General Machine Learning Classifiers and Data Fusion Schemes for Efficient Speaker Recognition

Siwar Zribi Boujelbene1, Dorra Ben Ayed Mezghani[2] and Noureddine Ellouze [3]

[1]Faculty of Humanities and Social Sciences of Tunis – Tunisia,
[2]High Institute of Computer Science of Tunis– Tunisia,
[3] National School of Engineer of Tunis– Tunisia,
BP – 37 Campus Universitaire 1002 Tunis – Tunisia
siwarboujelbene@yahoo.fr, Dorra.mezghani@isi.rnu.tn, N.Ellouze@enit.rnu.tn

***Abstract*:** Data fusion methods can take advantage of the concepts of diversity and redundancy to improve system performance. Diversity can be used to improve system performance through the incorporation of different information. Similarly, redundancy can achieve the same goals through the re-use of data. These concepts have been thoroughly applied on pattern recognition problems. The basic idea is that if several classifiers can be constructed, whose errors are mutually uncorrelated, then performance advantages can be obtained through the propel classifiers fusion.

The contribution of this paper is to study the fusion of several machine learning classifiers and to analyze data fusion schemes for text independent speaker identification. Feature spaces are defined by combining the Mel-scale Filterbank Cepstrum Coefficients (MFCC) and delta coefficient. Each feature is modelled using the gaussian mixture model (GMM) that constructs a speakers' models dictionary used later as inputs for classification. Then, four popular supervised machine learning classifiers are considered, namely the multilayer perceptrons classifier (MLP), the support vector machines classifier (SVM), the decision tree (DT) classifier and the radial basis function networks classifier (RBF). The scores (outputs) of classifiers are considered according to different scenario. Results showed that the best performance had been achieved by fusing the SVM, the MLP and the DT classifiers that reported a speaker identification rate equal to 94.15 %.

***Keywords*:** SVM, MLP, DT, RBF, data fusion, speaker recognition**.**

1. Introduction

Speaker recognition refers to the concept of recognizing a speaker by his/her voice. There are two main tasks within speaker recognition task: speaker verification and speaker identification [23].

The objective of speaker verification is to verify the claimed identity of that speaker based on the voice samples of that speaker alone. Speaker identification deals with a situation where the person has to be identified as being one among a set of speakers by using his/her voice samples. The speaker identification problem may be subdivided into closed set and open-set [4]. If the target speaker is assumed to be one of the registered speakers, the recognition task is a closed-set problem. If there is a possibility that the target speaker is none of the registered speakers, the task is called an open-set problem. In general, the open-set problem is much more challenging. In the closed-set task, the system makes a forced decision simply by choosing the best matching speaker from the speaker database. However, in the case of open-set identification, the system must have a predefined tolerance level so that the similarity degree between the unknown speaker and the best matching speaker is within this tolerance. Another distinguishing aspect of speaker recognition systems is that they can either be text-dependent or text-independent depending on the application. In the text-dependent case, the input sentence or phrase is fixed for each speaker, whereas in the text-independent case, there is no restriction on the sentence or phrase to be spoken.

As any speech recognition system, speaker recognition system consists on two stages; namely, feature extraction and classification (see figure 1).

Feature extraction consists on obtaining the characteristic patterns of the signal of a speaker. It can be considered as a data reduction process that attempts to capture the essential characteristics of the speaker with a small data rate. The feature extractor converts the digital speech signal into a sequence of numerical descriptors, called feature vectors. Features provide a more stable, robust, and compact representation than the raw input signal. Classifier uses these features as inputs.

State of the art speaker recognition systems are based on a cepstral feature extraction follow by a GMM [24] or a hybrid GMM/SVM classifier [15], [17], [13]. Nowadays, in classification stage, a new approach consists in fusing different systems is increasingly used. This technique can be divided in tree main categories: systems based on feature's diversity [8], [12], systems based on a classifier's diversity [31] and systems based on data fusion diversity [10]. Indeed, searchers are looking for the best set of feature, the best set of classifier and/or the best set of data fusion schemes.

In this paper, our study deals with the two last categories. It consists, and after extracting the feature vectors, on using four popular supervised machine learning classifiers for text independent closed-set speaker identification. This includes the MLP, the SVM, the DT and the RBF classifiers. These classifiers are evaluated according to different scenarios. In addition, several data fusion schemes are considered namely the majority voting, the mean rule and the product rule.

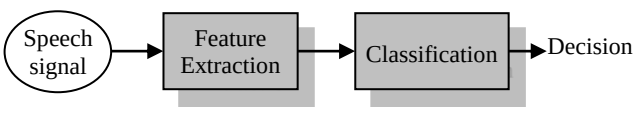

Figure 1. General diagram of speaker recognition system

This paper is organized as follows. Section 2 provides a description of the feature extraction and presents the features that we consider in this study. Afterward, in section 3, we introduce the machine learning theory for speaker recognition. Then, in section 4, we present briefly the MLP, the SVM, the DT and the RBF classifiers. This is followed in section 5 by a description of methods adopted for fusing the scores of different classifiers. Finally the experiments we made and the obtained results are drawn in section 6.

