

Addressing database mismatch in forensic speaker recognition with Ahumada III: a public real-casework database in Spanish

Daniel Ramos¹, Joaquin Gonzalez-Rodriguez¹,
Javier Gonzalez-Dominguez¹ and Jose Juan Lucena-Molina²

¹ATVS - Biometric Recognition Group, Escuela Politecnica Superior
C./ Francisco Tomas y Valiente 11, Universidad Autonoma de Madrid E-28049 Madrid, Spain

²Acoustics and Image Processing Department, Criminalistic Service
Direccion General de la Policia y de la Guardia Civil, Ministerio del Interior, Madrid, Spain

daniel.ramos@uam.es

Abstract

This paper presents and describes Ahumada III, a speech database in Spanish collected from real forensic cases. In its current release, the database presents 61 male speakers recorded using the systems and procedures followed by Spanish Guardia Civil police force. The paper also explores the usefulness of such a corpus for facing the important problem of database mismatch in speaker recognition, understood as the difference between the database used for tuning a speaker recognition system and the data which the system will handle in operational conditions. This problem is typical in forensics, where variability in speech conditions may be extreme and difficult to model. Therefore, this work also presents a study evaluating the impact of such problem, for which a corpus quoted as NIST4M (NIST MultiMic MisMatch) has been constructed from NIST SRE 2006 data. NIST4M presents microphone data both in the enrolled models and in the test segments, allowing the generation of trials in a variety of strongly mismatching conditions. Database mismatch is simulated by eliminating some microphone channels of interest from the background data, and computing scores with speech from such microphones in unknown testing conditions as usually happens in forensic speaker recognition. Finally, we show how the incorporation of Ahumada III as background data is useful to face database mismatch in real-world forensic conditions.

1. Introduction

The interest in using automatic systems for forensic speaker recognition has increased in the last years. The main reasons for this are the improvement of accuracy in the technology [1] and a more comprehensive study about the role of automatic speaker recognition in forensic science [2]. However, speaker recognition technology have still important challenges to address. First, speech data scarcity remains a problem for automatic systems. Many forensic cases involve *recovered* questioned recordings presenting a small amount of speech, often under unfavourable quality conditions. Recent research has focused on increasing the robustness and accuracy of speaker recognition technol-

ogy under short-duration conditions [3]. Second, session variability mismatch introduces variation to different utterances of the same speaker, typically due to factors such as transmission channel, speaking style, speaker emotional state, environmental conditions, recording devices, etc. This variability seriously degrades the performance of a system, and its compensation has been subject of abundant recent research [4, 5]. In fact, this has been a major topic of research in recent Speaker Recognition Evaluations (SRE) conducted by NIST [1]. Finally, we identify the database mismatch problem, understood as the variation in the conditions between the dataset used for tuning an automatic speaker recognition system (referred to as background or development database) and the data used in real-world operational conditions (known as evaluation or operational database).

Database mismatch is a typical situation in forensic speaker recognition, mainly because the limitation in the availability of real-casework databases for system tuning, and also because the conditions of the speech in real-world forensic recording are extremely variable. Due precisely to session variability and extreme conditions, there are several frequent situations where the database mismatch may constitute a serious problem. On the one hand, problematic incriminatory questioned speech may typically include:

- Telephone wire-taps. The development database may not have speech data recorded in the same conditions as the wire-tapping system. This problem may become serious if an effort has not been made by the forensic laboratory in order to collect field data using their recording system at hand. The acquisition of such database is usually expensive and time-consuming, as the final corpus should have a significant size and represent the session variability of telephone recordings used in casework.
- Distant or hidden microphone, in cases where such a device is present on an individual or a prepared room. In this case, the environmental conditions of the recordings are extremely variable depending on the room characteristics, the microphone position and layout, etc. It is virtually impossible to simulate such conditions for the purpose of recording a representative database, especially if it is desired to take all possible cases into account.

