F.S.A.I. Vol. 10, No. 3, pp. 63-78, 2004

Automatic Speaker Recognition Decision Tools

Mihaela Costin*, Tudor Barbu*, Anthony Grichnik**

*Institute for Theoretical Informatics of the Romanian Academy,
mccostin@iit.iit.tuiasi.ro, tudbar@iit.iit.tuiasi.ro
**Caterpillar Inc., Technology & Solutions Division
Peoria IL, USA, grichaj@cat.com

Abstract: Complex decision-making, as speaker recognition, often implies multiple tools weighted inference in order to augment the certainty coefficient on the person identity. Results tested by previous presented methods - using linear prediction coding (LPC), autoregression (AR) coefficients and Mel-frequency cepstral coefficients (MFCC) used in neural networks (NN) recognition, are compared with those obtained by technique of Hausdorff based distance measure (HDM) and vowel detection. High recognition coefficients had been obtained by our verification methods (ARNN - max. 89%, HDM - max. 95% and VD - 96%). Using these methods in parallel, in an aide-decision system with weighted certainty coefficients, we propose a schema of reinforced decisions on speaker identity. Further researches will be made on the opportunity of using HDM to conceive medical training systems dedicated to hearing-impaired patients.

Key words: Hausdorff based distance measure, vowel detection, autoregression coefficients, weighted inference, certainty coefficient, person identity.

1. Introduction

Automatic speaker recognition (ASR) contains different, particular aspects, to begin with noisy multi-speaker environment analysis to the high accuracy recognized identity of a claimed speaker from a pre-existent records database. Also termed as voice recognition, this area (including its sub-domains), is subject of different classifications.

Thus, text dependence: text-dependent, versus text-independent voice recognition, implies the use of a prescribed (against a free) text, during both training and testing.

In the recognition stage, speakers must pronounce words from the prescribed sets used in training. This supposes cooperative persons and discrete-word speech recognition. The system returns the name of the person most closely matching the registered pronunciations. As a remark, nowadays person validation systems are usually doubled, using, beside voice recognition, secondary pattern recognition system e. g. face, iris, or fingerprint recognition, to prevent against good imitation or tape use [1].

Text-independent systems require impressing volumes of training data ensuring that the entire vocal range is captured. It is useful for not cooperative subjects (in surveillance).

Recording environment: ideal versus noisy recording environment indicates the character of duly, laboratory conditions (the same devices for training and testing, high quality microphone, anechoïd room, and little or no background noise), versus spontaneous registrations. This domain concentrates much of the actual research.

Verification versus identification: speaker verification consists in determining whether the speaker is the person he claims to be (minimizing false rejection or approval error), while the identification process may be *closed set* or *open set* and establishes which is the most closed speaker matching the unknown.

Real-time versus off-line operating: determines the time of getting the decision in speaker recognition. Our studies concentrate on discrete-word speech recognition, in real, noisy conditions environment and real-time/off-line speaker verification.

2. Short overview

Voice recognition is complementary to speech recognition. Even if both techniques use similar methods of speech signal processing up to a point, speaker-independent recognition must resolutely ignore any idiosyncratic speech characteristics of the speaker and focus on those aspects of the speech signal richest in linguistic information. On the contrary, voice recognition must amplify those peculiar speech characteristics that individuate a person and suppress, if possible, linguistic characteristics, which have no bearing on the recognition of the individual speaker [2]. The border is unfortunately difficult to mark and the process depends on very typical aspects.

Recognition tasks are usually preceded by an acoustic analysis front-end, in order to extract significant parameters from the time signal [3], based on a model of the signal or of the production mechanism. Short-Time Fourier Transform (STFT) [4], Cepstrum, and miscellaneous related schemes [5], [6] were conceived starting strictly from the physical phenomena characterizing the speech waveform, based on the quasi-periodic model of the signal. Conversely, LPC technique was developed modelling the human speech production mechanism [7], [8]. But the focus in speech recognition is on perceived sound rather than on physical properties of the signal or of the production mechanism, so, these analysis schemes have been recently modified by also incorporating, perceptual-related phenomena [9]. STFT-derived auditory models, linear prediction on a warped frequency scale, perceptually based linear predictive analysis, are a few simple examples of how human auditory perceptual function is taken into account in designing signal representation algorithms [3], [10]. Thus, Mel-Frequency Cepstrum Coefficients (MFCC) [11], [12], that transforms the linear frequency domain into a logarithmic one, resembling to the human auditory sensation of tone height, are often used in ASR systems.

These schemes use a "short-time" analysis framework [13]. Dynamical changes of sound properties are tracked by short segments on the registered sound, called analysis frames, overlap one another. This framework is based on the assumption that, due to the mechanical characteristics of the generator, the properties of the signal change relatively slowly with time. Even if overlapped analysis windows are used, important fine dynamic characteristics of the signal are discarded. Just for that reason, but without solving completely the problem of correctly taking into account the dynamic properties of speech, "velocity"-type parameters (differences among parameters of successive frames) and "acceleration"-type parameters (differences of differences) [14] have been included in acoustic front end of almost all ASR systems. The use of temporal changes in speech spectral representation (ΔMFCC, ΔΔMFCC) resulted in a significant improvement in ASR systems.

