Biometric Identification through Speaker Verification over Telephone Lines

Dr. J. Gonzalez-Rodriguez, S. Cruz-Llanas and Dr. J. Ortega-Garcia
Speech and Signal Processing Group (ATVS)
DIAC-EUIT.Telecom.- Univ. Politecnica de Madrid
Campus Sur -28031 Madrid SPAIN
Email: jgonzalz@diac.upm.es
http://www.atvs.diac.upm.es

ABSTRACT

In this contribution, the identity of a remote user is verified through his voice by means of a simple telephone in order to gain access to a specific system or service. We have used state-of-the-art text-independent speaker modeling algorithms, likelihood normalization in the verification process, and channel normalization techniques. Several experiments are presented showing the negative effects of channel variability and temporal lapse between training and testing recordings, but we will show that using the appropriate parameterization, channel compensation and multisession training will allow us to obtain less than 1% of joint false acceptances and false rejections from single utterances of the speaker passport number.

INTRODUCTION(*)

The process of authentication of a person in order to gain access to a specific system or service has obtained a great increase in interest in the last years. In order to enhance the security level obtained by identification cards and personal identification numbers (PIN), a number of biometric patterns are being used, including fingerprint, hand geometry, iris, retina, and voice.

Different levels of accuracy can be achieved with these techniques, but maybe the one inducing greater problems is the human voice, due to its inherent variability. However, if we are able to cope with this variability using the appropiate modeling, updating and thresholding techniques, speech is one of the best ways of introducing biometric patterns into security systems because of the simplicity of the speech acquisition systems. Additionally, for remote applications, such as telephone banking, speech is the more appropiate way of performing secure transactions, not based on its reliability levels (we are relying today in touch-tone PINs), but in its easiness of access, because everywhere and at any time you have a telephone accessible, and moreover, the biometric measure is performed by the service provider with no need of a special biometric pattern acquisition system in the user side.

(*) This work has been supported by CICYT under project TIC-97-1001-C02-01

SYSTEM DESCRIPTION

Speaker Modeling

In this work, the speaker characteristics are obtained from MFCC (Mel-Frequency Cepstral Coefficients) vectors, including temporal information through delta and delta-delta coefficients. From these vectors, text-independent speaker verification is performed, where the speaker modeling is performed through state-of-the-art Gaussian Mixture Models (GMMs) [1][2], trained with Maximum Likelihood.

In GMM systems, each speaker model λ is given by:

$$\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\} \quad i = 1, .., M$$

with mean vector $\vec{\mu}_i$ and covariance matrix Σ_i; a gaussian mixture density is given by a weighted sum of component densities:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x})$$

where $\vec{x}$ is our L-dimensional cepstral vector, with mixture weights p_i and

0-7803-5247-5/99/$10.00 ©1999 IEEE

component densities $b_i(\vec{x})$ given by the equation:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{L/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)' \Sigma_i^{-1} (\vec{x}-\vec{\mu}_i)\right\}$$

The speaker recognition system models the speaker characteristics with a one state model per speaker with a discrete set of gaussian mixtures (M=8, M=16 or M=32) corresponding to the probabilistic distribution of the MFCC vectors obtained from the speaker database described below.

Speaker Verification

The task of automatic verification of a identity from a speaker voice is performed by the system from two different inputs: the test utterance and a claimed identity. With these inputs, the system computes the likelihood of the test utterance against the claimed model, and compares it with the claimed-speaker threshold, accepting the speaker as the correct *user*, or rejecting him as an *impostor*.

Then, as a function of the selected threshold, two different errors may appear:

- ✓ *false rejections* (FR), or type I error: a correct user is rejected.
- ✓ *false acceptances* (FA), or type II error: an impostor is accepted.

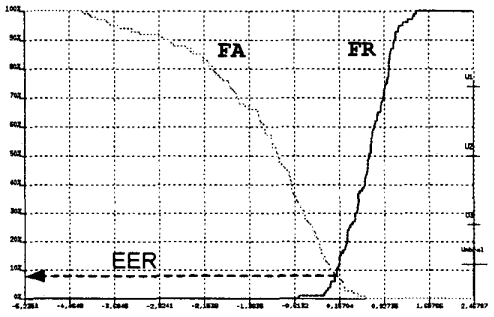

Figure 1.- False acceptance (FA) and false rejection (FR) curves as functions of the value of the possible thresholds.

In order to describe the performance of speaker verification systems, ROC (Receiver Operating Characteristic) and DET (Detection Error Tradeoff) curves are usually used [3]. However, we can obtain a simple estimate of the performance of the system from the point where the false acceptances equal the false rejections, known as the *equal error rate* (EER), which will be used in the following to describe and compare the performance of the system in different conditions.

