Reliability of voice comparison for forensic applications

Moez Ajili

▶ **To cite this version:**

Moez Ajili. Reliability of voice comparison for forensic applications. Signal and Image Processing. Université d'Avignon et des Pays de Vaucluse, 2017. English. NNT : . tel-01949669

HAL Id: tel-01949669
https://hal.science/tel-01949669

Submitted on 10 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ACADÉMIE D'AIX-MARSEILLE
UNIVERSITÉ D'AVIGNON ET DES PAYS DE VAUCLUSE

THESIS

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

In Computer Science

Doctoral School 536 «Sciences et Agroscience»
Laboratoire d'Informatique (EA 4128)

Reliability of voice comparison for forensic applications

Presented by
Moez AJILI

Defended on November 28, 2017 before the jury members:

M.	Didier Meuwly	Principal scientist, NFI, Netherlands	Reviewer
M*me*	Martine Adda-Decker	Professor, LPP, Paris	Reviewer
M.	Tomi Kinnunen	Associate Professor, Joensuu, Finland	Examiner
M.	Georges Linarès	Professor, LIA, Avignon	Examiner
M.	Jean-François Bonastre	Professeur, LIA, Avignon	Advisor
M.	Solange Rossato	Associate Professor, LIG, Grenoble	Co-advisor

Laboratoire d'Informatique d'Avignon

ACRONYMS

AFCP	Association Francophone de la Communication Parlée
ANOVA	ANalysis Of VAriance
ASpR	Automatic Speaker Recognition
BN	Background Noise
CLLR	Log Likelihood Ratio Cost
CMS	Cepstral Mean Subtraction
CMVN	Cepstral Mean and Variance Normalization
CPS	Classical Probability Scales
DCF	Decision Cost Function
DET	Detection Error Trade-off
DNN	Deep Neural Networks
DTW	Dynamic Time Warping
EER	Equal Error Rate
EFR	Eigen Factor Radial
EM	Expectation Maximisation
FA	Factor Analysis
FAR	False Acceptance Rate
FASR	Forensic Automatic Speaker Recognition
FVC	Forensic Voice Comparison
FRR	False Rejection Rate
FSC	Forensic Speaker Comparison
FSR	Forensic Speaker Recognition
GFCP	Groupe Francophone de la Communication Parlée

GMM	Gaussian Mixture Model
GMM-UBM	Gaussian Mixture Model-Universal Background Model
GMR	Gaussian Mixture Regression
HM	Homogeneity Measure
HMM	Hidden Markov Model
IAFPA	International Association for Forensic Phonetics and Acoustics
LDA	Linear Discriminant Analysis
LDC	Linguistic Data Consortium
LFCC	Linear Frequency Cepstral Coefficients
LLR	Log-Likelihood-Ratio
LPCC	Linear Predictive Cepstral Coefficients
LR	Likelihood Ratio
SNR	Signal to Noise Ratio
SRE	Speaker Recognition Evaluation
SV	SuperVector
JFA	Joint Factor Analysis
MAP	Maximum a Posteriori
MFCC	Mel Frequency Cepstral Coefficients
NAP	Nuisance Attribute Projection
NIST	National Institute of Standards and Technology
NRC	National Research Council
PAV	Pool Adjacent Violator
PLDA	Probabilistic Linear Discriminant Analysis
PLP	Perceptual Linear Prediction
ROC	Receiver Operating Curve
SITW	Speakers In The Wild
SVM	Support Vector Machine
TCE	Total Cross Entropy
TV	Total Variability
UBM	Universal Background Model

UKPS	United Kingdom Position Statement
VAD	Voice Activity Detection
VQ	Vector Quantisation
WCCN	Within Class Covariance Normalization

Acknowledgement

There is an ancient African proverb that says, "*It takes a village to raise a child*." A portion of that proverb -'it takes a village"- is now colloquially used to acknowledge the influence a group of people can have when contributing to something bigger than themselves.

