ELSEVIER

Speech Communication 31 (2000) 255–264

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

AHUMADA: A large speech corpus in Spanish for speaker characterization and identification ☆

Javier Ortega-Garcia [a,*], Joaquin Gonzalez-Rodriguez [a], Victoria Marrero-Aguiar [b]

[a] *Dpt. Ingenieria Audiovisual y Comunicaciones, EUIT Telecomunicacion, Universidad Politecnica de Madrid, Ctra. Valencia km 7, 23031 Madrid, Spain*

[b] *Dpt. Lengua Española, Universidad Nacional de Educacion a Distancia, Madrid, Spain*

Received 12 August 1998; received in revised form 10 March 1999

Abstract

Speaker recognition is an emerging task in both commercial and forensic applications. Nevertheless, while in certain applications we can estimate, adapt or hypothesize about our working conditions, most of the commercial applications and almost the whole of the forensic approaches to speaker recognition are still open problems, due to several reasons. Some of these reasons can be stated: environmental conditions are (usually) rapidly changing or highly degraded, acquisition processes are not always under control, incriminated people exhibit low degree of cooperativeness, etc., inducing a wide range of variability sources on speech utterances. In this sense, real approaches to speaker identification necessarily imply taking into account all these variability factors. In order to isolate, analyze and measure the effect of some of the main variability sources that can be found in real commercial and forensic applications, and their influence in automatic recognition systems, a specific large speech database in Castilian Spanish called AHUMADA (/aumáda/) has been designed and acquired under controlled conditions. In this paper, together with a detailed description of the database, some experimental results including different speech variability factors are also presented.

Keywords: Speech databases; Speaker characterization; Speaker recognition

1. Introduction

Speaker recognition is a biometric characterization process aimed at the identification of people by their voices. Fingerprint or iris analyses are good examples of other biometric approximations to person identification, where the test sample is directly matched with the known pattern. However, voice identification must be accomplished from a different point of view, in an analogous way to face recognition or graphological analysis of handwriting, as signal variability (written signs, facial features or speech characteristics) incorporates to the identification process an additional level of complexity (Champod and Meuwly, 1998).

In this context, coping with real commercial and forensic recognition implies dealing with

☆ This work has been supported by CICYT under Project TIC97-1001-C02-01.

* Corresponding author. Tel.: +34-91-3367796; fax: +34-91-3367784.

E-mail address: jortega@diac.upm.es (J. Ortega-Garcia).

0167-6393/00/$ - see front matter
PII: S0167-6393(99)00081-3

speech variability (Acero, 1993; Junqua and Haton, 1996). Regarding speaker identity, several factors of variability must be taken into account:

- Peculiar intra-speaker variability (manner of speaking, age, gender, inter-session variability, dialectal variations, emotional condition, etc.).
- Forced intra-speaker variability (Lombard effect, external-influenced stress, cocktail-party effect).
- Channel-dependent external influences (kind of microphone, bandwidth and dynamic range reduction, electrical and acoustical noise, reverberation, distortion, etc.).

Consequently, delimiting the problem of speech variability, together with analyzing the quantitative results of speaker recognition systems will lead to an integral and comprehensive approach to commercial and forensic speaker recognition.

Following this perspective, a speaker recognition-oriented large database (Boves et al., 1994; Godfrey et al., 1994; Naik, 1994; Gibbon et al., 1997) called AHUMADA, [1] has been designed and acquired, involving 104 male speakers. The resulting data comprises more than 15 GB of recorded material (Ortega-Garcia et al., 1998), and incorporates several speech variability factors.

The present contribution is organized as follows. Section 2 describes the AHUMADA speech corpus, enumerating the designed tasks, distribution of ages and time interval between sessions. Section 3 describes the technical features related to AHUMADA database: microphones and audio equipment, room acoustics and signal-to-noise ratio (recorded and enhanced), and speech intelligibility. Section 4 describes the automatic speaker verification system employed to validate the database. In Section 5, several speaker verification experiments are presented, analyzing the effect of some variability factors. Section 6 describes some of the database extensions that are being considered, and some perspectives regarding forensic applications of speaker verification.

[1] In honor of the founder of the Guardia Civil Corps, the Duke of Ahumada.