2. Feature Extraction

Feature extraction is a fundamental speech recognition stage. For speaker recognition, we are interested in the features that correlate with the physiological and behavioural characteristics of the speaker. These characteristics exist both in the supra-segmental features (voice source characteristics) of speech and in the spectral envelope (vocal tract characteristics). Although, it's impossible to separate these kinds of characteristics, signal measurements such as short term and long term spectra, and overall energy are easy to extract. These measurements provide the means for effectively discriminating among speakers [19].

2.1 Critera for feature selection

Features are numerical measurement used in computations that try to discriminate between classes. The selection of features depends largely on the application. For speaker recognition, Rose suggests that the optimal features should have the following properties [37]: easy to measure, high inter-speaker variation, low intra-speaker variation, robust against disguise and mimicry, robust against distortion and noise and maximally independent of the other features.
It is unlikely that a single feature would fulfil all these requirements. Fortunately, due to the complexity of speech signals, a vast number of complementary features can be combined to improve accuracy. In literature, a large number of features have been proposed for speaker recognition.

2.2 Spectral features

Feature extraction is usually computed by temporal methods like the Linear Predictive Coding (LPC) or frequencial methods like the Mel Frequency Cepstral coefficient (MFCC) or both methods like Perceptual Linear Coding (PLP). A nice property of spectral methods is that logarithmic scales (either amplitude or frequency), which mimic the functional properties of human ear, improve recognition rates [2]. In [27] Vergin et al. suggest that MFCC has been widely accepted as a features inputs for a typical speaker recognition system because it is less vulnerable to noise perturbation, gives little session variability and is easy to extract than others.
In a speech signal, the various type of information can change rapidly through time. For this reason, the signal $s[K]$ is divided into frames as in (1):

$$f[n] = \{s[nW + k] : k = 0,...W - 1\}, \quad (1)$$

each consisting of W samples. W must be large enough to include sufficient information, but it must also be small enough to ensure that the assumption of stationarity is reasonable. Frames normally overlap with their starting points following each other by L samples (L<W), because the signal does in fact change during the length of one frame..
For MFCC feature extraction, each feature vector is extracted from a frame. The frame is passed through a Humming filter and converted to the frequency domain using the discrete Fourier transform (DFT) [35]. Mel-scale frequency is related to linear frequency by the followed formula:

$$Mel(f) = 1127 \ln\left(1 + \frac{f}{700}\right) \quad (2)$$

The frequency range in Mel-scale is divided into a number of equal-sized bands. In linear frequency, triangle filters are positioned so that the width of each filter is equal to two bands in the Mel scale. Two successive filters also overlap each other by one of these Mel-scale bands. The value for energy in each band after filtering is called a Mel filter bank coefficient.
Cepstral analysis involves working with the spectrum of the spectrum. More specifically, the inverse Fourier transform is applied to the log-spectrum of the signal. Mel filter bank coefficients mi comes directly from the signal spectrum. They can be transformed into Mel-frequency cepstral coefficients (MFCCs) c_i by using the discrete cosine transform:

$$c_i = \sum_{j=1}^{B} m_j \cos\left(\frac{\pi i}{B}(j - 0.5)\right) \quad (3)$$

B is the number of Mel filter bank coefficients. The resulting MFCCs for each frame are grouped into a D-dimensional feature vector x.

2.3 Dynamic feature

While speaking, the articulators make gradual movements from a configuration to another one, and these movements are reflected in the spectrum. The rate of these spectral depends on the speaking style and speech context. Some of these dynamic spectral coefficients are clearly indicators of the speaker itself. According to Soong et al., delta features are the most widely used method for estimating feature dynamics [26]. Two kinds of delta features can be employed: first derivatives (delta) and second derivatives (delta-delta). These coefficients give an estimate of the time derivative of the features they are applied to.
The delta features coefficients are obtained by simply calculating the difference between two successive feature vectors. The resultant vector is appended to the second feature vector, making the procedure causal (i.e. only history is taken into account). The new feature vector then has double the dimension of the original.

3. Machine Learning

Machine learning is one of the hottest research areas. It has been widely adopted in real-world applications, including speech recognition, handwritten character recognition, image classification and bioinformatics.

3.1 Learning paradigms

In the literature, there are three main categories of machine learning classifiers, namely, those that are trained with supervised training classifiers, those that are trained with unsupervised training classifiers and those that are trained with hybrid training classifiers.

In supervised classifiers category, the data provided to the model training classifiers contains class information via a label [16]. However, unsupervised classifiers category does not require any information regarding class membership for the training data. Whereas, in hybrid classifiers category, training classifier combines both supervised and unsupervised learning. Part of the solutions (network weights, architecture, or computer programs) is determined through supervised training classifiers, while the others are obtained through unsupervised training classifiers.

In term of speaker recognition, unsupervised training classifier uses only data from a specific speaker to create a model. Whereas supervised training classifier assigns different labels to different speakers in a population. Examples of supervised training classifiers include multilayer perceptrons, support vector machines, decision trees and radial basis function networks. Supervised training classifiers have an advantage over unsupervised training classifiers in that they can better capture the differences between a particular speaker and other speakers in the population [10]. However, the amount of training data and hence, the computational effort in deriving a speaker model can be more than that for unsupervised training classifiers.

