FACING SEVERE CHANNEL VARIABILITY IN FORENSIC SPEAKER VERIFICATION CONDITIONS

Javier Ortega-Garcia, Santiago Cruz-Llanas and Joaquin Gonzalez-Rodriguez
DIAC-EUITT. Universidad Politécnica de Madrid
Ctra. Valencia, km. 7. E-28031 Madrid, Spain
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.

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

1. INTRODUCTION(*)

Coping with forensic identification implies dealing with speech variability [1]. 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 variability (kind of microphone, bandwidth and dynamic range reduction, electrical and acoustical noise, reverberation, etc). From all of these variability factors related, it is widely assumed that in forensic approaches, some of them highly degrade the speaker verification process, namely: channel influence, inter-session variability and emotional state [2, 3].

In this contribution we intend to explore the influence of channel and inter-session variability in the recognition process. For 'severe channel variability' we mean: mismatch in the telephone channels (kind of microphone and type of communication –internal/local/long distance calls), mismatch in the type of microphone employed for *in situ* recordings and, even, severe forensic cross-mismatching (training with telephone speech and testing with *in situ* speech).

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

Tests are accomplished through a text-independent GMM-based verification system, employing TelVoice [4] and AHUMADA/GAUDI [5] speech databases for these experiments. Results are provided making use of three sessions of telephone speech (internal routing, local and long distance sessions) and other three sessions of *in situ* stereo recordings (with four different microphones, one of them common for all sessions). For channel compensation purposes, Cepstral Mean Normalization (CMN) scheme will be applied. Severe cross-mismatching (telephonic speech vs. direct microphonic speech) will be also tested.

The paper is organized as follows: Section 2 quotes the speech databases employed in the experiments. In Section 3 the speaker verification system used is described. Section 4 shows the experimental results obtained, in both text-dependent and text-independent modes, while in Section 5 conclusions are presented.

2. SPEECH DATABASES

2.1. TelVoice Speech Database

TelVoice speech database [4], consists in 11 male and 9 female speakers, each of them pronouncing their passport number four times in each one of the 5 different telephonic sessions (local and long-distance calls). Experiments 1 and 2 will be accomplished with TelVoice data.

2.2. GAUDI/AHUMADA Speech Database

GAUDI/AHUMADA speech database [5], consists in 100 male and 100 female speakers, multi-session, multi-channel, microphonic/telephonic huge database. It 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. The rest of experiments (from Experiment 3 to Experiment 9) accomplished in this paper are making use of GAUDI/AHUMADA speech data.

3. SPEAKER VERIFICATION SYSTEM

In order to perform some speaker verification tests over the available data, a text-independent automatic speaker verification system, based in Gaussian Mixture Models (GMM) [6], has been employed. Tests have been

accomplished over a subset of (randomly selected) 25 speakers from the total number of 104 available speakers. All microphonic speech material used for training and testing has been down-sampled to 8 kHz (from the original sampling frequency of 16 kHz). 14 Mel-frequency cepstral coefficients (MFCC) plus 14 ΔMFCC have been used as feature vectors in all cases. Frames of 32 ms. taken every 16 ms., with Hamming windowing and pre-emphasis factor of 0.97 are used as input to the system.

Tests without normalization and with likelihood-domain normalization [7] 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 has been applied:

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

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.

4. EXPERIMENTAL RESULTS

4.1. Text-Dependent Speaker Verification

GMMs are statistical models in which temporal structure of speech is lost. This property makes them suitably for the text-independent recognition problem. Anyway, if training and testing is accomplished using the same phonetically-specific (short) utterance (PIN, password, passport code, etc), the model will be speaker *and* phonetically specific, rejecting in this way other non-specific utterances.

In order to determine this "text-dependent" (phonetically-specific) capability of GMMs, two different experiments are carried out [8]. These two experiments make use of TelVoice speech database.

Experiment 1 shows results (Table 1) when training with 1 session and testing with the 4 remaining sessions.

EER(%)	M	P1	P2	P3	P4	P5
NONE	8	**20.3**	**16.5**	**20.1**	**15.0**	**21.5**
	16	**19.0**	**15.4**	**19.5**	**15.2**	**25.7**
	32	**20.8**	**16.0**	**20.3**	**17.4**	**26.1**
CMN	8	**5.8**	**4.7**	**10.0**	**7.2**	**13.0**
	16	**6.4**	**5.3**	**10.3**	**7.7**	**13.0**
	32	**8.9**	**6.0**	**13.6**	**10.5**	**15.3**

Table 1. Verification results when training with session N (PN) and testing with the 4 remaining sessions in each case, varying the number of mixtures in the model (M), with score normalization, with CMN compensation or without channel normalization (NONE), expressed in terms of EER(%).

