SPEECH VARIABILITY IN AUTOMATIC SPEAKER RECOGNITION SYSTEMS FOR FORENSIC PURPOSES

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

EUIT Telecomunicación, Universidad Politécnica de Madrid
Ctra. Valencia, km. 7. Campus Sur. E-28031 Madrid, Spain
e-mail: jortega@diac.upm.es http://www.atvs.diac.upm.es

ABSTRACT

It is becoming increasingly usual to find audio physical traces (telephone calls, recorded tapes, security surveillance recordings, etc.) while committing crimes, forcing in consequence speech research community to find reliable methods that allow the association of an unknown voice sample with a determined person identity. Regarding speech variability in forensic approaches, some of these variability sources highly degrade the speaker verification process, namely: channel influence, inter-session variability and emotional state. In this contribution, channel and inter-session variability will be explored in order to accomplish real automatic systems for forensic speaker recognition. Results will be presented making use of 'GAUDI/AHUMADA' large speaker recognition-oriented database in Spanish.

Keywords: Forensic Acoustics, Speaker Verification, Channel Normalization, Inter-session Variability, Speech Databases.

1. INTRODUCTION(*)

Speaker Recognition is a biometric characterization process aimed at the identification of people by their voices. Fingerprint or iris analysis 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 98].

In this context, coping with real commercial and forensic recognition implies dealing with speech variability [Acero 93, Junqua 96, Gonzalez-Rodriguez 99]. 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 [Boves 94 , Godfrey 94, Naik 94 , Gibbon 97], a speaker recognition-oriented large database called 'GAUDI/AHUMADA', has been designed and acquired, involving 104 male and 99 female speakers. The resulting data comprises more than 35 GB of recorded material [Ortega-Garcia 98a], and incorporates several speech variability factors.

In this context, and in order to evaluate how speech variability affects speaker recognition systems, 'GAUDI/AHUMADA' speech database will be used. This corpus incorporates several variability factors, being the most relevant among all: *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. 'GAUDI/AHUMADA' will include in the near future other 300 more speakers (single-channel, single-session acquisition) to be used as impostors. The subset of 'GAUDI/AHUMADA' containing the 100 male users has been previously called 'AHUMADA'.

Due to the inherent non-cooperative nature of speakers in forensic applications, only text-independent automatic speaker recognizers should be used. In this sense, a GMM-based verification system is being used in order to obtain quantitative results. Maximum likelihood estimation of the models is performed, and Mel-Frequency Cepstral Coefficientes (MFCC) together with their first-

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

0-7803-5247-5/99/$10.00 ©1999 IEEE

order derivative features, namely delta-MFCC are used at the parameterization stage.

2. AHUMADA SPEECH SUBCORPUS

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. 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.

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 97], 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 85], 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 [Ortega-Garcia 98b]. As we wanted to evaluate text-independent verification results, Gaussian Mixture Models (GMM) have been used [Reynolds 92]. 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.). MFCC and their first-order derivatives (delta-MFCC) 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 [Furui 94]:

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

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.

4. EXPERIMENTAL RESULTS

4.1. Text-Independent Speaker Verification

All experiments included in this section make use of 'GAUDI/AHUMADA' database. Specifically, a subset of 25 male speakers are selected as users, and 25 other speakers operate as impostors. Tasks b (10 digit strings of 10 digit each, namely b01:b10) and c (10 fixed utterances, namely c01:c10) are used for training and/or testing the system. Microphonic sessions M1/M4, M2/M5, M3/M6 (where MJ/MK stands for same session, different microphone in stereo recording), and telephonic sessions T1, T2, and T3/M7 (simultaneous telephonic/microphonic recording) are used. Acquisition through M1, M2, M3, and M7 is accomplished using the same microphone. M4, M5, and M6 are different microphones among them.

Experiment 1 (Table 1) shows benchmark results for AHUMADA microphonic and telephonic data, when using same session and channel data, namely, training (c01:c05) and testing (b01:b10, c06:c10) with T3 and M7 separately. 1 utterance per impostor (from c06:c10) is used.

EER(%)	NO SCORE NORM.		SCORE NORM.	
CH. NORM. / TR/TS	NONE	CMN	NONE	CMN
a) T3/T3	**7.1**	**16.3**	**0.3**	**1.0**
b) M7/M7	**4.1**	**19.0**	**0.2**	**3.1**

Table 1. Verification results when training (TR) and testing (TS) with T3 and M7 data (separately), with/without score normalization, with CMN compensation or without channel normalization (CH. NORM.), expressed in terms of EER(%).

As it can be derived from Table 1, score normalization significantly improves verification results. When CMN is applied in these cases, where no channel variation exists, results slightly degrade, as some speaker information is removed.

4.1.1. Microphonic Speech

Experiment 2 (Table 2) shows results when M4 and M5 are used for testing, and session M7 is used for training, though varying microphone and session.