3.2 Classifiers structure for speaker recognition

The classifier consists of the speaker modelling, the pattern matching and the decision logic. Its operation constitutes two important steps (train and test). In the training step, feature vectors are used to obtain the M speaker models. For each speaker, a different model is obtained from his/her speech. In the testing step, feature vectors are first computed from an unknown speaker. For speaker identification, the feature vectors are compared with each of the M speaker models presented by the speakers' model dictionary in order to have the scores file: Score(1) to Score(M). These scores are used to bring up a decision. In a closed set scenario, the best score, Score(i), identifies the unknown speaker as speaker i. This means that speaker model i most likely generated the feature vectors. In an open set scenario, Score(i) is further compared against a threshold to decide if there is an adequate match between the unknown speaker and the best model i. If the match is deemed to be adequate, the speaker is identified. Otherwise, it is decided that no speaker model represents the unknown speaker. For speaker verification, the unknown speaker claims a certain identity j. Only Score(j) is calculated and compared against a threshold to verify or reject the claimed identity.

4. Supervised Machine Learning Classifiers for Speaker Recognition

Currently, several supervised training classifiers have been investigated for speaker recognition. These include multilayer perceptrons [22], [14], support vector machines [25], [28], [15], [29], [21], [32], decision trees [11], [30] and radial basis function networks [7]. Such classifiers are able to generate a model that can distinguish one speaker among M classes of speakers. In fact, during training, the supervision is affected to a label that is associated to each feature vector. This label determines the class membership of that vector (speaker to which it belongs). This partitioning of training data is illustrated in figure 2.

4.1 Multilayer perceptron

The multilayer perceptron (MLP) is a popular form of neural network that has been considered forvarious speech recognition [34]. Perceptrons use the basic architecture illustrated in figure 3.

The network functions through combining the various features vectors with some set of weights. This sum is then used as input for a single neuron's activation function. The output of the activation function is then taken to be the output of the network. Perceptrons with multiple outputs are composed of several independent perceptron networks each determining the value of a single output. The weights for MLPs are trained with the backpropagation algorithm such that they can associate a high output response with particular input patterns [38].

For speaker recognition, test vectors, from training data, should have a "one" response for that speaker's MLP for a specific speaker, whereas from different speakers, test vectors should have a "zero" response [14]. For speaker identification, all test vectors are applied to each MLP and the outputs of each vector are accumulated. The speaker is identified as a corresponding to the MLP with the maximum accumulated output. For speaker verification and in order to be verified, all test vectors are applied to the model of the speaker. The output is accumulated and then normalized. If the normalized output exceeds a threshold, the speaker is verified, else rejected.

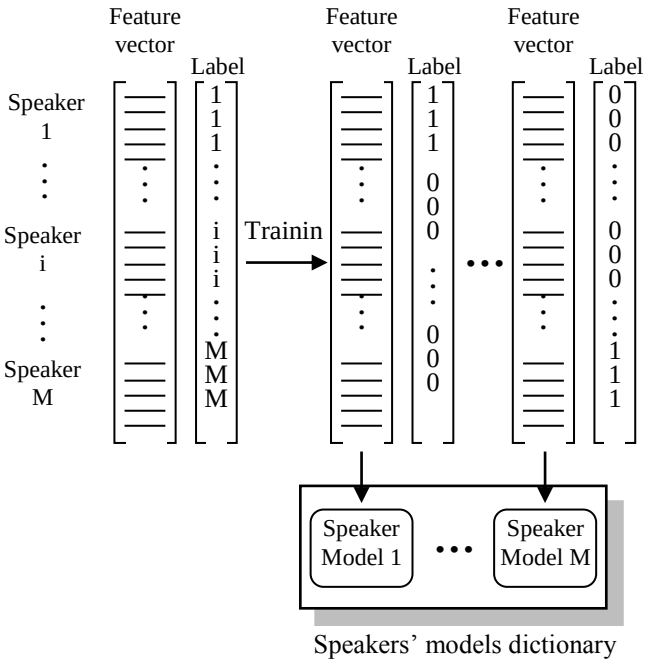

Figure 2. Supervised training data partitioning

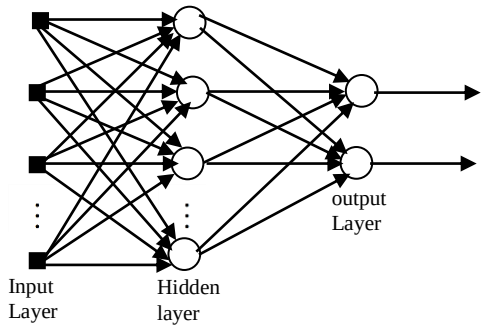

Figure 3. Multi layer perceptron architecture with one hidden layer

4.2 Decision tree

A decision tree (DT) describes a collection of rules, organized in a hierarchical fashion, that implement a decision structure. It consists of leaves and nodes. A leaf records an answer (often called a class) and a node specifies some test conditions to be carried out on a single feature value of an instance, with one branch and sub-tree for each possible result of the test [36]. For a given feature vector, a decision is made by starting from the root of a tree and moving through the tree determined by the outcome of a condition test at each node until a leaf is encountered [36]. The process of building a decision tree is a recursive partitioning of a training set.

