximum overall average membership function rule is stated. The theoretical analysis and the experimental results show that this rule can be used in both speaker identification and speaker verification. It is more effective than the minimum overall average distortion rule.

1. INTRODUCTION

Let **C** be a set of codebooks known as speaker models C_i, $i = 1,\ldots, c$. Codebooks were trained by a VQ algorithm using a set of acoustic vectors extracted from *training* utterances. Let $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ be a sequence of acoustic vectors extracted from a *test* utterance. This utterance is uttered by an *unknown* speaker whose codebook is in **C**. To design a classifier for speaker recognition, a set of discriminant functions for speakers need to be defined. Using these discriminant functions and the set of codebooks **C**, a decision rule is established to assign the sequence X to one of the codebook C_i. This classifier design scheme is called supervised learning [1].

In the non fuzzy VQ approach, to obtain codebooks, the most widely used method is the *k*-means algorithm, where the LBG algorithm is an commonly extended version, proposed by Linde, Buzo, and Gray [4]. A set of training vectors of each speaker is clustered into a set of codevectors referred to as a codebook, such that the overall average distortion is minimised. The acoustic feature space is partitioned into separate regions. To design a classifier, the discriminant functions are defined as overall average distortion functions $D(i)$, $i = 1,\ldots, c$. They are computed using distortions between the sequence of test vectors and codebooks defined as the distances between these vectors and the nearest codevectors in the codebooks (the *nearest neighbour selection* rule). Then the speaker i is recognised if the overall average distortion $D(i)$ is a minimum. This decision rule is known as the *minimum overall average distortion rule* in the VQ method .

Most of contributions done in the fuzzy VQ approach were training codebooks using the fuzzy *c*-means (FCM) clustering algorithm generalized by Bezdek [10]. In speech and speaker recognition, this algorithm is called the fuzzy VQ algorithm [9]. Several fuzzy based contributions for classifier design but not in the VQ approach were done by Pal and Majumder [2] for speaker recognition and by Bezdek [3] for pattern recognition. Pal and Majumder defined fuzzy membership functions associated with the sequence X for the speaker models. Bezdek used fuzzy mean vectors and the *nearest neighbour selection* rule to design a prototype classification.

A new contribution in the fuzzy VQ approach for speaker recognition is presented in this paper. Codebooks are trained by the non fuzzy LBG algorithm, but a set of new discriminant functions based on fuzzy *c*-means clustering is defined. For each vector **x** in the sequence X, a fuzzy membership function is defined. It describes the degree to which this vector is a member of codebook and is computed as in the FCM clustering method where the distance between the vector and the codebook is defined using the *nearest neighbour selection* rule. For a sequence X, an average fuzzy membership function $F(i)$ for the codebook C_i is computed. Then the speaker i is recognised if the average membership function $F(i)$ is a maximum. This decision rule can be named the *maximum overall average membership function* rule. Since $F(i)$ is computed as in the FCM clustering, this function itself is a normalised score. Therefore it can be used in speaker verification, where a claimed speaker i is accepted if the function $F(i)$ is greater than a given threshold.

2. THE MINIMUM OVERALL AVERAGE DISTORTION RULE IN VQ APPROACH

The overall average distortion *D(i)* of the codebook C_i is defined by

$$D(i) = \frac{1}{T}\sum_{t=1}^{T} d_t(i) \tag{2.1}$$

where $d_t(i)$ is the distance between the vector $\mathbf{x}_t$ and the codebook C_i. Using the *nearest neighbour selection* rule, this distance is defined as the distance between the vector $\mathbf{x}_t$ and the nearest codevector $\mu_k(i)$ in the codebook C_i

$$d_t(i) = \min_k d(\mathbf{x}_t, \mu_k(i)), \quad k = 1,\ldots, K \tag{2.2}$$

where *K* is the number of codevectors in the codebook. The most commonly used measure is the Euclidean distance. In many practical applications, the Mahalanobis distance

$$d^2(\mathbf{x}_t, \mu_k) = (\mathbf{x}_t - \mu_k)' \Sigma_k^{-1} (\mathbf{x}_t - \mu_k) \tag{2.3}$$

is used where $(\mathbf{x}_t - \mu_k)'$ is the transpose of $(\mathbf{x}_t - \mu_k)$ and Σ_k^{-1} is the inverse of the covariance matrix.

Using *D(i), i = 1,…, c* as discriminant functions, the decision rule known as the *minimum overall average distortion rule* in VQ method for speaker identification is stated as follows

Decide speaker *i* if

$$D(i) < D(k)$$

for all $k \neq i$ (2.4)

For speaker verification, a claimed speaker *i* is accepted if

$$D(i) < \varepsilon \tag{2.5}$$

where ε is a given threshold

A remark could be made about the above rules from their expressions. In (2.1) 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 [11] 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.