Despite appearing as the single author of this piece of work, I believe that it takes not just a single individual but "a village" to bring a PhD to fruition. To that extent, I would like to take the opportunity to thank my village.

Foremost, I am not sure I can express how thankful I am to my two main supervisors Professor Jean-François Bonastre and Dr. Solange Rossato. Thank you for introducing me to speaker recognition field and especially forensic speaker recognition as an undergraduate. There is no way I would have started a PhD without your initial inspiration and encouragement. Thank you for your advice and guidance which have been instrumental in shaping this thesis. Thank you for your continual support, friendship and the confidence you have always put in me. Thank you for all of your time, patience and energy.

I am also grateful to the members of my jury, Prof. Didier Meuwly, Prof. Martine Adda-Decker and Prof. Tomi Kinnunen, for reviewing my work and for their valuable comments and to Prof. Georges Linarès for acting as the president of the jury.

A thank you must also go to all the researchers in LIG especially GETALP team where I spent three months. Thank you Laurent Besacier, Michel Vacher, François Portet, Benjamin Lecouteux, Zied Elloumi, marwa and Javier serrano for their warm welcome, their support and help during my research stay.

A thank you is also extended to those who have helped me along the way through discussions, insight, and simply encouragement. Thank you to Juliette Kahn, Anthony Larcher and Olivier Galibert.

I would also like to say "thank you" to all my work mates at LIA who have shared with me tons of time, stress, achievements, friendship, bad moods and laughs, and who daily enrich my professional and personal life. I have to wholeheartedly thank (in non-perfect order of appearance):

- Waad, my best friend and biggest supporter, you know the details of this PhD almost as well as I do. Thank you for being always there when things got stressful,

lending your ear as I discussed a long day's worth of work, providing unyielding patience, and your continued support and unconditional love. I will not forget our phrase at the end of each challenge, "*life is really hard !*" Thank you my brother for making my years at the laboratory an enjoyable and unforgettable experience.

- Speaker Recognition Team : Driss, Pierre-Michel and Mikael. Thank you guys for all the effort we did together especially during the evaluation campaigns. A special thanks go to driss with whom I shared long discussions around many coffee break. I really loved your professional/personal charcaters, your stories and everything.

- All my office buddies over the years, Elvys, Luis, Thomas, Maxime, Mayeul, Sabrine: thank you for your unique sense of humour, advice and support.

- All my collegues : Mohamed Bouaziz, Imed, Titouan, Anthony Poujade, Mohamed Bouallegue, Mohamed Morchid, Richard, Fabrice, Zak, Xavier, Cedric, Nejat, Oussama, Etienne, Adrien, Abdelillah, Imen, Olfa, Majed, Fen, Mathias, Teva, Tesnim, Cyril, Bassam, Stephane, Simone, Yann, Corinne...

Now to those behind the scenes. I must thank my family: Mom, Dad's soul, my sisters and my brother. Mummy, I thank you for believing in me and helping me believe in myself, for giving me the dream and keeping the dream alive against all odds. It was not that easy to do all of that alone. Mohamed, you are a brother and a father, sounding board, counselor and best friend all rolled into one. Lastly, to my sisters, Faouzia, Moufida and sawssen- thank you for your advice and support. Thanks all for always being there.

My deepest gratitude goes to my lovely fiancée. Thank you Maha, for coming into my life, for all the love and support you have given me and for seeing this thesis through with me. This thesis would not have taken the shape it has now, without your help.

To my village - I thank you.

Above all, I thank God for guiding and taking care of me every step of the way.