For speaker recognition, the feature vectors are obtained from the training data for all speakers. Then, the data is labeled and a binary decision tree is trained for each speaker. The leaves of the binary decision tree identify the class label as follow: a one corresponds to the speaker and a zero corresponds to "not the speaker". For speaker identification, all feature vectors are applied to each decision tree for the test utterance. The labels are scored and the speaker having the maximum accumulated score is selected.

4.3 Support Vector Machines (SVM)

Support vector machines are, originally, introduced by Vapnik [39]. In a support vector machine, input vectors are mapped into a very high-dimension feature space through a non-linear mapping. Then a linear classification decision surface is constructed in the high-dimension feature space. This linear decision surface can take a non-linear form when it is mapped back into the original feature space.

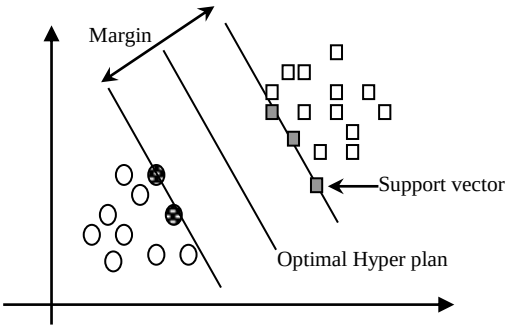

Figure 4. For SVM, a good separation is achieved by the hyperplane that has the largest distance to the neighboring data of two classes

Special properties of the decision surface ensure good generalisation ability of this learning machine [1]. While learning from data, SVM performs structural risk minimization (SRM) unlike the classical adaptation methods that minimize training error in a specific norm and maximize the geometric margin. Hence they are also known as maximum margin classifiers. The following figure presents the maximum-margin hyperplane and margins for a SVM trained with samples from two classes (linear separable case). Samples on the margin are called the support vectors. A critical aspect of using SVMs successfully is the design of the inner product, the kernel, induced by the high dimensional mapping.

For speaker recognition, the first approach in using SVM classifiers was implemented by Schmidt in [25]. Another approach became recently more popular, consists of using a combination of GMMs and SVMs. In [17], [18], [21], [29], [31], [13] several types of combination were proposed.

4.4 Radial basis functions

A Radial basis function (RBF) networks are embedded into a two layer feed forward neural network (see figure 5) [9]. Such a network is characterized by a set of inputs, a set of outputs and in between the inputs and the outputs there is a layer of processing units called hidden units. Each of them implements a radial basis function. The output units implement a weighted sum of hidden unit outputs. According to Park and al. [9], the RBF network classifier consists on: first specifying the hidden unit activation function, the number of processing unit, a criteria for modelling a given task and a training algorithm for finding the parameters of the network. Second and after having a set of input-output training data, RBF networks consists on clustering the training data into M clusters. The centroids of the set of clusters are used within the kernel functions, which are typically Gaussian kernels or sigmoids. The outputs of the kernel functions are used for training a single layer perceptron [7].

5. Data Fusion Schemes

The ultimate goal of designing pattern recognition systems is to achieve the best possible classification performance for one task. This led, traditionally, to the development of different classification methods.

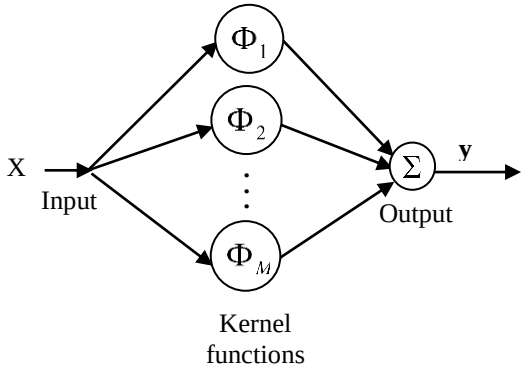

Figure 5. Radial basis function network

The results of an experimental assessment of the different

classifiers would then be the basis for choosing one of the classifiers as a final solution to the problem. In [20], Kittler et al., suggested that in such studies, it had been observed that although one of the classifiers would yield the best performance, the sets of patterns misclassified by the different classifiers would not necessarily overlap. This suggested that various classifier designs potentially offered complementary information about the patterns to be classified which could be harnessed to improve the performance of the selected classifier.

These observations motivated the relatively recent interest in classifiers fusion. Multiple expert fusion aims to make use of many different designs to improve the classification performance. Over the last few years a myriad of methods for fusing the outputs of multiple classifiers have been proposed [5]. Several fusion schemes have been devised and it has been experimentally demonstrated that some of them consistently outperform a single best classifier [20]. However, there is presently inadequate understanding: why some fusion schemes are better than others, in what circumstances and if the way of fusion influences on the results.

In [3] and [6], Kittler developed a common theoretical framework for fusing several classifiers which use different pattern representations. They presented a number of possible fusion schemes, namely product, sum, min, max, and majority vote rules, and compared their performance empirically using two different pattern recognition problems.