Recently, in speech analysis and recognition, the introduction of auditory models [9] which explicitly consider non-linear phenomena occurring in the perception mechanism, has given encouraging results especially when speech is degraded by noise.

3. Neural networks text dependent speaker recognition

Our previous studies in speaker [15] and speech recognition [7], revealed the following aspects:

- distinct prominence in recognition process of each spectral frequency band stated as a secondary result in [16];
- significant importance of the signal phase in voice recognition [15];
- vowel recognition preponderance in speaker finding is a prior indices [15], [16].

A direct conclusion is the necessity of using weighted inferences by certainty coefficients in the decision process. Autoregressive (AR) coefficients [17] were used to compute feature vectors in order to teach, in parallel, a Multilayer Perceptron (MLP) compared to a Radial Basis Function (RBF) structure of neural networks (NN) [18].

The basic schema of the system includes an acquisition module of the speech signal for speakers that claim to be recognized. The AR coefficients of the signal, passed by more filtering channels constitute the feature vectors for the neural networks recognition. Note that the double reversion of the signal is essential for the signal phase preserving [15], [13].

A low pass (LP), more band-pass (BP) and a high-pass (HP) filter are used [19], which have limits detected by a special procedure, improved comparing to the previous papers, based on the steepest decent procedure on the signal spectral density [20].

Two methods were until now alternately used in [15] to divide the spectrum of the pre-processed speech frame on frequency bands: the first algorithm chosen fix Mel frequencies, the second none took the "inflection" points detection on the curve resulted by the continuous summing of the spectrum amplitude.

Those values, refined, become the borders for the frequency bands necessary for the further zero phase band pass filtering. AR model is applied then to the signal obtained on each band and the feature vectors constitute inputs for the neural networks as we observe in the general structure of AR recognition schema.

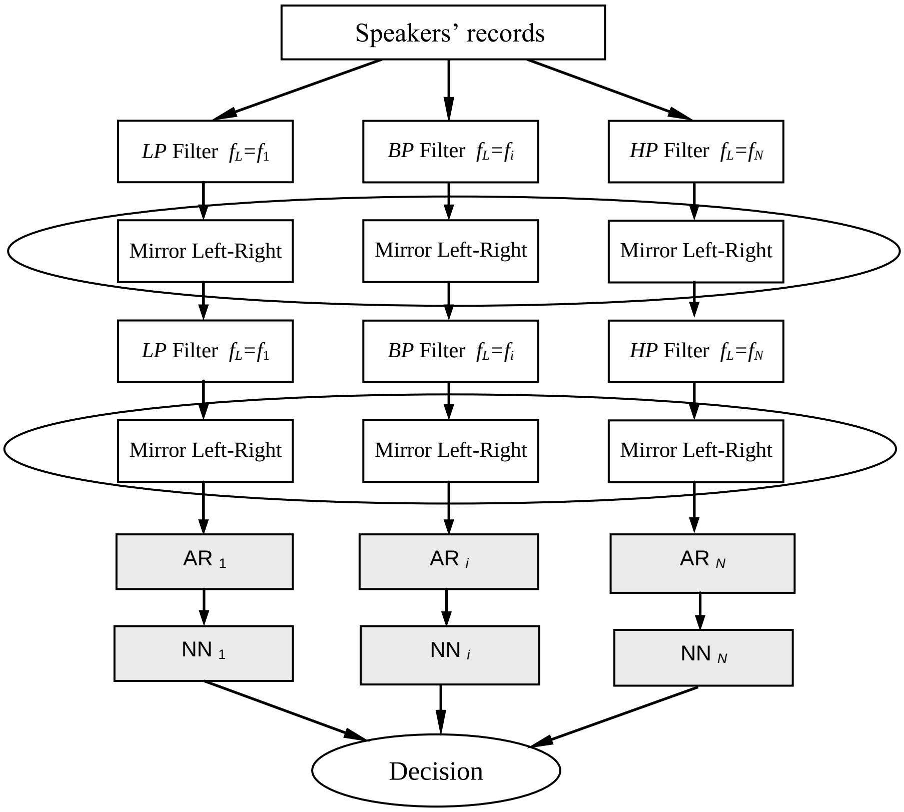

Fig. 1 - General structure of AR recognition

In the classification stage the band-signal AR coefficients are used to train the neural networks. On the trained MLP and RBF networks the medium percents of the word dependent speaker recognition finally obtained were 76% with the MLP network and 89% using the RBF structure [15].

Bands on frequencies > 2500 Hz are more important in ASR than lower bands [21] (related to voice characteristics, features on higher frequency bands are more selective compared to lower frequency bands).

Another result is remarkable: it is possible to obtain a good recognition coefficient on speaker verification, accurately enough, only by the vowels pronunciation [16] - easier than taking into account the consonant phonemes uttered, too.