2. AHUMADA large speech corpus

2.1. Design of the spoken tasks

The speech corpus has been designed regarding many sources of variability, allowing us to focus on them and to study their underlying effects in speaker recognition systems. Some examples included in AHUMADA corpus are:

- In situ recordings and telephone speech.
- Read texts at different speech rate.
- Read speech versus spontaneous speech.
- Different microphones and telephone handsets.
- Inter-session variability in six different recording sessions.
- Dialectal variations of speakers (which may be even different for one particular speaker when reading or naturally speaking).
- Fixed utterances for all speakers through all sessions versus specific utterances for each speaker in each session.

In order to obtain the referred intra-speaker variability factors, the enrolled speakers were requested to utter the following:

(a) 24 isolated digits, discarding the first and the last two of them due to prosodic considerations. The remaining 20 digits consist in two repetitions of isolated digits (from 0 to 9).
(b) 10 digit strings consisting of 10 digits each, being the first five strings identical to all speakers through all recording sessions, and the last five strings specific for each speaker for all sessions.
(c) 10 phonologically and syllabically balanced phrases of 8–12 word length. These utterances were identical to all speakers through all sessions.
(d) One phonologically and syllabically balanced text, of about 180 words (more than 1 minute of duration), read at a normal speaking rate. This text was fixed for all speakers through all sessions.
(e) Two repetitions of the previous fixed text, asking the speakers to read it at a fast and at a slow speaking rate. (This task was only requested in sessions 1, 3 and 5, where in situ studio recordings were accomplished.)
(f) One specific text, different from speaker to speaker and from session to session, for each

speaker. This text was randomly selected from novels and newspapers, and at least 1 minute of this kind of speech is available.

(g) More than 1 minute of spontaneous speech, asking every speaker to narrate something familiar to them, avoiding long pauses and hesitations, in a descriptive manner. Some paintings and pictures were available, and issues like "describe your last holidays", "describe the place where you live/were born", etc., were also suggested.

2.2. Phonological and syllabic balance

Tasks (c) and (d) have been specifically designed in order to reproduce the frequency of occurrence of phonemes and syllabic schemes mostly found in spoken Castilian Spanish. The selected lexicon corresponds to the most usual in Spanish. The 'standard' frequency of occurrence (from now on called "Reference") used in the design phase was measured over an oral corpus of more than 20 000 words (Juilland and Chang-Rodriguez, 1969; Quilis and Esgueva, 1980; Guerra, 1983). In task (c), the total number of phonemes is 409, while the correlation coefficient (Pearson test) between Spanish standard phonological occurrence and the designed utterances was 0.9966. In the same task, the total number of syllables was 185 with a syllabic correlation coefficient of 0.9963.

In task (d), a fixed text for all speakers with about 180 words, the total number of phonemes is 712. The correlation coefficient between Spanish standard phonological occurrence and the designed text was 0.9988. Moreover, the total number of syllables in it was 305, with a correlation coefficient in this case of 0.9960. In both tasks, the level of significance is 0.001 (the maximum attainable). For tasks (c) and (d), Figs. 1–3 show frequency of occurrence of phonemes, syllabic groups and stress patterns, respectively.

2.3. Distribution of ages

In order to determine an adequate age distribution of speakers in the database, sociological

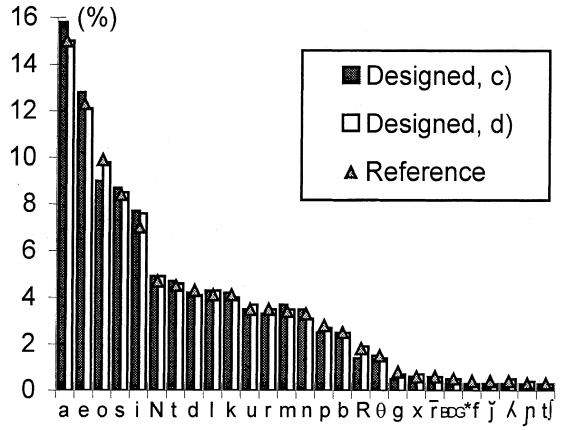

Fig. 1. Frequency of occurrence (%) of phonemes in designed tasks (c) and (d) compared to "Reference" distribution. BDG* stands for cumulative ocurrence of /B/, /D/ and /G/.

