Application of the MLVQ1 in Speaker Identification

S. OUAMOUR-SAYOUD*, H. SAYOUD and M. BOUDRAA
USTHB, Institut d'Electronique, BP 32 El-Alia, Bab-Ezzouar, Alger, Algérie.
***Email:** sayoud@ifrance.com,

ABSTRACT

In this paper we describe a new method, in automatic speaker recognition, based on modified LVQ1 (MLVQ1) and using 3 prosodic features: the mean of the pitch, the original duration and the low-frequency energy. For this purpose, we conceived a new metric, optimized in automatic speaker recognition, which we called ODHEF. The tests of speaker recognition are done in Arabic corpus with 2 different sets: a closed set and an open set. The results show that the prosodic features are relevant and that the modified LVQ1 (MLVQ1) method is interesting in text-dependent speaker identification.

1 INTRODUCTION

In 1972 ATAL proposed a speaker recognition method based on the pitch contour with a recognition rate of 97% for a learning time of 2 seconds [1].
BENNANI showed that the neural networks can provide a good performance in speaker identification if the number of speakers is limited [3].
Several investigations, in the Intra and inter-Speaker variability of the pitch [12] showed that the pitch average can represent a good feature in speaker identification.
Thus, in this work, we associate the pitch average with two other prosodic features, for the task of speaker recognition. In order to associate these three heterogeneous features in a LVQ approach, a new adapted metric is proposed.
The results show that the MLVQ algorithm provides an efficient speaker recognition in our database.

2 DATABASE

We use an Arabic speech database, where we find a lot of pertinent phonemes (a pertinent phoneme is a phoneme which has a high inter-speaker variability) like "Kaf", "Rra", "Aïn" and "Dtta" (see table 1). This choice is made after many experimentations. The average duration of an utterance is 5.5 seconds. Each utterance is repeated 6 times.

Table 1 Some Arabic pertinent phonemes used in our sentences.

Phoneme designation	Latin similarity
“Kaf”	K
“Rra”	Spanish R
“Aïn”	Nothing
“Dtta”	Nothing

There are two types of sets: a closed set and an open set.

In the closed set, all the speakers are referenced by the system.

But in the open set we have some new speakers who represent the impostors.

The speech signal is recorded at 16 kHz, with 16 bits and with a medium SNR.

3 VARIABILITY OF SOME PROSODIC FEATURES

3.1. Mathematical expression of the variability

We define below some technical words of the variability.

Inter-Speaker Variability (of a feature): which represents the variance of the means (of the feature) for the different speakers.

Intra-Speaker Variability: which represents the mean of the different intra-speaker variances (of the feature).

Wolf Ratio (WR): which represents the ratio of the **Inter-Speaker Variability** on the **Intra Speaker-Variability** [14] (see formula 1)**.**

$$WR(feature) = \frac{Var_{inter\_speaker}(feature)}{Var_{intra\_speaker}(feature)} \qquad (1)$$

where **Var** means the variance.

3.2. Physical meaning

If the Inter-Speaker Variability, of a feature, is high then the different speakers can be easily separated by this feature.

If the Intra-Speaker Variability, of a feature, is low: then every speaker can be represented by one reference based on this feature, with a high accuracy.

So, if the Wolf Ratio, of a feature, is high then we can say, according to formula 1, that this feature is relevant in speaker recognition.

3.3. Wolf Ratio of the mean of Fo

An estimation of this ratio in the closed set with 10 repetitions of the same sentence (5.5 s of length) shows that the Wolf Ratio for the mean of Fo is equal to 229.

$WR(Fo_{avr}) = 229$

Results show that this ratio is very high, according to seven other parameters tested simultaneously with Fo.

3.4. Recognition performance with Fo_{avr} alone

We tried to identify all the speakers by using only the mean of the pitch, with a nearest neighbour classification, in the case of a text-independent speaker recognition. The percentage of good identification was 54.5%. This result shows that this feature is interesting in speaker recognition.

4 THE DISTANCE "ODHEF" (a new distance adapted to heterogeneous features)

4.1. Problem

If the features used in speaker identification are similar (same kind), it is possible to use an uniform metric (eg. L^2). But, if the features are not similar or if they have different origins, it will be impossible and illogical to use the classic metrics.

