8 Conclusion

In this thesis we propose a new approach to estimate and compensate for the effects of mismatch arising in forensic automatic speaker recognition casework in a corpus-based Bayesian interpretation framework, due to the technical conditions of encoding, transmission and recording of the databases used. Mismatch in the recording conditions between databases used in analysis can lead to erroneous or misleading estimates of the strength of evidence, and it is of utmost necessity to quantify and reduce the uncertainty introduced in the likelihood ratios, in order to avoid possible miscarriages of justice.

We investigated two main directions in applying the Bayesian interpretation framework to problems in forensic speaker recognition, the first concerning the problem of mismatched recording conditions of the databases used in the analysis and the second concerning the Bayesian interpretation as applied to real forensic case analysis.

The main contributions related to handling mismatched recording conditions of the databases used in forensic speaker recognition, within a Bayesian interpretation framework, include a methodology for estimating and statistically compensating for the differences in recording conditions, guidelines for the creation of a forensic speaker recognition database that can be used in order to perform forensic casework in mismatched conditions and the analysis of mismatched technical conditions in training and testing phases of speaker recognition and their effect on human-aural and automatic forensic speaker recognition.

The contributions related to Bayesian interpretation applied to real forensic case analysis include the analysis of the variability of the strength of evidence using bootstrapping techniques, statistical significance testing and confidence intervals, handling

cases where the suspect data is limited, and using complementary information to the likelihood ratio about the risk of errors involved in choosing a certain hypothesis.

8.1 Handling mismatched recording conditions in the Bayesian interpretation framework

8.1.1 Handling mismatch in recording conditions in corpus-based forensic speaker recognition

If an existing mismatch between databases is undetected, it is likely that the uncompensated usage of the Bayesian interpretation framework gives erroneous results. After detecting a mismatch, the expert has a choice of selecting another database compatible with the questioned recording (if possible), deciding not to analyze the case in the Bayesian interpretation framework or performing statistical compensation for the mismatched conditions. Detecting and compensating mismatches between databases helps to reflect more accurately the similarity or dissimilarity of the voices in the recordings.

We have introduced a methodology to estimate and statistically deal with differences in recording conditions of the databases used at the level of scores. This statistical compensation for mismatch is applied at the level of scores, as opposed to the level of features or of models. This method of statistical compensation is thus universally applicable to the many different combinations of feature extraction and modeling techniques used in automatic speaker recognition.

Under the assumption that the recording conditions of the questioned recording as well as potential population database are known precisely, it is possible to use databases for scaling score distributions, named 'scaling databases'. These scaling databases contain a set of the same speakers in various different recording conditions, in order to estimate parameters for scaling distributions to compensate for mismatch. In our experiments, we have observed that the number of speakers in the scaling database does not need to be very high (in the order of tens) to estimate the shift in parameters. While statistical compensation allows us to calculate likelihood ratios that are closer to the likelihood ratios obtained in matched conditions than the corresponding likelihood ratios in mismatched conditions, they are not *exactly* the same as the likelihood ratios in matched conditions. The compensation method is an approximate transformation based on means and variances and there is still a certain extent of error, especially with distributions that deviate from the Gaussian distribution significantly.

Although primarily the Bayesian interpretation framework is used for the evalu-

ation of evidence in court, it is also a valuable tool for investigatory purposes. The methodology of compensating mismatch is equally important for both these purposes in order to avoid results that are affected by differences in recording conditions. When it is not possible to use suspect reference recordings that are recorded in exactly the same conditions as those of the potential population, compensation reduces the effect of mismatch, and estimates scores that could have been obtained if the suspect reference and potential population databases had been recorded in the same conditions.

8.1.2 Methodology for creating databases to handle mismatch

We have proposed a methodology to create a forensic speaker recognition database in order to estimate the mismatch in recording conditions that arises in forensic cases, to compensate for its effects and to quantify the uncertainty that is introduced due to changing conditions. This database contains:

- One or more databases in the most commonly encountered recording condition, which contains a sufficient number of speakers that can serve as a potential population database.
- A subset of speakers present in these databases, are used to record several smaller databases in different recording conditions which contains a sufficient number of speakers from which distribution scaling parameters representative of the difference in recording conditions can be calculated.

Forensically realistic databases for evaluation of forensic speaker recognition casework have been created according to the requirements of the Bayesian interpretation methodology in order to validate the methods proposed. Two of the databases used in this thesis, namely the IPSC-02 and the IPSC-03 [see Appendix B and C] were recorded during the course of this thesis and were created as specifically adapted for use in the Bayesian interpretation methodology. In order to perform statistical compensation, we require such a database with the same speakers in different recording conditions, from which these statistical compensation parameters can be estimated.