4. Vowel preponderance in voice recognition process

Some interesting results in the speaker classification process are obtained by a phonetic rule-based test detection. A good classification is possible, by a criterion selection process. For example, in order to differentiate five vowels pronounced with a Romanian sonority, we worked with *N* (e.g. 15) spectral values presented to the input. To find out the conditions we have to determine the elimination of those characteristics that are less important, and the “threshold” values are further searched in the classification, by a C5 structure [22]. These conditions may be further implemented by a rule-based programming and tried out in the ASR technique.

```
IF C1: (B7 > 0.04391)
THEN
       IF C2: (B9 <= 0.05173)
       THEN vowel = ‘o’
       ELSE vowel = ‘a’
ELSE
       IF C3: (B11 > 0.00967)
       THEN vowel = ‘e’
       ELSE
                IF C4: (B14 <= 0.0232)
                THEN vowel = ‘u’
                ELSE vowel = ‘i’
```

A training set consisting of 100 registered vowel utterances, pronounced by each of a number of four speakers, was used. Applying it to the test set we obtained a surprising recognition coefficient of 94%. The speaker's age and gender influence the results of tests, (fundamental frequencies are depending on) [23].

Tabel 1

Vowel pronounced	a	e	i	o	u
Detected vowel(decision)					
vowel: a	20				
vowel: e		19		1	
vowel: i			19		1
vowel: o	1			18	1
vowel: u			2		18
Total error: 6%					

An interesting aspect is that the important band (B*i*) in the recognition process is also selected, giving the clear information for the band which has an importance to be stressed [16], and thus, more knowledge on uttered vowels permits to select the speaker.

The values were normalized in amplitude in order not to overpass a maximum 1. Another observation, which is more a verification of an already uttered hypothesis [15] is that the upper bands (more than 2500 Hz) have more importance in speaker recognition than in speech.

For some vowels, *e*, *i* upper bands are important in recognition even if the higher bands signal amplitudes are very low:

```
Rule 1:       B9 > 0.05271

       => vowel a

Rule 2:        B7 <= 0.04439
               B11 > 0.00976
       => vowel e

Rule 3:        B7 <= 0.04439
       B11 <= 0.00976
               B14 > 0.0212
=> vowel i

Rule 4:        B7 > 0.04439
               B9 <= 0.05271
               => vowel o

Rule 5:        B7 <= 0.04439
               B11 <= 0.00976
               B14 <= 0.0212
=> vowel u
```

As a side remark, these upper bands on the sound harmonics over 2500 Hz are important even in musical sonority recognition. Such results enable us to realize a rule-based system to support the ASR.

The speakers have to pronounce a number of vowels and a preliminary segmentation is realized on the uttered words. One of the good segmentation methods in order to select the vocalic trend is an image processing selection method on the registered word spectrogram.

On the vocalic trends obtained, recognition process is practiced and a preliminary confidence coefficient is obtained in order to constitute a sort of meta-rule for further investigations.

5. An Automatic Speaker Recognition using a Hausdorff-based metric

We propose now an automatic text-dependent speaker recognition approach [24], which uses a special nonlinear metric in the classification stage. Let us consider the following speaker recognition task.

We assume that we have a sequence of spoken utterances, each of them representing the same spoken word. The recognition task consists of identifying the speaker of each spoken word. We propose a supervised technique for solving this text-dependent ASR problem, assuming that the speakers number is known and a training set is also available.

Let $s_1,...,s_n$ be the audio signals of the spoken utterances to be recognized. As mentioned before, they must have the same transcript. Let us assume that these vocal utterances are provided by *N* speakers. The training set is created by recording each speaker several times with the same word. Thus, we obtain the set of signal prototypes $\{S_1^1,...,S_{n_1}^1,...,S_1^N,...,S_{n_N}^N\}$, where S_j^i represents the *j*th vocal prototype of *i*th speaker number *i* and n_i is the number of prototypes related to that speaker.

Like any pattern recognition approach [25], our speaker recognition method consists of two parts: the speech feature extraction and the feature vector classification [26]. In the first stage we perform vector feature extraction operations for both the vocal signals to be recognized and the prototype signals.

We use a “delta delta Mel cepstral analysis” for vocal feature extraction, the Mel Frequency Cepstral Coefficients (MFCC) being the dominant features used for speech and speaker recognition [27]. Thus, a short-time signal analysis is performed on each of the involved vocal sounds.

Each signal is divided in overlapping segments of length 256 samples with overlaps of 128 samples. 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 DFT (Discrete Fourier Transform) to it. Converting this spectrum on the melodic scale, the Mel-spectrum is obtained. Then, the Mel cepstral acoustic vector is computed by applying first the logarithm, then the DCT (Discrete Cosinus Transform) to the Mel spectrum.

The MFCC vectors are good feature vectors but better results are obtained by further processing of these acoustic vectors. Thus, we compute delta Mel cepstral coefficients (ΔMFCC) as the first order derivatives of MFCC, and the “delta delta Mel frequency cepstral coefficients” (ΔΔMFCC), as the second order derivatives of MFCC.