Within the context of speaker recognition, data fusion comprises the fusion of scores from different classifiers trained for a speaker. These classifiers may be trained by considering different features data or different classifiers. It is desired that the errors of one classifier are corrected by the others and vice versa. Indeed, if all classifiers are in agreement upon an error (all classifiers make the same mistake), so no combination will rectify the error. However, as long as there is some degree of uncorrelation among the errors, performance can be improved with the proper combination [10].

6. Experiments and Results

The work presented here belongs to the category of combining the benefits based on classifiers' diversity and data fusion schemes diversity for text independent closed-set speaker identification (see figure 6). Different scenarios are implemented as follows:

- In the first scenario, the MLP, the SVM, the DT and the RBF classifiers were firstly evaluated individually.
- In the second scenario, the outputs of the MLP, the SVM and the DT classifiers are fused in different ways and with several data fusion schemes.
- In the third scenario, and in order to study the effect of the RBF classifier on such fusions, the output of the RBF is fused with the outputs of the MLP, the SVM and the DT fusions in different ways and with several data fusion schemes.

6.1. Common experimental setup

As a common experimental setup, feature extraction is the fundamental step that deals with the discriminative features used by the next stage of classification. Feature space forms the input to the classifier that recognizes the pattern. Features are extracted from the DR1 dialect (New England region) of TIMIT corpus through MATLAB Toolbox. They are extracted from the speech signal every 10 ms using a 25 ms window. In [31], Zribi Boujelbene and al. evaluated different combined features by using the MFCC, delta, delta-delta and energy coefficients and suggested that the combination of MFCC and delta yield significantly better than all other combination of latest coefficient for speaker identification task. Thus, we used for this study the combined MFCC and delta features. So, each feature vector contains 24 coefficients characterized by the middle frame of every utterance followed by the label.

As the second common experimental setup, the modelling step is assured by the gaussian mixture model and estimated through the EM algorithm (Expectation Maximization) that maximize the Likelihood criterion (ML). The aim of ML estimation is to find the model parameters, which maximizes the likelihood of the GMM given the training data. So, each speaker is modelled and referred by his/her model. Accordingly, we have a dictionary of speakers' models that constructs the inputs for different machine learning classifiers.

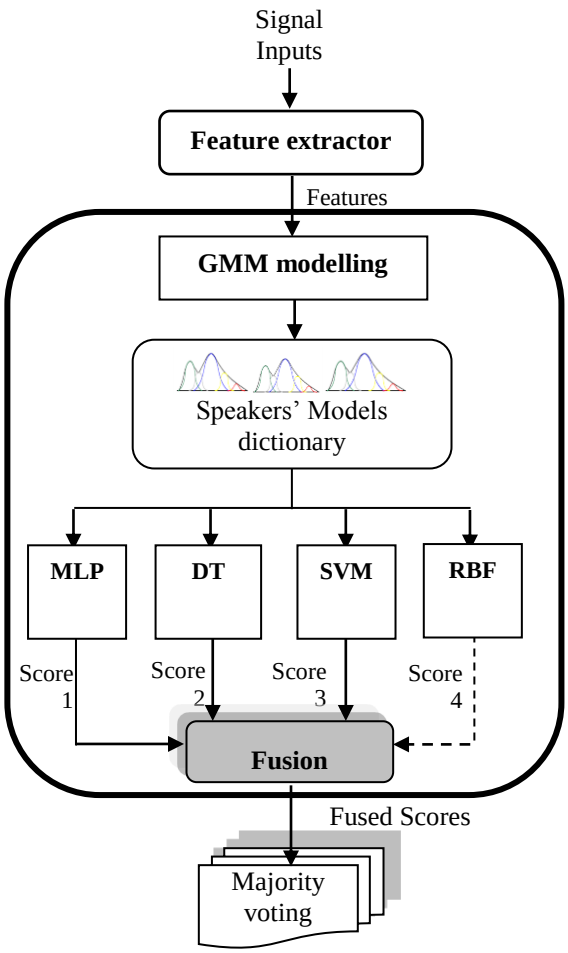

Figure 6. The main structure of different scenarios of fusions for efficient speaker recognition

6.2. Validation approach

One important aspect of recognition performance is the

generalization error. The generalization performance of a learning classifier relates to its prediction capability on the independent test data. In [12], [33] Zribi Boujelbene and al., found that the cross validation approach using ten-fold is the most powerful approach for estimating the error of generalization for speech recognition.

All experiments were carried out using a ten-fold cross validation approach. In fact, data were randomly partitioned into ten equally sized where 90% were used for training and the remaining 10% were used for testing. This technique was repeated for ten times, each time with a different test data. Training and testing data were normalized, as data normalization is required for some kernels due to restricted domain, and may be advantageous for unrestricted kernels.

6.3. Results and discussion

In the first scenario, table I shows the performance of the individual classifiers.

In the second scenario, we focus on studying the performance of fusing the MLP, the SVM and the DT classifiers in different ways of fusion and different data fusion schemes. In table II, we are limited to present six ways of fusions namely Fusion1, Fusion2, Fusion3, Fusion4, Fusion5 and Fusion6.

