AUTOMATIC SPEAKER VERIFICATION EXPERIMENTS USING A CONTINUOUS SPEECH RECOGNIZER

Doru MUNTEANU[*], Constantin PINTILIE[**], Laurențiu APOSTOL[**]

[*] Military Technical Academy, address: 81-83, Bvd. George Coşbuc, Sect. 5, Bucharest, Romania, phone: + 40.21.3354660, email: munteanud@mta.ro, web: www.mta.ro
[**] UTI SYSTEMS, address: 27C Cernauti Street, Sect. 3, Bucharest, Romania, phone: +40.21.2055923, email: {constantin.pintilie, laurentiu.apostol}@uti.ro, web: www.uti.ro

Corresponding author: Doru MUNTEANU

This paper focuses on a special issue of biometrics – automatic speaker verification (ASV). There is a great interesting in developing and performance increasing of ASV applications because of the advantages offered comparing to other biometrical methods. The most important aspect is that such a speech processing application has a low implementation cost. State-of-the-art speaker recognizer are based on statistical models such as VQ, GMM or HMM. This work reports experiments on prompted text speaker verification on a Romanian corpus previously built. First, the continuous speech recognizer architecture is built at monophone level, context independent, single mixture. Then, two models are trained using appropriate data: a) Client model consisting of speaker dependent phonemes (SD) trained with few minutes of client's speech; and b) World model consisting of speaker independent phonemes (SI) trained with all sentences available in the database. Each phone has a two state left-right HMM with diagonal covariance matrices. The speaker verification system is text-prompted as a sentence HMM is constructed for the key text by concatenating corresponding models. The normalized log-likelihood is computed and compared with a threshold to decide whether to accept or reject the speaker. In the verification stage, the normalized log-likelihood is computed by the difference between the log-likelihood obtained through Viterbi forced alignment of the client model and world model, respectively. Finally a procedure used to determine the verification system performances is presented, including FAR and FRR graphics vs. threshold, ROC curves and various criteria for threshold calibration.

Key words: Biometric, automatic speaker verification systems, acoustic modeling, hidden Markov models.

1. INTRODUCTION

Today there are many needs for biometrics (i.e. methods for automatic recognition of a person based on behavioral or physiological characteristics) in:

– homeland security: airport security, strengthening our national borders, in travel documents, visas;
– enterprise-wide network security infrastructures;
– secure electronic banking;
– investing and other financial transactions;
– retail sales, law enforcement;
– health and social services;
– counter-terrorism measures, etc.

Most of these are already benefiting from biometrics technologies. The experimental results presented in this paper have been obtained during a research project regarding the use of biometrics in fight against terrorism. A voice recognition system can be used in identifying an unknown voice recording intercepted by the authorities. The speech sample is then compared against other models stored in a database [12].

A biometric authentication system has to compare a registered biometric sample against a newly captured biometric sample. During *enrollment*, a sample of biometric is captured, processed and stored for later comparison. Recognition may be used in either *identification mode* (the system identifies the person with the best match from the entire enrolled population) or *verification mode* (the system authenticates a person's claimed identity).

Speaker recognition methods are considered today "classic" biometrics, dating for more than four decades. Speaker recognition systems extract acoustic features from the speech signal that are different between individuals. These features reflect:

a. *anatomy*: physical size and geometrical shapes of articulators – lungs, larynx (vocal cords), velum, tongue, teeth, lips;
b. *learned behavioral patterns*: speaking style.

Speaker recognition deals with both verification and identification and is part of the larger area of speech signal processing (figure 1). Speaker verification decides if a speaker is whom he claims to be. A speaker identification system decides if a speaker is a specific person or is among a group of persons. In speaker verification, a person makes an identity claim (e.g., by entering a number or presenting a smart card) In text-dependent recognition, the phrase is known to the system and can be fixed or prompted (visually or orally) while in text-independent recognition the system has to recognize any phrase, regardless of its phonetic content.

