IDENTIVOX©: A PC-WINDOWS™ TOOL FOR TEXT-INDEPENDENT SPEAKER RECOGNITION IN FORENSIC ENVIRONMENTS

Joaquín GONZALEZ-RODRIGUEZ[1], Javier ORTEGA-GARCIA[1],
J. J. LUCENA-MOLINA[2]

[1] *Speech and Signal Processing Group, Department of Audio-Visual and Communication Engineering, Technical University, Madrid, Spain*

[2] *Laboratorio de Acustica e Imagen, Servicio de Policia Judicial, Direccion General de la Guardia Civil, Madrid, Spain*

ABSTRACT: A state-of-the-art Gaussian-Mixture-Modelling (GMM) text-independent speaker recognition system has been developed. The system, perfectly suited to the Bayesian approach for Forensic Speaker Recognition, works in a user-friendly platform (Win95/98/NT/2000/me) in a fully user-configurable environment. Speaker modelling, identification and verification are the main functions of the system. The system optionally performs channel normalisation and/or likelihood normalisation, the latter through a Universal Background Model. The work of the system user is organised in sessions (files), where the user can configure, save, open, and print reports for each individual session. User-selected speech audio files, parameterised into LPCC or MFCC, are used to train the speaker models, and test speech segments are evaluated against the selected set of candidates (identification) or against a claimed model (verification). Some excellent results are provided with different speech databases. A free English-demo version for evaluation of the system is available to forensic and research institutes writing to identivox@atvs.diac.upm.es.

KEY WORDS: Voice identification; Speaker recognition; Forensic acoustics.

Problems of Forensic Sciences, vol. XLVII, 2001, 246–253
Received 13 November 2000; accepted 15 September 2001

INTRODUCTION

As a result of the continuous collaboration established between the Speech and Signal Processing Group ("Área de Tratamiento de Voz y Señal" in Spanish, ATVS) from Universidad Politécnica de Madrid (UPM), and the Acoustics and Image Laboratory of the Spanish Guardia Civil, a Windows™ based tool has been developed for text-independent voice identification in forensic conditions.

The basic technology is our own implementation of Gaussian Mixture Models (GMM) [7] and other well-known speech-related techniques, which

have been shown as the best present solution to the problem of text-independent speaker recognition [3], as have been shown in last NIST evaluations [2].

The paper is organised as follows. After a short introduction to the IdentiVox software and its components, the different windows and options are shown, to conclude with some experiments and results obtained by the Speech and Image Laboratory of Guardia Civil with our system.

IDENTIVOX 2000

The IdentiVox 2000 software is a multitask MDI (MultiDocument Interface) Windows application developed with Microsoft Visual C++. We have developed a classes library, programmed in ANSI C++ intended to the development of automatic speaker recognition (and also other biometric) applications.

These classes can be divided into three levels, as can be seen in Figure 1:

1. Basic classes: basic signal processing functions and input/output communications.
2. Specific classes: speaker recognition functions (modeling, identification, threshold establishment and verification).
3. Session classes: those are the communication interface with the specific classes easing the development of applications.

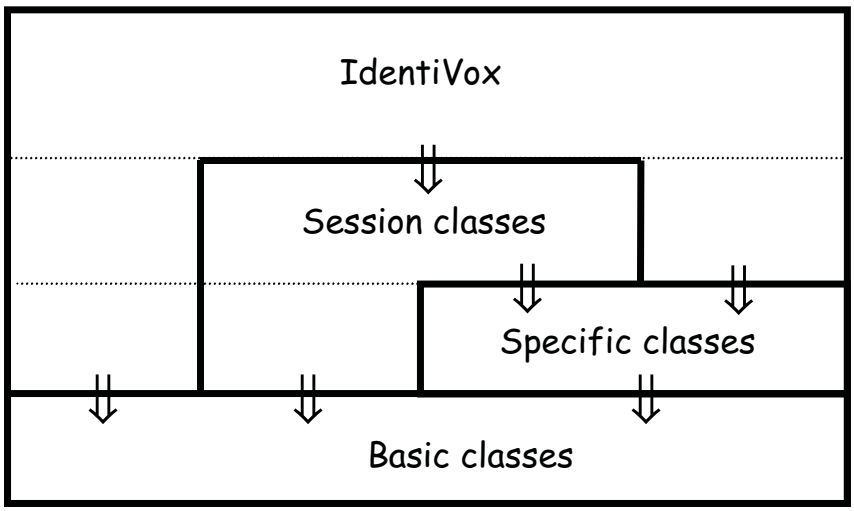

Fig. 1. IdentiVox software structure over ATVS classes library.

