QUANTITATIVE INFLUENCE OF SPEECH VARIABILITY FACTORS FOR AUTOMATIC SPEAKER VERIFICATION IN FORENSIC TASKS

J. Ortega-García, S. Cruz-Llanas and J. González-Rodríguez

DIAC, EUIT Telecomunicación, Universidad Politécnica de Madrid, Spain

ABSTRACT(*)

Regarding speaker identity in forensic conditions, several factors of variability must be taken into account, as peculiar intra-speaker variability, forced intra-speaker variability or channel-dependent external influences. Using 'AHUMADA' large speech database in Spanish, containing several recording sessions and channels, and including different tasks for 100 male speakers, automatic speaker verification experiments are accomplished.

Due to the inherent non-cooperative nature of speakers in forensic applications, only text-independent recognizers are likely to be used. In this sense, a GMM-based verification system has been used in order to obtain quantitative results. Maximum likelihood estimation of the models is performed, and LPC-cepstra, delta- and delta-delta-LPCC, are used at the parameterization stage.

With this baseline verification system, we intend to determine how some variability sources included in 'AHUMADA' affect speaker identification. Results including speaking rate influence, single- and multi-session training and cross-channel testing are presented when likelihood-domain normalization is applied.

1. INTRODUCTION

It is becoming increasingly usual to find audio physical traces (telephone calls, recorded tapes, security surveillance recordings, etc.) in situations in which people commit a crime. In this sense, it is essential to find reliable methods that allow the association of an unknown voice sample with a known person identity. Speaker Recognition is a characterization process in which people claim to be identified by their voices. Anyway, voice identification, specially in forensic approaches, must take into account signal variability, which incorporates to the identification process an additional level of complexity [1].

In this context, coping with forensic identification implies dealing with speech variability [4, 6]. Regarding speaker identity, several factors of variability must be taken into account: *i)* Peculiar intra-speaker variability (manner of speaking, age, gender, inter-session variability, dialectal variations, emotional condition, etc.). *ii)* Forced intra-speaker variability (Lombard effect, external-influenced stress, cocktail-party effect). *iii)* Channel-dependent external influences (kind of microphone, bandwidth and dynamic range reduction, electrical and acoustical noise, reverberation, etc).

In this sense, delimiting the problem of speech variability, together with analyzing the quantitative influence of this speech variability on the results of speaker recognition systems may lead to an indispensable and comprehensive approach to forensic speaker recognition.

For evaluating the influence of some of these variability factors, 'AHUMADA' speech database [3, 5, 6] has been used. Some examples of the variability factors 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, or inter-session variability in six different recording sessions.

The present paper is organized as follows. Section 2 describes 'AHUMADA' speech corpus. Section 3 presents the recognition system employed and the five different verification experiments that have been carried out. Section 4 analyzes the results previously shown. And, finally, some conclusions are reached in Section 5.

2. 'AHUMADA' SPEECH CORPUS

2.1. Design of the Database

Tasks. The enrolled speakers were requested to utter the following: a) 24 isolated digits; b) 10 digit strings consisting of ten digits each; c) 10 phonologically and syllabically balanced utterances of 8-12 word length; d) 1 phonologically and syllabically balanced text, of about 180 words (more than 1 minute of duration), read at a normal speaking rate; e) Two repetitions of the previous fixed text, asking the speakers to read it at a fast and at a slow speaking rate; f) 1 specific text, different from speaker to speaker and from session to session, for each speaker; g) More than 1 minute of spontaneous speech, asking every speaker to describe (avoiding long pauses and hesitations) whatever they wanted.

Phonological and Syllabic Balance. Tasks c) and d) have been specifically designed in order to reproduce the frequency of appearance of phonemes and syllabic schemes, mostly found in spoken Castilian Spanish [7]. 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 was measured over an oral corpus of more than 20,000 words.

(*) This work has been supported by the CICYT under Project TIC97-1001-C02-01

Recording sessions. 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 (microphones) were simultaneously used. The notation used to specify both microphones in each case is MIC*n*_1 and MIC*n*_2, were *n* corresponds to one of the three possible sessions.

Time Interval between Sessions. Following, it can be found the time intervals between the first *in situ* session and the rest of them: a) *Session 2 (telephone):* 73% of recordings were done within 15 days interval from session 1. b) *Session 3 (in situ):* 80% of recordings were done between 20 and 40 days after session 1. c) *Session 4 (telephone):* 73% of recordings were accomplished in a time interval of 15 to 50 from session 1. d) *Session 5 (in situ):* 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. e) *Session 6 (telephone and microphone):* 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.

