A NOVEL TECHNIQUE FOR THE COMBINATION OF UTTERANCE AND SPEAKER VERIFICATION SYSTEMS IN A TEXT-DEPENDENT SPEAKER VERIFICATION TASK

L. Rodríguez-Liñares, C. García-Mateo

E.T.S.E. Telecomunicación
University of Vigo - SPAIN
leandro@tsc.uvigo.es, carmen@tsc.uvigo.es

ABSTRACT

In this paper we present a novel technique for combining a Speaker Verification System with an Utterance Verification System in a Speaker Authentication system over the telephone.

Speaker Verification consists in accepting or rejecting the claimed identity of a speaker by processing samples of his/her voice. Usually, these systems are based on HMM's that try to represent the characteristics of the talkers' vocal tracts [1].

Utterance Verification systems make use of a set of speaker-independent speech models to recognize a certain utterance and decide whether a speaker has uttered it or not. If the utterances consist of passwords, this can be used for identity verification purposes [2][3].

Up to now, both techniques have been used separately. This paper is focused on the problem of how to combine these two sources of information. A new architecture is presented to join an utterance verification system and a speaker verification system in order to improve the performance in a text-dependent speaker verification task.

1. INTRODUCTION

Continuous HMM (Hidden Markov Models) based systems are presently the state of the art for speaker recognition purposes [1][4]. They perform a stochastic matching that can be formulated as measuring the likelihood of a collection of vectors given models of the speakers. These vectors are obtained from the voice of the speakers and try to represent the speakers' vocal-tract characteristics during the production of distinct sounds.

Such a Speaker Recognition system does not take into account another important information present as well in the utterance: the message. In prompted-text of password based Speaker Recognition systems, the speakers are addressed to pronounce personal utterances that identify them. These utterances are matched against a set of models that represent the vocal tract characteristics of the different sounds regardless of the speaker identity with the purpose of validating the message. Besides, prompted-text systems can be improved by changing the utterances the speakers are addressed to pronounce. This prevents the systems against recordings being used by impostors trying to gain access.

If stochastic matchings of the utterances against both speaker models and phone models are performed, we obtain two probabilities: a speaker probability and a message probability. It can be expected that these probabilities are somehow uncorrelated and that the combination of them yield better results that any of them separately. The problem is how to combine the outputs of both sub-systems in order to improve the performance of the final system. In this paper we present two different methods for combining the Speaker and Utterance Verifiers in order to improve the overall performance.

The rest of this paper is organized as follows: Section 2 presents the database and the Speaker and Utterance Verifiers we used and Section 3 the architectures of the dual recognizer when the speakers share thresholds and when the thresholds are speaker-dependent. Finally, in Section 4 we present some conclusions and guidelines for further work.

2. EXPERIMENTAL CONDITIONS

2.1. The Database

The experiments were conducted using our own database, called "TelVoice" [1]. It has been designed for Speaker Recognition purposes and consists of 59 speakers with 10 telephone calls each. The recording time is variable across speakers, ranging from three weeks to more than one year.

We have made some choices about recording conditions and speech parametrization. The voice was sampled at 8KHz and off-line filtered to remove the 50 Hz electric-supply noise. Energy and 12 Mel-cepstrum coefficients were computed using a Hamming window with frame length of 25 ms and a frame period of 10 ms. Preemphasis (k=0.97) and liftering (parameter 22) were also used. First and second derivatives of the energy and the Mel-cepstra were appended to the parameters of each frame. This makes a total of 39 parameters per vector.

We conducted the experiments presented in this paper with a subset of this database consisting of 20 speakers (10 males and 10 females) with 5 sessions each one. Each session consists of four repetitions of the Spanish Identity Card number made up of 8 digits. The speakers were addressed to pronounce it naturally (digit by digit, grouping digits or as a whole, as they usually do) but always the same way across sessions.

One of the sessions was used for training models and for calculating thresholds, while the other four sessions were used for testing: three to simulate clients and one to simulate impostors. That is, the four utterances of the first three sessions where assumed as from "true" speakers (we tested each file against the true identity) and the first utterances of the fourth session were tested against the 20 possible identities of the database. This makes up a total of 260 tests of clients and 380 tests of impostors.

2.2. Independent Verifiers

The Speaker Verifier was built up by training a GMM for each speaker with one recording session (approximately 20 seconds) using a Voice Activity Detector (VAD) to identify the noise segments [1]. The GMMs are covariance-tied and the number of gaussian mixtures is 16. In the testing phase, for each verification test we normalized the obtained probability by the probabilities of 6 (3 far and 3 close) speaker's cohorts [4].