Personal bibliography

Research related to this Thesis

- (Ajili et al., 2017a) Ajili, M., Bonastre, J. F., Kheder, W. B., Rossato, S., & Kahn, J. (2017). Homogeneity Measure Impact on Target and Non-target Trials in Forensic Voice Comparison. Proc. Interspeech 2017, 2844-2848.
- (Ajili et al., 2017b) Ajili, M., Bonastre, J. F., Kheder, W. B., Rossato, S., & Kahn, J. (2017, March). Phonological content impact on wrongful convictions in forensic voice comparison context. In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on (pp. 2147-2151). IEEE.
- (Ajili et al., 2016a) Ajili, M., Bonastre, J. F., Rossato, S., & Kahn, J. (2016, March). Inter-speaker variability in forensic voice comparison: a preliminary evaluation. In Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on (pp. 2114-2118). IEEE.
- (Kheder et al., 2016a) Kheder, W. B., Ajili, M., Bousquet, P. M., Matrouf, D., & Bonastre, J. F. (2016). LIA System for the SITW Speaker Recognition Challenge. In INTERSPEECH (pp. 848-852).
- (Ajili et al., 2016b) Ajili, M., Bonastre, J. F., Kahn, J., Rossato, S., & Bernard, G. (2016). FABIOLE, a Speech Database for Forensic Speaker Comparison. In International Conference on Language Resources, Evaluation and Corpora (LREC).
- (Ajili et al., 2016) Ajili, M., Bonastre, JF., Ben Kheder, W., Rossato, S., & Kahn, J. (2016, December). Phonetic content impact on forensic voice comparison. In Spoken Language Technology Workshop (SLT), 2016 IEEE (pp. 210-217). IEEE.
- (Ajili et al., 2015b) Ajili, M., Bonastre, J. F., Rossato, S., Kahn, J., & Lapidot, I. (2015). An information theory based data-homogeneity measure for voice comparison. In Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH).
- (Ajili et al., 2015a) Ajili, M., Bonastre, J. F., Rossato, S., Kahn, J., & Lapidot, I. (2015, November). Homogeneity measure for forensic voice comparison: A step forward reliability. In Iberoamerican Congress on Pattern Recognition (pp. 135-142). Springer.

- (Bonastre et al., 2015) Bonastre, J. F., Kahn, J., Rossato, S., Ajili, M. Forensic Speaker Recognition: Mirages and Reality. In Fuchs, S., Pape, D., Petrone, C., Perrier, P. (2015). Individual Differences in Speech Production and Perception.

Other works

- (Ajili et al., 2018) Ajili, M., Bonastre, J. F., & Rossato, S. (2018). Voice Comparison and Rhythm: Behavioral Differences between Target and Non-target Comparisons. Proc. Interspeech 2018, 1061-1065.
- (Ajili et al., 2018b) Ajili, M., Rossato, S., Zhang, D., & Bonastre, J. F. (2018). Impact of rhythm on forensic voice comparison reliability. In Proc. Odyssey 2018 The Speaker and Language Recognition Workshop (pp. 1-8).
- (Ajili et al., 2018a) Ajili, M., Bonastre, J. F., Ben Kheder W., Rossato, S., & Kahn, J. Comparaison des voix dans le cadre judiciaire: influence du contenu phonétique.
- (Rossato et al., 2018) Rossato, S., Zhang, D., Ajili, M., & Bonastre, J. F. (2018, June). Suivre le rythme de tes paroles. In XXXIIe Journées d'Études sur la Parole.
- (Kheder et al., 2018) Kheder, W. B., Matrouf, D., Ajili, M., & Bonastre, J. F. (2018). A Unified Joint Model to Deal With Nuisance Variabilities in the i-Vector Space. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(3), 633-645.
- (Lee et al., 2017) Lee, K. A., Sun, H., Sizov, A., Wang, G., Nguyen, T. H., Ma, B., ..., Ajili, M., ... & Halonen, M. (2017). The I4U Mega Fusion and Collaboration for NIST Speaker Recognition Evaluation 2016. In Proc. Interspeech.
- (Kheder et al., 2017) Kheder, W. B., Matrouf, D., Bousquet, P. M., Bonastre, J. F., & Ajili, M. (2017). Fast i-vector denoising using MAP estimation and a noise distributions database for robust speaker recognition. Computer Speech & Language, 45, 104-122.
- (Rouvier et al., 2016) Rouvier, M., Bousquet, P. M., Ajili, M., Kheder, W. B., Matrouf, D., & Bonastre, J. F. (2016). LIA system description for NIST SRE 2016. In Proc. NIST SRE 2016 Workshop.
- (Lee et al., 2016) Lee, K. A., Sun, H., Sizov, A., Wang, G., Nguyen, T. H., Ma, B., ..., Ajili, M., ... & Halonen, M. (2016). The I4U submission to the 2016 NIST speaker recognition evaluation. In Proc. NIST SRE 2016 Workshop.
- (Kheder et al., 2016b) Kheder, W. B., Matrouf, D., Ajili, M., & Bonastre, J. F. (2016). Iterative Bayesian and MMSE-based noise compensation techniques for speaker recognition in the i-vector space. In Proc. Speaker Lang. Recognit. Workshop (pp. 60-67).

