# A Supervised Text-Independent Speaker Recognition Approach

Tudor Barbu

***Abstract*—**We provide a supervised speech-independent voice recognition technique in this paper. In the feature extraction stage we propose a mel-cepstral based approach. Our feature vector classification method uses a special nonlinear metric, derived from the Hausdorff distance for sets, and a minimum mean distance classifier.

***Keywords*—**Text-independent speaker recognition, mel cepstral analysis, speech feature vector, Hausdorff-based metric, supervised classification.

## I. Introduction

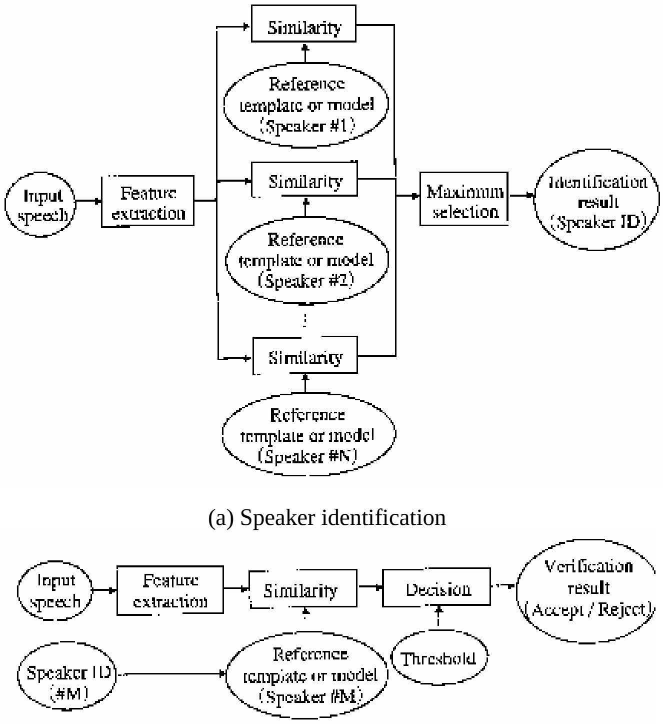
THIS paper approaches the speaker recognition field, an important biometric domain, providing a supervised text-independent recognition system. Speaker recognition, or voice recognition, represents the process of automatically recognizing who is speaking on the basis of individual features included in speech waves. It makes it possible to use the speaker's voice to verify their identity and control access to various services [1].

Voice recognition methods can be divided into *text-dependent* and *text-independent* techniques. The former approaches discriminate the users based on the same spoken utterance [2], while the latter do not rely on a specific speech [3]-[5].

The most successful speech-independent recognition methods are based on Vector Quantization (VQ) or Gaussian Mixture Model (GMM). The VQ-based methods are parametric approaches which use VQ codebooks consisting of a small number of representative feature vectors [4], while the GMM-based methods represent non-parametric techniques using *K* Gaussian distributions [5].

Also, voice recognition encompasses both identification and verification of speakers [1]. The basic structures of speaker identification and verification systems are represented in the next figure.

The speaker identification system proposed in this paper uses the melodic cepstral analysis in the feature extraction stage and a minimum mean distance classifier in the classification part. We utilize a threshold –based approach for speaker verification.


Similarity
Reference template or model (Speaker #1)
Input speech
Feature extraction
Similarity
Reference template or model (Speaker #2)
Similarity
Reference template or model (Speaker #N)
Maximum selection
Identification result (Speaker ID)
Input speech
Feature extraction
Similarity
Decision
Verification result (Accept / Reject)
Speaker ID (#M)
Reference template or model (Speaker #M)
Threshold

(a) Speaker identification

(b) Speaker verification

Fig. 1 Speaker identification and verification scheme

Our feature extraction approach is described in the next section. We propose DDFMCC–based matrices as vocal feature vectors. In the third section we provide a special nonlinear metric, obtained from the Hausdorff metric for sets, which is able to measure the distance between these feature vectors [2], [6].

An extended version of the minimum distance classifier is proposed in the fourth section of our paper. It uses mean distance values and the previously described nonlinear metric [2], [6], [7].