Likelihood Normalization

One of the biggest problem in speaker verification is the intra-speaker variability in the likelihood scores for different repetitions of the same utterance. Then, instead of using fixed thresholds for every speaker, we can make use of likelihood-ratios, obtaining in this way an utterance-dependent threshold. Several possibilities for this likelihood ratios exist, from cohort-speakers to universal background models. In this work, we have used the following equation for the likelihood ratio [4]:

$$\log L = \log p(X \mid \lambda = \lambda_q) - \max_{\substack{j \\ j \neq q}} \log p(X \mid \lambda = \lambda_j)$$

where the test-utterance is matched against every model in the database, and the one obtaining the highest likelihood is used to normalize the score obtained against the claimed model.

Channel Normalization

As we will see later, one of the degrading effects that mostly influences the performance of automatic speaker recognition systems is the channel variability between recordings. As we are using cepstral coefficients, this influence of the channel consists in a channel-dependent additive cepstral vector, different for different communications, but constant along each utterance.

In these conditions, performing channel normalization through CMN (Cepstral Mean Normalization) or RASTA log-spectral filtering [5][6], will improve the verification results, performing channel-independent comparisons between utterances. In CMN the mean of the cepstral vectors is estimated and subtracted in order to high-pass filter the original cepstral coefficients:

$$y[n] = x[n] - \frac{1}{N}\sum_{i=1}^{N} x_i[n]$$

RASTA processing of speech again high-pass filter the cepstral coefficients with the following IIR difference equation:

$$y[n] = x[n] - x[n-1] + 0.97\, y[n-1]$$

removing in a different way the estimated mean of each coefficient.

EXPERIMENTS

Speaker Database

In order to perform a real evaluation of the system, we have used *TelVoice* [7], a multisession telephone-speech database recorded at University of Vigo (Spain). It includes at least five recording sessions in different actual phone calls from each one of 59 speakers (39 male and 20 female). Each recording session consists in digit strings, fixed phrases, 15 seconds of spontaneous speech and 4 repetitions of their passport numbers, which will be used as passwords for each of the users of the verification system.

Evaluation

The described system has been tested with a subset of 20 speakers (11 male and 9 female) from the TelVoice database, using just the four repetitions of the passport number in each one of five recording sessions. Each 30 ms. frame of speech is parameterized with 12 MFCC+logE with their respective velocity (Δ), which gives a 26 coefficient vector. Speaker models will be obtained for different number of mixtures, in order to obtain the optimum size of the model as a function of the amount of training data available.

In each test, we will characterize the performance of the system with its EER (equal error rate) from the computed false acceptance and false rejection curves.

Multisession Variability

The first experiment shows the influence of the time lapse between training and testing sessions. In this case, we train the speaker models with speech from just one session, as we can see in the following table.

Test #	P11	P12	P13	P14	P15
Training session	1	2	3	4	5
Test sessions	2,3, 4,5	1,3, 4,5	1,2, 4,5	1,2, 3,5	1,2, 3,4

Table 1.- Description of the training and testing material for the analysis of the multisession effects.

In table 2 we can see the results of this experiments, for different number of mixtures in the speaker model. Different effects are present together here, as telephone-multisession (time lapse and different channel), and small amount of speech material to train the models. As a consequence, the resulting *equal error rates* (EER) are quite high, and a real working system would not be accepted with this performance.

EER (%)						
M	P11	P12	P13	P14	P15	Average
8	20.3	16.5	20.1	15.0	21.5	18.7
16	19.0	15.4	19.5	15.2	25.7	19.0
32	20.8	16.0	20.3	17.4	26.1	20.1

Table 2.- Equal error rates (EER) when we train the models with speech from one session and test the system with the other four sessions.

In order to reduce these effects, we can perform multisession training, which means to train the models with speech from different sessions, as we can see in tables 3 and 4.

Test #	P21	P22	P23	P24
Training	1	1,2	1,2,3	1,2,3,4
Test	2,3,4,5	3,4,5	4,5	5

Table 3.- Description of the training and testing material for multisession training.

EER (%)				
M	P21	P22	P23	P24
8	20.3	15.8	10.4	8.1
16	19.0	16.2	9.3	7.5
32	20.8	15.5	8.7	6.7

Table 4.- Equal error rates (EER) for multisession training.

We can observe from table 4 that the EER's of the system are greatly reduced, which is justified by the fact that we are including in the model characteristics that remain constant in different sessions, the channel charactesistics have being averaged, and finally, we are including more speech from the speakers in the models.

Channel Variability

We can also analyze the advantages of using channel normalization techniques for these experiments. If we repeat the former experiments (P1# and P2#), but training different models for the different normalization techniques into analysis, we obtain the following verification results:

EER (%)							
Ch.N.	M	P11	P12	P13	P14	P15	Av.
None	8	20.3	16.5	20.1	15.0	21.5	18.7
	16	19.0	15.4	19.5	15.2	25.7	19.0
	32	20.8	16.0	20.3	17.4	26.1	20.1
CMN	8	5.8	4.7	10.0	7.2	13.0	8.1
	16	6.4	5.3	10.3	7.7	13.0	8.5
	32	8.9	6.0	13.6	10.5	15.3	10.9
RASTA	8	5.1	4.2	7.8	5.9	10.7	6.8
	16	4.4	3.2	8.1	6.3	10.8	6.6
	32	5.2	3.8	9.4	8.6	12.5	7.9

Table 5.- Effect of channel normalization with single-session training.