- (Kheder et al., 2016d)Kheder, W. B., Matrouf, D., Ajili, M., & Bonastre, J. F. (2016). Probabilistic Approach Using Joint Clean and Noisy i-Vectors Modeling for Speaker Recognition. In INTERSPEECH (pp. 3638-3642).
- (Ben Kheder et al., 2016) Kheder, W. B., Matrouf, D., Ajili, M., & Bonastre, J. F. (2016). Probabilistic Approach Using Joint Long and Short Session i-Vectors Modeling to Deal with Short Utterances for Speaker Recognition. In INTERSPEECH (pp. 1830-1834).
- (Kheder et al., 2016c) Kheder, W. B., Matrouf, D., Ajili, M., & Bonastre, J. F. (2016). Local binary patterns as features for speaker recognition. Odyssey 2016, 346-351.
- (Kheder et al., 2015) Kheder, W. B., Matrouf, D., Bonastre, J. F., Ajili, M., & Bousquet, P. M. (2015, April). Additive noise compensation in the i-vector space for speaker recognition. In Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on (pp. 4190-4194). IEEE.
- (Matrouf et al., 2015) Matrouf, D., Kheder, W. B., Bousquet, P. M., Ajili, M., & Bonastre, J. F. (2015, August). Dealing with additive noise in speaker recognition systems based on i-vector approach. In Signal Processing Conference (EUSIPCO), 2015 23rd European (pp. 2092-2096). IEEE.
- (Kheder et al., 2014) Kheder, W. B., Matrouf, D., Bousquet, P. M., Bonastre, J. F., & Ajili, M. (2014, October). Robust speaker recognition using map estimation of additive noise in i-vectors space. In International Conference on Statistical Language and Speech Processing (pp. 97-107). Springer, Cham.
- (Vacher et al., 2015) Vacher, M., Lecouteux, B., Romero, J. S., Ajili, M., Portet, F., & Rossato, S. (2015, October). Speech and speaker recognition for home automation: Preliminary results. In Speech Technology and Human-Computer Dialogue (SpeD), 2015 International Conference on (pp. 1-10). IEEE.