It can be seen in Table 1 the effectiveness of CMN channel compensation scheme in telephonic tests, due to the inherent variability of these kind of channels. Verification results vary slightly from test to test, confirming than the small amount of training data (4 utterances) produces better results when using only 8 mixtures, while 32 mixtures can produce an overestimated models.

Experiment 2 shows multi-session training results, when varying channel compensation scheme (CMN or none), number of mixtures per model and number of training sessions. Table 2 shows these results.

EER(%)	M	TR1	TR2	TR3	TR4
NONE	8	**20.3**	**15.8**	**10.4**	**8.1**
	16	**19.0**	**16.2**	**9.3**	**7.5**
	32	**20.8**	**15.5**	**8.7**	**6.7**
CMN	8	**5.8**	**3.6**	**3.3**	**2.7**
	16	**6.4**	**3.1**	**2.8**	**1.9**
	32	**8.9**	**3.0**	**2.2**	**1.5**

Table 2. Verification results when cumulative multi-session training with 1 (TR1), 1+2 (TR2), 1+2+3 (TR3), or 1+2+3+4 (TR4) training sessions, and testing with the remaining session(s) in each case, varying the number of mixtures (M), with score normalization, with CMN compensation or without channel normalization (NONE), expressed in terms of EER(%).

Experiment 2 shows some conclusive results, as using CMN and increasing the number of training sessions directly improves verification results. In this case, as multi-session training process implies more training data, increasing the number of mixtures also improves the results, achieving (in the best case) 1.5% EER.

4.2. Text-Independent Speaker Verification

All experiments included in this section make use of AHUMADA/GAUDI 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 M*J*/M*K* stands for same session, different microphone 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 3 (Table 3) 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 3. 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 3, 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.

4.2.1. Microphonic Speech

Experiment 4 (Table 4) 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 4. 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 3, 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 5 (Table 5) 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 5. 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 6 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 6.b) with respect to 6.a). Results of Experiment 6 are shown in Table 6.

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 6. 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 6 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%.

4.2.2. Telephonic speech

Experiments 4 through 6 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 7 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 7 shows the results of Experiment 7.

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 7. 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 7 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 8 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 8 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 8. 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 8.a) with respect to 7.c), in which EER decreases from 24.2% to 9.8%.

4.2.3. Telephonic/Microphonic Speech Mismatch

A frequent situation in forensic cases, specially in those in which no technical staff is involved in the known speech sample recording at Court, is to find that unknown speech samples usually come from telephone calls, while known speech is microphonic-quality.

Experiment 9 shows results (Table 9) when cross-channel mismatch (telephonic/microphonic) is present, namely, when training procedure is accomplished from telephonic session T3, and testing is carried out regarding microphonic session M7 (It should be reminded that M7 and T3 are acquired simultaneously). These results also include, in some cases, band-pass filtering (300-3400 Hz.) of the original data, in order to approach better to telephonic band-limited speech.

EER(%)	NO SCORE NORM.		SCORE NORM.	
CH. NORM. / TR/TS	NONE	CMN	NONE	CMN
a) T3/M7	**24.4**	**33.3**	**20.2**	**19.2**
b) T3/M7f	**31.1**	**31.3**	**28.8**	**13.0**
c) T3f/M7f	**24.6**	**31.4**	**20.9**	**14.2**

Table 9. Verification results when training (TR) with original telephonic speech T3 (and band-pass filtered speech, T3f), and testing (TS) with M7 (and band-pass filtered speech, M7f), with/without score normalization, with CMN compensation or without channel normalization (CH. NORM.), expressed in terms of EER(%).

Results in Table 9 show that, at least, microphonic speech M7 must be band-pass filtered in order to achieve (in the best case) 13% EER. When filtering also training data, 9.c), results are similar to 9.a), but CMN is more effective, achieving 14.2% EER as best.

5. CONCLUSIONS

All experiments demonstrate that score normalization is essential in speaker verification tasks. Benchmark Experiment 3 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 1, 4, 7, 8 and 9). It also produces EER decreasing when mismatch between sessions is encountered (Experiment 5), and when multi-session training is accomplished (Experiments 2, 6, and 8). Increasing training data is also a relevant issue (Experiment 2, 6.a, and 8), specially in forensic approaches, where speakers show low degree of cooperativeness but where large amounts of data are usually available.

6. ACKNOWLEDGEMENTS

We wish to thank our students Daniel Garcia-Romero, Raul Dorado, Oscar Ledesma, Miguel A. Chacon and Esteban Martinez for their hard work in obtaining the verification results shown in this contribution.

7. REFERENCES

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

[2] C. Champod and D. Meuwly (1998), "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.

[3] 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).

[4] L. Rodriguez-Liñares and C. Garcia-Mateo (1998), "On the Use of Acoustic Segmentation in Speaker Identification", *Proc. EUROSPEECH'97*, Rhode (GR), pp. 2315-2318.

[5] J. Ortega-Garcia *et al.* (1998), "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.

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

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

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