Some results of our numerical experiments are presented in the fifth section. The work ends with a conclusion section. The main contributions of this paper are the bidimensional voice feature vectors, the new Hausdorff-based metric and the proposed classifier.

Manuscript received April 29, 2007. Tudor Barbu has a PhD degree in Computer Engineering. He is with the Institute for Computer Sciences of the Romanian Academy, Iasi Branch (e-mail: tudbar@iit.tuiasi.ro).

## II. Speech Feature Extraction Approach

The vocal feature extraction represents the first part of the voice identification process. The Mel Frequency Cepstral Coefficients (MFCCs) are the dominant features used for speech and speaker recognition [2], [4], [6]. Thus, we propose a mel cepstral analysis for the vocal sound feature extraction operation.

Other voice recognition methods, such as those based on Vector Quantization, utilizes MFCC-based unidimensional feature vectors [4]. The speech feature extraction approach proposed by us creates bidimensional feature vectors.

Thus, a short-time analysis is performed on the sound signal to be featured [2], [6]. The speech signal is divided in overlapping frames having the length 256 and overlaps of 128 coefficients.

Then, each resulted segment is windowed, by multiplying it with a Hamming window of length 256. The spectrum of each windowed sequence is then computed, by applying FFT (*Fast Fourier Transform*) to it. The cepstrum of each windowed frame $s[n]$ is then computed as:

$$C[n] = \mathrm{IFFT}(\log|\mathrm{FFT}(s[n])|)) \tag{1}$$

where IFFT represents the Inverse FFT. Next we use the mel-scale, which translates regular frequencies to a scale that is more appropriate for speech. It is described as:

$$mel\,(f) = 2595 \cdot \log_{10}(1 + f / 700) \tag{2}$$

Thus, a sequence of mel-frequency cepstral coefficients (MFCCs) are obtained for each frame. Each such MFCC set represents a melodic cepstral acoustic vector. Next, a derivation process is performed on these MFCC acoustic vectors.

*Delta mel cepstral coefficients* (DMFCC) are computed as the first order derivatives of mel cepstral coefficients. Then, the *delta delta mel frequency cepstral coefficients* (DDMFCC) are obtained as the second order derivatives of MFCCs. These derivative processes are used because of the intra-speaker variability. Therefore, they tell us how fast a speaker's voice is changing.

Thus, a set of DDMFCC acoustic vectors result for the initial voice signal. Each of them is composed of 256 samples, but the speech information is codified mainly in the first 12 coefficients. Therefore, each acoustic vector is truncated at its first 12 samples and then it is positioned as a column of a matrix.

The resulted DDMFCC acoustic matrix constitutes a powerful speech discriminator which works successfully as a feature vector for the processed vocal sound. Let us note $V(S)$ the feature vector of the speech signal $S$.

Each feature vector has 12 rows and a number of columns depending on the length of the speech signal. Therefore, because of their different dimensions, these speech feature vectors cannot be compared using linear metrics, such as the most known Euclidean distance. For this reason, a special nonlinear metric is introduced by us in the next section.

## III. A Nonlinear Metric for Voice Feature Vectors

In this section we propose a special nonlinear metric which is able to compute the distance between different sized matrices having a single common dimension, like the acoustic matrices representing our speech feature vectors. It derives from the Hausdorff metric for sets [2], [6].

The *directed Hausdorff metric* is given by the following relation:

$$h(A,B) = \{\max_{a \in A} \min_{b \in B} \{dist\,(a,b)\} \tag{3}$$

where *dist* is any proper metric between the points of sets *A* and *B* (for example, the Euclidean distance). It is is termed also as *forward Hausdorff distance*, while $h(B,A)$ represents the *backward Hausdorff distance* for sets. Thus, we obtain the general definition for the Hausdorff distance for sets as follows:

$$H\,(A,B) = \max\{\, h(A,B), h(B,A)\} \tag{4}$$

From the relations (3) and (4), the next Hausdorf distance formula is obtained:

$$H\,(A,B) = \{\max_{a \in A} \inf_{b \in B}\, dist\,(a,b), \max_{b \in B} \inf_{a \in A}\, dist\,(a,b)\} \tag{5}$$