Résumé

Dans les procédures judiciaires, des enregistrements de voix sont de plus en plus fréquemment présentés comme élément de preuve. En général, il est fait appel à un expert scientifique pour établir si l'extrait de voix en question a été prononcé par un suspect donné (*prosecution hypothesis*) ou non (*defence hypothesis*). Ce prosessus est connu sous le nom de "*Forensic Voice Comparison (FVC)*" (comparaison de voix dans le cadre judiciaire). Depuis l'émergence du modèle *DNA typing*, l'approche Bayesienne est devenue le nouveau "*golden standard*" en sciences criminalistiques. Dans cette approche, l'expert exprime le résultat de son analyse sous la forme d'un rapport de vraisemblance (LR). Ce rapport ne favorise pas seulement une des hypothèses ("*prosecution*" ou "*defence*") mais il fournit également le poids de cette décision. Bien que le LR soit théoriquement suffisant pour synthétiser le résultat, il est dans la pratique assujetti à certaines limitations en raison de son processus d'estimation. Cela est particulièrement vrai lorsque des systèmes de reconnaissance automatique du locuteur (ASpR) sont utilisés. Ces systèmes produisent un score dans toutes les situations sans prendre en compte les conditions spécifiques au cas étudié. Plusieurs facteurs sont presque toujours ignorés par le processus d'estimation tels que la qualité et la quantité d'information dans les deux enregistrements vocaux, la cohérence de l'information entre les deux enregistrements, leurs contenus phonétiques ou encore les caractéristiques intrinsèques des locuteurs. Tous ces facteurs mettent en question la notion de fiabilité de la comparaison de voix dans le cadre judiciaire. Dans cette thèse, nous voulons adresser cette problématique dans le cadre des systèmes automatiques (ASpR) sur deux points principaux.

Le premier consiste à établir une échelle hiérarchique des catégories phonétiques des sons de parole selon la quantité d'information spécifique au locuteur qu'ils contiennent. Cette étude montre l'importance du contenu phonétique: Elle met en évidence des différences intéressantes entre les phonèmes et la forte influence de la variabilité intra-locuteurs. Ces résultats ont été confirmés par une étude complémentaire sur les voyelles orales basée sur les paramètres formantiques, indépendamment de tout système de reconnaissance du locuteur.

Le deuxième point consiste à mettre en œuvre une approche afin de prédire la fiabilité du LR à partir des deux enregistrements d'une comparaison de voix sans recours à un ASpR. À cette fin, nous avons défini une mesure d'homogénéité (NHM) capable d'estimer la quantité d'information et l'homogénéité de cette information entre les deux enregistrements considérés. Notre hypothèse ainsi définie est que l'homogénéité

soit directement corrélée avec le degré de fiabilité du LR. Les résultats obtenus ont confirmé cette hypothèse avec une mesure NHM fortement corrélée à la mesure de fiabilité du LR. Nos travaux ont également mis en évidence des différences significatives du comportement de NHM entre les comparaisons cibles et les comparaisons imposteurs.

Nos travaux ont montré que l'approche "force brute" (reposant sur un grand nombre de comparaisons) ne suffit pas à assurer une bonne évaluation de la fiabilité en FVC. En effet, certains facteurs de variabilité peuvent induire des comportements locaux des systèmes, liés à des situations particulières. Pour une meilleure compréhension de l'approche FVC et/ou d'un système ASpR, il est nécessaire d'explorer le comportement du système à une échelle aussi détaillée que possible (le diable se cache dans les détails).

Mots clés— Reconnaissance du locuteur, apprentissage automatique, paradigme Bayesien, comparaison des voix dans le cadre judiciaire, fiabilité, contenu phonémique, mesure d'homogénéité.

Abstract

It is common to see voice recordings being presented as a forensic trace in court. Generally, a forensic expert is asked to analyse both suspect and criminal's voice samples in order to indicate whether the evidence supports the prosecution (same-speaker) or defence (different-speakers) hypotheses. This process is known as Forensic Voice Comparison (FVC). Since the emergence of the DNA typing model, the likelihood-ratio (LR) framework has become the new "golden standard" in forensic sciences. The LR not only supports one of the hypotheses but also quantifies the strength of its support. However, the *LR* accepts some practical limitations due to its estimation process itself. It is particularly true when *Automatic Speaker Recognition* (ASpR) systems are considered as they are outputting a score in all situations regardless of the case specific conditions. Indeed, several factors are not taken into account by the estimation process like the quality and quantity of information in both voice recordings, their phonological content or also the speakers intrinsic characteristics, etc. All these factors put into question the validity and reliability of FVC. In this Thesis, we wish to address these issues.

