Speaker Recognition-Oriented 'AHUMADA' Large Speech Corpus


(1) DIAC, EUTI Telecommunicación, Univ. Politécnica de Madrid, Spain
Ctra. Valencia, km. 7. 28031 Madrid. jortega@diac.upm.es
(2) Dpt. Lengua Española, UNED, Spain
(3) Servicio de Policía Judicial, D. G. Guardia Civil, Spain

Abstract

A specific speech database called AHUMADA (aurmáda/) has been designed and collected for speaker recognition tasks in Castilian Spanish. AHUMADA incorporates six different recording sessions, including both in situ studio and telephone speech recordings. A total of 104 male speakers uttered isolated digits, digit strings, phonologically and syllabically balanced short utterances, phonologically and syllabically balanced read text and more than one minute of spontaneous speech, so about 15 GB of speech material is available. Some examples of the variability factors included in AHUMADA corpus can be: 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); or fixed utterances for all speakers through all sessions versus specific utterances for each speaker in each session.

Introduction

Speaker Recognition is a biometric-based feature characterization task in which people claim to be identified by their voices. Other biometric approximations to personal identification can be DNA, fingerprint, or retina analysis. Anyway, voice identification must be accomplished from a different point of view, more alike to face recognition or graphologist analysis of handwriting, in which signal (written signs or facial features) variability makes identification being a process with additional high-level of complexity (Champod, 1998). 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-depending external influences (kind of microphone, bandwidth and dynamic range reduction, electrical and acoustical noise, reverberation, etc).

In order to develop automatic identification systems, well suited to these variability factors already addressed, a speaker recognition-oriented large database (Boves, 1994; Godfrey, 1994; Naik, 1994; Gibbon, 1997) called 'AHUMADA' (aurmáda/, in honour to the founder of the Guardia Civil Corps, the Duke of Ahumada) has been designed and acquired, and more than 15 GB of speech material is now available (Ortega-García et al., 1998).

Design of the Speech Corpus

The speech corpus has been designed to include many of the speaker variability sources, allowing us to focus on them and study their underlying effects in speaker verification systems. In this sense, 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 ten digits each, being the first five strings identical for all speakers through all recording sessions, and the last five strings specific for each speaker for all sessions.
c) 10 phonologically and syllabically balanced utterances of 8-12 word length. These utterances were identical for all speakers through all sessions.
d) 1 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) 2 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) 1 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 describe (avoiding long pauses and hesitations) whatever they wanted. There were available some paintings and pictures, and subjects like "describe your last holidays", "describe the place where you live/were born", etc., were also suggested.

(*) This work has been supported by CICYT under Project TIC97-1001-C02-01.
Phonological and Syllabic Balance

Tasks c) and d) have been specifically designed in order to reproduce the frequency of appearance of phonemes, syllabic schemes and stress patterns, mostly found in spoken Castilian Spanish. The selected lexicon corresponds to the most usual in Spanish. The 'standard' frequency of appearance (from now on called "Reference") used in the design phase has been measured over an oral corpus of more than 20,000 words (Huilland, 1969; Quilis, 1980; Guerra, 1983). Referred to task c), frequency of appearance of phonemes (Figure 1), syllabic groups (Figure 2) and stress patterns (Figure 3) are shown.

![Figure 1: Frequency of appearance (%) of phonemes in designed task c) compared to "Reference" distribution.](image1)

![Figure 2: Frequency of appearance (%) of syllabic groups in designed task c) compared to "Reference" distribution.](image2)

Figure 3: Frequency of appearance (%) of stress patterns in designed task c) compared to "Reference" distribution.

Data Collection and Acquisition

Recording Microphones and 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 MICn_1 and MICn_2, were n corresponds to one of the three possible sessions. Consequently, MIC1_1, MIC3_1 and MIC5_1 were the same microphone, namely SONY ECM-66B, lavalier 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 soundcard applications.

In sessions 2, 4 and 6, telephone line was used to collect the data. In session 2, every speaker was making a phone call from the same telephone in an internal-routing call. In session 4, speakers were requested to make a local call from its own home telephone, 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 different standard handsets (Reynolds, 1997). In this last telephone recording session, simultaneous microphone acquisition was performed (MIC6_2), using the same lavalier 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 fed to a professional DAT device (Tascam DA-30 MKII), where digital recording at 44.1 kHz was accomplished.

Recording-room Acoustics

A quiet room was selected to make the recordings of sessions 1, 3, 5 and 6 (simultaneous telephone and microphone input). No anechoic chamber or acoustic
cabin was used, as we wanted to have real-environment (quiet) recording conditions. To avoid undesired room reverberation, several acoustic panels were placed around the desk where recordings were made.

Measurements done with acoustic specific equipment showed 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 is 0.48 sec. Reverberation time variation with frequency is presented in figure 4.

**Speech Intelligibility**

For speech intelligibility objective measures, it is widely assumed that Speech Transmission Index (STI) is an excellent approximation. Relationship between modulation indexes emitted and received measure the reduction of the modulation index, expressed in terms of the modulation frequency, and varying from 1 to 0. This function is called Modulation Transfer Function (MTF), and must be specified for each carrier frequency.