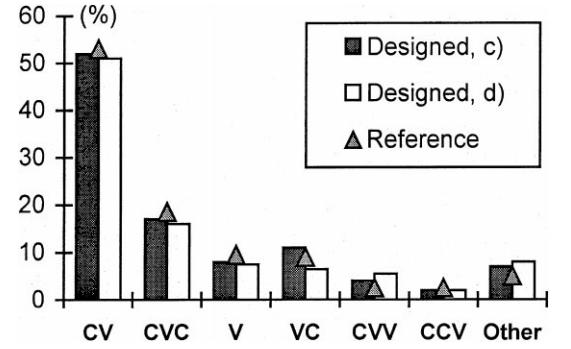

Fig. 2. Frequency of occurrence (%) of syllabic groups in designed tasks (c) and (d), compared to Reference distribution.

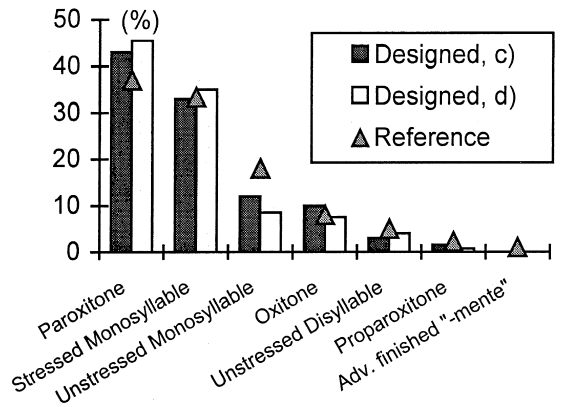

Fig. 3. Frequency of occurrence (%) of stress patterns found in tasks (c) and (d), compared to Reference distribution.

implications of technology should be taken into account, as equi-distribution of ages may not respond to a real age distribution of users in a specific commercial application. On the other hand, in forensic applications criminals are also unequally distributed in age.

In our case, age distribution of speakers in AHUMADA database fits real police data of people under arrest. This is the reason why more weight has been applied to the range of ages between 28 and 42 years, as Fig. 4 shows.

2.4. Time interval between sessions

As inter-session variability is a crucial factor to cope with in speaker recognition-oriented databases, at least a time interval separation of 11 days between equivalent sessions (on one side, microphone sessions 1, 3 and 5, and on the other side, telephone sessions 2, 4 and 6) was disposed.

Recordings began in June 1997, with microphone session 1. Fig. 5 shows time intervals between in situ (microphone) sessions. In relation to the first telephone session (session 2), 73% of recordings were done within 15 days interval from session 1. Specifically, 36% were accomplished the same day of session 1, with a maximum time interval (100% of recordings) of 40 days, where time intervals between telephone sessions are shown in Fig. 6.

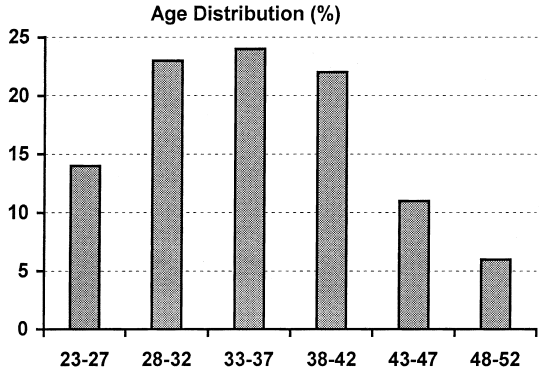

Fig. 4. Age distribution of the male population of AHUMADA.

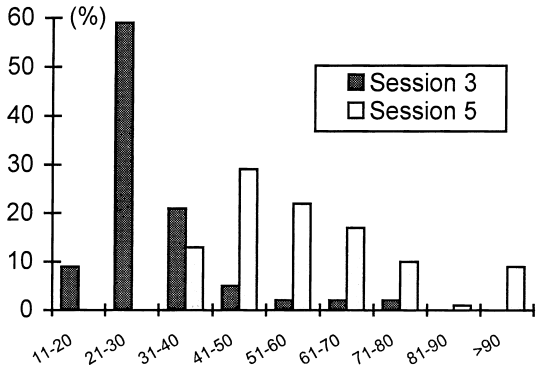

Fig. 5. Time interval of microphone sessions 3 and 5 related to first microphone session (session 1). Results are presented in intervals of 10 days, and in % referred to the total number of recordings within each session.