In this case, we propose a new distance, for automatic speaker recognition, which is adapted to the heterogeneous features.

We call it **"ODHEF"** or Optimal Distance for HEterogeneous Features.

4.2. Description of the metric

This hybrid metric is adapted to heterogeneous features used in speaker recognition.

In this approach, three parameters are significantly important to estimate a distance between two speakers:

- the Wolf Ratio (symbolized by: *Effectiv*)
- the normalization by the mean (symbolized by: *Norm*)
- the Euclidean distance L^2

Thus, a judicious way to define this distance is given by the formula 2 (which represents the distance between the vectors of two speakers: X and Xr).

$$Dist^2(X, Xr) = \sum_{i=1}^{N} (Effectiv_i)(Norm_i).(Xr_i - X_i)^2 \quad (2)$$

The product (*Effectiv*)(*Norm*) is called coefficient of adaptation or "*Adapt*".

X is a N-dimensional vector representing the features of an utterance uttered by the speaker *X* and Xr represents the features of an utterance uttered by the reference speaker *Xr*.

The heterogeneous components (features) of X are X_1, X_2, ... X_N; and the heterogeneous components (features) of Xr are identically Xr_1, Xr_2 ... Xr_N.

The first term ($Effectiv_i$), after the symbol Σ, is the effectiveness coefficient represented by the Wolf Ratio, in order to give more importance to the most relevant parameters (in speaker recognition), because a uniform association of a relevant parameter with a non-relevant parameter, in a uniform metric, will reduce the effectiveness of the first parameter.

The second term ($Norm_i$) is a normalization by the mean, in order to give a reasonable association for all the parameters. For example, the associated of a parameter of about 10^3 with a parameter of about 10^{-1} is not equitable, if the parameters are not normalized.

Finally, the third term is the standard Euclidean distance.

We called this distance (formula 2): Optimal Distance for HEterogeneous Features (*ODHEF*).

5 METHOD OF SPEAKER RECOGNITION

5.1. Introduction

The system of speaker recognition is based on three prosodic features: the mean of the pitch, the original duration and the low-frequency energy. It uses a MLVQ algorithm for the training.

5.2. Features

This method is based on the mean of the pitch (Fo_{avr}), the original duration (D_{orig}) and the low frequency energy (E_{lf}). Each utterance is represented by a 3-dimensional vector (Fo_{avr} , D_{orig} , E_{lf}).

We can see on the figure 3 the representation of some speakers with their 3 prosodic parameters.

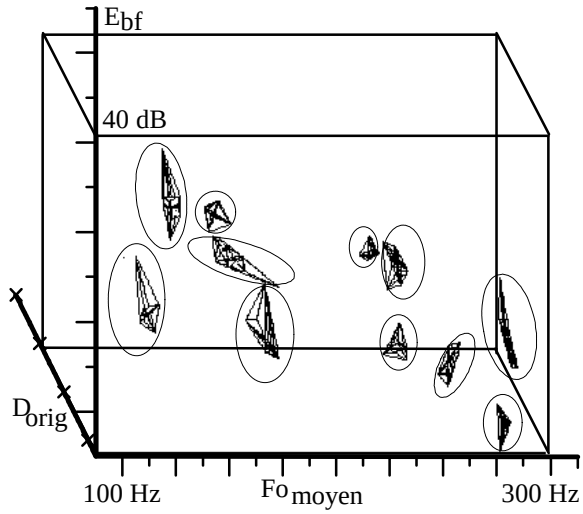

Fig. 1 Representation of some speakers areas by the 3 prosodic dimensions: Fo_{avr}, D_{orig} and E_{lf}.