2.2. Technical Features and Audio Equipment

Recording Microphones. The relation of microphones is a s follows: 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.

Telephone Handsets. In sessions 2, 4 and 6, conventional telephone line was used to collect the data. In session 2, every speaker was making a phone call 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 (Reynolds, 1997a), T6_0 to T6_9.

Recording-Room Acoustics. A quiet room was selected to accomplish the recordings of sessions 1, 3, 5. 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. 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 sec.

Signal-to-Noise Ratio. We have specifically calculated Signal-to-noise ratio (SNR) as the logarithmic ratio between RMS power of the speech signal and RMS power of the noise. For noise, here we understand the non-speech part of the analyzed segment. For speech, continuously-speaking segments of at least 3 sec. have been selected in order to calculate the RMS power of the whole segment as RMS power of speech. After the application of the high-pass FIR filter designed to reject the low components (under 65 Hz.) of the noise present, we get an average SNR value of 40.1 dB, for 10 randomly selected speakers and tasks through all the microphone and telephone speech.

Speech Intelligibility. In our study, Rapid STI, namely RASTI (Steeneken, 1985), has been measured. RASTI measure reduces to 9 values the original 98 STI values. These 9 values are 4 modulation frequencies for the octave band centered at 500 Hz. and 5 modulation frequencies for the octave band centered at 2 kHz. It is assumed that RASTI values over 0.75 are equivalent to excellent intelligibility. Six different points of the room were randomly selected in order to determine RASTI; the values obtained cover a range from 0.73 to 0.81. RASTI values were obtained using a Brüel & Kjær RASTI type 3361 measuring equipment.

3. THE OVERALL VERIFICATION SYSTEM

3.1. System Description

In order to perform some speaker recognition tests over the available data, a speaker verification system has been used [6]. As we wanted to evaluate text-independent verification results, Gaussian Mixture Models (GMM) have been used [8]. 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 features and their derivatives have been used taking analysis frames of 30 ms. every 15 ms., with Hamming windowing and pre-emphasis factor of 0.97 are used as input to the system. For both training and testing, silences longer than 0.8 s. have been removed. All 25 speakers were used as claimants for their corresponding models and as impostors for the rest of speaker models.

Likelihood-Domain Normalization of Scores. 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 [2]:

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

where S_c means claimed speaker model.

Speaker verification rates. 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.

3.2 Speaker Verification Experiments

Experiment 1: Channel effect varying parameterization. In this first test, 40 secs. of speech from task d) (fixed read text) for each speaker have been used in the training stage. All

speech used in this training stage has been acquired from channel MIC1_1. Different feature vectors have been also used, namely cepstral coefficients derived from LPC analysis (LPCC), and their first and second derivatives, Δ- and ΔΔ-cepstral coefficients. In this way, 3 different models have been trained for each speaker: 1 model with 10 LPCC, 1 model with 10 LPCC+ΔLPCC and 1 model with 10 LPCC+ΔLPCC+ ΔΔLPCC.

In the testing stage, speech utterances from task d) have also been used, but using now channel MIC1_2 (same session, second channel). Results in Figure 1 show how microphone mismatch and derivative coefficients affect speaker verification results as a function of test utterance duration.

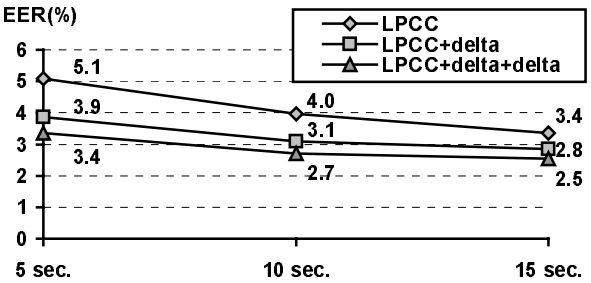

Figure 1: Speaker verification results in terms of EER. Models trained with 40 s. read speech from MIC1_1. Tests with 5, 10 and 15 s. of read speech from MIC1_2. Results also show the improvements when using first and second derivative coefficients.