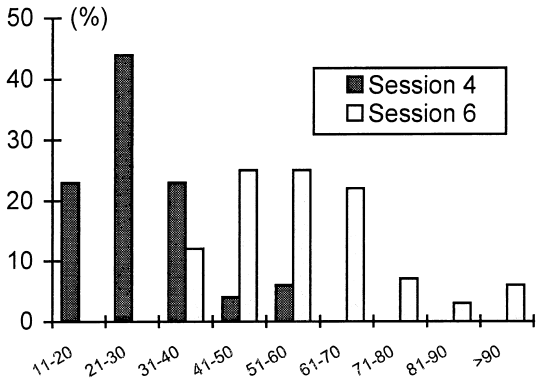

Fig. 6. Time interval of telephone sessions 4 and 6 referred to first telephone session (session 2). Results are presented in intervals of 10 days, and in % related to the total number of recordings within each session.

3. Speech acquisition

In this section, the most relevant technical features specified in the acquisition of AHUMADA speech corpus are presented. The description of these technical features will give a complete idea about the audio conditions and characteristics of AHUMADA. The description comprises the type of microphones and audio equipment used; the acoustics of the recording room (basically, reverberation time and equivalent noise level); the signal-to-noise ratio of the signal acquired; the high-pass discrete-time filter used to remove

the low-frequency noise components; and, finally, the speech intelligibility measured in the recording room.

3.1. Recording microphones and audio equipment

As it has been previously mentioned, six recording sessions were established. Sessions 1, 3 and 5 were in situ recorded in a quiet studio-like room and supervised by a trained operator. In each of these in situ recordings, two different input channels were simultaneously used: in one of them, the same microphone was used for all sessions; in the other, different microphones were used from session to session.

The notation used to specify both microphones in each case is MIC*n*_1 and MIC*n*_2, where *n* corresponds to one of the three possible sessions. The list of the used microphones is the following:

- MIC1_1, MIC3_1 and MIC5_1 correspond to the same microphone, namely SONY ECM-66B, lapel unidirectional electret type, at about 10 cm from the speaker mouth.
- MIC1_2 is an AKG D80S dynamic cardioid microphone, placed on a desk at about 30 cm from speaker.
- MIC3_2 is an AKG C410-B head-mounted dynamic microphone.
- MIC5_2 is a low-cost Creative Labs desk microphone for PC sound-card applications.

In sessions 2, 4 and 6, conventional telephone line was used to collect the data. In session 2, every speaker was calling from the same telephone, namely T2_1, in an internal-routing call. In session 4, speakers were requested to make a local call from their own home telephone, T4_1, trying to search a quiet environment (they were asked to be alone in a closed room). In session 6, a local call was made from a quiet room, using 10 randomly selected standard handsets, T6_0 to T6_9 (Reynolds, 1997a). In this last telephone recording session, simultaneous high-quality microphone acquisition was performed (MIC6_2), using the same lapel type SONY microphone as in MIC1_1, MIC3_1 and MIC5_1.

In each session, both microphones (connected through a high-quality Behringer MIC502 preamplifier) and telephone lines (connected through a specific adapter) were supplied to a professional DAT device (Tascam DA-30 MKII), where digital recording at 44.1 kHz was accomplished.

3.2. Recording-room acoustics

A quiet room was selected to accomplish the recordings of sessions 1, 3, 5 and 6 (where session 6 stands for simultaneous telephone and microphone speech acquisition). No anechoic chamber or acoustic cabin was used, as it was desired to have real-environment recording conditions (in terms of reverberation), although maintaining low noise levels. To avoid undesired room reverberation, several acoustic panels were placed around the desk where recordings were performed.

Specific equipment have been used in order to accomplish several acoustic measurements, showing good acoustic conditions for the speech recording sessions. An equivalent noise level of only 27 dBA was measured, and the upper limit for the reverberation time in a third-octave band analysis was 0.48 s. Reverberation time variation with frequency is presented in Fig. 7.

3.3. SNR

Signal-to-noise ratio (SNR) is one of the most important features in the process of acquiring and characterizing a speech database, as this parameter represents an objective relation between desired and undesired signal variances in the log domain,

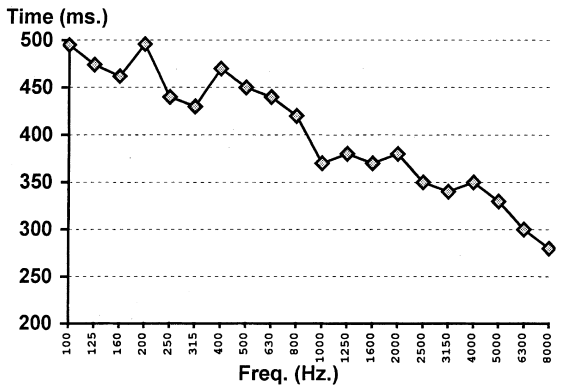

Fig. 7. Reverberation time versus frequency in the recording room.

expressed in decibels (dB). The referred AHUMADA tasks have been acquired under 'clean' conditions, delimiting thus a range of 35–45 dB SNR values.