Each resulted ΔΔMFCC acoustic vector has a 256 samples dimension. For reducing this size, we truncate each acoustic vector to the first 12 coefficients, which we consider to be sufficient for speech featuring. Then we create a 12 row matrix by positioning these truncated acoustic vectors as columns.

The obtained ΔΔMFCC-based matrix represents a satisfactory speech feature vector for speaker recognition. Thus, for each signal s_k and each S_j^i, a ΔΔMFCC - based feature vector $V(s_k)$ or $V(S_j^i)$ is then computed. These feature vectors $\{V(S_1^1),...,V(S_{n_1}^1),...,V(S_1^N),...,V(S_{n_N}^N)\}$ represent the training vectors of the recognition system.

The second level of our speaker recognition approach is the feature vector classification. We propose an extended variant of the minimum distance classifier. The classical variant of this classifier consists of a set of prototypes and an appropriated metric. The pattern to be recognized is inserted in the class corresponding to the closest prototype [29].

Our vector classification approach considers not only one but a set of prototype vectors for each class. The vector to be classified is compared with each class, the mean value of the distances between it and the training vectors of that class being computed. It is inserted in the class corresponding to the smallest mean distance.

For our recognition problem each class is related to a known speaker and its feature training set contains the prototype vocal signals belonging to that speaker. Thus, $\{V(S_1^i),...,V(S_{n_i}^i)\}$ represents the training feature set of the *j*th class. Our classifier has to compute distances between feature vectors like $V(s_k)$ and $V(S_j^i)$, therefore it must compare different sized vectors. This means that a linear metric, like the Euclidean distance, cannot be used in this case.

For this reason we propose a nonlinear metric which is able to compare matrices having a single common dimension, like the matrices representing our speech feature vectors. We introduce a distance which derives from Hausdorff metric for sets [28]. If *A* and *B* are two different-sized sets $(|A| \neq |B|)$, the Hausdorff metric is defined as the *maximum distance of a set to the nearest point in the other set*. Thus, we have the relation:

$$h(A,B) = \max_{a\in A}\{\min_{b\in B}\{dist(a,b)\}\} \qquad (1)$$

where *h* represents the Hausdorff distance between sets and *dist* is any metric between points (for example the Euclidean distance).

In our case we have to compare two matrices, different sized but with one common dimension, instead of two sets of points. So, let us consider $A = (a_{ij})_{n\times m}$ and $B = (b_{ij})_{n\times p}$ the two matrices having the same first dimension. We use notation *n* for the rows number, although we already used it for the number of vocal utterances. Let us assume that $m \neq p$.

We introduce two more vectors, $y=(y_i)_{p\times 1}$ and $z=(z_i)_{m\times 1}$, then compute $\|y\|_p = \max\limits_{0\le i\le p}|y_i|$ and $\|z\|_m = \max\limits_{0\le i\le m}|z_i|$. With these notations we create a new nonlinear metric *d*, having the following form:

$$d(A,B)=\max\left\{\sup_{\|y\|_p\le 1}\inf_{\|z\|_m\le 1}\|By-Az\|,\sup_{\|z\|_m\le 1}\inf_{\|y\|_p\le 1}\|By-Az\|\right\} \tag{2}$$

This restriction-based distance represents the Hausdorff distance between the sets $B\left(y:\|y\|_p\le 1\right)$ and $A\left(z:\|z\|_m\le 1\right)$ in the metric space R^n. Therefore, it results:

$$d(A,B)=h\left(B\left(y:\|y\|_p\le 1\right),A\left(z:\|z\|_m\le 1\right)\right) \tag{3}$$

Obviously, the metric d depends on y and z.

Trying to eliminate these terms, we have found a new distance that doesn’t depend on these vectors anymore. This is not a Hausdorff distance, but a new obtained “Hausdorff-based” metric, defined as:

$$d(A,B)=\max\left\{\sup_{1\le k\le p}\inf_{1\le j\le m}\sup_{1\le j\le n}|b_{ik}-a_{ij}|,\sup_{1\le j\le m}\inf_{1\le k\le p}\sup_{1\le i\le n}|b_{ik}-a_{ij}|\right\} \tag{4}$$

The nonlinear function *d* verifies the distance properties: positivity ($d(A,B)\ge 0$), symmetry ($d(A,B)=d(B,A)$) and triangle inequality ($d(A,B)+d(B,C)\ge\ d(A,C)$). The distance that we propose constitutes a satisfactory discriminator between sound feature vectors in the classification process [26]. Therefore, each speech signal s_k must be inserted in the class indicated by $id=\arg\min\limits_i \frac{\sum_{j=1}^{n_i} d(V(s_k),V(S_i^j))}{n_i}$, where *d* represents the nonlinear distance given by the relation (4). This means that the vocal utterance corresponding to s_k, was generated by the speaker with number *id*.

Our experiments show that the proposed automatic speaker recognition approach gives good results. Thus, let us consider three speakers which provide same-transcript vocal utterances to be recognized. We want to test our proposed minimum distance classifier, so, we will consider a training set and a set of speech signals to be recognized, for each speaker. First, the spoken word *zero* is produced by the speakers. We consider a training set containing two vocal signals for each speaker. The six training feature vectors $V(S_j^i)$ corresponding to these signals are depicted as gray scale images in Fig. 2. Each column is related to a speaker and its mark (1, 2 or 3) represents the speaker number.