EER(%)	NO SCORE NORM.		SCORE NORM.	
CH. NORM. / TR/TS	NONE	CMN	NONE	CMN
a) M7/M4	**24.7**	**19.3**	**21.7**	**8.5**
b) M7/M5	**21.1**	**24.4**	**14.4**	**8.1**

Table 2. Verification results when training (TR) with M7 mic. speech and testing (TS) with M4 and M5 data, with/without score normalization, with CMN compensation or without channel normalization (CH. NORM.), expressed in terms of EER(%).

In this case, results get worse with respect to Table 1, as different channels and sessions are used for training. When score normalization and, specially, when CMN technique is applied, we achieve about 8% EER. As it can be seen in this case, CMN is highly effective, despite inter-session variability remains as the main degradation factor.

Experiment 3 (Table 3) concentrates on inter-session variability, as microphones M1, M2 and M3 used for training, and microphone M7 used for testing are all the same microphone, but

corresponding all of them to different acquisition sessions.

EER(%)	NO SCORE NORM.		SCORE NORM.	
CH. NORM. / TR/TS	NONE	CMN	NONE	CMN
a) M1/M7	**16.7**	**28.8**	**12.0**	**10.6**
b) M2/M7	**19.1**	**28.0**	**13.4**	**8.7**
c) M3/M7	**17.5**	**27.4**	**10.6**	**7.0**

Table 3. Verification results when training (TR) with the same microphone in different recording sessions (M1, M2 and M3) and testing (TS) with M7 data (also the same microphone), with/without score normalization, with CMN compensation or without channel normalization (CH. NORM.), expressed in terms of EER(%).

As it can be derived from the previous table, CMN is also effective for coping with inter-session variability, compensating also slight same-channel variations among sessions.

Experiment 4 concentrates on multi-session training, varying the number of utterances for training each model. In M1+M3, M1c01:M1c03+M2c04:M2c05 (5 utterances), have been used, while in M1+M2+M3, M*N*c01:c10 (30 utterances) have been used, so 6 times more training speech is used in 4.b) with respect to 4.a). Results of Experiment 4 are shown in Table 4.

EER(%)	NO SCORE NORM.		SCORE NORM.	
CH. NORM. / TR/TS	NONE	CMN	NONE	CMN
a) M1+M3/M7	**16.7**	**29.4**	**7.3**	**8.2**
b) M1+M2+M3/M7	**14.9**	**25.1**	**7.0**	**5.3**

Table 4. Verification results when training (TR) with M1+M3 (5 utterances per speaker) or M1+M2+M3 (30 utterances per speaker), and testing (TS) with M7 data, with/without score normalization with CMN compensation or without channel norm., expressed in terms of EER(%).

Results in Table 4 confirm that we can take advantage of the amount of training data, in the sense that multi-session training with 30 utterances (10 from M1, 10 from M2, and 10 from M3) instead of multi-session training with only 5 utterances (3 from M1 and 2 from M3), enables EER to be reduced to 5.3%.

4.1.2. Telephonic speech

Experiments 2 through 4 describe results when telephonic speech is used. For this purpose, data from T1, T2 and T3 telephonic sessions have been used. Anyway, T1, T2 and T3 do not exhibit complete telephonic consistency, as in T1, every speaker was calling from the same telephone, in an internal-routing call; in T2, speakers made a local call from their own home telephone; and in T3, a local call was made from a quiet room, using 10 different standard handsets.

In this sense Experiment 5 makes use, in order to verify the telephonic consistency of T1, T2 and T3 data, of real telephonic speech from 'GAUDI/AHUMADA' female subcorpus, namely T4, T5 and T6. T4, T5 and T6 are all obtained in a real local-call acquisition process. Table 5 shows the results of Experiment 5.

EER(%)	NO SCORE NORM.		SCORE NORM.	
CH. NORM. / TR/TS	NONE	CMN	NONE	CMN
a) T1/T3	**26.3**	**25.8**	**17.8**	**17.6**
b) T2/T3	**40.4**	**27.3**	**36.7**	**20.3**
c) T1+T2/T3	**29.6**	**32.4**	**21.6**	**24.2**
d) T4/T6	**21.5**	**34.3**	**13.9**	**14.4**
e) T5/T6	**22.7**	**33.6**	**15.2**	**14.3**
f) T4+T5/T6	**23.8**	**34.4**	**13.7**	**15.3**

Table 5. Verification results when training (TR) with T1, T2, T1+T2, T4, T5, and T4+T5 (5 utterances per speaker in every case), and testing (TS) with T3 (male speakers) or T6 (female speakers), with/without score normalization, with CMN compensation or without channel normalization (CH. NORM.), expressed in terms of EER(%).

Table 5 confirms the inconsistency of telephonic situations among T1, T2, and T3 in comparison with real telephonic data T4, T5 and T6, as average EER of 20.7% (T1, T2 and T3) decreases to 14.6% when using female data. Anyway, these results, even in the female case, are not satisfactory enough.

