A Bayesian network approach combining pitch and spectral envelope features to reduce channel mismatch in speaker verification and forensic speaker recognition

Mijail Arcienega, Anil Alexander, Philipp Zimmermann† *and Andrzej Drygajlo*

Signal Processing Institute, Swiss Federal Institute of Technology (EPFL)
†Institut de Police Scientifique, University of Lausanne (UNIL)
Lausanne, Switzerland
mijail.arcienega@epfl.ch, alexander.anil@epfl.ch,
philipp.zimmermann@unil.ch, andrzej.drygajlo@epfl.ch

Abstract

The aim of this paper is to reduce the effect of mismatch in recording conditions due to the transmission channel and recording device, using conditional dependencies of prosodic and spectral envelope features. The developed system is based on a Bayesian network framework which combines statistical models of the pitch and spectral envelope features. This approach is applied to forensic automatic speaker recognition, where mismatched recording conditions pose a serious problem to the accurate estimation of the strength of voice evidence. The method is evaluated using a forensic speaker recognition database that contains three different recording conditions typical to forensic tasks. The performance of the system is evaluated using both speaker verification as well as forensic speaker recognition measures.

1. Introduction

Human recognition of speakers is affected by recording and environmental conditions, and human judgments on the identity of speakers become less reliable in adverse conditions [1]. Differences in the transmission channel, recording devices, and environmental conditions can introduce distortion in speech that is the main source of mismatch in recording conditions. Automatic speaker recognition systems are also vulnerable to this problem of channel mismatch which is a significant problem in applications involving modern communication networks, such as forensic speaker recognition.

Conditions typical to forensic cases, in which recordings are made by the police (anonymous calls and wiretapping), cannot be controlled and are far from ideal for automatic speaker recognition. Mismatch due to differences in the phone handset and the transmission channel, and background noise, can affect the estimation of the strength of evidence in forensic speaker recognition [2].

The Gaussian mixture models (GMMs) have been successfully applied to text-independent speaker recognition systems where they have been used to model the spectral envelope [3]. The effect of channel distortions and noise on the performance of such systems is a serious concern. Although prosodic features are known to be less affected by these impairments than spectral envelope features, interest in using these features had diminished over the years, as these features alone did not give the accuracy required by automatic systems. Prosodic features are worth re-examining for speaker recognition systems, in the context of mismatch due to channel distortions. Mismatch problems are of increasing significance in tasks such as forensic automatic speaker recognition.

In this paper, we present a method using Bayesian networks [4, 5] to combine prosodic features with those of the spectral envelope in order to reduce the effects of channel mismatch. The performance of the system is evaluated using both speaker verification as well as forensic speaker recognition measures. We compare the performance of the GMM-based system using only spectral envelope features, with a Bayesian network approach, capable of taking advantage of the dependencies between the pitch and spectral envelope features. We evaluate both systems using data from a forensic speaker recognition database containing three different channel conditions.

2. The Bayesian Network based system

The Bayesian network (BN) used in this paper, is built on the same principles as the one presented in [6]. The main idea is to build conditional models (for the pitch ϱ as well as for the spectral envelope features $\vec{x}$) given an auxiliary variable, the voicing status s.

The voicing status s is introduced in order to better capture the variations and to better model the distributions of voiced and unvoiced features. At time t, the voicing status can either be $s_t = 1$ (voiced) or $s_t = 2$ (unvoiced).

Spectral envelope features aim at modeling the envelope of the spectrum, excluding, as much as possible, the characteristics of harmonics that are related to the pitch. Therefore, the pitch and the spectral envelope carry complementary and uncorrelated information, and one can assume that given the voicing status (s_t), $\vec{x}_t$ and ϱ_t are conditionally independent, i.e.:

$$p(\vec{x}_t|\varrho_t, s_t) = p(\vec{x}_t|s_t). \tag{1}$$

The Bayesian network associated to the features at time t is shown in Figure 1.