In our study, Rapid STI (RASTI), has been measured. RASTI measure reduces to 9 values the original 98 STI values. This 9 values are 4 modulation frequencies at 500 Hz. and 5 modulation frequencies at 2 kHz. band. It is assumed that RASTI values over 0.75 produce excellent intelligibility. Figure 5 shows results obtained using a Brüel and Kjær RASTI type 3361 measuring equipment.

**Distribution of Ages**

The distribution of ages was designed in order to model a possible distribution of users of a certain speaker recognition application. Equi-distribution of ages may not respond to real user distribution, and more weight has been applied to the range of ages between 28 and 42 years. Figure 6 shows actual distribution.

![Age Distribution](image)

**Figure 6 : Age distribution of the male population of AHUMADA.**

**Time Interval between Sessions**

As inter-session variability is an important factor to be taken into account in speaker recognition-oriented databases, at least a time interval separation of 15 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 meant to. Anyway, the enrollment availability of speakers may have caused some deviations from these initial requirements.

Recordings began in June 1997, with session 1. Following, it can be found the time intervals between session 1 and the rest of the sessions:

- **Session 2:** 73% of recordings were done within 15 days interval from session 1. Specifically, 36% were accomplished the same day of session 1. The maximum time interval (100% of recordings) is 40 days.
- **Session 3:** 80% of recordings were done between 20 and 40 days after session 1, and 92% between 15 and 50 days. The minimum interval is 10 days after session 1, and the maximum is 80 days.
- **Session 4:** 73% of recordings were accomplished in a time interval of 15 to 50 from session 1. 19% were done between 40 and 80 days after session 1.
- **Session 5:** The minimum interval between session 1 and session 5 is 30 days. 77% of them were acquired between 40 and 80 days after session 1, while 10% were separated in time from 80 to more than 90 days.
- **Session 6:** The minimum time interval of session 6 recordings is 30 days after session 1. 78% of speech material was recorded between 40 and 80 days after session 1. The last 9% of recordings were done between 80 and more than 90 days after session 1.

**Speaker Recognition System**

In order to perform some speaker recognition tests over the available data, a speaker verification system has been
used. As we wanted to evaluate text-independent verification results, Gaussian Mixture Models (GMM) (Reynolds, 1992; Ortega-Garcia, 1994; 1997) have been used. 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 120 samples with Hamming windowing and pre-emphasis factor of 0.97 are used as input to the system. For training the system, the first 40 s. of the read fixed text (task 2.1.d) from session 1 has been used, generating one model per speaker. All verification tests have been done 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 speaker models. Tests without normalization and with likelihood-domain cohort normalization (Rosenberg, 1992; Furui, 1994; Matsui, 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, we have applied the following normalization criterion:

$$\log L(X) = \log p(X|S = S_c) - \max_{S_{ref}, S = S_c} \log p(X|S)$$

where $S_c$ means claimed speaker model. Balance between false rejection error and false alarm errors is searched, so equal error rate (EER) for each speaker is computed, and average EER through all speakers for each case is presented in the next section.

**Speaker Verification Benchmark Results**

As it has been already mentioned, model training has been performed using about 40 s. of read speech from a fixed text, equal for all speakers (task 2.1.d), corresponding to session 1 and 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 establish some baseline results for the rest of testing experiments. Figure 7 shows these results.

Baseline results in Figure 7 do not include normalization. When likelihood-domain normalization was applied, EERs less than 0.5% were found in all referred cases. Figure 8 shows verification results when testing was accomplished with spontaneous speech (task 2.1.g) from session 1 using MIC1_1.

**Figure 8**: EER for different duration of testing sequences of spontaneous speech, session 1, MIC1_1.

In Figure 9, 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).

**Figure 9**: Verification results due to using different microphones in training and testing phases.

Finally, Figure 10 presents EER for testing sequences of spontaneous speech (task 2.1.g) of session 5 with MIC5_1.

**Figure 10**: Inter-session variability between sessions 1 and 5, testing with spontaneous speech.

**Analysis of Results**

Speaker verification experiments described show excellent results when same session, same microphone, same task, and enough amount of testing data (15 s.) is used: normalizing the results of Fig. 7 gives less than 0.5% EER. 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 (Fig. 8), EER increases to (in the best case) 2.1% which is still an acceptable limit. Nevertheless, if we use read speech to test but we change the microphone used (Fig. 9) we get a best EER of 4.3%. If we focus on inter-session variability (Fig. 10) with spontaneous testing speech, 8.5% EER is obtained as best.

Conclusions and Future Work

Speaker Identification Results
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 $\Lambda$ and $\Delta\Lambda$ cepstra, $\Lambda$ and $\Delta\Lambda$ energy; the use of channel compensating techniques like CMN and RASTA; the use of multi-session and multi-task training; the use of more sophisticated normalization schemes, including general population (background) models, etc., and the testing results for all 104 speakers, will focus the work to be done in the near future over the multi-variability data of AHUMADA corpus.

AHUMADA Extension and Sub-Corpora
At the present time, also about 100 female speakers are being recorded. AHUMADA male speakers together with these last 100 female speakers will constitute the corpus of target-speakers for automatic verification systems. In order to have also impostors for this automatic system, 150 male and 150 female speakers will be added to the initial corpus described in this paper. Also, it is important to mention that some special sub-corpora will be acquired: these will include emotional variations, GSM cellular telephone speech, commercial broadcasting transmissions, bilingual speaker (castilian spanish and catalan/basque/galician languages), brothers and twins, Lombard and noisy corpuses, etc.

References