Experiment 2: Influence of changes in speaking rate. In this experiment, the influence of speaking rate on our recognition system is measured. In the training phase, models were generated with 40 sec. of read speech at a normal speaking rate (task d), using 10 LPCC+ ΔLPCC.

At the testing stage, speech utterances from task e) were selected, which means that read speech at both a fast and a slow speaking rate has been used. Figure 2 shows the results of this verification experiment, showing the effect of mismatch speaking rate between training and testing.

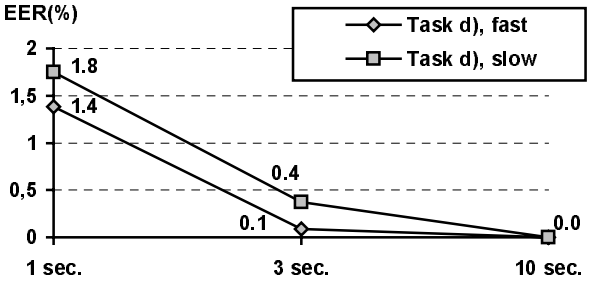

Figure 2: Speaker verification results, when testing is accomplished using read speech at both fast and slow speaking rate, and training considers only normal read speaking rate.

Experiment 3: Effect of multi-session training. This experiment concerns to the evaluation of the influence of single-session training versus multi-session training when multiple sessions are available in speaker identification tests. In this sense, results when only session 1 (40 sec.) is used for training and session 5 used for recognition are compared with testing in these same conditions when training is accomplished in sessions 1 and 3 (20 sec. from each session). It is important to note that all the microphones involved in this experiment 3 are always the same (MIC1_1, MIC3_1 and MIC5_1).
Figure 3 shows the results of this comparative experiment, when 10 LPCC+ ΔLPCC are used as input features.

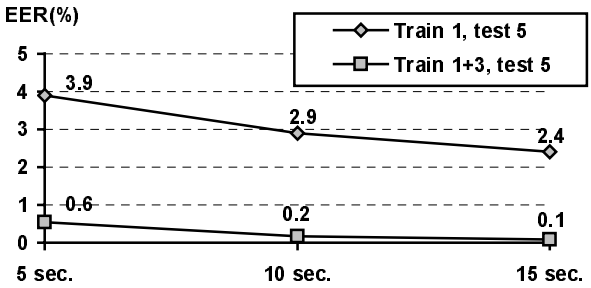

Figure 3: Speaker verification rates showing comparative analysis between single-session trained models (session 1) and multi-session trained models (session 1 and 3). Testing is accomplished over the same speech utterances, selected from session 5.

Experiment 4: Channel compensation through multi-session training. In this experiment, the training phase is identical to that accomplished in *Experiment 3*, thus obtaining single-session (session 1) and multi-session (sessions 1 and 3) models for each speaker.
For testing, in this case, speech utterances from session 5 have been used. Nevertheless, in this case, these testing utterances were obtained from MIC5_2, while training utterances came from MIC1_1 and MIC3_1. Results presented in Figure 4 show single- and multi-session training behavior facing microphone changes in testing phase.

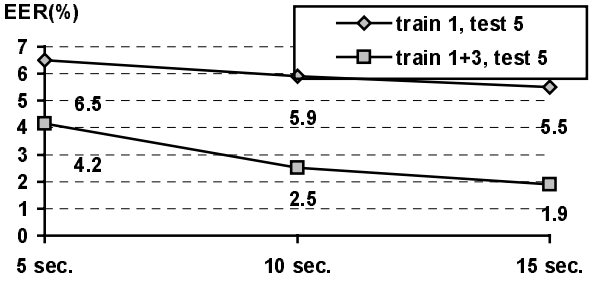

Figure 4: Verification results in terms of EER when training is done in a single- (session 1, MIC1_1) or multi-session manner (session 1 and 3, MIC1_1 and MIC3_1), and testing is realized with speech material from session 5 and microphone MIC5_2.

Experiment 5: Multi-session and cross-channel training. The train phase has been accomplished as follows: 20 sec. from

MIC1_1 plus 20 sec. from MIC3_2, both using task d) utterances, and 10 LPCC+ ΔLPCC features.
For the verification process, results have been obtained from utterances using MIC5_2 and also from MIC1_2. Figure 5 shows the scores obtained in terms of the EER.