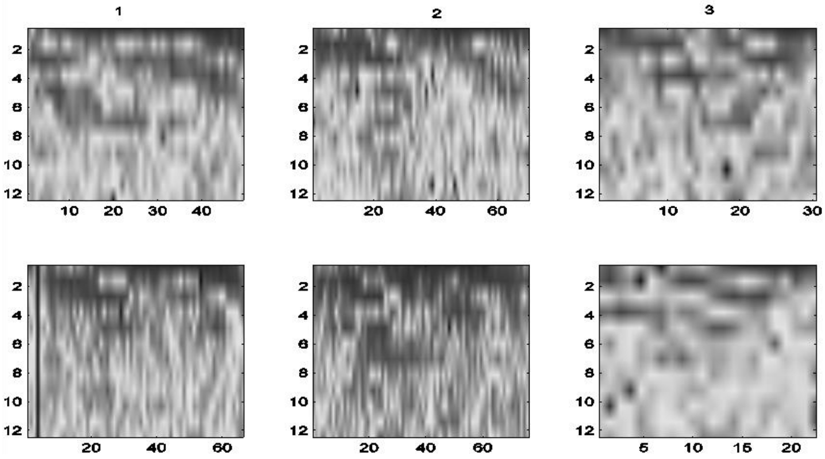

Fig. 2 - The Training Feature Set

Also, we have four speech utterances s_k representing the same word for each speaker. Their ΔΔMC feature vectors $V(s_k)$ are represented in Fig. 3.

The intensity images on each column correspond to the same speaker, which is marked by number 1, 2 or 3.

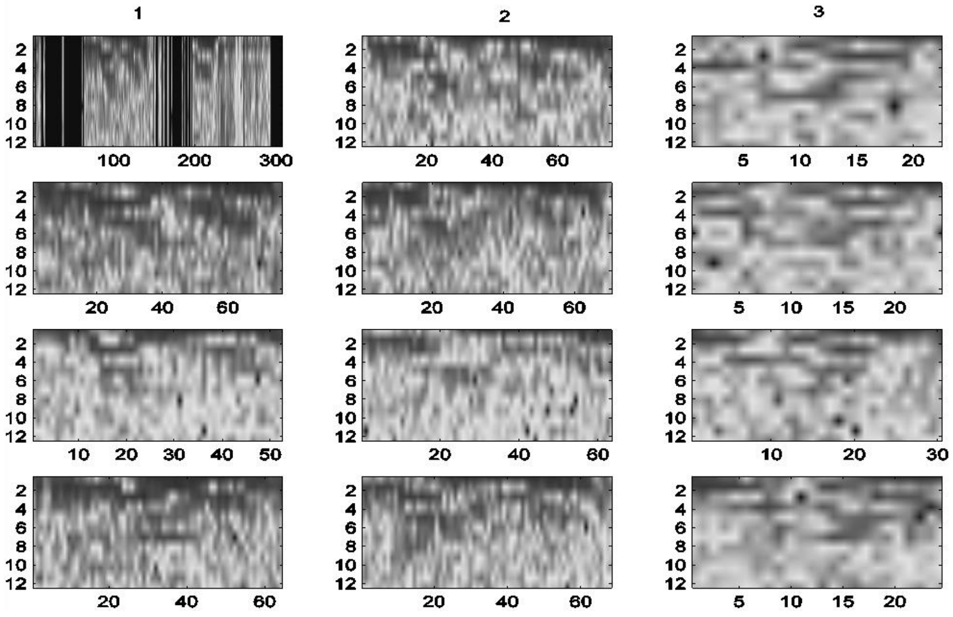

Fig. 3 - The Feature Vector Set

The distances (4) between the training vectors in Fig. 2 and the feature vectors from Fig. 3 are then computed. The computing results are displayed in the next table.

Table 2

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}	s_{11}	s_{12}
1	2.59	2.36	2.81	3.17	2.69	2.68	2.38	2.65	3.93	3.67	3.53	3.97
2	2.63	2.56	3.17	2.50	2.54	2.67	2.81	2.13	3.79	3.75	3.68	4.10
3	4.21	4.50	4.12	4.44	4.11	4.13	3.89	4.08	2.30	2.86	2.19	2.46

In the column corresponding to each s_k there are registered the mean distance values $\dfrac{\sum_{j=1}^{n_i} d(V(s_k),V(S_i^j))}{n_i}$ to each class (speaker), referred by 1, 2 or 3.

The minimum distance value must be located on the row corresponding to the speaker producing s_k. As we know, s_{1-4} are produced by 1, s_{5-8} are produced by 2, s_{9-12} are produced by 3. Thus, we distinguish a single classification error in this table: the value on the first row, fourth column, should be minimum but it is not, the minimum being located on the second row (2.50).