First, we propose to analyse how the phonetic content of a pair of voice recordings affects the FVC accuracy. We show that oral vowels, nasal vowels and nasal consonants bring more speaker-specific information than averaged phonemic content. In contrast, plosive, liquid and fricative do not have a significant impact on the LR accuracy. This investigation demonstrates the importance of the phonemic content and highlights interesting differences between inter-speakers effects and intra-speaker's ones. A further study is performed in order to study the individual speaker-specific information for each vowel based on formant parameters without any use of ASpR system. This study has revealed interesting differences between vowels in terms of quantity of speaker information. The results show clearly the importance of intra-speaker variability effects in FVC reliability estimation.

Second, we investigate an approach to predict the LR reliability based only on the pair of voice recordings. We define a homogeneity criterion (NHM) able to measure the presence of relevant information and the homogeneity of this information between the pair of voice recordings. We are expecting that lowest values of homogeneity are correlated with the lowest LR's accuracy measures, as well as the opposite behaviour for high values. The results showed the interest of the homogeneity measure for FVC reliability. Our studies reported also large differences of behaviour between FVC genuine and impostor trials. The results confirmed the importance of intra-speaker variability

effects in FVC reliability estimation.

The main takeaway of this Thesis is that averaging the system behaviour over a high number of factors (speaker, duration, content...) hides potentially many important details. For a better understanding of FVC approach and/or an ASpR system, it is mandatory to explore the behaviour of the system at an as-detailed-as-possible scale (The devil lies in the details).

Index terms— Speaker recognition, machine learning ,Bayesian paradigm, forensic voice comparison, reliability, phonemic content, speaker factor, homogeneity measure.

Contents

Chapter 1

Introduction

Contents

1.1 Scientific context

Forensic science is defined as the body of scientific knowledge and technical methods used to solve questions related to criminal, civil and administrative law (Tistarelli et al., 2014). One of the main branches of forensic science, namely criminalistics, is the profession and scientific discipline oriented to the recognition, identification, individualisation and evaluation of physical evidence by the application of natural science to law-science problems (Gialamas, 2000). The main aim in this branch is the inference of the identity of an unknown source from the scientific analysis of the evidence presented in a trial. Therefore, criminalistics is often considered as the science of individualisation (Kirk, 1963), understood as the process "to reduce a pool of potential sources of forensic trace to a single source". In this process, the role of forensic expert is the evaluation of the forensic evidence in order to help the trier-of-fact to answer this specific question, "Does the incriminatory piece of evidence of the unknown origin come from a given known source?". The debate about the presentation of forensic evidences in a court of law is a hot topic (Kennedy, 2003; Giannelli et al., 1993; Saks et Koehler, 2005; National Research Council, 2009; Jackson et al., 2006). Forensic expertises for traditional forensic science are based on categorical decision, either "individualisation" or "exclusion". So, given a test specimen (for example, a fingerprint recovered from the crime scene) and a reference specimen (for example, fingerprint of the suspected individual), the final result of the evaluation procedure is then a binary decision: Either the piece of evidence and the control material come from the same source (prosecution hypothesis)

or from different sources (defence hypothesis). The main drawback of this approach is that the forensic scientist jumps from reasoning under uncertainty to absolute conclusions, taking therefore the decision regarding only the piece of evidence at hand and ignoring several other sources of information that can be conclusive. Moreover, in such approach, the expert usurps the role of the court in taking decision.

Mainly impelled by several scientific and legal forces, traditional forensic identification is moving toward a new paradigm, "*Bayesian paradigm*" (Saks et Koehler, 2005). In this new paradigm, the report expertise is no longer categorical or deterministic but of probabilistic nature (Stoney, 1991; Champod et al., 2001; Inman et Rudin, 2000; Champod et al., 2016; Sallavaci, 2014). Indeed, the Bayesian paradigm, denoted as the logically and theoretically sounded framework to model and represent forensic evidence reports, has become the new "*golden standard*". In this framework, the expert should only provide the court with the strength-of-the-evidence summarized by a likelihood ratio (LR) value and should strictly avoid to form a view or make an influence to the final decision. According to (Saks et Koehler, 2005) three main key driving forces are responsible for this shift:

- Emergence and outgrowth of "DNA typing" as a model for scientifically sounded science.
- The considerable changes in the legal admissibility standards for expert testimony namely Daubert rules which are demanding more transparent procedures and scientific framework for a logical and testable interpretation of the forensic evidence.
- The high number of "wrongful conviction" cases revealed by the DNA test.