On the other hand, control speech recordings from a given suspect may be problematic in the following usual situations:

- Telephone conversations for which the suspect recognizes to be the author, either recorded at police depen-

This work has been supported by the Spanish Ministry of Education under project TEC2006-13170-C02-01. J. G.-D. also thanks Spanish Ministry of Education for supporting his doctoral research. We also thank Ana Rosa Gonzalez-Sanz and people from the Acoustics and Image Processing Department from Guardia Civil for their important effort in collecting data for forensic purposes.

dencies or wire-tapped. This is exactly the same case as for a recovered wire-tap.

- Speech recorded using a microphone. Although the recording conditions are much more controlled than in the hidden microphone case for questioned recordings, variability in the room layout, environmental conditions and microphone types makes extremely difficult to predict which kind of speech will be needed for a representative background database.

Although database mismatch is commonly seen as a problem in the community and among forensic experts, to the authors' knowledge, rigorous experimental studies regarding this important issue in forensic speaker recognition have been mainly focused on telephonic databases. With the release of NIST multi-microphone databases in SRE 2005 and 2006 [1], a richer dataset has been available for research in the topic. In fact, realistic databases for forensic speaker recognition are important for two main reasons: first, more data will be available for research in database mismatch; and second, forensic laboratories will be able to improve the robustness of their speaker recognition systems in casework conditions.

Driven by these needs, this paper presents the Ahumada III database, a real-casework publicly available corpus in Spanish, which has been acquired by the Acoustics and Image Processing Department of Spanish Guardia Civil. In its current release, Ahumada III Release 1 (Ah3R1), it includes speech data from real forensic cases recovered using one of the typical recording systems from Guardia Civil, namely analog magnetic tapes containing GSM tapplings. Moreover, the database is being extended with more material under this platform and also using SITEL, a Spanish nationwide digital tapping system. Ah3R1 includes variability in conditions such as noise, environmental characteristics, emotional state, country and region of origin and dialect of speakers, etc. Next releases will significantly increase the amount of data presenting strong variability from real cases.

This work also explores the database mismatch problem using the presented Ahumada III database and a corpus generated from multi-microphone data from NIST SRE 2006, namely the NIST4M database (NIST MultiMic MisMatch). This paper is organized as follows. First, the Ahumada III database is described in the context of Guardia Civil operative procedures for forensic speaker recognition. The experimental section then presents results which illustrate the impact of database mismatch in system performance in two different ways: database mismatch using NIST multi-microphone corpora, and robustness in real-casework conditions using Ahumada III. Results show the importance of the database mismatch problem and encourages research in the field and data collection. Finally, conclusions are drawn.

2. Addressing database mismatch in real casework: the Ahumada III database

In the last years, Spanish Guardia Civil has done a significant effort on the application of automatic speaker recognition systems to forensic voice evidences [2]. Following a Bayesian framework, and pursuing transparency in their reports, much of this effort has concentrated in the assessment of their system in the sake of testability, as a demanding requirement in the new paradigm of forensic identification sciences [6]. Most voice evidences arriving to Guardia Civil laboratories have two possible origins. First, digitized analog magnetic recordings

from GSM mobile calls are typical in cases between 1995 and 2004. From those recordings of this type received in the last ten years, those authorized (case by case) by the corresponding judge after a trial, have been added to a database registered in the Spanish *Ministerio del Interior*, known as *Base de Datos de Registros Acústicos* (BDRA)¹. Second, nationwide digital interception system (SITEL) has been used since 2005 by the two Spanish State Police forces. This system records digital wire-taps directly connected to all mobile telephone operators.

With the purpose of robustness in real-case conditions and proper assessment of systems, Guardia Civil has recorded several speech databases in the last decade. Back in 1998, the Ahumada I corpus was collected [7] containing 100 male speakers in telephone and microphone recordings. A subcorpus of Ahumada I spontaneous telephone speech was used in NIST Speaker Recognition Evaluations both in 2000 and 2001. A complementary database known as Gaudi with 100 female speakers was recorded in 2001². From 2004 to 2006, the Baeza database was recorded, including 3 GSM and microphone spontaneous conversational speech sessions recorded at the same time, from 761 males and 72 females. Baeza has shown, up to this moment, the most relevant reference population to be used in real cases, as it was recorded through the same Spanish GSM network as in actual cases. In 2008, all 100 male speakers from Ahumada I are still available and are to be again recorded through GSM and SITEL. In this way, Ahumada II (as is to be known) will constitute a major contribution for the analysis of long-term stability and degradation of speaker features after ten years. As most new cases come through SITEL recordings and are to be, by this time, evaluated with Baeza as reference population, also in 2008 almost 100 speakers from Baeza (different from Ahumada speakers) are to be recorded again through SITEL. This database will be known as Ahumada IV and will be used for system assessment.