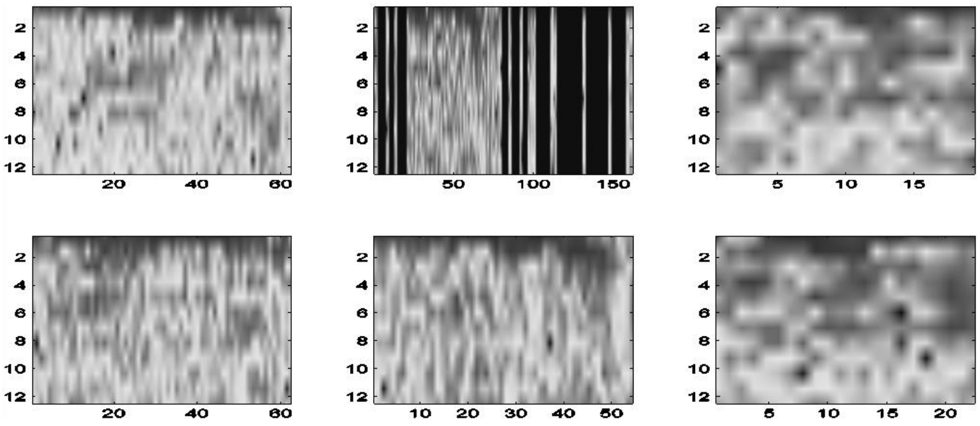

Fig.4 - The Training Feature Set

Let us consider another experiment, using *seven* as the spoken utterance. The training feature vectors $V(S_j^i)$ are displayed in Fig. 4, and the ΔΔMC feature vectors $V(s_k)$ are represented in Fig. 5. Like in the previous case, each column corresponds to a speaker.

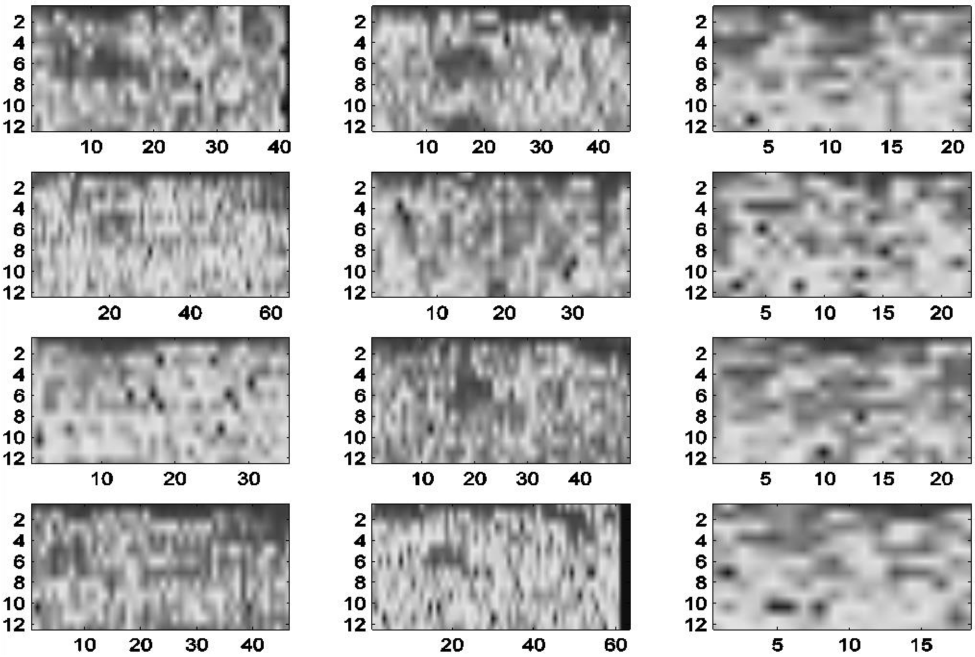

Fig. 5 - The Feature Vector Set

The computed mean distance values are displayed in Table 3. A single classification error is produced in the speaker recognition process. As in the previous case, in the fourth column, the minimum value (3.29) is located in the last row and not in the first row where it should be. The recognition results are about 95%, the variation depending on the degree of noise in the environment.

Table 3

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}	s_{11}	s_{12}
1	2.41	2.26	3.19	4.73	2.91	2.88	2.74	2.52	5.78	3.44	4.76	3.11
2	2.45	3.35	4.54	7.99	2.36	2.70	2.06	2.24	7.03	3.63	6.02	3.52
3	3.41	4.31	3.29	3.47	4.90	5.63	3.84	4.12	2.44	2.93	2.29	2.78

6. Automatic Speaker Recognition Decision System

An ASR Decision System (ASR-DS) and an aggregation criterion are proposed to reinforce the recognition coefficient obtained after a M number of essays by each method. Thus, instead of using the results of a single proposed method, knowing the "inter" and "intra" speaker variability, source of inevitable error possibility, we propose a multi-decision aggregation system. It uses a F_0 (fundamental frequency) detection module (F_0D), meant to do a-priory F_0 classification in male/female classes. A second module run with the AR - NN proposed method. The HDM and VD (vowel-based detection) modules are parallel implemented. Computed results in the parallel methods are weighted and aggregated in a decision system (Fig. 6. where w_i are weighted coefficients).