EER (%)					
Ch.Norm.	M	P21	P22	P23	P24
None	8	20.3	15.8	10.4	8.1
	16	19.0	16.2	9.3	7.5
	32	20.8	15.5	8.7	6.7
CMN	8	5.8	3.6	3.3	2.7
	16	6.4	3.1	2.8	1.9
	32	8.9	3.0	2.2	1.5
RASTA	8	5.1	4.0	2.4	3.1
	16	4.4	2.6	1.8	2.1
	32	5.1	2.0	1.4	1.0

Table 6.- Effect of channel normalization with multi-session training.

The advantages of channel normalization are clear from these experiments, where we have obtain reductions greater than 10 absolute points for single-session training, and EER's close to 1% for multisession training.

Speech Parameterization

We can also try to improve the speaker characteristics extraction with different types of parameterization. For this set of experiments, we have selected a 'good' configuration from the previous experiments, and in this way we will use RASTA channel normalization, and speaker models with 16 mixtures (M=16).

The different combinations include:

- ✓ LPCC.- Linear Prediction Cepstral Coefficients.
- ✓ MFCC.- Mel-Frequency Cepstral Coefficients.
- ✓ LogE.- logarithm of the energy of the frame
- ✓ Δ.- velocity of the coefficient (stands for how the coefficient varies with time)
- ✓ ΔΔ.- acceleration of the coefficient (stands for how the coefficient velocity varies with time)

M=16 RASTA	P31	P32	P33	P34
Training	1	1,2	1,2,3	1,2,3,4
Recog.	2,3,4,5	3,4,5	4,5	5

Table 7.- Description of the training and testing material for comparison of different parameterization alternatives.

EER (%)				
Parameterization	P31	P32	P33	P34
12 LPCC + 12 ΔLPCC	10,0	4,4	3,8	2,4
12 LPCC+logE+ 12 ΔLPCC+ΔlogE	7,4	3,3	3,5	2,6
12 LPCC+ 12 ΔLPCC+ 12 ΔΔLPCC	11,5	5,2	4,0	3,2
12 MFCC + 12 ΔMFCC	2,7	1,8	1,3	2,1
12MFCC+logE+ 12ΔMFCC+ΔlogE	4,3	2,5	1,8	2,1
12MFCC+ 12ΔMFCC+ 12ΔΔMFCC	1,6	1,5	1,3	0,9

Table 8.- Verification results for different parameterization alternatives.

From table 8 we observe that practical systems can be implemented with these techniques, with EERs close to 1% in several cases. Moreover, the EER (FA=FR) is not usually the optimum working point of a verification system, because usually we will look for high security systems (higher thresholds, FA<<FR) or systems that not disturb the user (lower thresholds, FR<<FA), obtaining in both ways systems that meet our requirements sctrictly (e.g. FA<0.5% or FR<0.5%).

CONCLUSIONS

To conclude, we have presented a system where less than 1% of false acceptances and false rejections are obtained together. If we consider that nowadays we are using home banking services relying just in PIN numbers, speaker verification through existing telephone links appears as the best way to include biometric identification for remote access with minimum cost, because the verification system will be installed in the receiver and no extra biometric pattern reading interface will be needed, obtaining universal access to the designed system.

REFERENCES

[1] D. A. Reynolds and R.C. Rose, "Robust Text-independent Speaker Identification using Gaussian Mixture Speakers Models", IEEE Trans. on Speech and Audio Procesing, Vol. 3, No. 1, January 1995.

[2] J. Ortega-Garcia, J. Gonzalez-Rodriguez et al. "AHUMADA: A Large Speech Corpus in Spanish for Speaker Identification and Verification", Proc. of the IEEE Intnl. Conf. on Acoustics, Speech and Signal Processing (ICASSP'98), pp. 773-776, Seattle (USA), 1998.

[3] G.R. Doddington, "Speaker Recognition Evaluation Methodology – An Overview and Perspective", Proc. of ESCA-IEEE Workshop on Speaker Recognition and its Commercial and Forensic Applications (RLA2C), pp. 60-66, Avignon (France), 1998

[4] A. Higgins et al., "Speaker Verification Using Randomized Phrase Prompting", Digital Signal Processing (Academic Press), vol. 1, pp. 89-106, 1991.

[5] J.C. Junqua and J.P. Haton, *Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Norwell, MA., 1996.

[6] J. Gonzalez-Rodriguez and J. Ortega-Garcia, "Robust Speaker Recognition through Acoustic Array Processing and Spectral Normalization", Proc. of the IEEE Intnl. Conf. on Acoustics, Speech and Signal Processing (ICASSP'97), vol. 2, pp. 1103-1106, Munich, Alemania, Abril 1997.

[7] L. Rodríguez-Liñares and C. García-Mateo, "On the Use of Acoustic Segmentation in Speaker Identification", Proc. Eurospeech'97, pp. 2315-2318, Rodas (Grecia), 1997.