In the third scenario, we focus on evaluating the effect of fusing the output of the RBF classifier with the outputs of the SVM, the DT and the MLP fusions. In table III, we are limited to study only two ways of fusions namely Fusion7 and Fusion8 characterised by the fusion of the output of the RBF classifier with Fusion3 and Fusion4.

Table I shows that the identification rate accuracy is varied between 12.76% and 93.88%. As the experimental results suggest, the RBF classifier presents the highest accuracy when compared to all other classifiers. Results show also that the MLP classifier performs slightly worse than the SVM classifier. We note that the RBF classifier performs well in higher dimensional spaces and it has the advantage over other classifiers, such DT, MLP and SVM.

Table II shows that the identification rate accuracy is varied between 90.42% and 94.15%. It shows also that fusions of the MLP, the SVM and the DT classifiers yield significantly better than the individual classifiers. This can be improved by adjusting the classifier weights on a speaker by speaker basis as opposed to using the same weights across all speakers. We note that the fusions way beginning by the SVM classifier (Fusion3 and Fusion4) have the same behavior whatever the data fusion schemes. That can be explained by the fact that SVM perform well in higher dimensional spaces. Moreover, SVM has the advantage over MLP and DT, that their training always reaches a global minimum.

Table 1. Identification rate accuracy for individual classifiers

Individual classifiers	Accuracy (%)
MLP	**57.71**
SVM	**60.64**
DT	**12.76**
RBF	**93.88**

Table 2. Identification rate accuracy using the fusion of the MLP, SVM and DT classifiers

Fusions classifiers		Data fusion schemes		
Label	Way	Majority voting	Mean rule	Product rule
Fusion$_1$	**MLP** - SVM - DT	90.95	90.42	94.15
Fusion$_2$	**MLP** – DT - SVM	92.82	94.15	94.15
Fusion$_3$	**SVM** – MLP - DT	94.15	94.15	94.15
Fusion$_4$	**SVM** – DT - MLP	94.15	94.15	94.15
Fusion$_5$	**DT** – MLP - SVM	92.82	92.82	94.15
Fusion$_6$	**DT –** SVM - MLP	92.82	94.15	94.15

It can be seen also, from table II, that the product rule outperformed others fusion schemes, what is a theoretical assumption apparently stronger than others rule. In fact, the identification rate accuracy is equal to 94.15% whatever the way of fusion. We conclude that the product rule is the most resilient to estimate the identification rate, which almost certainly explains its superior performance.

Table III shows that the identification rate accuracy is varied between 93.62% and 94.15%. It shows also that the majority voting scheme outperformed others fusion schemes. It can be seen, from table III, that the fusions classifiers have the same behaviour whatever the way of fusion. By using the RBF classifier, we waited for improving the fusions of the SVM, the DT and the MLP classifiers, but the experimental studies show that this scenario couldn't influence on our results.

7. Conclusion

In this paper, we focus on combining the benefits based on classifiers' diversity and data fusion schemes diversity for text independent closed-set speaker identification. In fact, our motivation consists on evaluating the fusion of the multilayer perceptrons, the support vector machines, the decision trees and the radial basis function networks classifiers and analysing the use of the majority voting, the mean rule and the product rule fusion schemes.

Table 3. Identification rate accuracy using the fusion of the MLP, SVM, DT and RBF classifiers

Fusions classifiers		Data fusion schemes		
Label	Way	Majority voting	Mean rule	Product rule
Fusion$_7$	**RBF** - SVM - DT - MLP	94.15	93.88	93.62
Fusion$_8$	SVM - DT - MLP - **RBF**	94.15	93.88	93.62

For this, feature spaces are defined by combining the Mel-

scale Filterbank Cepstrum Coefficients (MFCC) and delta coefficient. Each feature is modelled using the gaussian mixture model that constructs a speakers' models dictionary presenting inputs for classification. Afterward, three scenarios were proposed. The classifiers are firstly used individually. Then a comparative study based on different ways of fusing the MLP, the SVM and the DT classifiers is provided. After that, the RBF classifier is added to the set of fused classifier in order to improve our results. A comparison study is also done between data fusion schemes according to different scenarios.

Experiments show that fusions of the MLP, the SVM and the DT classifiers yield significantly better than the individual classifier. They show also that the best performance is that obtained by Fusion3 and Fusion4 ways characterised by them stability for all data fusion schemes. They show also that the introduction of the RBF classifier in our fusions do not improve our results.

As a previous research works, we will try to combine multiple modalities such as face, voice, and fingerprint.

References

[1] Cortes and Vapnik V., "Support vector networks," Machine Learning, vol. 20, pp. 273–297, 1995.

[2] Marcos Faundez-Zanuy, Mohamed Chetouani, "Nonlinear predictive models: Overview and possibilities in speaker recognition," Progress in Nonlinear Speech Processing Springer Verlag, publisher, pp. 170-189, 2007.