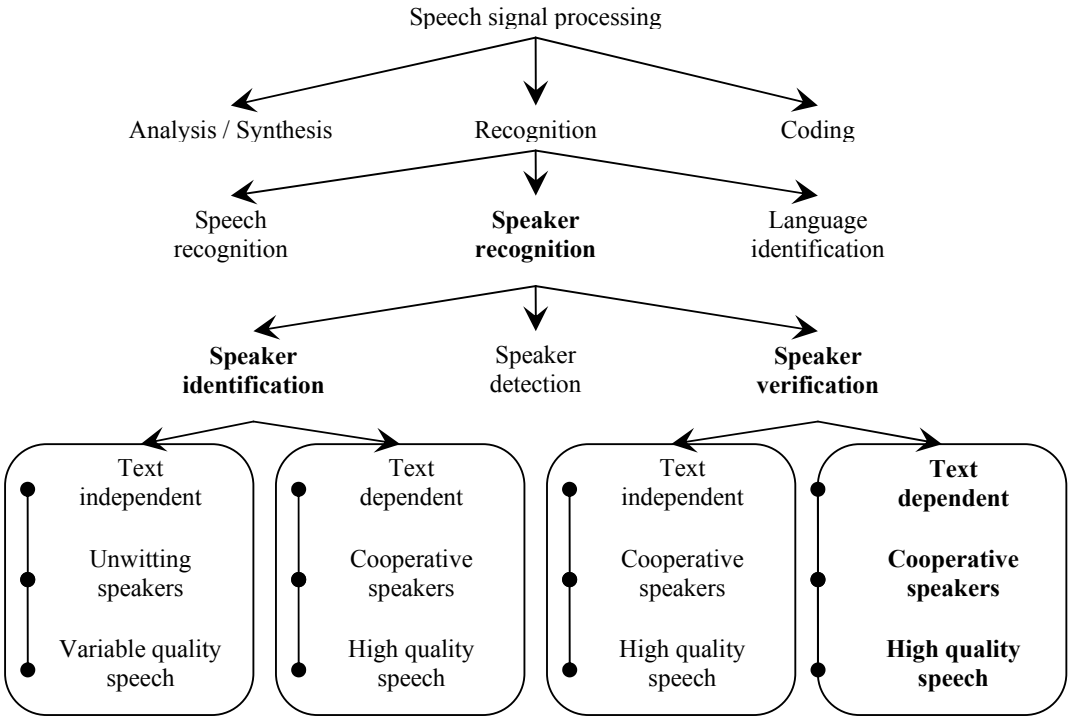

Figure 1. Speech signal processing [7].

In this paper we have used a continuous speech recognizer previously built [4], [10], [11] in order to build a prompted text automatic speaker verification (ASV) system, implying text dependency, cooperative speakers and high quality speech signal. Usually there are few simple steps in speaker verification: the claimant speaks the phrase into a microphone, the speech signal is processed by the system that makes the decision to accept or reject the user's identity claim or possibly to report insufficient confidence and request more speech signal before making the decision.

Although in laboratory conditions (high quality speech, clear phrases, etc), state of the art ASV systems achieve very good results, these performances are degrading in real-time conditions due to both human and environmental factors:

– misspoken phrases;
– emotional state;
– time varying microphone placement;
– poor or inconsistent room acoustics (additive noise, reverberations, etc.);
– channel mismatch (different microphones, analog-to-digital converters for enrollment and verification);
– sickness, etc.

These factors are very important because no matter how good is the recognition method used by the ASV systems, human errors (such as misspeaking) or environmental noise ultimately limits their performances.

2. TEXT-DEPENDENT ASV SYSTEM

The focus of this paper is the text-dependent ASV system based on the continuous speech recognition (CSR) system previously built. The details of the Romanian Language CSR are presented in section 3. It is based on hidden Markov models (HMM) at monophone level, context-independent, single mixture.