In this case, SNR has been specifically calculated as the log ratio between rms power of the speech signal and rms power of the noise. Here, it is considered 'noise' the non-speech part of the analyzed segment. The problem arises when trying to calculate the power of the speech signal, due to the implicit non-stationary nature of speech. This non-stationarity leads to rapid changes of signal levels, often in a dynamic range of 30–40 dB. Continuously-spoken segments of at least 3 s have been selected in order to calculate the rms power of the whole segment as rms power of speech.

Setting input speech peaks to $-5/-10$ dB (referred to 0 dB, maximum input level of the acquisition board), rms speech power can be measured in the margin of $-12/-20$ dB. After the application of the high-pass FIR filter (see Section 3.4), rejecting therefore the low components of the noise present, an average SNR value of 40.1 dB is obtained, for 10 randomly selected speakers and tasks through all the microphone and telephone speech. Anyway, this is an average value, and Fig. 8 shows the exact value for every different input. It also shows comparative values when no high-pass filter is applied to the input signal.

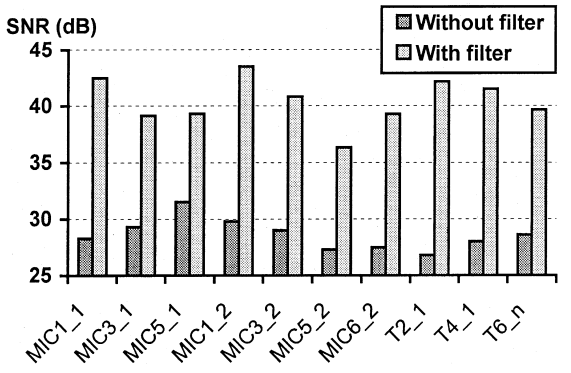

Fig. 8. Average SNR obtained at the different input channels involved in AHUMADA, with and without high-pass filtering of the signal. T6_*n* stands for an average value from T6_0 to T6_9.

3.4. High-pass filtering

The high-pass discrete-time filter employed in order to remove the low frequency components of undesired signals and noise is a linear phase, finite impulse response (FIR) causal filter, with cutoff frequency of 65 Hz. It has been designed through the window method, using Hamming windowing. In this way, the result imposes odd number of coefficients with even symmetry ($h[n] = h[N-n-1]$), obtaining a linear phase, constant group delay type I FIR filter. 1001 coefficients have been used, obtaining 20 dB of attenuation at the always problematical frequency of the main power supply (50 Hz at the recording room), and more than 40 dB below 40 Hz.

3.5. Speech intelligibility

It is widely assumed that Speech Transmission Index (STI) is an excellent approximation for determining objective speech intelligibility measures. In order to evaluate STI, it is necessary to calculate previously Modulation Transfer Function (MTF).

MTF measures the relation between emitted and received modulation indexes of a set of signals. This set consists in 7 different octave-band filtered pink noise signals from 125 Hz to 8 kHz, amplitude modulated by 16 different tones (modulation frequencies), ranging from 0.5 to 16 Hz in third-octave band separation. 98 values are thus obtained, each of them varying from 0 to 1. STI is then directly calculated from these MTF values (Steeneken and Houtgast, 1985).

In our study, Rapid STI, namely RASTI, has been used. RASTI is based in only 9 MTF values rather than the complete 98 MTF values. These 9 values correspond to 4 modulation frequencies for the octave band centered at 500 Hz and 5 modulation frequencies for the octave band centered at 2 kHz. Fig. 9 shows these 9 MTF measures. The tendency in this figure shows only a little degradation of intelligibility, primarily due to reverberation conditions and not to noise.

RASTI values over 0.75 are equivalent to excellent intelligibility. Table 1 shows RASTI values measured in six different points of the recording room. Both RASTI and MTF were obtained using

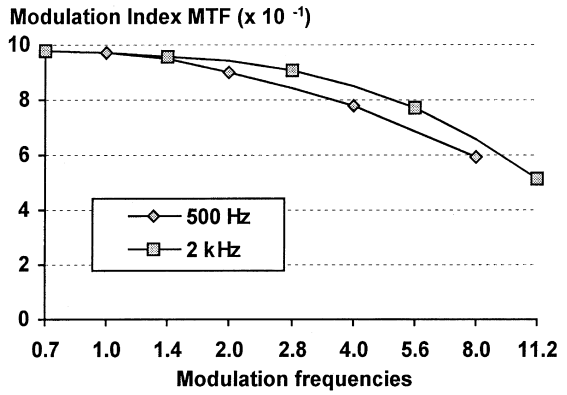

Fig. 9. Measure of intelligibility in terms of MTF.