Let us consider now matrices having a single common dimension (the number of rows), instead of sets. Thus, $A = (a_{ij})_{n \times m}$ and $B = (b_{ij})_{n \times p}$, *dist* being the Euclidean metric. Let us introduce two more helping vectors, $y = (y_i)_{p \times 1}$ and $z = (z_i)_{m \times 1}$, then compute $\|y\|_p = \max_{0 \le i \le p} |y_i|$ and $\|z\|_p = \max_{0 \le i \le m} |z_i|$, respectively. With these notations the following metric results:

$$d\,(A,B) = \max\left\{ \sup_{\|y\|_p \le 1} \inf_{\|z\|_m \le 1} \|By - Az\|, \sup_{\|z\|_m \le 1} \inf_{\|y\|_p \le 1} \|By - Az\| \right\} \tag{6}$$

This restriction based metric represents the Hausdorff distance between the sets $B(y : \|y\|_p \le 1)$ and $A(z : \|z\|_m \le 1)$ in the metric space $R^n$, therefore it can be written using the following formula:

$$d\,(A,B) = H\,(B(y : \|y\|_p \le 1), A(z : \|z\|_m \le 1)) \tag{7}$$

Next, after eliminating the terms *y* and *z* from the above formula, we finally obtain the following Hausdorff-based distance:

$$d\,(A,B) = \max\left\{ \sup_{1 \le k \le p} \inf_{1 \le i \le m} \sup_{1 \le i \le n} |b_{ik} - a_{ij}|, \sup_{1 \le i \le m} \inf_{1 \le k \le p} \sup_{1 \le i \le n} |b_{ik} - a_{ij}| \right\} \tag{8}$$

The resulted nonlinear function *d* verifies main distance properties:

- Positivity: $d(A,B) \geq 0$
- Simetry: $d(A,B) = d(B,A)$
- Triangle inequality: $d(A,B) + d(B,C) \geq d(A,C)$.

While not representing a Hausdorff metric anymore, the Hausdorff-based distance *d* given by formula (8) constitutes a satisfactory discriminator between the vocal feature vectors, therefore it could be successfully used in the next classification process.

## IV. Speaker Classification and Verification

We propose a supervised classifier for our voice (speaker) recognition system, providing a minimum mean distance classification approach [2], [6]. The training set of our classifier contains a collection of spoken utterances, generated by the registered (advised) speakers.

Each vocal utterance from the training set constitutes a vocal prototype and represents the same speech. We consider a large spoken text which contains all the English language phonemes. Each registered speaker should provide this speech several times.

Therefore, the resulted training set receives the form $P = \{P_1,..., P_N\}$, where each $P_i = \{s_1^i,..., s_{n(i)}^i\}$ represents the set of signal prototypes corresponding to the $i^{th}$ speaker, *N* being the number of advised speakers. For each $s_j^i$, where $i = \overline{1,N}$, $j = \overline{1,n(i)}$, the previously described vocal feature extraction is then performed, the feature training set $\{\{V(s_1^1),..., V(s_{n(1)}^1)\},..., \{V(s_1^N),..., V(s_{n(N)}^N)\}\}$ thus being obtained.

We provide a minimum mean distance classification approach, representing an extended variant of the minimum distance classifier [7]. There are *N* classes, each of them corresponding to a different advised speaker. Our classification procedure inserts each input vocal signal in the class of the closest registered speaker, which is the speaker corresponding to the smallest mean distance between the feature vector of the input signal and the prototype vectors of the speaker.

Therefore, the mean distance between the input $S_i$ and the training subset $P_j$, related to the $j^{th}$ speaker, is computed as $\frac{\sum_{k=1}^{n(j)} d(V(S_i), V(s_k^j))}{n(j)}$. So, we identify the $p^{th}$ speaker as being the closest to $S_i$, where

$$p_i = \arg\min_j \frac{\sum_{k=1}^{n(j)} d(V(S_i), V(s_k^j))}{n(j)}, \forall i \in [1, n] \quad (9)$$

This classification result, the *N* classes of vocal utterances, represents also the result of the speaker identification process. The next stage of the recognition process, speaker verification, has to decide if an identified speaker represents a registered user of the system.