2.1. The Conditional Models

Two Gaussian mixture models (GMMs) are used for representing the spectral envelope features. One of them ($\lambda_1^{\vec{x}}$) models the voiced part of speech while the second ($\lambda_2^{\vec{x}}$) models the un-

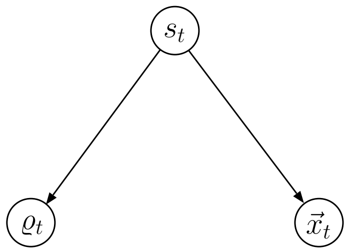

Figure 1: Bayesian network associated to the voicing status s, the pitch ϱ and the spectral envelope $\vec{x}$ at time t.

voiced part:

$$\lambda_i^{\vec{x}} = \left\{ c_{i,m}^{\vec{x}}, \mu_{i,m}^{\vec{x}}, \sigma_{i,m}^{\vec{x}} \right\}, \tag{2}$$

where, $i = 1, 2$ represents the voicing status and $m = 1, ..., M_i^{\vec{x}}$; $M_i^{\vec{x}}$ being the number of mixtures associated to each GMM.

The pitch modeling also depends on the voicing status. In voiced zones, one GMM is used for modeling the statistical properties of the pitch values, i.e., $p(\varrho|s = 1)$ is defined by a GMM with parameters

$$\lambda^{\varrho} = \{c_m^{\varrho}, \mu_m^{\varrho}, \sigma_m^{\varrho}\} \ ; \tag{3}$$

where $m = 1, ..., M_1^{\varrho}$; M_1^{ϱ} being the number of mixtures used for modeling the distribution of the pitch.

In unvoiced regions, a value for the pitch does not physically exist; nevertheless, a table of discrete probabilities can be still used to represent it. If we set $\varrho_t = 0$ in these regions, the probability $p(\varrho = 0|s = 2)$ will always equal 1. The pitch can therefore be characterized by $p(\varrho = 0|s = 2) = 1$ and $p(\varrho \neq 0|s = 2) = 0$.

Finally, the voicing status s probabilities are defined by two weights, w_1 and w_2, that represent the probabilities of being in a voiced zone, $p(s = 1)$, and the probability of being in a unvoiced zone, $p(s = 2)$, respectively.

The complete set of training data that belongs to an utterance, $O = \{\eta_1, ..., \eta_T\}$, where $\eta_t = \{\varrho_t, \vec{x}_t\}$, $l = 1, ...L$, and the sequence of states, $S = \{s_1, ..., s_T\}$, are therefore completely modeled by the Bayesian network (represented in Figure 1) with parameters λ:

$$\begin{array}{rcl} p(s = i) & = & w_i, \\ p(\vec{x}|s = i) & & \text{defined by } \lambda_i^{\vec{x}}, \\ p(\varrho|s = 1) & & \text{defined by } \lambda^{\varrho}, \quad \text{and} \\ p(\varrho = 0|s = 2) = 1 & ; & p(\varrho \neq 0|s = 2) = 0. \end{array} \tag{4}$$

2.2. Training

The sequence of state S is extracted at the same time as the pitch estimates with the reliable voiced/unvoiced decision algorithm presented in [7]. Vectors $\vec{x}$ are then separated into to voiced and unvoiced groups. The multivariate probability density function of the vectors in each group is trained using the Expectation-Maximization (EM) algorithm. The parameters for the model of the pitch in voiced zones are also calculated with the EM algorithm.

2.3. Likelihood Estimation

Let $O = \{\eta_1, ..., \eta_T\}$ be a test sequence and $S = \{s_1, ..., s_T\}$ the corresponding voicing status sequence.

Following [6], the likelihood measure, $p(O|S, \lambda)$, is equal to

$$p(O|S, \lambda) = p(X|S, \lambda) \cdot p(P|S, \lambda) \ ; \tag{5}$$

where X is the set of spectral envelope feature vectors and P the set of pitch values.

X can be further separated into X_V, the voiced part and X_U, the unvoiced part.

$$p(O|S, \lambda) = p(X_V|\lambda_1^{\vec{x}}) \cdot p(X_U|\lambda_2^{\vec{x}}) \cdot p(P_V|\lambda^{\varrho}) \ ; \tag{6}$$

where P_V represents the set of pitch estimates. One can see that the likelihood measure is the multiplication of likelihoods obtained separately for the voiced and unvoiced parts of the spectral envelope and the pitch.