For a particular speaker (client), the ASV system assumes 5 steps, as follows:

a. *Training the world model* all the available phrases of the database (training set) are used for training a speaker independent automatic speech recognizer (SI-ASR). Client phrases are excluded from the training pool. This procedure is performed once, regardless the enrollment of new clients.
b. *Training the client model* all the phrases captured from the client (training set) are used for training a speaker dependent automatic speech recognizer (SD-ASR). This enrolment procedure is mandatory for each new client.
c. *Viterbi alignment* for all test phrases (including both client and imposture phrases) from the test set)\, Viterbi forced alignment procedure is performed; for each phrase (observation), two acoustics scores are computed:
   – $\log P(O|\lambda_{SD})$ log-likelihood obtained by the SD-ASR
   – $\log P(O|\lambda_{SI})$ log-likelihood obtained by the SI-ASR

   where λ_{SI} and λ_{SD} are speaker specific HMMs and O is the observation sequence of the phrase.
d. *Acoustic score normalization* - normalized score is computed according to the formula:
$$L(O) = \log P(O|\lambda_{SD}) - \log P(O|\lambda_{SI}) \quad (1)$$
e. *Performance evaluation* – false acceptance rate (FAR) and false rejection rate (FRR) are determined for a specific range of the decision threshold T on the test phrase set.

The ASV system architecture is revealed by figure 2. In the training stage, both client and world model parameters are estimated using the speech from the training set. Then, all the phrases from the test set are aligned by the Viterbi procedure and the two acoustic scores are computed. Finally, these scores are differentiated and a normalized acoustic score is obtained for each test phrase.

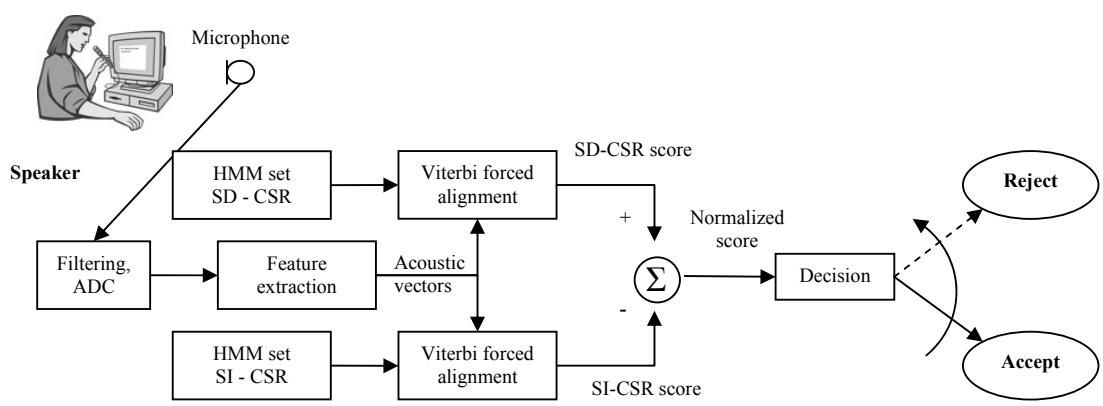

Figure 2. The ASV system architecture.

According to the "hypothesis testing", the ASV system has to choose one of the hypothesis:

a. H_0 – user is impostor

b. H_1 – user is the claimed speaker (client)

In figure 3 one may observe the match scores of the observations from two probability density functions (PDF) according to whether the user is an imposter or the client. The names of the probabilities are given in table 1.

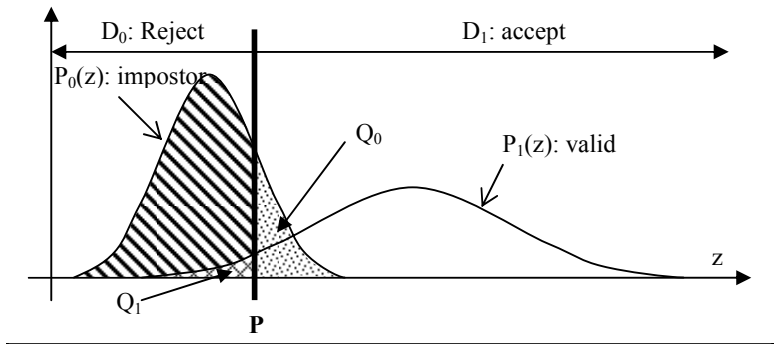

Figure 3. Client and impostor densities.