[3] J. Kittler, A. Hojjatoleslami and T. Windeatt, "Strategies for combining classifiers employing shared and distinct pattern representations," Pattern Recognition Letters, Vol. 18, No. 11-13, pp. 1373-1377, November 1997.

[4] S. Furui. "Recent advances in speaker recognition" Pattern Recognition Letters, Vol.18, No 9,pp.859–872, 1997.

[5] Kittler, J., Hancock, E.R., "Combining Evidence in Probabilistic Relaxation," International Journal of Pattern Recognition and Artificial Intelligence, Vol. 3, pp. 29-51, 1989.

[6] Kittler, J., "Combining Classifiers: A Theoretical Framework," Pattern Analysis and Applications, pp. 18-27, 1998.

[7] Han Lian, Zheng Wang, Jianjun Wang, Liming Zhang, "Speaker Identification Using Reduced RBF Networks Array," Springer Berlin / Heidelberg Editor, vol. 3173, pp. 924-929, 2004.

[8] Monte-Moreno, E. and Chetouani, M. and Faundez-Zanuy, M. and Sole-Casals, J., "Maximum Likelihood Linear Programming Data Fusion for Speaker Recognition," Speech Communication . Vol. 51, No. 9, pp 820-830, 2009.

[9] Park, J., Sandberg, J. W., "Universal approximation using radial basis functions network", Neural Computation, Vol. 3, pp. 246-257, 1991.

[10] Ramachandran, R., Farrell, K., Ramachandran, R., and Mammone, R., "Speaker recognition - general classifier approaches and data fusion methods," Pattern Recognition, Vol. 35, pp. 2801–2821, 2002.

[11] Yamagishi Jun'ichi, Masuko Takashi, Tokuda Keiichi, Kobayashi Takao, "Speaker adaptation using context clustering decision tree for HMM-based speech synthesis," Journal IEIC Technical Report , Vol. 103, No. 264, pp. 31-36, 2003.

[12] S. Zribi Boujelbene, D. Ben Ayed Mezghani, and N. Ellouze, "Improved Feature data for Robust Speaker Identification using hybrid Gaussian Mixture Models - Sequential Minimal Optimization System," The International Review on Computers and Software (IRECOS), Vol. 4.3, ISSN: 1828-6003, pp. 344-350, May 2009.

[13] S. Zribi Boujelbene, D. Ben Ayed Mezghani, and N. Ellouze, "Robust Text Independent Speaker Identification Using Hybrid GMM-SVM System", Journal of Convergence Information Technology – JDCTA, Vol. 3.2, ISSN: 1975-9339, pp. 103-110, June 2009.

[14] Bennani Y., P. Gallinari, "A connectionist approach for speaker identification," IEEE International Conference on Acoustic, Speech and Signal Processing, Albuquerque, New Mexico, pp. 265–268, April 1990.

[15] Fine, S., Navrátil, J., Gopinath, R.A., A hybrid GMM/SVM approach to speaker identification," the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'01), vol.1, pp.417-420, 2001.

[16] C. Ji and S. Ma, "Performance and efficiency: Recent advances supervised in learning," Proc. IEEE, vol. 87, pp. 1519–1535, Sept. 1999.

[17] Kharroubi, J., Petrovska-Delacretaz, D., Chollet, G., "Combining GMMs with support vector machines for textindependent speaker verification," in: Eurospeech, pp. 1757–1760, 2001.

[18] Kharroubi, J., Petrovska-Delacretaz, D., Chollet, G., "Text-independent speaker verification using support vector machines," in: Proceedings of Speaker Odyssey, pp. 51–54, 2001.

[19] Tomi Kinnunen, "spectral features for automatic text-independent speaker recognition", thesis, University of Joensuu, Department of computer science, Finland, December 2003.

[20] Kittler, J., Hatef, M., Duin, R.P.W., Matas, J., "On Combining Classifiers," the IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 20, pp. 226-239, 1998.

[21] P. Moreno and P. Ho, "A new approach to speaker identification and verification using probabilistic distance kernels," European Conference Speech Communication and Technology (EUROSPEECH 2003), Geneva, Switzerland, pp. 2965-2968, 2003.

[22] J. Oglesby, J.S. Mason, "Optimization of neural models for speaker identification," IEEE International Conference on Acoustic, Speech and Signal Processing, Albuquerque, New Mexico, pp. 261–264, April 1990.

[23] Reynolds, D, " An overview of automatic speaker recognition technology," the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002), Orlando, Florida, USA, pp. 4072–4075, 2002.

[24] [D. Reynolds and R. Rose, "Robust text independent speaker identification using gaussian mixture models," IEEE Trans. Speech and Audio Process., Vol. 3, No. 1, pp. 72-83, Janvier 1995.

[25] M. Schmidt, "Identifying speaker with support vector networks," In Interface '96 Proceedings, Sydney, 1996.