3. Description of database used in experiments

In this study, the *EPFL-IPSC03* database was used. This database for forensic speaker recognition is being recorded by the Institut de Police Scientifique (IPS), University of Lausanne, and the Signal Processing Institute, Swiss Federal Institute of Technology, Lausanne (EPFL). It contains speech from over 60 male speakers in three different recording conditions and several different controlled and uncontrolled speaking modes. The male speakers, aged between 18 and 50, are all university educated individuals speaking in Swiss French. The recording conditions of this database include speech transmitted through a public switched telephone network (PSTN), global system for mobile communications (GSM) network, and directly recorded in a calling room using a digital recorder.

These recordings were made in controlled conditions in a quiet room. The two telephones used were a telephone connected to a fixed line and a mobile telephone. The cues were presented to the subjects in the form of a printed handout (text and images). A SONY electret condenser microphone (CARDIO ECM-23) was placed at a distance of about 30 cm from the mouth of the speaker, and was connected to SONY portable digital recorder (ICD-MS1).

In order to study the effects of the telephone channel on the voice, all the telephone calls were made from the recording room to a remotely located ISDN server. The ISDN transmission standard used was the European ISDN (DSS1), and an answering machine application was used to record the telephone calls.

Six segments of speech for 20 speakers in PSTN, GSM and direct room recordings, three of which were used for the mock questioned recordings (between 15 and 40 seconds each) and three longer recordings were used as reference recordings (between 30 seconds and 180 seconds) for the mock suspected speaker. Reference data from 10 speakers in three different conditions was used to train universal background models (UBM) for each condition. The test data chosen contained spontaneous and read speech, as well as simulated dialogue.

4. Experiments

For these experiments, all sources of signal were downsampled to 8 kHz. 10 speakers of the database, described in the previous section, were used to build background models for each condition, and 20 other speakers were taken to be mock suspects for the purpose of performing the speaker recognition. Approximately two minutes of speech per speaker were used to train the background model, and one minute of speech was used to adapt the client models. Six different utterances of approximately 30 seconds each, were used for the tests. In the background model, 512 mixtures were used for the spectral envelope GMM and 64 mixtures for the pitch GMM. The results, presented in this section aim at comparing the performances of the Bayesian network (BN) system proposed here with the classical GMM-UBM based system [3]. The spectral envelope features used in the experiments are the Mel-frequency cepstral coefficients (MFCCs).

In the first experiment, the training data for the background model as well as the client models is speech recorded through a PSTN. Mismatched channel conditions were simulated using a) speech recorded through a PSTN, b) speech recorded through a cellular-telephone (GSM) and c) speech recorded in the calling room (Room). Figure 2 and Table 1 show the equal error rates (EERs) of a classical GMM-UBM based speaker verification system compared to the Bayesian network system.

Table 1: EERs when the training data is speech recorded through PSTN

Speech used for Tests	GMM-UBM EER [%]	BN system EER [%]
PSTN	4.8	3.3
GSM	42.3	31.9
Room	37.5	22.5

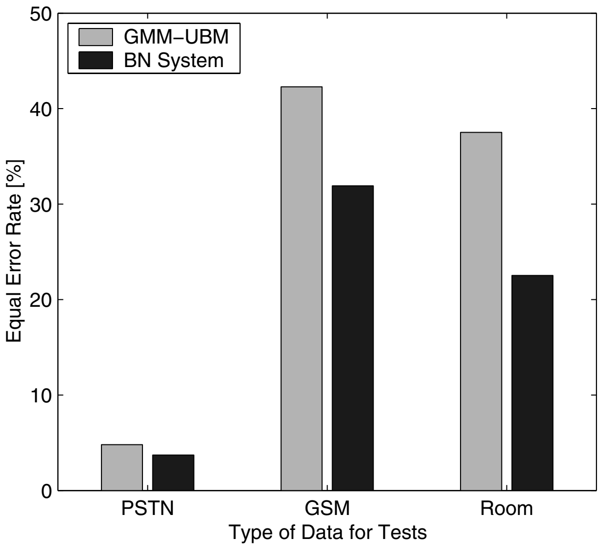

Figure 2: The performances of a classical GMM-UBM based speaker verification system as compared to the Bayesian network system. The training data is speech recorded through PSTN

In the second experiment, the background model as well as the client models are trained with speech recorded in the calling room. The tests for mismatch are performed, as in the first experiment, with all the three kinds of speech sources. Figure 3 and Table 2 show the EERs obtained.