8.1.3 Mismatched recording conditions and their effect on aural and automatic forensic speaker recognition

We have analyzed mismatched technical conditions, and their effect on forensic aural and automatic speaker recognition. With perceptual speaker recognition tests performed with laypersons as well as a baseline automatic speaker recognition system, it was observed that in matched recording conditions of suspect and questioned

recordings, the automatic systems showed better performance than the aural recognition systems. In mismatched conditions, however, the baseline automatic systems showed a comparable or slightly degraded performance compared to the aural recognition systems. The extent to which mismatch affected the accuracy of human aural recognition in mismatched recording conditions was similar to that of the automatic system, under similar recording conditions. Thus, the baseline automatic speaker recognition system should be adapted to each of the mismatched conditions in order to increase its accuracy. The adapted automatic system shows comparable or better performance than aural recognition in the same conditions. The perceptual cues that human listeners rely upon, in order to identify speakers, were analyzed. It was suggested that the accuracy of automatic systems can be increased using these perceptual cues that remain robust to mismatched conditions.

8.2 Applying Bayesian interpretation methodology to real forensic conditions

8.2.1 Scoring method and direct methods for the evaluation of the likelihood ratio

In addition to the univariate evaluation of the likelihood ratio, using the scores, we have shown that the multivariate approach of directly using the likelihood of observing features, given the statistical models can also be used for the estimation of the strength of evidence in forensic automatic speaker recognition, especially when the available suspect data is limited. These two approaches, named the Direct Method and the Scoring Method, differ in that one directly uses the likelihoods returned by the Gaussian Mixture Models (GMMs) and the other models the distribution of these likelihood scores and then derives the likelihood ratio on the basis of these score distributions. Statistical representations using probability distributions like the Tippett plots to evaluate the strength of evidence and to compare the two methods were also presented.

8.2.2 Bayesian interpretation in cases with sufficient and insufficient suspect reference data

A general interpretation framework to handle forensic cases using automatic speaker recognition, both in situations where there is a limited duration of recordings for the suspected speaker as well when the length of the recordings is sufficient to estimate the within-speaker variability, has been used. In many cases only one recording of

the suspected speaker is available due to the nature of the investigation. When suspect data is limited, the within-source variability of individual speakers can be approximated by the average within-source variability from databases that are similar in recording conditions to the databases used in the case.

8.2.3 Analysis of the variability of the strength of evidence

The strength of evidence, or the likelihood ratio, in forensic speaker recognition was shown to depend on various influences in the analysis, such as the approximation in the mathematical modeling of the score distributions, the choice of speakers in the potential population and the choice of the recordings to estimate the within-source variability of suspected speaker's voice. The likelihood ratio is seen to be variable and is often better represented using a range of possible values than a single number. This variability has been analyzed using a statistical significance analysis and a subset bootstrapping technique. Confidence intervals for the likelihood ratio estimates have been derived, and these intervals have been presented in the context of equivalent verbal scales used for reporting likelihood ratios to the courts. Considering the effect of this variability in the likelihood ratios for a given case, the expert should evaluate the following:

- The statistical significance of the evidence score obtained with respect to each of the hypotheses.
- The likelihood ratio, accompanied by confidence interval which gives a range of possible values of the strength of evidence.
- The equivalent of the range of the likelihood ratio on the verbal scale, and an explanation of the meaning of the scale.

8.2.4 Complementary measures to the strength of evidence

While the likelihood ratio provides an estimate of the strength of the evidence with respect to the two hypotheses, it does not consider the risk of errors in choosing either one. We have used a complementary measure, the Error Ratio (ER), to the likelihood ratio for interpretation of the evidence, which takes into consideration the relative risk of error for an evidence score in choosing either of the hypotheses. It is the proportion of cases for which recordings from the same source would be wrongly considered to come from different sources, divided by the proportion of cases in which recordings from different sources were wrongly considered to be from the same source, if the evidence score E is used as a threshold in a hypothesis test for match or non-match. Additionally, the error ratio is less sensitive than the likelihood ratio to artifacts of

the modeling each of the distributions as it is a ratio of areas and not a ratio of heights. While likelihood ratio measures the strength of the evidence, the error ratio gives complementary information, about the quality of the match in a score-based framework (given the trace and the databases), to interpret the observed value of the evidence. The likelihood ratio and the error ratio cannot be used instead of each other as they have different meanings and different evolutions.

8.3 Future directions

The work presented in this thesis can be extended in several directions:

- Combining the strength of evidence using aural-perceptive and acoustic-phonetic approaches (aural-instrumental) of trained phoneticians with that of the likelihood ratio returned by the automatic system. In combining the strength of evidence, it is necessary to take into consideration, whether the characteristics considered in aural-instrumental and automatic methods are statistically independent, and if this is not the case, this interdependence should be taken into account when evaluating the overall likelihood ratio.

- One of the assumptions made in the statistical compensation methodology, is that information regarding conditions of recording of the databases used in the analysis is known to the expert or can be requested for. However, this may not always be the case, and it can be necessary for him to determine the recording conditions by himself. In such situations, methods for automatically determining the recording conditions of the recordings in a case can prove very useful.

- Databases for both casework and research in forensic speaker recognition are necessary in the data-driven Bayesian approach. While we have presented guidelines for the creation of a database to handle the problem of mismatched recording conditions, there is a need for protocols for creating more general, forensically realistic databases, to handle the various kinds of situations encountered in forensic speaker recognition casework.

- In this work, the mismatched conditions considered were those of the recording conditions. However, mismatch in languages, and the linguistic content can also affect the strength of evidence, and it is necessary to estimate and compensate for the uncertainty they introduce in the strength of evidence.