So, a verification operation should be performed within each previously obtained speaker class. Let these classes be $C_1,..., C_N$. We propose a threshold-based approach, setting a threshold value *T* and then compare the resulted minimum mean distance values with it. Therefore, the following condition has to be tested:

$$\forall i \in [1, N], \forall S \in C_i \mid \frac{\sum_{k=1}^{n(j)} d(V(S_i), V(s_k^j))}{n(j)} \leq T, \quad (10)$$

where the threshold *T* is chosen from the numerical experiments. If condition (10) becomes true for a voice sequence *S* and a class $C_i$, then the vocal utterance *S* is accepted by the recognition system as an advised vocal input generated by the $i^{th}$ registered speaker. Otherwise, *S* is rejected by our system, and labeled as being provided by an unregistered user.

## V. Numerical Experiments

We performed a lot of numerical experiments using this speech-independent voice recognition system and obtained satisfactory results. The high recognition rate, approximately 85%, which has been obtained, proves the effectiveness of our system.

Let us describe now a simple speaker recognition example using our recognition approach. We consider three registered speakers and a long sequence of words, containing all the English phonemes, to be spoken by each of them. The chosen sequence of words is: *“bake flat head fix gas sky jet lamp no low hot quick sir use about voice wash box yes zoo boot put toy out car saw ship catch the sing measure”*.

The training set contains four vocal utterances, each of them recorded at 22050 Hz and having that sequence of words as text. We got one recording for the first advised user, two recordings for the second user and one recording for the last one.

The prototype speech signals and the corresponding training feature vectors, computed as DDMFCC-based matrices and represented as RGB color images, are displayed in Fig. 1.

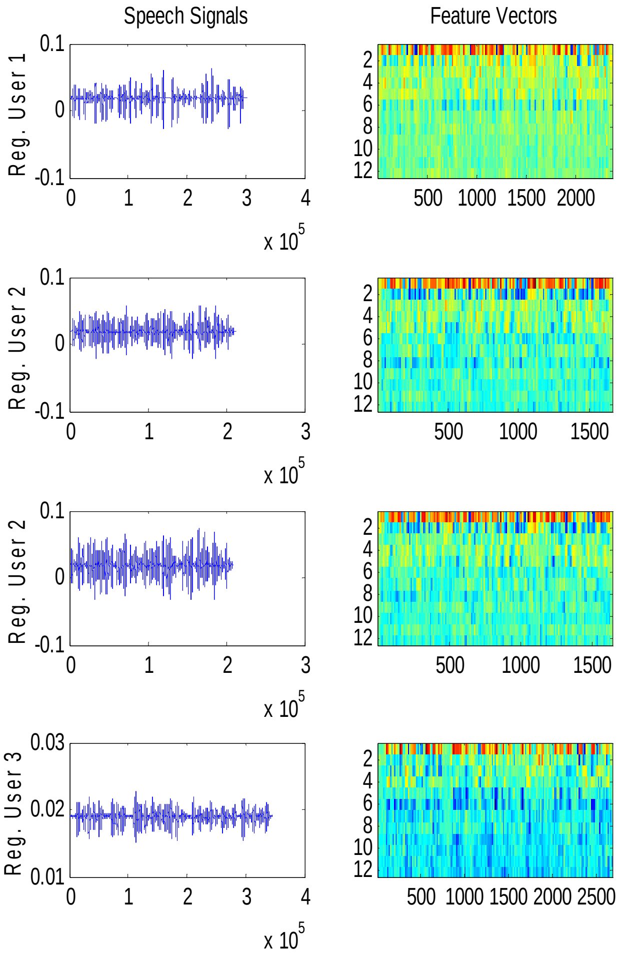


Fig. 1 The prototype speech signals and their feature vectors

Then, we consider a sequence of nine input speech utterances to be recognized, each of them having a different spoken text. The spoken texts are: *country*, *house*, *hello*, *apple*, *rabbit*, *recognition*, *tomorrow*, *window*, *car*. Their signals, $\{S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8, S_9\}$, are represented in the second figure.

Next, the feature vectors $V(S_i)$ are computed, using the technique described in the fourth section. They are displayed as color images in Fig. 3.