As we can see in Figs 2 and 3, all EERs are reduced when incorporating the pitch. The improvement is small in matched conditions (PSTN vs. PSTN, or Room vs. Room), where the spectral envelope features are not affected, but is significant in mismatched conditions (all other comparisons), where the channel degrades the spectral envelope features.

Table 2: EERs when the training data is speech recorded in the calling room.

Speech used for Tests	GMM-UBM EER [%]	BN System EER [%]
Room	1.8	1.0
GSM	22.8	18.9
PSTN	25.8	20.4

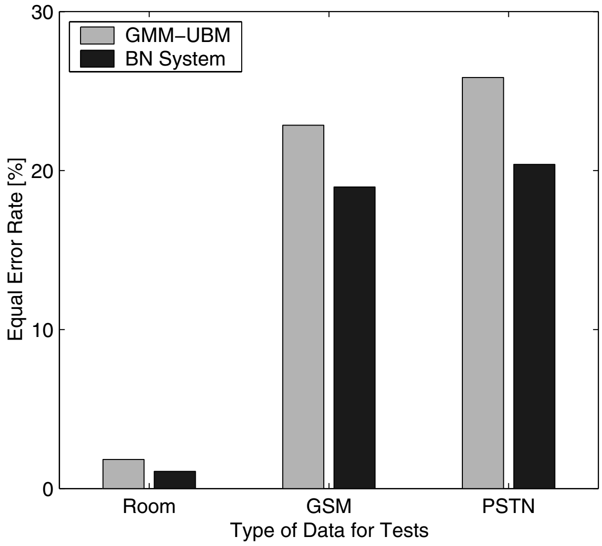

Figure 3: The performances of the proposed system as compared to a classical GMM-UBM system. The training data is speech recorded in the calling room.

5. Forensic speaker recognition evaluation

In forensic speaker recognition, there are cases where only one recording of the suspect is available due to the nature of the investigation, e.g., when it is not possible to have additional recordings of the suspect's voice, as it may alert him to the fact that he is being investigated, it is often necessary to perform one-to-one comparisons of the questioned recording and the recordings of the suspect's voice. The log-likelihood score obtained on comparing the questioned recording and suspected speaker's voice is called the evidence score (E). The strength of this evidence, expressed by the likelihood ratio, is evaluated with respect to two competing hypotheses: H_0 - two recordings have the same source, and H_1 - two recordings have different sources [8].

The framework is similar to the speaker verification domain where the task is to compare two recordings and conclude whether they have the same or different sources. Normally, a threshold is used in the verification domain to decide whether the two recordings come from the same source. In the forensic domain, it is not acceptable to use such a threshold, and measures such as the detection error tradeoff (DET) curves and the equal error rates (EER) (presented in the previous section) can only be used to measure the performance of the speaker verification systems.