To find a given performance probability area, the hypothesis determines over which PDF to integrate, and the threshold determines which decision region forms the limits of integration.

Table 1. Probability terms and definition

Probabilities	Decision (D)	Hypothesis (H)	Decision result
Q_0	1	0	False acceptance
Q_1	0	1	False rejection
$Q_d = 1-Q_1$	1	1	True acceptance
$1-Q_0$	0	0	True rejection

Let $p_A(z|H_0)$ be the conditional density function of the score z obtained by other speakers than the claimed one, and $p_A(z|H_1)$ the conditional density function of the score z obtained by the claimed speaker. If both conditional density functions are known, then the Bayes test with equal misclassification costs for speaker A is based on the likelihood ratio $\lambda_A(z)$:

$$\lambda_A(z) = \frac{p_A(z|H_0)}{p_A(z|H_1)} \tag{2}$$

The error probability, which is minimized according to Bayes rule, is determined by the overlap of the two PDFs. The smaller this surface is, the smaller the ASV system probability of error.

If the two density functions (of client speaker and other speakers) are unknown, they can be estimated from experimental samples:

a. Conditional pdf given the client speaker A, $p_A(z|H_1)$ is estimated from the scores obtained from the acoustic scores obtained by the client speaker, using its own model;

b. Condition pdf given for impostors, $p_A(z|H_0)$ is determined from the scores obtained by other speakers using the model of the speaker A.

Then, the likelihood ratio for speaker, $\lambda_A(z)$ may be now computed. Classification resumes to choosing the threshold T so that the decision rule becomes:

$$\lambda_A(z) = \begin{cases} \geq T, & \text{choose } H_0 \\ < T, & \text{choose } H_1 \end{cases} \tag{3}$$

Threshold T may be determined according to various criteria:

1) setting T to approximate minimum error performance, according to the a priori probabilities that the user is an impostor and that the user is the true speaker;
2) choosing T to satisfy a fixed FA or FR criterion (Neyman – Pearson);
3) varying T to find a desired FA/FR ratio.

Threshold T may be specific to each client registered in a database or may be a dynamically changed according to various conditions.

Both types of errors, FA and FR, can be increased or decreased. For instance, while the threshold increases, the false alarms increase too and the false rejections are decreasing. Usually, there is a compromise between FA and FR errors which depends on the threshold value. The ROC curve depicted in figure 4 is a graphical representation of the sensitivity of a binary classifier, obtained by varying the threshold. For an ASV system, the ROC curve is the plot of the FR probability vs. FA probability.

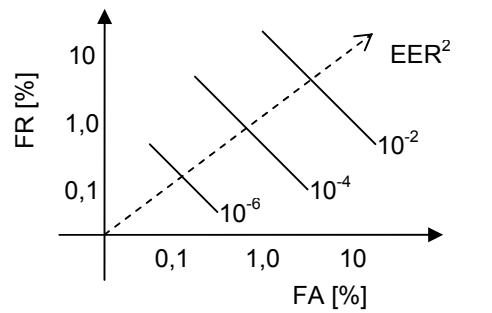

Figure 4. Hypothetic ROC curve for 3 ASV systems.

In figure 4, the dotted line represents equal error probability. The family of lines noted with 10^{-2}, 10^{-4}, 10^{-2} are systems with different FA x FR products. The best system is the closest to the origin, system with an equal error rate (ERR) of 10^{-6}.

3. CONTINUOUS SPEECH RECOGNIZER

The ASV system is based on a continuous speech recognizer previously built for Romanian language [4], [10], [11].

Speech database. We have used a database recorded in office environment. It covers more than 10 hours of read speech and 11 speakers (7 males and 4 females). Texts used are phonetically balanced and they contain 4100 phrases and more than 3000 common words from various fields like education, sport, politics, etc. For all speakers there are two sets: one used for training and the other used for testing.