[26] Soong, F., and Rosenberg, A., "On the use of instantaneous and transitional spectral information in speaker recognition," IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 36, No. 6, 871–879, 1988.
[27] R. Vergin, B. O' Shaughnessy and A. Farhat, "Generalized Mel frequency cepstral coefficients for large-vocabulary speaker idenpendent continuous-speech recognition," IEEE Trans. On ASSP, Vol. 7, No. 5, pp. 525-532, Sept. 1999.
[28] Wan, V., Renals, S., "SVMSVM: support vector machine speaker verification methodology," the International Conference on Acoustics Speech and Signal Processing, pp. 221–224, 2003.
[29] Wan, V., Campbell, W.M., "Support vector machines for verification and identification," Neural Networks for Signal Processing X, Proceedings of the 2000 IEEE Signal Processing Workshop, pp. 775–784, 2000.
[30] Wang, W., Lv, P., Zhao, Q.W., Yan, Y.H., "A Decision-Tree-Based Online Speaker Clustering Source Lecture Notes," In Computer Science; Vol. 4477. the 3rd Iberian conference on Pattern Recognition and Image Analysis, Girona, Spain, pp. 555 – 562, 2007.
[31] S. Zribi Boujelbene, D. Ben Ayed Mezghani and N. Ellouze, "Applications of Combining Classifiers for Text-Independent Speaker Identification," the 16th IEEE International Conference on Electronics, Circuits and Systems ICECS 09, pp. 723-726, Hammamet-Tunisia, December 2009.
[32] S. Zribi Boujelbene, D. Ben Ayed Mezghani, and N. Ellouze, "Support Vector Machines approaches and its application to speaker identification", IEEE International Conference on Digital Eco-Systems and Technologies DEST-09, Turkey, pp. 662-667, Jun 2009.
[33] Zribi Boujelbene, D. Ben Ayed Mezghani and N. Ellouze, "Vowel Phoneme Classification Using SMO Algorithm for Training Support Vector Machines", the IEEE International Conference on Information and Communication Technologies: from Theory to Applications ICTTA-08, Syria, pp. 1-5, 2008.
[34] D.P. Morgan, C.L. Scofield, "Neural Networks and Speech Processing," Kluwer Academic Publishers, Dordrecht, 1991.
[35] Poakis, J.G. and Manolakis, D.G, "Digital signal processing: principles, algorithms and applications". 3rd edn. Prentice-Hall, 1996.
[36] Quinlan J. R., "C.4.5: programs for machine learning," Morgan kaufmann, 1993.
[37] Rose, P. "Forensic Speaker Identification," Taylor & Francis, London, 2002..
[38] D.E. Rumelhart, J.L. McClelland, "Parallel Distributed Processing," MIT Cambridge Press, Cambridge, MA, 1986.
[39] V. Vapnik, "The nature of statistical learning theory," Springer, N.Y., 1995

Author Biographies

S. Zribi Boujelbene received a university diploma in computer science in 1999 from the High Institute of Management of Tunis (ISG-Tunisia), the MS degree in electrical engineering (signal processing) in 2004 from the National School of Engineer of Tunis (ENIT-Tunisia) and prepared a Ph. D. degree in ing (signal processing) from ENIT-

She is currently an associate professor in the computer science department at the Faculty of Humanities and Social Sciences of Tunis (FSHST-Tunisia). Her research interests include data mining approaches such as decision tree, support vector machines, neural networks, fuzzy logic, and pattern recognition such as speech recognition and speaker recognition.
E-mail: siwarboujelbene@yahoo.fr, zribi.siwar@planet.tn

D. Ben Ayed Mezghani received computer science engineering degree in 1995 from the National School of Computer Science (ENSI-Tunisia), the MS degree in electrical engineering (signal processing) in 1997 from the National School of Engineer of Tunis (ENIT-Tunisia), the Ph. D. degree in electrical engineering (signal processing) in 2003 from (ENIT-Tunisia).

She is currently an associate professor in the computer science department at the High Institute of Computer Science of Tunis (ISI-Tunisia). Her research interests include fuzzy logic, support vector machines, artificial intelligence, pattern recognition, speech recognition and speaker identification.
E-mail:Dorra.mezghani@isi.rnu.tn, DorraInsat@yahoo.fr

N. Ellouze received a Ph.D. degree in 1977 from l'Institut National Polytechnique at Paul Sabatier University (Toulouse-France), and Electronic Engineer Diploma from ENSEEIHT in 1968 at the same University.
In 1978, Dr. Ellouze joined the Department of Electrical Engineering at the National School of Engineer of Tunis (ENIT–Tunisia), as assistant professor in statistic, electronic, signal processing and computer architecture. In 1990, he became Professor in signal processing; digital signal processing and stochastic process. He has also served as director of electrical department at ENIT from 1978 to 1983. General Manager and President of the Research Institute on Informatics and Telecommunication IRSIT from 1987-1990, and President of the Institut in 1990-1994. He is now Director of Signal Processing Research Laboratory LSTS at ENIT, and is in charge of Control and Signal Processing Master degree at ENIT.
Pr Ellouze is IEEE fellow since 1987; he directed multiple Masters and Thesis and published over 200 scientific papers both in journals and proceedings. He is chief editor of the scientific journal Annales Maghrébines de l'Ingénieur. His research interest include neural networks and fuzzy classification, pattern recognition, signal processing and image processing applied in biomedical, multimedia, and man machine communication.
E-mail:N.Ellouze@enit.rnu.tn