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 [13]. Matsui and Furui proposed normalisation methods [12, 13] based on the *a posteriori* probability in speaker verification to reduce these variations.

3. THE MAXIMUM OVERALL AVERAGE FUZZY MEMBERSHIP FUNCTION RULE

Since a sequence of test vector *X* is extracted from an utterance and a classifier design for speaker recognition must decide to which speaker the sequence *X* belongs, therefore a decision made for the sequence *X* cannot be fuzzy, but at the lower level, for each vector $\mathbf{x}_t$ in the sequence, the fuzzy approach can be applied. We assume that a test vector $\mathbf{x}_t$ can belong to several codebooks and this belonging can be described by a fuzzy membership function. More generally, a classifier of *T* test vectors $\mathbf{x}_t$ in the sequence *X* into *c* codebooks can be described by a *c x T* matrix *U*, whose *i, t*th entry, u_{it} is the fuzzy membership of the vector $\mathbf{x}_t$ with the codebook C_i and satisfies

$$0 \le u_{it} \le 1 \qquad \text{and} \qquad \sum_{i=1}^{c} u_{it} = 1 \tag{3.1}$$

This classifier can be interpreted geometrically, in that the membership is used to identify the concentration of the test vector $\mathbf{x}_t$ in the codebook C_i. Moreover, if we define a fuzzy conditional risk function $h_{it} = 1 - u_{it}$, similar to the conditional risk function using the *a posteriori* probability in Bayesian classifier, then in order to achieve the minimum error rate, we should make a decision that the speaker *i* is correct if the membership function u_{it} is a maximum. More generally, the sequence *X* is assigned to the codebook C_i if the concentration of vectors $\mathbf{x}_t$, *t = 1,…, T* in this codebook is highest. This concentration can be computed as the overall average of fuzzy membership function of the codebook C_i with the degree of fuzziness *m* as follows

$$F(i) = \frac{1}{T}\sum_{t=1}^{T} u_{it}^m \tag{3.2}$$

With (3.2) we need not define a distortion intersection measure (DIM) as in section 2. If a test vector is out of the scope of VQ codebook vectors, corresponding to a

long distance, its contribution denoted by u_{it} is very small.

To compute matrix *U*, we can define a fuzzy objective function, which denotes the dissimilarity between the sequence *X* and the set of codebooks **C**

$$J_m(U;X) = \sum_{t=1}^{T}\sum_{i=1}^{c} u_{it}^m d_t^2(i) \qquad (3.3)$$

where $d_t(i)$ is the distance from $\mathbf{x}_t$ to the codebook C_i defined as the distance in (2.2).

Since the fuzzy objective function $J_m(U;X)$ denotes the dissimilarity between the sequence *X* and the set of codebooks **C**, so if this sequence *X* is extracted from an utterance uttered by an unknown speaker whose codebook is *not* in the set **C**, the value of $J_m(U;X)$ must be greater than the value corresponding to the codebook is in **C**. In speaker recognition, it is always assumed that the sequence *X* is extracted from an utterance uttered by an unknown speaker whose codebook is one of codebooks being considered. Therefore, the fuzzy objective function value of $J_m(U;X)$ must be a minimum. According to the well-known FCM method, this minimum value is obtained when

$$u_{it} = \left[\sum_{k=1}^{c} (d_{it}/d_{kt})^{2/(m-1)}\right]^{-1} \qquad (3.4)$$

Using *F(i) i = 1,…, c* in (3.2) as discriminant functions, the decision rule can be named the *maximum overall average membership function* rule and is stated as follows

Decide speaker *i* if

$$F(i) > F(k)$$

for all $k \neq i$ (3.5)

Since u_{it} is computed as in (3.4), the function *F(i)* itself is a normalised score, therefore it can be used in speaker verification as a normalisation method. The amount of calculation for (3.4) is enormous in the case of the large number of speakers, so the summation in (3.4) can be approximately computed by using a collection of background codebooks taken from the set **C** [14].

For speaker verification, a claimed speaker *i* is accepted if

$$F(i) > \varepsilon \qquad (3.6)$$

where ε is a given threshold

4. EXPERIMENTAL RESULTS

According to the above remarks, in this paper we present the results of VQ-based and FCM-based speaker recognition experiments using the short utterances. 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 μ = 0.9. For each frame, 46 mel-spectral bands of a width of 110 mel and 20 mel-frequency cepstral coefficients (MFCC) were determined [15]. In the training phase, 100 training tokens (10 utterances x 1 training session x 10 repetitions) of each speaker were used to train 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 test sessions x 2 repetitions) using (2.4) and (3.5). The experimental results are as follows:

Codebook size	Identification error results for	
	VQ-based rule	FCM-based rule
32	15.02 %	9.64 %
64	11.00 %	6.66 %
128	8.73 %	4.54 %