The utterance verifier is a speaker-independent speech recognizer that makes use of 25 context-independent phone models and a noise model [2][3]. The phone models consist of 3 states HMMs with 16 mixtures/state. A forced alignment between the utterance and the chain of models of the expected text is performed using the Viterbi algorithm and the segment likelihoods are normalized using the phone model of the closest competitor as antimodel. The normalized segment probabilities are accumulated and normalized by its lengths.

3. DUAL VERIFICATION

In this section, we address the problems of how to set up the thresholds in both the Speaker Verifier and the Utterance Verifier and how to combine these Verifiers to improve the overall performance.

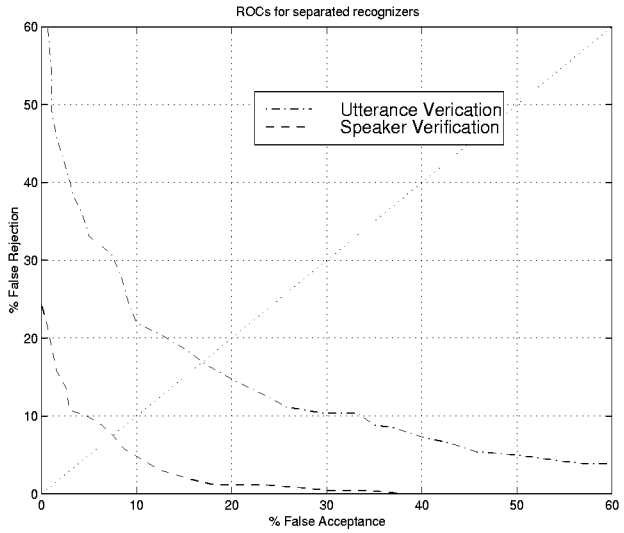

Figure 1: Receiver Operating Characteristics (ROC) curves for both recognizers with shared thresholds across speakers.

3.1. Shared Thresholds

The first possibility is to use two shared thresholds for the speakers: one for the speaker verification likelihoods and another for the utterance verification likelihoods. Varying these thresholds *a posteriori* we obtain two Receiver Operating Characteristic (ROC) curves that can be seen in figure 1.

We combined both tests in one with the criterium that both likelihoods have to be greater than their respective thresholds for a verification test to be passed. Now we have two thresholds, so the ROC is no longer a curve, but a hyperplane in a 4-dimension space. However, for representation purposes, we can create a 3-dimension figure with the evolution of the False Acceptance (FA) and the False Rejection (FR) Rates over a plane defined by the variation of the thresholds. Such representation can be seen in figure 2. The EER points of the individual sub-systems and the combined system can be seen in table 1. The difference in the FA and FR numbers are because of numerical reasons. It can be observed in the last line of this table that the improvement in the combined system relative to the Speaker Verification System is quite considerable.

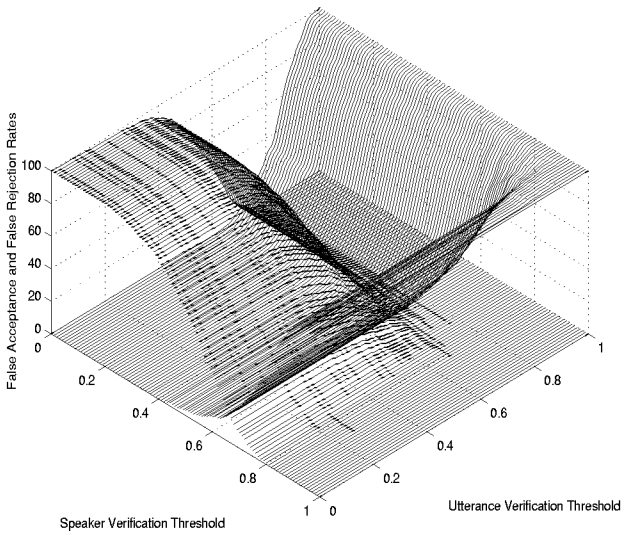

Figure 2: Surfaces of False Acceptance and False Rejection Rates for combined recognizers with shared thresholds across speakers.

Verification	False Acceptance Rate	False Rejection Rate
Speaker	7.632%	7.308%
Utterance	17.368%	16.538%
Dual	6.579% (-13.8%)	5.769% (-21.1%)

Table 1: Results obtained with Shared Thresholds and the rate improvement relative to a GMM-based Speaker Verification System.

3.2. Individual Thresholds

In case individual thresholds are used, we decided to calculate two thresholds per speaker (one per likelihood) just using the training session. We think that this is a more realistic point of view.