2.1. The Ahumada III database

Ahumada III consists of authorized conversational speech from real cases both from BDRA and SITEL. The expected size of the database in number of speakers and variety of conditions addressed is huge both in terms of number of available calls and amount of data. However, as conditions are not uniform, and speech recordings have to be authorized one by one, different releases of the database will be progressively available.

Ahumada III Release 1 (Ah3R1) consists of 61 speakers from a number of real cases with GSM BDRA calls across Spain, with a variety of country of origin of speakers, emotional and acoustic conditions, and dialects in the case of Spanish speech. The unique low variability dimension is gender, as all of them are male speakers. All 60 speakers in Ah3R1 have two minutes of speech available from a single phone call to be used as unquestioned (control) recording, with the purpose of model training or voice characterization. Additionally, ten speech segments for 31 speakers and five segments for 30 speakers are included for testing issues, each one from a different call. Such fragments present between 7 and 25 seconds of speech, with an average duration of 13 seconds. An evaluation protocol, equivalent to that of NIST SRE [1], is available and ready to be used. In this way, experienced NIST SRE par-

¹With reference *public scientific file number 1981420003 from Spanish Guardia Civil, Orden Ministerial INT/3764/2004 de 11 de noviembre*.

²A 25 Ahumada male subcorpus and an equivalent 25 Gaudi female subcorpus are freely available in <http://atvs.ii.uam.es/databases.jsp>.

ticipants can run and assess immediately their systems on the Ah3R1 data. Ah3R1 is publicly available for research purposes under a license agreement³ to be signed with Guardia Civil (contact: crim-acustica@guardiacivil.es). Sample speech segments are ready to be directly heard in ATVS website in order to perceive the quality and diversity of Ah3R1 recordings.

3. Experiments on database mismatch

In this section we present results illustrating the database mismatch problem, and how Ahumada III will improve accuracy in real forensic casework. For score computation, the ATVS GMM system has been used, where speech data known to come from a given speaker is represented using Gaussian Mixture Models (GMM) adapted from an Universal Background Model (UBM). GMM of 1024 mixtures have been used for modelling MFCC-based features with channel compensation based on feature warping and factor analysis [4].

3.1. Database mismatch with NIST4M database

An evaluation corpus, namely NIST4M (NIST MultiMic Mismatch), has been constructed using NIST SRE 2006 data. This corpus aims at simulating unfavourable session variability. Moreover, utterances in NIST SRE 2006 from all speakers having microphone conversations have been combined in NIST4M for trial generation, having multimicrophone data both in the enrolled models and in the test segments. Database mismatch has been simulated by eliminating all the data for a given microphone from the background data, using such microphone only in NIST4M evaluation data. Thus, trials for that microphone will be performed without considering such kind of data in the background set, and therefore under database mismatch. This situation is usual in forensic speaker recognition.

For the construction of NIST4M, half of the telephone utterances of each speaker were selected as testing segments and the rest as training data. For microphone data, 3 microphones in one session per speaker has been used for training data, and 6 microphones from each of the rest of sessions are used for testing. The microphone types changed with different speakers, in such a way that the data for all microphone types tend to be balanced. Trials were generated by performing all possible combinations between the same speaker utterances for target trials, and the next 12 speakers for non-target trials, avoiding different-gender comparisons.

Development data has been selected from past NIST SRE. Microphone data from NIST SRE 2005 has been divided in three sub-sets, separately used for training factor analysis parameters, UBM and T-Norm cohorts. This data has been complemented with NIST SRE 2005, 2004 and 2002 telephone data, as well as speech from Switchboard I⁴.