Feature extraction. The waveforms were recorded in an office room (SNR= 30 dB) with a cardioid desktop microphone (30Hz ÷ 16kHz) and an usual 32bit PC sound card with 16kHz / 16bit sampling. The waveform was first pre-emphasized (with a coefficient of 0.97), then parameterized by cepstral analysis on a 25,6ms Hamming window (overlapping 40%). A number of 12 basic Mel-frequency cepstral coefficients (MFCC) were extracted and first and second derivatives (36 parameters in total) were added. This parameters are used somehow surprisingly in speech recognition too, and are almost standardized [1], [2].

Acoustic models. Each Romanian phoneme (from the 34 phoneme set) was modeled with a three-state HMM in a left-right topology, using single mixture Gaussian continuous distribution. The covariance matrices are diagonal in order to reduce the resources required for the output probability computation. The HMM initialization is based on making all the models identical initially and then performing the well-known Baum-Welch procedure, also called embedded training. The global speech (all training utterances) mean and covariance are computed and these values are used to initialize the entire set of HMMs, consequently all these being identical. Further, embedded training is used in order to differentiate models. For each training phrase, a composite model is build by concatenating individual models according to the phrase labels and phonetic dictionary.

The main steps of the training procedure for the context independent models are:

1. HMM initialization - all models are identical;
2. Baum-Welch parameter re-estimation: 3 to 5 iterations (a threshold of 0.01 in log-likelihood was used for convergence) for composite models;
3. Viterbi forced alignment: when the training dictionary contain multiple pronunciations, the one with the best alignment score is selected;
4. Baum-Welch parameter re-estimation: 3 to 5 iterations.

4. EXPERIMENTAL RESULTS

Automatic speech recognizer, as well as ASV system was built using a very well known toolkit [5], [6].

Case study. The experiments were performed on speaker #3 within the database – who was chosen random from the speakers set. For this speaker, a client model was built, based on a speaker – dependent continuous speech recognizer (SD-CSR$_3$) – as depicted in section 3. Then, for the other speakers within the database, a world model was built, based on a speaker – independent continuous speech recognizer (SI-CSR). Both client and world models have the same structures and number of acoustical parameters, the differentiation being made by the training data.

In the verification stage, a Viterbi forced alignment procedure is applied to the phrase being tested, for both systems and the normalized score is computed (equation 1). Then, according to the ASV architecture, depicted in figure 2, a decision is taken based on the difference between the threshold and the normalized score.

For example, in table 2, there are listed few acoustic scores obtained for the test phrase ***'S0002'*** uttered by the client (speaker #3) and an impostor (speaker #1) with the same gender (male). One may see that the normalized scores are greater than zero for the client phrase, while the impostor phrase gets only negative acoustic score.

For all the test phrases, the normalized score is computed in order to determine the ASV system performances. Figure 5 is a plot of the normalized score obtained by the test phrases. Speaker phrases are in groups of 300 phrases. For instance, the speaker #3 (client) phrases indexes are in the range 601-900. One may observe that client phrases get clearly better scores than other speakers, allowing us to set a threshold for accept/reject decision.

System performances are evaluated following the steps:

a. Estimate probability function $p_3(z|H_1)$ based on the score histogram (figure 6) obtained by the ASV system of speaker #3 on the same speaker phrases;
b. Estimate probability function $p_3(z|H_0)$ based on the score histogram (figure 7) obtained by the ASV system of speaker #3 on the other speaker phrases.

Table 2. Acoustic scores obtained for the same phrase uttered by client and an impostor

Phrase idx. / Word	**Acoustic score z (log-likelihood)**					
	Speaker #1 (impostor)			**Speaker #3 (client)**		
	SI score	**SD Score**	**Norm. score**	**SI score**	**SD Score**	**Norm. score**
S0002						
STEAUA	-61,63	-75,30	-13,67	-66,21	-62,12	4,09
ESTE	-61,56	-64,88	-3,32	-67,35	-61,23	6,12
ECHIPA	-64,27	-66,73	-2,46	-69,82	-65,02	4,80
MEA	-61,27	-67,83	-6,56	-62,41	-59,21	3,20
DE	-56,59	-57,67	-1,08	-65,62	-61,25	4,37
FOTBAL	-65,22	-67,87	-2,65	-70,14	-62,43	7,71
PREFERATA'	-64,23	-66,37	-2,14	-66,66	-63,39	3,27