For each speaker, the variation of the False Acceptance and False Rejection Rates against the threshold were calculated both for the speaker verifier and for the utterance verifier. In case the False Acceptance Rate and False Rejection Rate cross, the threshold corresponding to the point where the False Rejection Rate goes to zero was taken. If the rates don't cross, the used criterium was to take the mean of the thresholds where the rates go to zero. Examples of these cases can be seen in figure 3. This figure corresponds to the Utterance Verifier for two speakers; the FA and FR Rates are represented against their normalized thresholds.With these criteria, the obtained FA and FR Rates for the Speaker and the Utterance Verifiers can be seen in the first two lines of table 2. It can be observed that, while for the FA rates the improvement is considerable, the FR rates are a little worse than with shared thresholds. That means that the values of the estimated thresholds are greater than their optimal values.

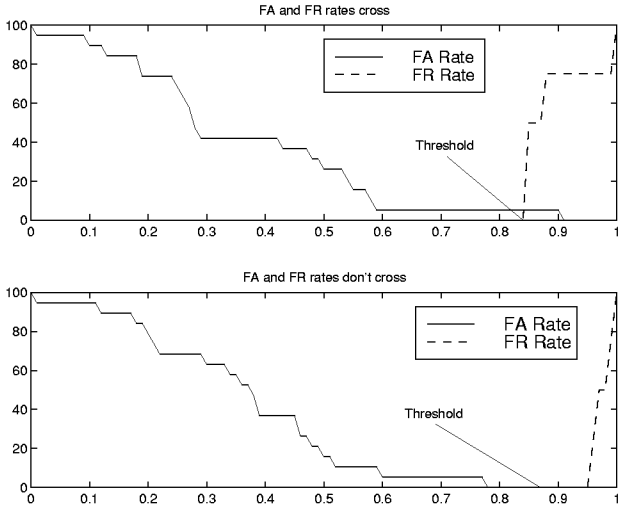

Figure 3: Example of the criteria for taking the values of the thresholds: FA and FR rates obtained with the Utterance Verifier for two speakers.

Now we have two thresholds per speaker and two likelihoods to perform a dual verification test. There are two extreme criteria to accept the claimed identity in the dual tests: to accept the speaker in case one of the probabilities exceeds its threshold (permissive test) or to demand that both probabilities exceed their respective thresholds simultaneously (restrictive test). The results obtained with these tests correspond to the last lines of table 2.

Verification	FA Rate	FR Rate
Speaker	3.684%	8.077%
Utterance	12.105%	21.923%
Dual (Permissive)	15.789%	3.846%
Dual (Restrictive)	0.000%	26.154%

Table 2: Results obtained with Individual Thresholds.

In real world applications, a system should work somewhere in between this extreme cases. The adopted solution was to build a continuous function that varies between this extremes with a control parameter α. Let's suppose that Ls and Lu are the speaker and utterance likelihoods, respectively. First, we normalize them making use of the speaker and utterance thresholds Ts and Tu:

$$\tilde{L}_x = \frac{L_x - T_x}{|T_x|} \quad x = \{s,u\}$$

A sigmoid function was applied for smoothing purposes:

$$S_x = \frac{1}{1+e^{-a\tilde{L}_x}}$$

with a=5. The scores Ss and Su take values between 0 and 1 depending on if they pass the test or they don't. Making use of these scores we implemented the restrictive and permissive tests:

$$O_{perm} = Sig_2(S_a + S_u - 0.1) - 0.5$$
$$O_{rest} = Sig_2(2S_aS_u - 0.5) - 0.5$$
$$Sig_2(x) = \frac{1}{1+e^{-bx}}$$

with b=2. Now, Operm is –0.5 when Ls<Ts and Lu<Tu and 0.5 if one of the likelihoods is greater than its correspondent test. Orest is –0.5 or 0.5 if one of the likelihoods is lesser that its threshold or if both likelihoods are greater that their correspondent threshold, respectively.

At last, the verification test is performed as:

$$V = \alpha\, O_{rest} + (1-\alpha)\, O_{perm}$$
$$V > 0 \;\Rightarrow\; Accepted$$

The variation of False Acceptance and False Rejection Rates with α can be seen in figures 4 and 5. In the latter, the horizontal lines represent the error probabilities using speaker likelihoods with individual thresholds. The working point would be placed somewhere in the area where the rates are below their respective horizontals simultaneously (α between 0.6 and 0.62 approximately).

Verification	α	FA Rate	FR Rate
Speaker	--	3.684%	8.077%
Dual 1	0.59	5.000% (+35.7%)	5.000% (-38.1%)
Dual 2	0.62	1.842% (-50.0%)	8.077% (±0.0%)
Dual 3	0.60	3.421% (-7.1%)	6.154% (-23.8%)

Table 3: Results obtained with the combined system.