For experiments presented here microphones 3 (*m3*) and 5 (*m5*) as quoted by NIST [1] have been chosen. Both microphones present different characteristics, being *m3* a distant microphone and *m5* a higher-quality device. Two different testing conditions have been defined. On the one hand, we define a *same-microphone* condition where trials containing only a given microphone both in the training and testing speech data have been selected. For this condition, in the case of *m3* there are 33 target and 585 non-target scores, and in the case of *m5* there are 30 target and 540 non-target scores. Such numbers are quite small, and therefore results obtained for this condition

should be carefully interpreted. However, we have considered them here in order to illustrate database mismatch when session variability is controlled. On the other hand, we define a *microphone-model* condition, where trials containing a given microphone in the models and every possible channel in the test segment have been considered. For this condition, when *m3* is selected there are 498 target and 6095 non-target scores; whereas 515 target and 5898 non-target scores have been generated for *m5*.

For database mismatch simulation, on the one hand, the scores for the *m3* microphone-model condition in a database matching situation will have data from NIST SRE 2005 *m3* microphone in the UBM, factor analysis data and T-Norm cohort. On the other hand, in database mismatch conditions data recorded over NIST SRE 2005 *m3* microphone will be eliminated from the UBM, factor analysis data and T-Norm cohort. The same procedure is applied to *m5*. It is needed to remark that the data from a given microphone represents only a 12, 5% of the whole data used for training UBM and factor analysis, and only 30 up to 340 models in the whole T-Norm cohort. This is important in order to evaluate the impact of the elimination of such a relatively small amount of data from the background set.

3.1.1. Results

Table 1 shows the effect of database mismatch for the same-microphone condition described above. Performance is measured in the form of equal error rate (EER) and Detection Cost Function (DCF) as defined by NIST [1]. Results show that database mismatch leads to a significant degradation in performance, especially for the DCF performance measure. However these results should be carefully interpreted, as the number of scores in each condition is scarce.

	DB match		DB mismatch	
	EER	DCF	EER	DCF
<i>m3</i> vs. <i>m3</i>	15.2137	0.0391	17.094	0.0472
<i>m5</i> vs. <i>m5</i>	1.4815	0.0033	2.2222	0.0143

Table 1: Performance of same-microphone condition under database mismatch.

Figure 1 illustrates the effects of database mismatch for the microphone-model condition. This aims to simulate a typical forensic situation, where speech from the suspect may have been recorded in particular microphone conditions for which a background database is not available. DET curves show a significant degradation in discrimination performance when the microphone of the considered model is eliminated from the background data.

3.2. Experiments with Ahumada III

In this section, the use of Ahumada III for compensating database mismatch in real-casework data is illustrated. The simulation has been conducted by adding Ahumada III data to the UBM trained with NIST SRE data described in Section 3.1. Although more accurate results can be obtained by using Baeza or BDRA databases from Guardia Civil, we selected a NIST-SRE-based background dataset for the illustration of database mismatch, and also for transparency, because any laboratory having the NIST SRE databases can test the same results in Ahumada III. The same ATVS GMM system has been used for performing the experiments, except for session variability compensa-

³ Available at <http://atvs.ii.uam.es/databases.jsp>

⁴ See LDC website <http://www ldc.upenn.edu/>

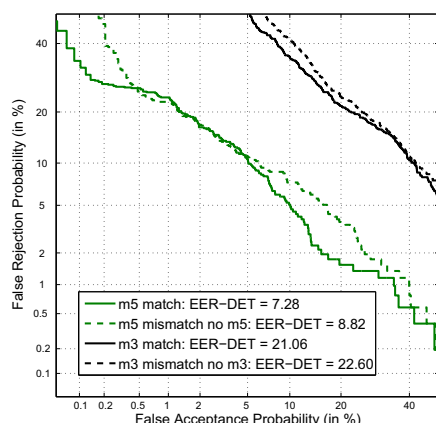


Figure 1: DET curves for the microphone-model condition, illustrating the effect of database mismatch.