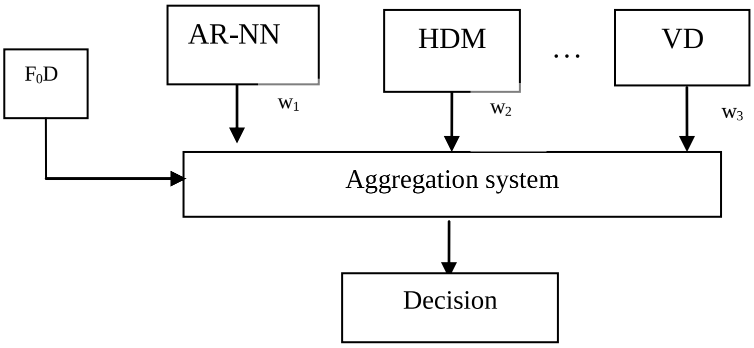

Fig. 6 - Aggregation in ASR aide-decision system

We realize the weighted aggregation situating ourselves in the context of the belief functions [29]. Let Δ be frame of discernment (also called the universe of discourse or the domain of reference), on which evidences induce some beliefs, constituted of a number of speakers, S_i, $i=1,...N$, and let Ω be the Boolean algebra of propositions derived from Δ, (Ω contains the conjunctions, disjunctions and negations of any set of propositions from Δ). On this frame a-priori evidence is induced by F_0 detection, on some so-called focal elements on which our belief is focalized in order to group them on gender and age. So, to identify a male speaker, normally we have to look for him in the detected group with all the three exposed methods (playing here the role of experts) and if a different result is obtained then a conflict have to be managed.

Depending on statistical previous results, the importance of each of the three "expert modules" is different regarding to the expected result. A total mass of belief is calculated for each detected speaker according to the Dempster's rules [29].

In the belief and plausibility models each time one of the functions *m (mass), bel (belief), pl (plausibility)* or *q (communality function)*, is introduced, it is preferable to abstain from defining each one in relation to the others. The simple declaration of one of them automatically implies the others, the supplementary symbols being sufficient to know ones are interrelated. Given two belief functions bel$_1$ and bel$_2$ induced by two distinct pieces of evidence on the event A, the belief function bel$_{12}$ that results from their combination is obtained by Dempster's rules of combination and expressed with communality functions they becomes:

$$q_{12}(A) = q_1(A)q_2(A) \tag{5}$$

a relation whose simplicity explains the advantage of these functions. In the previous relation, the communality function is a function q : Ω -> [0,1], such that, for two events A and B we have:

$$q(A) = \sum_{B \to \neg A} m(A \vee B) \tag{6}$$

In this decision schema, studies on two types of structures and the appropriated aggregating technique have to be further elaborated:

a) a parallel structure (the same priority for all the decision criteria) and

b) a hierarchic algorithm of detection (vowel pronunciation detection first and then the AR-NN, the HDM weighting detection, etc).

7. Conclusions and future work

High recognition coefficients had been obtained by our methods (ARNN - max. 89%, HDM - max. 95% and VD - 96%) but in a security system unfortunately this is not enough.

Results obtained by one method have to be reinforced in a decision system, final recognition percent being the result of a special inference. In order to acquire a higher reliability in an ASR system, we proposed an inference schema to aggregate the certainty coefficients obtained by each method.

As a further direction, decision system using weighted aggregation to reinforce recognition probity might be situated in the context of belief functions and Dempster's rules.

Particularly future possible applications of the presented recognition methods are to be tested. Hausdorff distance method may be the basis of a system conceived for medical therapy. Unsupervised systems based on these measures might be envisaged in hearing and speech training of the hearing impaired subjects, or even for the recovering period of the cochlear implanted patients [30]. The system has to identify, based on these distances, and to show the good or bad pronunciations of the patients suffering of cophosis (deep perception deafness).

References

[1] BIGUN, B.D., SMERALDI, F., FISCHER, S., MAKAROV, A., *Multi-Modal Person Authentication*, NATO ASI Series, Springer, Vol. **F-163**, pp. 26-50, 1997.

[2] MARKEL, J.D., GRAY, A.H., *Linear Prediction of Speech*, Springer Verlag, New York, 1976.

[3] COSI, P., *Auditory Modeling for Speech Analysis and Recognition*. In M. Cooke, S. Beet and M. Crawford Eds., Visual Representation of Speech Signals Cichester, John Wiley & Sons, pp. 205-212, 1993.

[4] STOLOJANU, G., PODARU, V., CETINĂ, F., *Numeric Processing of Speech*, Military Ed., Bucharest, 1984.

[5] CALLIOPE, *La parole et son traitement automatique*, Paris, Masson, 1989.

[6] RABINER, L., JUANG, B. H., *Fundamentals of Speech Recognition*, Englewood Cliffs, Prentice Hall, New York, 1993.