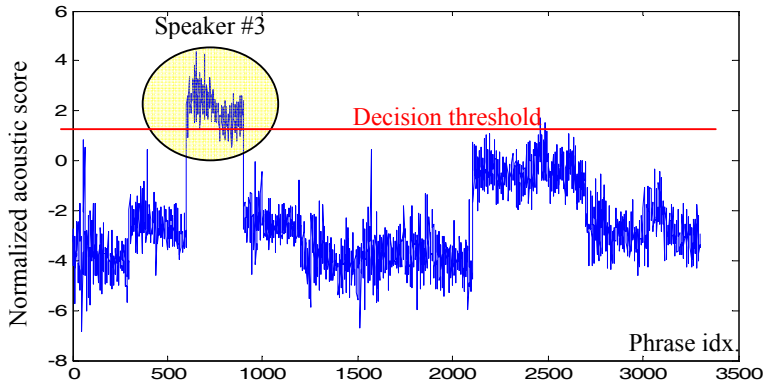

Figure 5. Normalized score obtained for all the test-phrases.

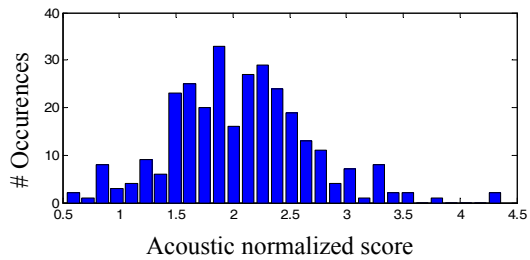

Figure 6. Histogram for scores of the client phrases.

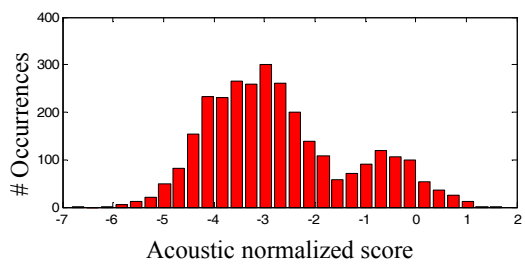

Figure 7. Histogram for scores of the other speaker phrases.

We may now determine the two probability function distributions (pdf) $p_3(z|H_0)$ and $p_3(z|H_1)$ corresponding to the histograms. For the particular case of speaker #3, functions are plotted in figure 8;

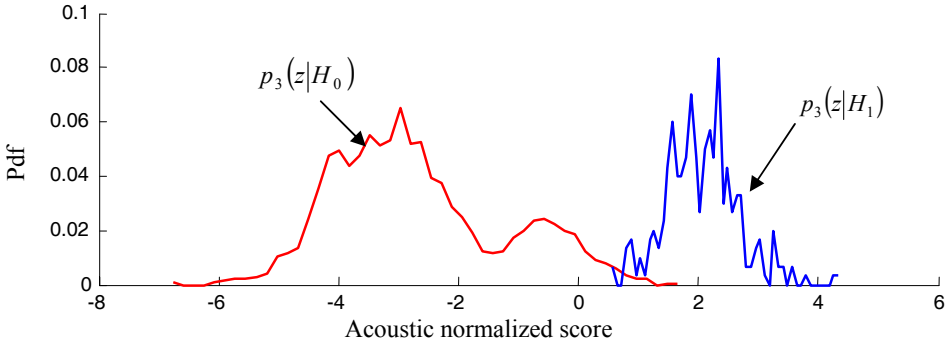

Figure 8. Distributions of the impostor and client phrases

c. Compute the false rejection rate (FRR) depending on the decision threshold T:

$$\mathrm{FRR}(T) = \frac{N_C(T)}{N_{TC}} \tag{4}$$

where $N_C(T)$ is the number of the client phrases accepted by the ASV system, given a threshold T, and N_{TC} is the total number of the client phrases;

d. Compute the false acceptance rate (FAR) depending on the decision threshold T:

$$\mathrm{FAR}(T) = \frac{N_I(T)}{N_{TI}} \tag{5}$$

where $N_I(T)$ is the number of the impostor phrases accepted by the ASV system, given a threshold T, and N_{TI} is the total number of the impostor phrases;

e. Graphics of $\mathrm{FAR}(T)$ and $\mathrm{FRR}(T)$ are obtained (figure 9).