We can sum up the main characteristics of the IdentiVox 2000 system as the following:

- Fully visual environment: all functions, as speech parameterisation from "wav" files, speaker modelling, and identification and verification tests are performed in usual Windows™ actions.
- Working sessions: all the work related to a session (typically a forensic case or a system validation/optimisation experiment) is saved in a single file, which can be stored or later opened for further work in the same case or experiment. Additionally, full reports of every action in a session (properties of a session, audio files used to build each model, audio files used in tests, results, etc.) are automatically saved to text files, which can be printed or imported to a word processor.
- Speech parameterisation: in a fully configurable environment, the user can select all the options for parameterisation, as number and type of coefficients (LPCC – Linear Predictive Cepstral Coefficients, MFCC – Mel-Frequency Cepstral Coefficients), window length, windows type (rectangular, hamming, hanning), or the optional use of Δ (velocity) and $\Delta\Delta$ (acceleration) coefficients for co-articulation effects. Channel compensation techniques (CMN, RASTA) [5] can be optionally used to cope with different recording channels (e.g., different telephone calls) in training and test conditions.
- Speaker modelling: text-independent speaker models are obtained with Maximum Likelihood (ML) training of Gaussian Mixture Models (GMM), where the number of Gaussian mixtures of the model can be selected by the user, which must be properly selected as a function of the training-speech length (typical values are from 8 to 64 or 128 mixtures). The models are trained with audio files selected by the user in a usual Windows Explorer action.
- Speaker identification: once speaker models are available, we can perform identification tests of audio files with a selection of competing models. Results are provided visually, with a highest-likelihood ordered list of models with their correspondent likelihood. A results text file is also automatically generated or extended with each new experiment.
- Threshold establishment: in order to perform speaker verification, a threshold is needed for each speaker model. As no hard decision is needed in forensic acoustics, we work by default with three different thresholds per model (permissive, EER – equal error rate, and restrictive). For the estimation of the False Acceptance (FA) and False Rejection (FR) curves of each model, different options are available (same set of impostor files for all speakers, different sets for each speaker, and impostors different or the same and the remaining speakers of the

session). The FA and FR curves of each model, jointly with its thresholds (which can be graphically modified), can be optionally seen after their computation.

– Speaker verification: once speaker models and thresholds are available, verification tests can be performed directly. In order to improve the robustness of the verification process, likelihood normalisation can be optionally used, through the use of a Universal Background Model, obtained from the Ahumada speech database [6]. Single or multiple audio files can be tested with one or multiple speaker models. Different graphical results are shown for each pair audio file-speaker model.
– Additional tools: two extremely useful tools are included with the IdentiVox software, consisting in a speech splitter, which allows the user to obtain length-defined (optionally overlapped) files from a longer one (or from a list of audio files), and the audio format converter, which allows easy conversion from any audio format to "wav" mono files, which is the input audio format to the system.

A 30-days free English-demo version for evaluation of the system is available to forensic and research institutes under application to the main author of this paper, or directly to the email address identivox@atvs.diac.upm.es.

Finally, we have to note that IdentiVox 2000 is perfectly suited for performing semi-automatic interpretation of the evidence in the Bayesian approach [1], taking care of the design of the experiment and with a simple recombination and computation over IdentiVox 2000 likelihood outputs. However, a "Bayesian" version of IdentiVox will be available through year 2001, providing likelihood ratios just from the test speech (the "criminal" recording), some reference speech (the suspected person speech), and the selection of some reference population. The software will include tools for reference population management and selection (modelling, labelling, etc.), which will allow obtaining reference populations in each language, dialect, or any specific user defined subset, from speech databases.

SYSTEM DESCRIPTION

From now on, we are going to show different windows and options of the system in order to give an idea of its graphical appearance and possibilities.

In Figure 2 we can observe the different bar and menu tools, with three simultaneous sessions opened, with a just finished modelling process in the active session.

Figure 3 shows different options in the definition of a new session. Of course, once defined, they cannot be changed in the same session.

Finally, Figure 4 shows an example of rejection in a verification test, where the False Acceptance and False Rejection curves of the speaker are observed together with its thresholds and the likelihood of the incoming speech, which is rejected.