- Fo_{avr} is the mean of the pitch computed for the entire utterance (in Hz). The choice of this parameter is justified by the high Wolf ratio obtained for Fo_{avr}. **$WR(Fo_{avr})=229.$**

- D_{orig} is the duration, in ms, of an original word placed inside a sentence. The original word is an unfamiliar word to the dialect of the population. Because, we show, experimentally, that the duration of a word, which is uttered for the first time by a speaker, is very characteristic of this speaker. Its Wolf coefficient is: **$WR(D_{orig})=26.9.$**

- E_{lf} is the low-frequency energy corresponding to the output of the 3rd (or the 4th) canal of the filter bank in dB, in the case of 24 outputs in the filter bank (with a Mel scale) as described by Bimbot [4]. The choice of the low frequency energy is justified by the results of Bonastre [5]. The filter 3 (or 4) was proposed after several experiments. **$WR(E_{lf})=13.9$**.

5.3 Algorithm

The method of speaker identification proposed is based on a MLVQ1 algorithm (Modified Learning Vector Quantization). This algorithm is nicely adapted for the speaker recognition tasks.

The learning time is about 30 seconds and the distance used is the ODHEF Distance (see section 4). The statistical computation of the adaptation matrix for this distance gives the following results (the adaptation matrix contains the coefficients of adaptation):

$Adapt(1)=Adapt(Fo_{avr})=0.24.$

$Adapt(2)=Adapt(D_{orig})=0.21.$

$Adapt(3)=Adapt(E_{lf})=0.54.$

Note that

$Adapt(Fo_{avr})+Adapt(D_{orig})+Adapt(E_{lf}) =1$. (3)

5.3.1 The LVQ1

Assume that a number of 'codebook vectors' mi (free parameter vectors) are placed into the input space to approximate various domains of the input vector x by their quantized values. Usually several codebook vectors are assigned to each class of x values, and x is then decided to belong to the same class to which the nearest mi belongs. Let

$$c = \arg\min\{\text{length}(x - mi)\} \quad (4)$$

define the nearest mi to x, denoted by mc.

Values for the mi that approximately minimize the misclassification errors in the above nearest-neighbor classification can be found as asymptotic values in the following learning process. Let x(t) be a sample of input and let the mi(t) represent sequences of the mi in the discrete-time domain. Starting with properly defined initial values, the following equations [6] define the basic LVQ1 process:

$$mc(t + 1) = mc(t) + alpha(t)[x(t) - mc(t)] \quad (5)$$
if x and mc belong to the same class,

$$mc(t + 1) = mc(t) - alpha(t)[x(t) - mc(t)] \quad (6)$$
if x and mc belong to different classes,

$$mi(t + 1) = mi(t) \text{ for i not equal to c.} \quad (7)$$

Here $0 < alpha(t) < 1$, and alpha(t) may be constant or decrease monotonically with time. In the above basic LVQ1 it is recommended that alpha should initially be smaller than 0.1.

5.3.2. Modified LVQ1 (MLVQ1)

The classification decision in this algorithm is identical with that of the LVQ1.

In learning, however, only the false neighbor mc_{false} (wrong class) is updated and moved away from x.

The basic modified LVQ1 process is:

$$mc(t + 1) = mc(t) \quad (8)$$
if x and mc belong to the same class,

$$mc(t + 1) = mc(t) - alpha(t)[x(t) - mc(t)] \quad (9)$$
if x and mc belong to different classes,

$$mi(t + 1) = mi(t) \text{ for i not equal to c.} \quad (10)$$

where the function $\alpha(t)$ ought to be a monotically decreasing function of time: $0< \alpha(t) <1$ [7].

6 RESULTS

The results of this work are summarized in table 2.

In the closed set and for a text-dependent speaker identification, we obtain a recognition score of 98% (percentage of good identification) by using the classic LVQ1 and we obtain a recognition score of 100% (percentage of good identification) by using the MLVQ1.

If the parameter Fo_{avr} is not used, this score becomes only 66.7%: this result proves that the mean of the pitch is important in speaker identification.

Furthermore, when impostors are introduced into the speakers set, the identification score decreases to 93.3% for both the LVQ1 and the MLVQ1, and then the recognition in the open set is less accurate.

If the parameter Fo_{avr} is not used, this score becomes only 62.2%: this result shows that the mean of the pitch is important in speaker verification too (rejection of impostors).

Finally, we note that the MLVQ1 represents a good approach in speaker recognition (compared with the classic LVQ1 algorithm).