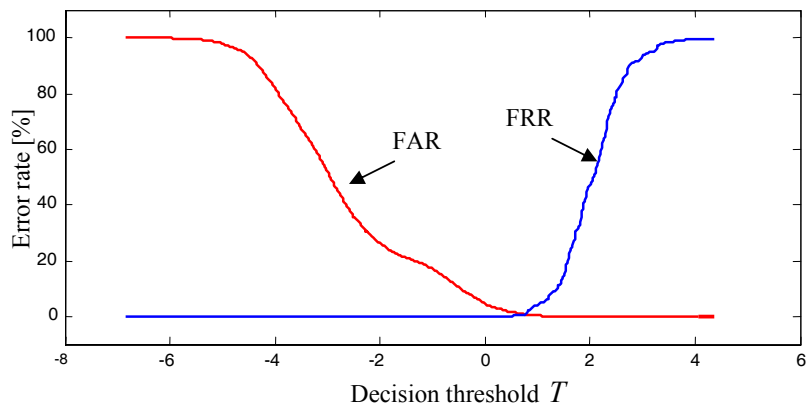

Figure 9. FAR and FRR vs. decision threshold.

Variation range of the threshold T is delimited by the minimum and maximum score obtained over the all phrase test set. The ROC curve and the ERR can be now determined. One may observe (figure 10), that the ERR of this particular ASV system is 0,6 % for the threshold T =0,77;

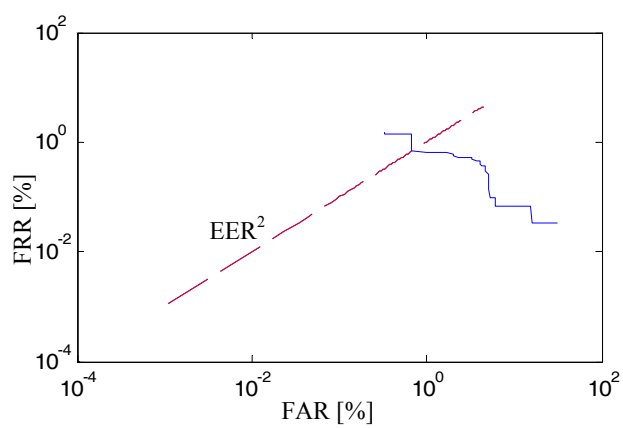

Figure 10. The ROC curve for the ASV system.

f. Determine threshold T according to a criteria regarding either FAR, FRR or both. In table 3, are listed several criteria and their appropriate thresholds and error rates.

Table 3. Threshold, FAR and FRR for different decision criteria

Nr. crt.	Criterion	Decision threshold (T)	FAR [%]	FRR[%]
1.	Minimum FAR x FRR	0,77	0,6	0,6
2.	Minimum FAR	1,76	0,0	31
3.	Minimum FRR	0,52	1,46	0,0

One may observe that in order to have a 0 false acceptance rate, a higher threshold is needed ($T = 1{,}76$) leading to a great rate of the false rejections! ($\mathrm{FRR} = 31\%$). For access applications this could be the case, as the system should not allow impostors to pass. In order to decrease the FRR, the number of the attempts could be increased.

For each client, a baseline GMM was trained and tested using in both stages the same data as the proposed system. In order to have the same number of parameters for both systems, a number of 104 Gaussians have been considered for the baseline system, as the proposed system has 34 phonemes with three states each.

In order to have a cross validation of the performance obtained, each speaker model was tested against the phrases of the remaining speakers. The world model did not include the client phrases. EER for the other speakers of the dataset is also presented in table 4.