Experiment 6 is accomplished in order to establish whether better results can be obtained increasing the number of utterances per speaker involved in the training process. In this sense, Table 6 shows results when 20 utterances (c01:c10 per session) per speaker are used.

EER(%)	NO SCORE NORM.		SCORE NORM.	
CH. NORM. / TR/TS	NONE	CMN	NONE	CMN
a) T1+T2/T3	**29.2**	**25.1**	**19.5**	**9.8**
b) T4+T5/T6	**20.5**	**30.8**	**10.3**	**8.7**

Table 6. Verification results when training (TR) with T1+T2, and T4+T5 (20 utterances per speaker in each case), and testing (TS) with T3 (male speakers) or T6 (female speakers), with/without score normalization, with CMN compensation or without channel normalization (CH. NORM.), expressed in terms of EER(%).

In this case, with respect to the previous experiment, in which only 5 utterances per speaker were used, the fact of employing 20 utterances per speaker (10 from T1/T4, and 10 from T2/T5) significantly improves verification rates, allowing to obtain EERs around 9%. Specially meaningful is

case 6.a) with respect to 5.c), in which EER decreases from 24.2% to 9.8%.

5. CONCLUSIONS

All experiments demonstrate that score normalization is essential in speaker verification tasks. Benchmark Experiment 1 show that in good conditions (single session, same channel), EER can be lowered to less than 0.5%. The combination of score normalization and CMN techniques decreases ERR significantly. This decreasing is specially relevant when mismatch among channels is found (Experiments 2, 5 and 6). It also produces EER decreasing when mismatch between sessions is encountered (Experiment 3), and when multi-session training is accomplished (Experiments 4 and 6). Increasing training data is also a relevant issue (Experiments 4.a, and 6), specially in forensic approaches, where speakers show low degree of cooperativeness but where large amounts of data are usually available.

6. REFERENCES

[Acero 93] A. Acero, *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Dordrecht (NL), 1993.

[Boves 94] L. Boves *et al.*, "Design and Recording of Large Data Bases for Use in Speaker Verification and Identification", *ESCA Workshop on Automatic Speaker Recognition*, pp. 43-46, Martigny (CH), 1994.

[Junqua 96] J.-C. Junqua and J.-P. Haton, *Robustness in Automatic Speech Recognition - Fundamentals and Applications*, Kluwer Academic Publishers, Dordrecht (NL), 1996.

[Champod 98] 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*, Avignon (FR), pp. 125-134, 1998.

[Furui 94] S. Furui, "An Overview of Speaker Recognition Technology", *ESCA Workshop on Automatic Speaker Recognition*, Martigny (CH), pp. 1-9, 1994.

[Gibbon 97] 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.

[Godfrey 94] J. Godfrey, D. Graff and A. Martin, "Public Databases for Speaker Recognition and Verification", *ESCA Workshop on Automatic Speaker Recognition*, pp. 39-42, Martigny (CH), 1994.

[Gonzalez-Rodriguez 99] J. Gonzalez-Rodriguez, *Influence and Compensation of the Acoustical Environment in Automatic Speaker Recognition Systems*, (in Spanish), Ph. D. Thesis, Universidad Politécnica de Madrid, 1999.

[Junqua 96] J.-C. Junqua and J.-P. Haton, *Robustness in Automatic Speech Recognition - Fundamentals and Applications*, Kluwer Academic Publishers, Dordrecht (NL), 1996.

[Naik 94] J. Naik, "Speaker Verification over the Telephone Network: Databases, Algorithms and Performance Assessment", *ESCA Workshop on Automatic Speaker Recognition*, pp. 31-38, Martigny (CH), 1994.

[Ortega-Garcia 98a] J. Ortega-Garcia *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, 1998.

[Ortega-Garcia 98b] J. Ortega-Garcia, S. Cruz-Llanas and J. Gonzalez-Rodriguez (1998), "Quantitative Influence of Speech Variability Factors for Automatic Speaker Verification in Forensic Tasks", *5th Intl. Conf. on Spoken Language Processing, ICSLP-98*, Sydney (AUS), 1998.

[Reynolds 92] D. Reynolds, *A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification*, Ph. D. Thesis, Georgia Institute of Technology, 1992.

[Reynolds 97] D. Reynolds, "HTIMIT and LLHDB: Speech Corpora for the Study of Handset Transducer Effects", *IEEE Intl. Conf. on Acoust. Speech and Signal Proc. ICASSP-97*, pp. 1535-1542, Munich (D), 1997.

[Steeneken 85] H. J. M. Steeneken and T. Houtgast, "RASTI : A Tool for Evaluating Auditoria", in *RASTI, Brüel & Kjær Technical Review*, No. 3, pp. 13-30, 1985.