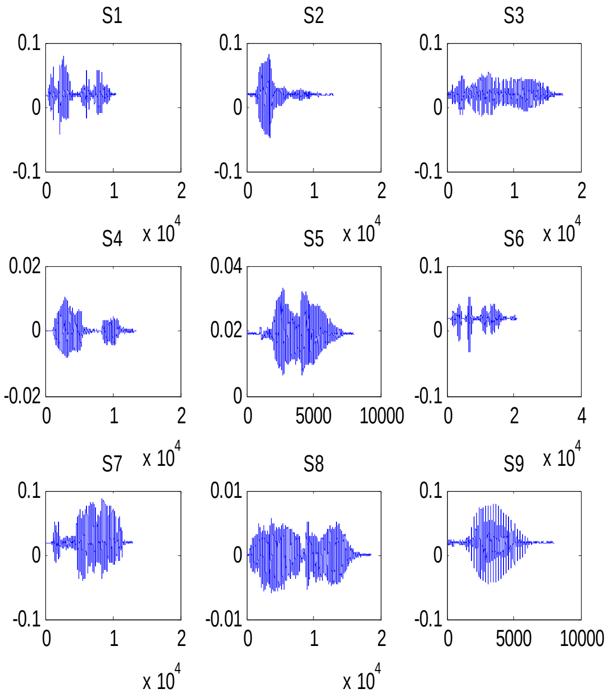


Fig. 2 The input speech signals

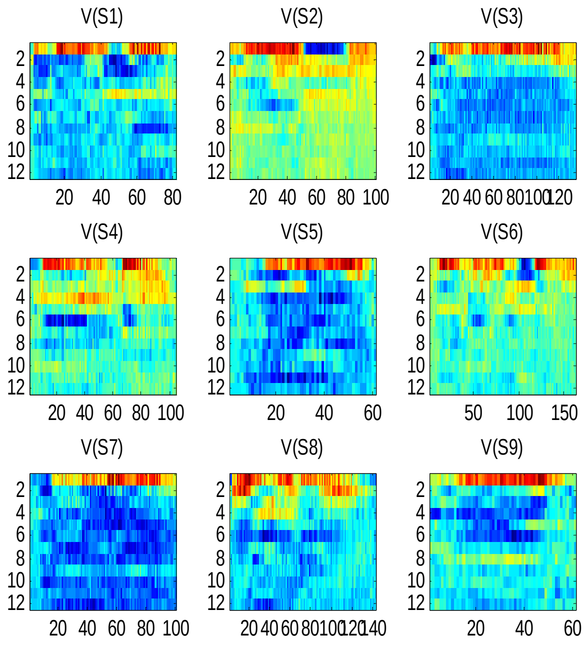


Fig. 3 The speech feature vectors

Next, the mean distance values between the input feature vectors and the training feature subsets are computed. The obtained values, calculated as $\frac{\sum_{k=1}^{n(j)} d(V(S_i), V(s_k^j))}{n(j)}$, are registered in the Table I.

TABLE I
MEAN DISTANCE VALUES

| | Speaker 1 | Speaker 2 | Speaker 3 |
|---|---|---|---|
| Input 1 | 5.6678 | 3.4676 | 6.2476 |
| Input 2 | 4.1606 | 7.2581 | 6.4327 |
| Input 3 | 5.9543 | 3.8976 | 4.3671 |
| Input 4 | 6.0946 | 4.7342 | 2.9857 |
| Input 5 | 10.6853 | 9.7366 | 8.6545 |
| Input 6 | 5.1522 | 5.7855 | 6.3879 |
| Input 7 | 7.5031 | 4.8871 | 10.8775 |
| Input 8 | 8.7624 | 7.9964 | 5.8976 |
| Input 9 | 3.2465 | 8.0245 | 4.9082 |

First, the identification procedure is performed on this dataset. Let us analyze these distance values, registered in the table above.

On the second row of the table, corresponding to $S_1$ (Input 1), the minimum distance value is 3.4676. It corresponds to the second registered speaker, therefore the input signal 1 must be associated to that speaker.