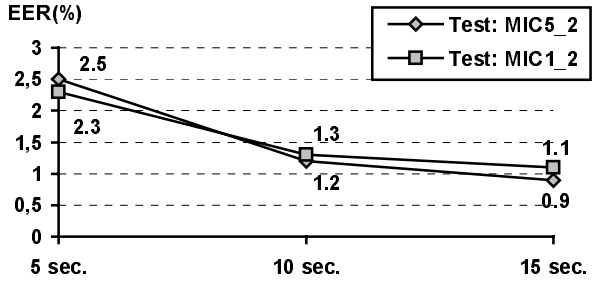

Figure 5: Speaker verification results, when multi-session and cross-channel training is done (MIC1_1 and MIC3_2). Testing phase has been carried out with utterances from MIC5_2 and MIC1_2.

4. EVALUATION OF SPEAKER VERIFICATION RESULTS

As *Experiment 1* illustrates, training and testing in channel mismatch conditions provoke degradation in verification results. The length of the testing utterance improves results significantly, as well as the use of first and second derivative cepstral feature vector. From the worse case (5 sec. length, only 10 LPCC) to the best one (15 sec., 10 LPCC+ ΔLPCC+ ΔΔLPCC) EER is reduced by 50%, which is a remarkable gain, particularly taking into account that these are factors that in many cases can be voluntarily selected.

Experiment 2 shows that speaking rate is not a decisive factor causing degradation in verification results, and only a small influence is observed.

On the contrary, *Experiment 3* shows the enormous influence of multi-session training on verification results. When only Session 1 is used for training, and testing is accomplished in Session 5, EERs varying from 3.9% to 2.4% are found. However, when training is done considering both Session 1 and 3, results are drastically reduced to 0.6% to 0.1%.

Experiment 4 also shows interesting results, specifically that multi-session training is effective not only in multi-session testing, but also in channel compensation: EER decreases from 5.5% to 1.9% on MIC5_2 when training includes not only MIC1_1, but also MIC3_1.

Finally, *Experiment 5* confirms the effectiveness of multi-session training in channel normalization, as similar results are obtained in channel MIC1_2 as in MIC5_2, when training is done using MIC1_1 and MIC3_2. Training with different session and microphones consistently improves results on other sessions and other microphones.

5. CONCLUSIONS

Speech variability is one of the dominant questions involved in speaker identification, specially when applied to the forensic field. In this contribution, some of these variability factors have been quantitatively analyzed, determining their objective influence over automatic systems.

Roughly, it can be said that channel compensation and multi-session training are two of the most outstanding factors, in terms of their effect over speaker verification results. Specifically, some of the experiments conducted show that multi-session training has beneficial effects on channel compensation.

On the other hand, factors like speaking rate (fast / normal / slow) do not seem to be specially important in terms of their influence on the recognition rates.

6. REFERENCES

1. C. Champod and D. Meuwly, "The Inference of Identity in Forensic Speaker Recognition", *ESCA Workshop on Speaker Recognition and its Commercial and Forensic Applications, RLA2C*, pp. 125-134, Avignon (FR), 1998.

2. S. Furui, "An Overview of Speaker Recognition Technology", *ESCA Workshop on Automatic Speaker Recognition*, pp. 1-9, Martigny (CH), 1994.

3. D. Gibbon, R. Moore and R. Winski, eds., *Handbook of Standards and Resources for Spoken Language Systems*, EAGLES Spoken Language Working Group, Mouton de Gruyter, 1997.

4. J. Ortega-García and J. González-Rodríguez, "Robust Speech Modeling for Speaker Identification in Forensic Acoustics", *ESCA Workshop on Automatic Speaker Recognition*, pp. 217-220, Martigny (CH), 1994.

5. J. Ortega-García *et al.*, "AHUMADA : A Large Speech Corpus in Spanish for Speaker Identification and Verification", *IEEE Intl. Conf. on Acous. Speech and Signal Proc., ICASSP-98*, vol. II, pp. 773-776, Seattle (Wa, USA), 1998.

6. J. Ortega-García, J. González-Rodríguez and V. Marrero-Aguiar, "An Approach to Forensic Speaker Verification Using 'AHUMADA' Large Speech Corpus in Spanish", *Speech Communication, submitted for publication*, 1998.

7. A. Quilis and M. Esgueva, "Frecuencia de Fonemas en el Español Hablado", *LEA*, 2, pp. 1-25, 1980.

8. D. Reynolds, *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification*, Ph. D. Thesis, Georgia Institute of Technology, 1992.