Table 4. EER [%] for all database clients

Speaker	1	2	3	4	5	6	7	8	9	10	11	Overall
Gender	F	M	M	M	F	F	F	M	M	M	M	-
Proposed system	0,0	2,3	0,6	0,0	0,3	1,6	1,33	1,0	1,3	2,0	2,6	**1,18**
GMM system	0,3	1,6	1,0	0,0	0,3	1,3	0,3	0,3	0,6	1,3	1	**0,72**

One may observe that baseline and proposed systems have similar performances. Although the baseline performs slightly better, the proposed system has the advantage of multiple usage of the speaker-dependent automatic speech recognizer that may be also applied in speech recognition tasks. The main objective of this paper was to obtain a state of the art speaker verification system using a speech recognizer.

5. CONCLUSIONS

In this paper an automatic speech verification system was proposed based on a continuous speech recognizer. In the training stage, are trained two recognizers: one speaker-dependent for the client model and one speaker-independent for the world model. Both recognizers have the same structure and number of parameters. In the verification stage, for the input phrase, they are performing a forced alignment Viterbi procedure. A normalized acoustic score is obtained and compared with a threshold for acceptance decision. The experimental results reveal that the method proposed here can obtain error rates less than 1%.

Taking into account the constraints in the search space of the forced alignment procedure, the recognition time is greatly reduced, leading to a low computation cost. Experiments were performed on a 1.5 GHz processor, the mean processing time for one phrase being smaller than 1 second (less than 0.02 real time!). This could imply low cost for implementing the verification device, but in the enrollment phase still important resources are needed to train the speaker dependent recognizer of the client from scratch. A solution for this disadvantage could be to adapt parameters of the world models to the new speaker by classical adaptation techniques such as MAP or MLLR.

REFERENCES

1. ETSI- STQ, *Aurora group: Distributed Speech Recognition*, URL: http://portal.etsi.org/fixed/Quality/SpeechRecognition.asp.
2. Speech Processing, *Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms*, URL: http://webapp.etsi.org/WorkProgram/Report_WorkItem.asp?WKI_ID=18820.

3. TAKASHI M. K. *et al.*, *Imposture Using Synthetic Speech against Speaker Verification Based on Spectrum and Pitch*, URL= citeseer.ist.psu.edu/444076.html.
4. MUNTEANU D., *Methods for Romanian Language Speech Recognition*, Ph.D. Thesis, 2006.
5. YOUNG S. *et al.*, *The HTK Book (for HTK version 3.3)*, Cambridge University Engineering Department, 2005.
6. YOUNG S., *The HTK hidden Markov model toolkit: Design and philosophy*, Cambridge University Engineering Department, doc. nr. CUED/F-INFENG/ TR.152.
7. CAMPBELL J. P., *Prolog to Speaker Recognition: A Tutorial, Proceedings of IEEE*, **Vol.** 85, No. 9, p. 1436-1462, 1997.
8. FURUI S., *Recent Advances in Speaker Recognition*, Pattern Recognition Letters, **Vol.** 18, No. 9, 1997, pp. 859-872.
9. HUANG X., ACERO A., HON H., *Spoken Language Processing, A Guide to Theory, Algorithm, and System Development*, Prentice Hall, 2001.
10. MUNTEANU D., DUMITRU O., "Romanian Language Continuous Speech Recognition by Context-Dependency Modeling", *Int. conf. DOGS 2004*, Sombor, Serbia and Montenegro, pp. 9 – 12, Sep. 2004.
11. OANCEA E., GAVĂT I., DUMITRU O., MUNTEANU D., "Continuous Speech Recognition for Romanian Language Based on Context Dependent Modeling", *Proc. of International Conference COMMUNICATIONS 2004*, pp. 221-224, **Vol.** I, Bucharest, June 2004.
12. PINTILIE C. *et. al.*, *Voiceprint Identification as a Support Against Terrorism*, National Program for Research and Development "SECURITATE 2006", No. S182.