Table 1. Speaker identification error results for VQ-based rule and FCM-based rule

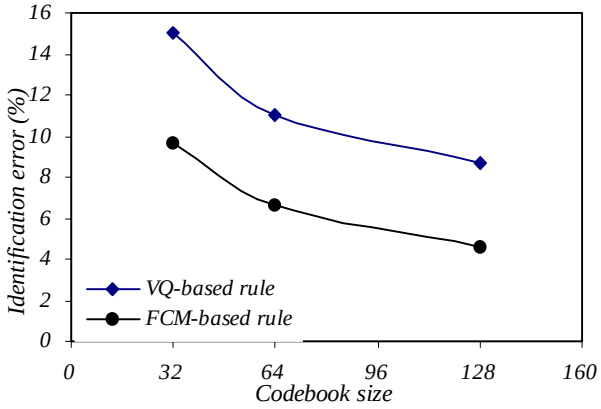

Figure 1. Speaker identification error results for VQ-based rule and FCM-based rule

Speaker verification in the text-dependent mode was carried out by testing 160 tokens for each codebook (10 short utterances x 8 test sessions x 2 repetitions) and the background speaker set includes all the eight same-gender speakers, using (2.5) and (3.6) . The experimental results are as follows:

Codebook size	Equal error rate results for	
	VQ-based rule	FCM-based rule
32	10.95 %	3.80 %
64	9.22 %	2.81 %
128	8.21 %	2.38 %

Table 2. Equal error rate results for VQ-based rule and FCM-based rule

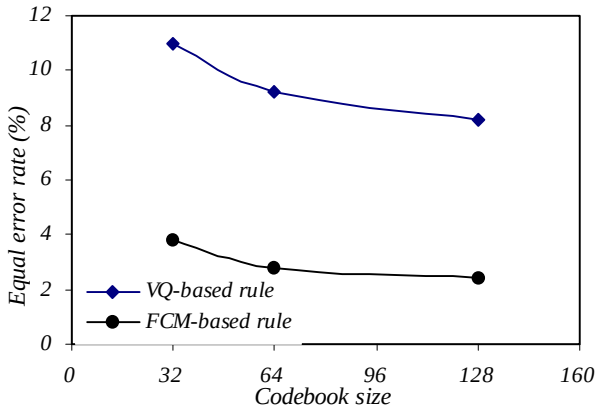

Figure 2. Equal error rate results for VQ-based rule and FCM-based rule

6. CONCLUSION

In this paper, the maximum overall average membership function rule has been proposed for speaker recognition. This rule has been compared with the well-known the minimum overall average distortion rule in the VQ method. Results show an error reduction for the new rule and show that the maximum overall average membership function rule is applicable in speaker recognition.

7. REFERENCES

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

[2] Sankar K.Pal and Dwijesh Dutta Majumder (1977), *"Fuzzy sets and decision making approaches in vowel and speaker recognition"*, IEEE Trans. Syst., Man, Cybern., pp. 625-629.

[3] James C. Bezdek and Patrick F. Castelaz (1977), *"Prototype classification and feature selection with fuzzy sets"*, IEEE Trans. Syst., Man, Cybern., vol. SMC-7, no. 2, pp. 87-92.

[4] Yoseph Linde, Andres Buzo and Robert M. Gray (1980), *"An algorithm for Vector quantiser Design"*, IEEE Trans. on Communications, vol. COM-28, no. 1, pp. 84-95.

[5] Michael P. Windham (1983), *"Geometrical fuzzy clustering algorithms"*, Fuzzy sets and Systems, vol. 10, pp. 271-279

[6] James C. Bezdek (1987), " *Pattern Recognition with Fuzzy Objective Function Algorithms*", Plenum Press, New York and London.

[7] P. Chou, T. Lookabaugh, and R. Gray (1989), "*Entropy-constrained vector quantisation*", IEEE Trans. Acoustic, Speech, and Signal Processing, vol. ASSP-37, pp. 31-42.

[8] P. Chou, T. Lookabaugh, and R. Gray (1989), "*Entropy-constrained vector quantisation*", IEEE Trans. Acoustic, Speech, and Signal Processing, vol. ASSP-37, pp. 31-42.

[9] X.D. Huang, Y. Ariki, and M.A. Jack (1990), "*Hidden Markov Models For Speech Recognition*", Edinburgh University Press.

[10] James C. Bezdek and Sankar K. Pal (1992), "*Fuzzy Models for Pattern Recognition*", IEEE Press.

[11] 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.

[12] 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.

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

[14] Douglas A. Reynolds (1995), *"Speaker identification and verification using Gaussian mixture models"*, Speech Communication, vol. 17, pp. 91-108

[15] 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.