In table 3, some results obtained with this system are represented. These results are compared with the system that

makes use of the speaker likelihood with individual thresholds. It can be observed that, in the best case (labelled dual 3), the improvement in the FA and the FR rates are 7.1% and 23.8%, respectively.

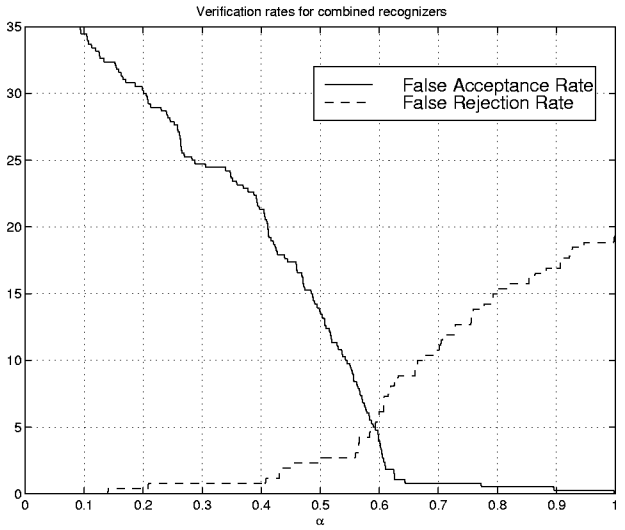

Figure 4: False Acceptance and False Rejection Rates obtained with the combined recognizer when α varies between 0 and 1.

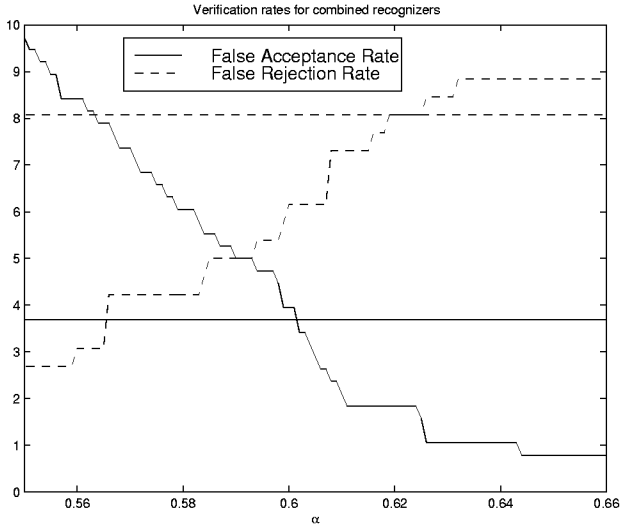

Figure 5: Zoom of figure 3 including the Verification Rates using Speaker Likelihoods with individual Thresholds.

4. CONCLUSIONS AND FURTHER WORK

For comparison reasons, we include in table 4 the results presented all along this paper.

Verification	FA Rate	FR Rate
Shared Thresholds		
Speaker	7.632%	7.308%
Utterance	17.638%	16.538%
Dual	6.579%	5.769%
Indivual Thresholds		
Speaker	3.684%	8.077%
Utterance	12.105%	21.923%
Dual 1	5.000%	5.000%
Dual 2	1.842%	8.077%
Dual 3	3.421%	6.154%

Table 4: Results obtained with the presented systems.

As it was expected, the performance of the classical GMM-based Speaker Verification system can be improved with a Dual Verification system. We think that this is a simple and efficient way of overcoming the performance limits of the GMM architectures. Besides, with such a system, the obtained architectures are:

- more configurable: as it was explained, the working point of the system is easily controlled by a parameter α instead of having to adjust the whole set of thresholds.
- more reliable: an additional level of security is added to the system as long as it is safer against the use of recordings by impostors.

Regarding future lines of investigation, if the definition of *Operm* is examined, it can be seen that its formulation is very similar to the output of a single-layer perceptron. We are obtaining promising results combining the recognizers with a Neural Network.

5. REFERENCES

1. Rodríguez-Liñares L. and García-Mateo M., "On the Use of Acoustic Segmentation in Speaker Identification", *Proc. of EuroSpeech'97* 5: 2315, 1997.
2. Li Q., Juang B., Zhou Q. and Lee C., "Verbal Information Verification", *Proc. of EuroSpeech'97* 2: 839, 1997.
3. García-Mateo C. and Lee C., "A Study on Subword Modelling for Utterance Verification in Mexican Spanish", *Proc. of 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, 614, 1997.
4. Reynolds, D. "Speaker identification and verification using Gaussian mixture speaker models", *Speech Communication* 17: 91-108, 1995.

ACKNOWLEDGMENTS

This work has been partially supported by Spanish CICYT under the project TIC96-0964-C04-02 and Xunta de Galicia