Table 1
RASTI values measured at 6 different points of the recording room

Point	#1	#2	#3	#4	#5	#6
RASTI	0.80	0.81	0.79	0.73	0.75	0.75

a Brüel & Kjær RASTI type 3361 measuring equipment.

4. Speaker verification system

4.1. Description of the overall system

In order to perform some speaker recognition tests over the available data, a text-independent automatic speaker verification system, based in Gaussian Mixture Models (GMM) (Reynolds, 1992; Ortega-Garcia and Gonzalez-Rodriguez, 1994, 1997), has been employed. Tests have been accomplished over a subset of (randomly selected) 25 speakers from the total number of 104 available speakers. All studio-recorded speech material used for training and testing has been down-sampled to 8 kHz (from the original sampling frequency of 16 kHz). Cepstral coefficients derived from LPC analysis (LPCC) of order 10 have been used as feature vectors. Frames of 30 ms taken every 15 ms, with Hamming windowing and pre-emphasis factor of 0.97 are used as input to the system.

In order to train the system, the first 40 s of the read fixed text (task 2.1.d) from session 1 have been used, generating one model per speaker. All verification tests have been performed using these 25 models. For both training and testing, silences longer than 0.8 s have been removed. As in some cases there was not enough remaining speech material for the testing phase, overlapping between consecutive testing sequences has been forced: 0% for 5 s sequences, 50% for 10 s sequences and 66.6% for 15 s sequences. All 25 speakers were used as claimants for their corresponding models and as impostors for the rest of models.

4.2. Likelihood-domain normalization of scores

Tests without normalization and with likelihood-domain normalization (Rosenberg et al., 1992; Furui, 1994; Matsui and Furui, 1994) have been accomplished. As the density at point X (input sequence) for all speakers other than the true speaker, S, is frequently dominated by the density for the nearest reference speaker, nearest reference speaker normalization criterion (1) has been applied:

$$\log L(X) = \log p(X \mid S = S_c) - \max_{S \in \text{ref},\ S \neq S_c} \log p(X \mid S), \quad (1)$$

where S_c means claimed speaker model. Balance between false rejection error and false alarm errors is required in order to calculate equal error rate (EER) for each speaker. Average EER through all speakers for each case is presented in the next section.

5. Speaker verification benchmark results

5.1. Results obtained

As it has been already mentioned, model training has been performed using about 40 s of read speech per speaker from task 2.1.d, using MIC1_1. The remaining speech from this task (same session, same microphone) has been used for initially testing the verification system, in order to

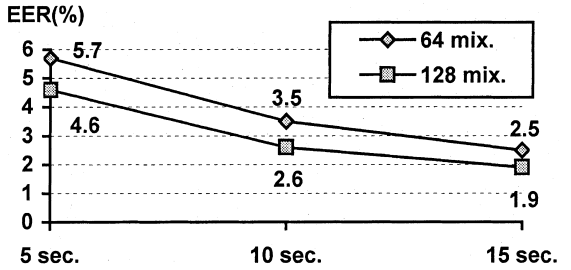

Fig. 10. Baseline verification results with both 64 mixture and 128 mixture GMMs when no normalization is applied.

Table 2
Speaker verification results in terms of EER (%)[a]

EER (%)	5 s		10 s		15 s	
	No norm	Norm	No norm	Norm	No norm	Norm
Test 1	10.1	5.4	6.7	3.0	4.9	2.1
Test 2	15.1	6.1	13.0	4.8	12.7	4.3
Test 3	15.0	8.6	14.8	8.6	14.6	8.5

[a] All experiments show results for test sequences of 5, 10 and 15 s, with/without likelihood normalization scheme.

establish some baseline results for the rest of testing experiments.

Baseline results in Fig. 10 do not include normalization. When likelihood-domain normalization was applied, EERs less than 0.5% were found in all referred cases.