[7] COSTIN, M., ZBANCIOC, M., *Hints about some basic indispensable elements in speech recognition*, Computer Science Journal of Moldova, v.10, n.**2**., pp.169-196, 2002.

[8] ZBANCIOC, M., COSTIN, M., *Using Neural Networks and LPCC to Improve Speech Recognition*, International IEEE Conference SCS Proc., Iaşi, Vol. **1**, pp. 445–448, 2003.

[9] SENEFF, S., *A computational model for the peripheral auditory system: application to speech recognition research*, Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, New York, IEEE Press, pp. 37.8.1-37.8.4, 1986.

[10] DE POLI, G., PRANDONI, P., TONELLA, P., *Timbre clustering by self-organizing neural networks*, Proceedings of the **X**th Colloquium on Musical Informatics, University of Milan, pp. 102-107, Milan: AIMI, 1993.

[11] LEE, K., HON, H., REDDY, R., *An Overview of the SPHINX Speech Recognition*, IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. **38**, Nr. 1, January 1990.

[12] COSTIN, M., ZBANCIOC, M., *Improving Cochlear Implant Performances by MFCC Technique,* International IEEE Conf. SCS Proceedings, Vol. **1**, pp. 449-452, 2003.

[13] MITRA, K.S., *Digital Signal Processing*, **2**nd Ed. McGraw-Hill, pp. 531, 2001.

[14] FURUI S., *Speaker Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum*, IEEE Trans. on Acoustics, Speech and Signal Processing, ASSP **34(1),** 52-59, 1986.

[15] COSTIN, M., GRICHNIK, A., ZBANCIOC, M., *Tips on Speaker Recognition by Autoregressive Parameters and Connectionist Methods,* International IEEE Conference SCS, Proceedings, Vol. **1**, pp. 169 – 172, 2003.

[16] COSTIN, M., ZBANCIOC, M., CIOBANU, A., BERGER-VACHON, C., *Some Attempts in Improving Cochlear Implanted Patients Performances*, Modeling and Automatic Methods, **IX**-th IPMU International Conference, Annecy France, pp 711-718, Vol. **II**, 2002.

[17] BROCKWELL, P.J., DAVIS, R.A., *Introduction to Time Series and Forecasting,* Springer, **2**nd Edition, pp. 45-174, 2002.

[18] HAYKIN, S., *Neural Networks,* MacMillan College Publ. Co. Inc., NY, pp. 236-262, 1994.

[19] HAYKIN, S., *Adaptive Filter Theory,* Prentice Hall, Englewood Cliffs, NJ, **2**nd Ed., pp.186-244, 1991.

[20] PRESS, H. W., TEUKOLSKY, A.S., VETTERLING, T.W., FLANNERY, P.B., *Numerical Recipes in C, The Art of Scientific Computing,* **2**nd Edition, Cambridge University Press, pp. 421, 813, 1994.

[21] BOËX, C., PELIZZONE, M., *Speech Recognition with a CIS Strategy for the Ineraid Multi-channel Cochlear Implant,* American Journal of Otology **17**, pp. 61-68, 1996.

[22] QUINLANN, R., C5, http://www.ph.tn.tudelft.nl/PRInfo/software/ available in November 1997.

[23] KLEVANS, R., RODMAN, R., *Voice Recognition,* Artech House (Telecommunications Library), Boston, London, pp. 50-170, 1997.

[24] MINH, N.DO, *An Automatic Speaker Recognition System,* http://lcavwww.epfl.ch/~minhdo/, Swiss Federal Institute of Technology, Lausanne, 2000.

[25] DUDA, R.O., *Pattern Recognition for HCI*, Department of Electrical Engineering, San Jose State University, 1997, visited March 2004,
http://www.engr.sjsu.edu/~knapp/HCIRODPR/PR_home.htm

[26] BARBU, T., *Discrete Speech Recognition Using a Hausdorff based Metric*, Proceedings of the 1st International Conference of E-Business and Telecommunication Networks, ICETE 2004, Setubal, Portugal, Vol. **3**, pp.363-368, Aug. 2004.

[27] LOGAN, B., *Mel Frequency Cepstral Coefficients for Music Modelling*, Compaq Computer Corporation, Cambridge, 2000, and in Proc Int. Symposium on Music Information Retrieval (ISIMIR). Plymouth, MA, 2000.

[28] NORMAND, G., BOUILLOT, M., *Hausdorff distance between convex polygons*, Web project for CS 507, http://cgm.cs.mcgill.ca/~godfried/teaching/cg-projects/98/normand/main.html, Computational Geometry, McGill University, 1998.

[29] SMETS, PH., MAMDANI, E.H., DUBOIS, D., PRADE, D., *Logics for Automated Reasoning*, Academic Press London, printed in England, Harcourt Brace Jovanovich - Publishers, pp. 253-286, 1988.

[30] LOIZOU, P., *Signal Processing for Cochlear Prosthesis: A Tutorial Review*, Proceedings of the 40th Midwest Symposium on Circuits and Systems, 3-6 Aug 1997**,** Sacramento, CA, USA, Vol. **2**, pp. 881-885, 1997.