Table 2 Scores of speaker recognition.

Algorithm	Good identification (%)	
	Closed set	Open set
LVQ1 with Fo_{avr}	98	93.3
MLVQ1 with Fo_{avr}	100	93.3
MLVQ1 without Fo_{avr}	66.7	62.2
Improvement brought by Fo_{avr} (using the MLVQ1)	+33.3	+31.1

7 CONCLUSION

In this work, we test the efficiency of three prosodic features: the mean of the pitch, the original duration and the low-frequency energy.

In order to get a valid association between these three heterogeneous features a new metric (ODHEF) is proposed and we used a new modified LVQ1 (MLVQ1) algorithm to recognize the different speakers.

The results show that this method provides a good speaker identification performance and that the prosodic association is very efficient in the closed set: the percentage of good identification is 100%. But the recognition accuracy decreases if some impostors are introduced into the speakers set, because these impostors cause some confusions.

We also note that the mean of the pitch is very important in speaker recognition: its importance was predictable because the Wolf ratio for this parameter was extremely high.

Finally, the results obtained with MLVQ1, in this particular speech database, show that this method represents a good approach in speaker recognition.

REFERENCES

[1] B.S. ATAL 1972, Automatic Speaker Recognition Based on Pitch Contours. J.A.S.A. Vol. 52, pp 1687-1697.

[2] B.S. ATAL 1974, Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J.A.S.A. Vol. 55, No 6, pp 1304-1312.

[3] Y. BENNANI et al. 1995, Neural Networks for Discrimination and Modelization of Speaker. Speech Communication, Vol. 17, Nos. 1-2, August 1995.

[4] F. BIMBOT et al. 1995, Second-Order Statistical Measures for Text-Independent Speaker Identification. Speech Communication, Vol. 17, Nos. 1-2, August 1995.

[5] F. BONASTRE et al. 1997, Traitement Indépendant de Sous-bandes Fréquentielles par des méthodes Statistiques du Second Ordre pour la Reconnaissance du Locuteur. Proceedings of the 4th Congress on Acoustics, pp 357-360, Marseille, France 14-18 April 1997.

[6] T. KOHONEN 1992, The Learning Vector Quantization Program Package Version 2.1 (October 9, 1992). Helsinki University of Technology.

[7] K. KONDO et al. 1994, Speaker-Independent Spoken Digits Recognition Using LVQ. Proceeding of the IEEE Workshop, Neural Network for Signal Processing IV, 1994, pp 4448-4451.

[8] D. A. REYNOLDS 1994, Experimental Evaluation of Features for Robut Speaker Identification. IEEE Trans. on Speech and Signal Processing, Vol. 2, No 4, pp 639-643, 1994.

[9] H. SAYOUD et al. 1994, PDA à Ambiguité Modifiée. Robustesse sur la Parole Corrompue, Proc. of the International Conference on Signal and Systems, ICSS'94, n°2, IV-16 à IV- 20, Alger, Sept. 1994.

[10] H. SAYOUD et al. 1995, Détecteurs de Pitch à Convergence de Fréquence. Mediteranean conference on electronics and automatic control MCEA 95, Grenoble - France 13-15 Sept. 1995.

[11] H. SAYOUD et al. 1997, Un Nouvel Analyseur d'Erreur pour PDA: L'Erreur Spectrale 3D. Proceedings of the 4th Congress on Acoustics, pp. 377-380, Marseille, France 14-18 avril 1997.

[12] H. SAYOUD et al. 1997, Internal Thesis Supervised by SAYOUD, Etude Statistique de la Variabilité Inter et Intra-Locuteur des Paramètres Prosodiques et Acoustiques de la Parole, Thèse d'ingéniorat, Institut d'électronique, USTHB, Alger.

[13] H. SAYOUD et al. 1998, Error Correction Algorithms for PDAs, Proposed in CESA'98 Tunisia, April 1-4 1998.

[14] J. WOLF 1972, Efficient Acoustic Parameters for Speaker Recognition, J.A.S.A. Vol 51, pp 2045-2056. 1972.