Table 2 shows three different verification experiments, namely Test 1, 2 and 3. Test 1 shows verification results when testing was accomplished with spontaneous speech (task 2.1.g) from session 1 using MIC1_1. In Test 2, same training text used for training and testing (task 2.1.d) of session 1, but considering the effect of using the second microphone (MIC1_2). Finally, Test 3 presents EER for testing sequences of spontaneous speech (task 2.1.g) of session 5 with MIC5_1.

5.2. Analysis of speaker verification rates

The speaker verification experiments described show excellent results when same session, same microphone, same task, and enough amount of testing data (at least 15 s) is used: with score-domain normalization, less than 0.5% EER has been obtained. These two last mentioned parameters, namely testing sequence length and likelihood-domain normalization, produce, with no doubt, significant improvements in all cases. When just the kind of speech is changed, from read speech to spontaneous descriptive speech (Test 1), EER increases up to (in the best case) 2.1% which is still an acceptable limit. Nevertheless, if we use read speech for testing but we change the microphone used (Test 2) we get a best EER of 4.3%. If we focus on inter-session variability (Test 3) with spontaneous testing speech, 8.5% EER is obtained as best.

The results addressed may only give a certain initial idea of the possibilities that AHUMADA database can offer in speaker recognition tasks. In this sense, the use of more efficient features, including Δ and $\Delta\Delta$ cepstra, Δ and $\Delta\Delta$ energy; the use of channel compensating techniques, like CMN, RASTA and others; the use of multi-session and multi-task training; the use of more sophisticated normalization schemes (Gonzalez-Rodriguez and Ortega-Garcia, 1997); the use of general population (background) models (Reynolds, 1997b); and the testing results for all 104 speakers, will focus the work to be carried out over the multi-variability data of AHUMADA corpus.

6. Conclusions and perspectives

6.1. AHUMADA extensions and sub-corpora

At the moment, also about 100 female speakers have been recorded through the same multi-session procedure. Shortly, 150 male and 150 female additional speakers will be recorded in a single-session procedure in order to be used as impostors, and added to the initial corpus described in this paper.

The underlying general idea in designing and acquiring all this speech material in a particular speaker recognition perspective has been to allow a better and more accurate understanding and evaluation of the techniques for identifying people by their voices. Inside this complex process, AHUMADA database will contribute to concentrate on several specially relevant issues. Some of

these, in both commercial and forensic applications, can be enumerated: the use of different speaker and background models for training and testing; the use of different channel-compensation schemes; the availability of target-speakers and impostors in automatic verification systems; the in-depth analysis of forensic-specific speech variability, or the ability to select certain populations to prepare "voice line-ups".

It is important to mention that some special sub-corpora are being acquired in the same project, namely, GSM mobile telephone speech, and bilingual speakers (Castilian Spanish and Catalan languages). It is also foreseen to acquire some special sub-corpora in the near future: these will include different emotional conditions, commercial broadcasting AM/FM/TV simultaneous transmissions, other bilingual speakers (Castilian Spanish and Basque/Galician languages), brothers and twins, and Lombard and noisy speech.

6.2. Forensic perspectives

Real forensic scenarios cannot be forced or even simulated. Those situations in which people commit a crime cannot be substituted by any kind of laboratory or controlled environment. However, it is becoming increasingly usual to find audio physical traces (telephone calls, recorded tapes, security surveillance recordings, etc.). Hence, automatic systems for speaker identification in forensic tasks constitute remarkable scientific analysis tools, as far as they provide an objective measure concerning identity resemblance.

However, speaker verification in forensic tasks is still an open field, in the sense that for many real cases speech technology may not lead to absolute identification certainty. Anyway, this certainty is not always a must in forensic cases, as automatic scores may serve for being aware or concentrate efforts in some reliable direction. In consequence, forensic cases may not always require 'hard' identification decisions (accepted/rejected), but also 'soft' decisions derived from careful analysis of the scores provided by the system.

To summarize, it can be affirmed that the application of state-of-the-art automatic identification systems, with their ability to treat non-specific or neglected massive speech utterances in a general probabilistic manner, without a priori discarding unclear or ambiguous evidences (as is sometimes done in human non-automatic supervised processes) constitute an essential and indispensable practice in modern forensic applications of speaker recognition.

Acknowledgements

We wish to acknowledge all the Guardia Civil Corps for their valuable help in recording all the speech material involved in the AHUMADA corpus. Specially, we would like to mention Civil Guard J.J. Díaz-Gómez, Capt. R. García-Jiménez, Maj. J.J. Lucena-Molina, and Lt. Col. J.A. García Sánchez-Molero, together with all the staff at 'Laboratorio de Acústica e Imagen', from 'Servicio de Policía Judicial' at 'Dirección General de la Guardia Civil'.