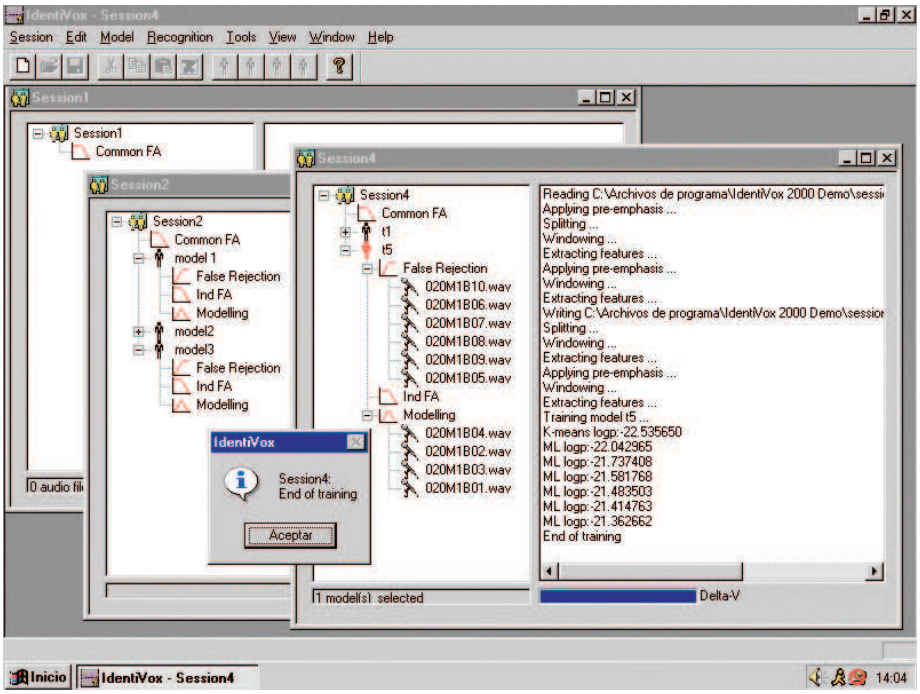

Fig. 2. IdentiVox workspace with three simultaneous working session opened.

We can simultaneously observe the FA and FR curves for a speaker, its three associated thresholds, and the likelihood of test speech with this model.

EXPERIMENTS AND RESULTS

The speaker recognition performance of Gaussian Mixture Models has been extensively shown [2, 3, 7], as happens with our own implementation of this technology [4, 6]. As an example, IdentiVox obtain joint values of FA and FR smaller than 0.5% for single recording-session speech. In order to know the performance of IdentiVox in much more complex situations, we will report a simple example.

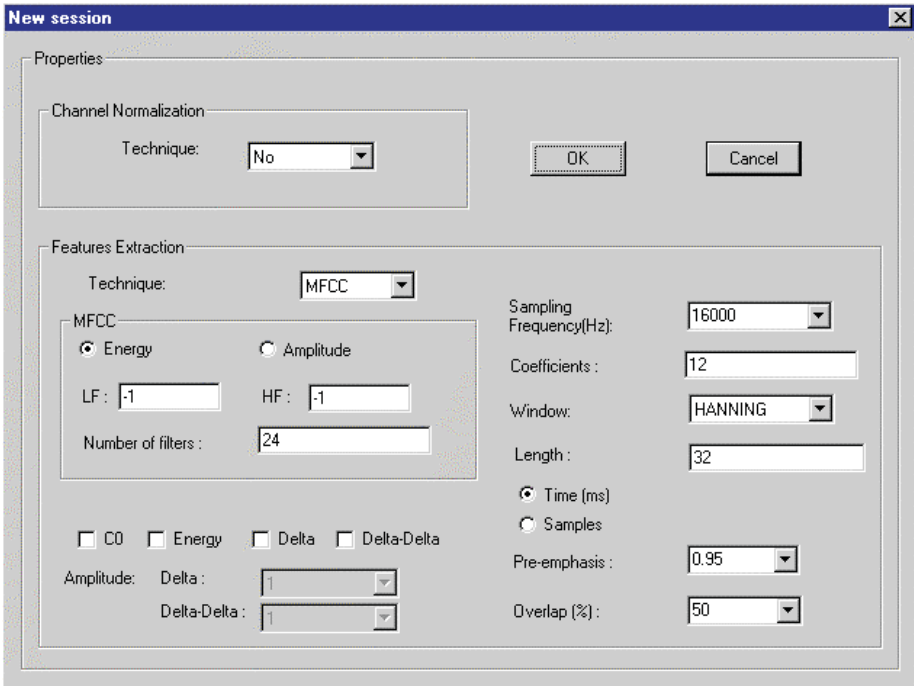

Fig. 3. User configurable options in a new session.