On the row corresponding to $S_2$, the minimum distance is 4.1606, that corresponds to the first advised speaker. On the row corresponding to $S_3$, the minimum value is 3.8976, that corresponds to the second speaker.

On the row of $S_4$, the minimum value is 2.9857, corresponding to the third speaker. On the row of $S_5$, the minimum value is 8.6545, corresponding to the third speaker, too.

On the row of $S_6$, the minimum distance is 5.1522, which corresponds to the first speaker. On the row of $S_7$, the minimum distance is 4.8871, that corresponds to the second speaker. On the row of $S_8$, the minimum distance is 5.8976, that corresponds to the third speaker. On the row of $S_9$, the minimum distance is 3.2465, that corresponds to the first speaker.

Therefore, we got the following identification result. The input signals 2, 6 and 9 belong to the first speaker, the input speeches 1, 3 and 7 are associated to the second registered speaker, and the input signals 4, 5 and 8 belong to the third speaker.

The second part of the recognition process is the speaker verification. Thus, from this verification operation, it results that the fifth input signal cannot be associated to any speaker class. The third advised speaker is closest to it, but the corresponding distance value is too large.

Using our experiments, we set the threshold $T$ =7.5. Then, the relation (10) is applied. Of course, we get $T$ < 8.6545, which means that the signal $S_5$ does not belong to Speaker 3, and has to be classified as being produced by an unregistered user. The other minimum distance values are less than $T$, so all the other speaker identifications are correct.

Therefore, the voice recognition result is given by the final speaker classification: Speaker 1 =>$\{S_2, S_6, S_9\}$, Speaker 2 =>$\{S_1, S_3, S_7\}$, Speaker 3 =>$\{S_4, S_8\}$ and finally, Unregistered Speaker =>$\{S_5\}$.

## VI. CONCLUSION

A text-independent voice recognition system has been proposed in this paper. This work has brought important contributions both in the feature extraction stage and also in the classification stage of the vocal pattern recognition process.

Most speaker recognition techniques, especially those based on Vector Quantization, use unidimensional feature vectors, therefore our representation of the vocal feature vectors as truncated acoustic matrices with DDMFCC coefficients constitutes an novel element in voice recognition. Also, using this mel-cepstral analysis, we can compute the pitch frequency values of a speech signal and develop other feature extraction methods (*pitch-based* techniques).

The minimum mean distance classifier, proposed by us as an extension of the most used supervised classification approach, represents another contribution of this paper. Obviously, the main contribution of this work is the proposed Hausdorff-based metric, used in the speech feature vector classification process.

Our speaker recognition system produces high recognition rates, thus being able to perform a proper identification of any human person. Also, it gives more speech freedom to its users, a person being able to provide any vocal utterance as an input to the system. This voice recognition system can be included as a subsystem in a more complex biometric system, which may contain additional features, like fingerprint or facial recognition.

## REFERENCES

[1] R. A. Cole, J. Mariani, H. Uszkoreit , A. Zaenen, V. Zue, *Survey of the State of the Art in Human Language Technology*. Cambridge University Press, 1997.

[2] T. Barbu, “Speech-dependent voice recognition system using a nonlinear metric”, *International Journal of Applied Mathematics*, Volume 18, No. 4, 2005, pp. 501-514.

[3] H. Gish, M. Schmidt, “Text-Independent Speaker Identification”, *IEEE Signal Processing Magazine*, IEEE,oct. 1994, pp. 1437-62.

[4] N. Bagge, C. Donica, “ELEC 301: Final Project Text Independent Speaker Recognition”, *ELEC 301 Signals and Systems Group Projects*, 2001.

[5] D. A. Reynolds, R. C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models”, *IEEE Trans. Speech Audio Processing*, vol. 3, no. 1, 1995, pp. 72-83.

[6] T. Barbu, “Discrete speech recognition using a Hausdorff-based metric”, In *Proceedings of the 1st Int. Conference of E-Business and Telecommunication Networks, ICETE 2004*, Setubal, Portugal, Vol. 3, Aug. 2004, pp.363-368.

[7] R. Duda, P. Hart, D. G. Stork, *Pattern Classification*, John Wiley & Sons, 2000.