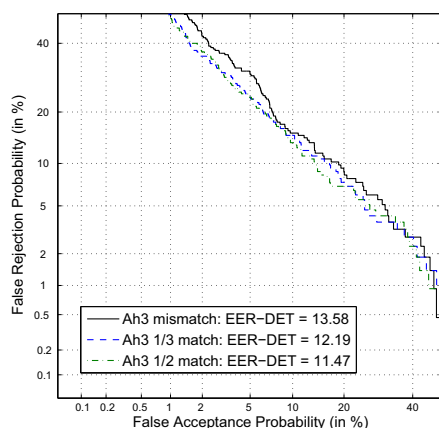


Figure 2: DET curves for experiments using Ahumada III. Database mismatch is reduced by increasingly adding Ahumada III data to the NIST-SRE-based UBM.

tion, which has not considered factor analysis. Moreover, due to data scarcity in Ah3R1, we will not explore effects in factor-analysis-based session variability compensation or T-Norm cohorts, but it will be considered for future work as next releases of Ahumada III will be available.

In order to simulate the effect of using Ahumada III in real casework, we have used a cross-validation procedure, where we rotate subsets of the speakers used for UBM training (30 or 20 from the 60 speakers considered in the test) and compute scores with the rest of available speakers (30 or 40). We call both conditions the 1/2 and 1/3 match respectively. Thus, we obtain a number of 300 target trials for both conditions, 9700 non-target trials for the 1/2 match condition and 5700 non-target trials for the 1/3 match condition. In both situations, the proportion of Ahumada III speech data with respect to NIST SRE data in the UBM does not exceed 15% after silence removal.

Figure 2 shows the DET plots of the proposed experiments. It is shown that the inclusion of Ahumada III in the UBM increases discrimination accuracy, as expected. Moreover, the

increase is highly significant considering the small percentage (less than 15%) of Ahumada III data with respect to NIST SRE data in the UBM.

4. Conclusions

This paper has presented the first release of Ahumada III, a speech corpus in Spanish from real police investigations. The database is being built as a tool for aiding forensic laboratories to assess and tune their speaker recognition systems in the most possible realistic conditions. In this sense, the important problem of database mismatch has been explored using the presented Ahumada III database and also other multichannel corpora such as NIST4M, a database constructed from NIST SRE 2006 multichannel data. In this work, database mismatch has been simulated in UBM modelling, channel compensation based on factor analysis and T-Norm cohorts. As a result, it has been observed that database mismatch has a critical impact in the performance of systems, even if the background data matching the operational speech conditions is relatively small with respect to the whole dataset used for tuning. It is also shown how databases such as Ahumada III help to compensate for this important problem. There have been many topics which have not been explored in this work, such as the influence of database mismatch in session variability, its impact in the calibration and *LR* computation and a more rigorous study for the different microphone channels in the NIST4M corpus. However, with Ahumada III new releases, scientists will be able to deeply explore such problems in a realistic forensic framework.

5. References

- [1] M. A. Przybicki, A. F. Martin, and A. N. Le, "NIST speaker recognition evaluations utilizing the Mixer corpora-2004, 2005, 2006," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1951–1959, 2007.
- [2] J. Gonzalez-Rodriguez, Phil Rose, D. Ramos, Doroteo T. Toledano, and J. Ortega-Garcia, "Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.
- [3] Benoît Fauve, Nicholas Evans, and John Mason, "Improving the performance of text-independent short duration SVM- and GMM-based speaker verification," in *Proc. of Odyssey*, Stellenbosch, South Africa, 2008.
- [4] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in gmm-based speaker verification," *IEEE Transactions on Audio, Speech and Signal Processing*, vol. 15, no. 4, pp. 1448–1460, 2007.
- [5] R. Vogt and S. Sridharan, "Explicit modelling of session variability for speaker verification," *Computer Speech and Language*, vol. 22, no. 1, pp. 17–38, 2007.
- [6] M. J. Saks and J. J. Koehler, "The coming paradigm shift in forensic identification science," *Science*, vol. 309, no. 5736, pp. 892–895, 2005.
- [7] J. Ortega-Garcia, J. Gonzalez-Rodriguez, and V. Marrero-Aguilar, "AHUMADA: A large speech corpus in Spanish for speaker characterization and identification," *Speech Communication*, vol. 31, no. 2-3, 2000.