We would also like to acknowledge our students S. Cruz, E. Martínez, O. Ledesma and M.A. Chacón, who worked hard in any of the tasks that were carried out to acquire the AHUMADA database and to obtain the speaker verification results on it.

References

Acero, A., 1993. Acoustical and Environmental Robustness in Automatic Speech Recognition. Kluwer Academic Publishers, Dordrecht.

Boves, L. et al., 1994. Design and recording of large databases for use in speaker verification and identification. In: ESCA Workshop on Automatic Speaker Recognition, Martigny, pp. 43–46.

Champod, C., Meuwly, D., 1998. The inference of identity in forensic speaker recognition. In: ESCA Workshop on Speaker Recognition and its Commercial and Forensic Applications, RLA2C, Avignon, pp. 125–134.

Furui, S., 1994. An overview of speaker recognition technology. In: ESCA Workshop on Automatic Speaker Recognition, Martigny, pp. 1–9.

Gibbon, D., Moore, R., Winski, R. (Eds.), 1997. Handbook of Standards and Resources for Spoken Language Systems. EAGLES Spoken Language Working Group. Mouton, Berlin.

Godfrey, J., Graff, D., Martin, A., 1994. Public databases for speaker recognition and verification. In: ESCA Work-

shop on Automatic Speaker Recognition, Martigny, pp. 39–42.

Gonzalez-Rodriguez, J., Ortega-Garcia, J., 1997. Robust speaker recognition through acoustic array processing and spectral normalization. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-97, Munich, pp. 1103–1106.

Guerra, R., 1983. Recuento Estadístico de la Sílaba en Español. In: Estudios de Fonética, 1, Collectanea Phonetica, VII, CSIC, Madrid, pp. 9–112.

Juilland, A., Chang-Rodriguez, E., 1969. Frequency Dictionary of Spanish Words. Mouton, The Hague.

Junqua, J.-C., Haton, J.-P., 1996. Robustness in Automatic Speech Recognition – Fundamentals and Applications. Kluwer Academic Publishers, Dordrecht.

Matsui, T., Furui, S., 1994. Similarity normalization method for speaker verification based on a posteriori probability. In: ESCA Workshop on Automatic Speaker Recognition, Martigny, pp. 59–62.

Naik, J., 1994. Speaker verification over the telephone network: databases, algorithms and performance assessment. In: ESCA Workshop on Automatic Speaker Recognition, Martigny, pp. 31–38.

Ortega-Garcia, J., Gonzalez-Rodriguez, J., 1994. Robust speech modeling for speaker identification in forensic acoustics. In: ESCA Workshop on Automatic Speaker Recognition, Martigny, pp. 217–220.

Ortega-Garcia, J., Gonzalez-Rodriguez, J., 1997. Providing single- and multi-channel acoustical robustness to speaker identification systems. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-97, Munich, pp. 1107–1110.

Ortega-Garcia, J. et al., 1998. AHUMADA: a large speech corpus in Spanish for speaker identification and verification. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-98, Seattle, WA. Vol. II, pp. 773–776.

Quilis, A., Esgueva, M., 1980. Frecuencia de Fonemas en el Español Hablado. In: LEA, 2, pp. 1–25.

Reynolds, D., 1992. A gaussian mixture modeling approach to text-independent speaker identification. PhD Thesis. Georgia Institute of Technology.

Reynolds, D., 1997a. HTIMIT and LLHDB: speech corpora for the study of handset transducer effects. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-97, Munich, pp. 1535–1542, 773–776.

Reynolds, D., 1997b. Comparison of background normalization methods for text-independent speaker verification. In: Proceedings of IVth European Conference on Speech Communication and Technology, EUROSPEECH-97, Rhodes, pp. 963–966.

Rosenberg, A.E., DeLong, J., Lee, C.H., Juang, B.H., Soong, F.K., 1992. The use of cohort normalized scores for speaker verification. In: Proceedings of International Conference on Spoken Language Processing, ICSLP-92, Banff, Canada, pp. 599–602.

Steeneken, H.J.M., Houtgast, T., 1985. RASTI: a tool for evaluating auditoria. In: RASTI, Brüel & Kjær Technical Review, No. 3, pp. 13–30.