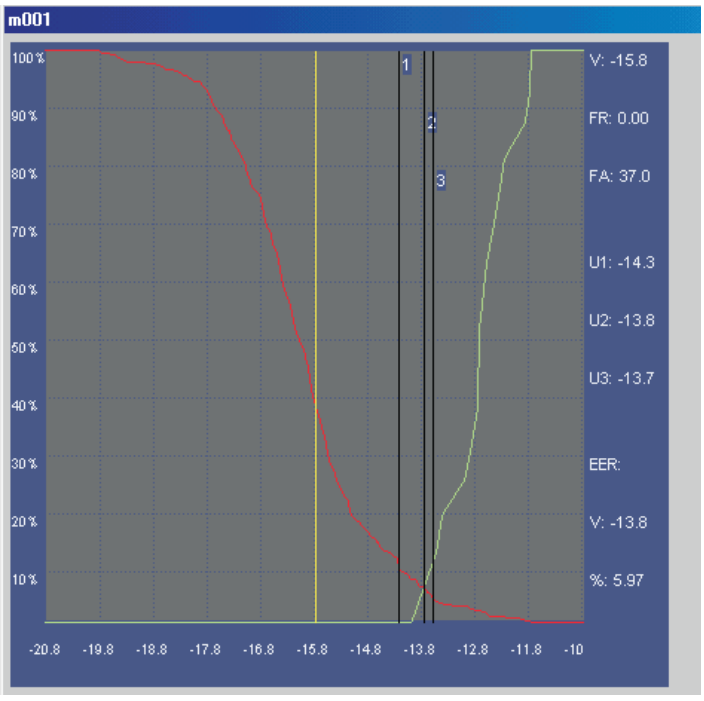

Fig. 4. Example of rejection in a verification test (white line at –15.8).

Guardia Civil has performed with IdentiVox 2000 the test we are going to report with speech data from the Ahumada database [6] (this Ahumada database has been selected by NIST in order to test in Spanish every speaker recognition system competing in NIST 2000 and 2001 evaluations).

In this experiment, 50 speakers have been randomly selected out of 103 male speakers from the telephone subcorpus of Ahumada. Each speaker performs three different telephone calls, separated two weeks and two months respectively. Speaker models are obtained with 30 seconds of spontaneous telephone speech from the fist and second recordings sessions (total of 60 seconds). The test speech is obtained from three 10-seconds trials from the spontaneous telephone speech of the third recording session.

Every test speech file is tested with every model, which gives a total number of 25500 verification trials with real multisession telephone speech. In this situation, for a forensic working point, the system jointly obtains a 0.3% of False Acceptance and a 15% of False Rejection.

CONCLUSIONS

In this paper, a text-independent speaker recognition Windows-based system has been described. Using the IdentiVox tool, as has been shown, it is extremely easy to use state-of-the-art technology in speaker identification and verification. Excellent results with these techniques have been reported and referenced. Moreover, the possibility of performing Bayesian interpretation of the evidence with this tool gives it an added value. The IdentiVox software has been extensively and in-depth tested by the Audio and Image Laboratory of Spanish Guardia Civil, with excellent performance shown in simulated and real tests. In case you are interested in its evaluation, please email me to identivox@atvs.diac.upm.es.

ACKNOWLEDGEMENTS:

Authors wish to thank all people from Speech and Signal Processing Group (ATVS) of the Universidad Politecnica de Madrid and Speech and Image Laboratory of Guardia Civil, and remark specially those directly involved in the development of the IdentiVox tool, namely Marta Garcia-Gomar and Oscar Garcia-Ledesma from ATVS-UPM, and in its extensive testing and improvement suggestions, specially Juan-Jesus Diaz-Gomez from Guardia Civil.

References:

1. Champod C., Meuwly D., The inference of identity in forensic speaker recognition, ESCA Workshop on Speaker Recognition and its Commercial and Forensic Applications, RLA2C, Avignon 1998, pp. 125–134.
2. Doddington G. R., Przybocki M. A., Martin A. F. [et al.], The NIST speaker recognition evaluation – overview, methodology, systems, results, perspective, *Speech Communication* 1992, vol. 31, pp. 225–254.
3. Furui S., An overview of speaker recognition technology, ESCA Workshop on Automatic Speaker Recognition, Martigny 1994, pp. 1–9.
4. González-Rodríguez J., Cruz-Llanas S., Ortega-García J., Biometric identification through speaker verification over telephone lines, Proceedings of IEEE Carnahan Conference on Security Technology, Madrid 1999, pp. 238–242.
5. González-Rodríguez J., Ortega-García J., Robust speaker recognition through acoustic array processing and spectral normalization, IEEE Intl. Conf. on Acous. Speech and Signal Proc., ICASSP-97, Munich 1997, pp. 1103–1106.
6. Ortega-Garcia J., González-Rodríguez J., Marrero-Aguiar V., Ahumada: A large speech corpus in Spanish for speaker characterization and identification, *Speech Communication* 2000*,* vol. 31, pp. 255–264.
7. Reynolds D., A Gaussian mixture modelling approach to text-independent speaker identification, Ph.D. Thesis, Georgia Institute of Technology 1992.