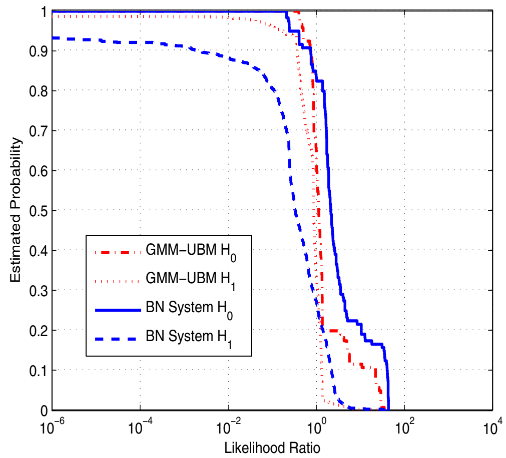

Figure 4: Tippett plots PSTN-Room

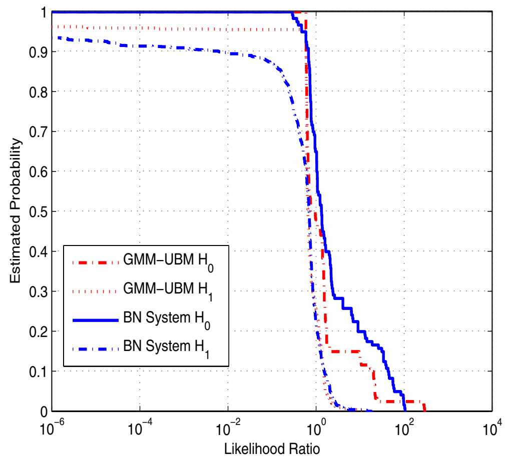

Figure 5: Tippett plots PSTN-GSM

The performance of forensic speaker recognition systems can be represented using probability distribution plots such as the Tippett plots $P(LR(H_i) > LR)$ (Figures 4 and 5). The integration of the probability distribution of LRs, which can be used to represent how many cases are above a given value of likelihood ratio with respect to each hypothesis, is called the Tippett Plot. This representation has been used in the interpretation of the results of forensic DNA analysis [9]. The extent of separation between the curves of the H_0 and H_1 score distributions is an indication of how well the system differentiates between cases where the suspect is indeed the source of the questioned recording and cases where the suspect is not the source of the questioned recording in terms of likelihood ratios.

In Figure 4 we observe that when the BN system is applied to mismatched conditions (using PSTN recording for training and room recording for testing), there is a considerable increase in the separation between the two curves, indicating a better performance. Similarly, in Figure 5), when the PSTN recording is used for training and the GSM recording for testing, there is an improvement in the performance, although it is not as significant as in the case presented in Figure 4. The mismatch due to PSTN vs. GSM training and testing conditions has a more pronounced effect on the strength of evidence than PSTN vs. Room training and testing conditions.

6. Conclusions

The Bayesian network approach presented in this paper has proved its capacity to exploit the information carried by the additional features (pitch and voicing status) in order to improve the recognition scores in mismatched conditions. The pitch, carrying information about the speaker's identity, has already been proved to be strongly robust to noise [6], and here we show that is also robust to channel distortions. Convolutional modifications in speech such as the ones introduced by PSTN or GSM channels may severely affect spectral envelope features but have almost no influence on the pitch. Incorporating these prosodic features, using a Bayesian network, has been shown to improve the performance of both speaker verification and forensic speaker recognition systems in mismatched training and testing conditions.

7. References

[1] A. Alexander, F. Botti, D. Dessimoz, and A. Drygajlo, "The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications," *Forensic Science International*, vol. 146, pp. S95–S99, December 2004.

[2] W. Campbell, D. Reynolds, J. Campbell, and K. Brady, "Estimating and evaluating confidence for forensic speaker recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, (Philadelphia, USA), pp. 717–720, 2005.

[3] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, January/April/July 2000.

[4] F. V. Jensen, *Bayesian Networks and Decision Graphs*. Springer-Verlag, 2001.

[5] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann Publishers, 1988.

[6] M. Arcienega and A. Drygajlo, "A Bayesian network approach for combining pitch and reliable spectral envelope features for robust speaker verification.," in *AVBPA*, vol. 2688 of *Lecture Notes in Computer Science*, pp. 78–85, Springer, 2003.

[7] M. Arcienega and A. Drygajlo, "Robust voiced/unvoiced decision associated to continuous pitch tracking in noisy telephone speech," in *International Conference on Spoken Language Processing*, vol. 4, (Denver, Colorado USA), pp. 2433–2436, September 2002.

[8] F. Botti, A. Alexander, and A. Drygajlo, "An interpretation framework for the evaluation of evidence in forensic automatic speaker recognition with limited suspect data," in *Proceedings of 2004: A Speaker Odyssey*, (Toledo, Spain), pp. 63–68, 2004.

[9] I. W. Evett and J. S. Buckleton, "Statistical analysis of STR data : Advances in forensic haemogenetics," *Springer-Verlag*, vol. 6, pp. 79–86, 1996.