These radical changes in criminalistics made inroads into many other forensic sciences. Forensic Voice Comparison (FVC[1]), which refers to the comparison of a recording of an unknown criminal's voice (the evidence or the trace) and a recording of a known suspect's voice (the comparison piece), is one of the fields affected by this change. Indeed, nowadays, establishing the identity of a perpetrator based on a piece of voice recording is under the spotlight as there is an increasing demand for this kind of expertise in courts. Moreover, it is becoming rare to see a law case in which there is no mention of the use of a smartphone or some other modern communication tool.

Forensic voice comparison is undergoing a paradigm shift (Gonzalez-Rodriguez et Ramos, 2007; Morrison, 2009a) towards emulating other forensic disciplines in quantifying the value of the evidence based on transparent, reliable and testable methods. Forensic speaker comparison will constitute the core study in this thesis.

1.2 Issues in forensic voice comparison based ASpR approach

The Likelihood ratio (LR) framework is being increasingly used by the experts and quite often required by "best practice guides" issued by the expert's associations (Meuwly

[1] Also known as Forensic Speaker Recognition (FSR)

et al., 1998; Gonzalez-Rodriguez et al., 2007; Morrison, 2009a; Aitken et Taroni, 2004). Automatic Speaker Recognition (ASpR), wherein a system verifies a speaker's identity using a sample of speech, is considered as one of the most appropriate solutions when LR framework is involved (Gold et French, 2011). Even though ASpR systems have achieved significant progresses in the past two decades and have reached impressive low error rates ($\approx$ 1% (Dehak et al., 2011; Bousquet et al., 2014; Hansen et Hasan, 2015)), the forensic scenario is still a very challenging one for ASpR for several reasons:

Real-life *LR* approximation/estimation processes As said before, the *LR* provides a theoretically founded value of the relative strength of its support to the prosecutor or the defence hypothesis. So, it appears to be self-sufficient and does not need any further processing or confidence measure to take into account the characteristics of a specific voice comparison trial. But, in real world, the *LR* is *approximated* using a specific process and this process accepts several limitations. It is particularly true when automatic FVC is considered, as the ASpR systems are outputting a *score* in all situations regardless of the case specific conditions. Moreover, the ASpR FVC systems use different normalization steps to see their scores as *LR* including the so-called "calibration" (Brummer et van Leeuwen, 2006; Gonzalez-Rodriguez et al., 2006; Gonzalez-Rodriguez et Ramos, 2007; Gonzalez-Rodriguez et al., 2007; Kinoshita et al., 2014; Nautsch et al., 2016). Several other normalization steps are used, like at the acoustic parameterization level (Pelecanos et Sridharan, 2001; Ganapathy et al., 2011) , at the iVector level (Bousquet et al., 2012, 2014; Garcia-Romero et Espy-Wilson, 2011) or at the score level (Auckenthaler et al., 2000; Glembek et al., 2009; Swart et Brümmer, 2017). A "reference population" is also often used to evaluate the "typicality" (Drygajlo et al., 2015; Champod et Meuwly, 2000; Hughes, 2014; Ishihara et al., 2008; Kinoshita et al., 2014). A large majority of the involved normalization approaches are based on training data and represent a potential source of biases as a mismatch between the training material and a given forensic trial could be large. Moreover, the amount of mismatch is often unknown as the trial conditions could be partially or largely unknown and as the training conditions are not always well defined.

Trial conditions The speech recordings may be recorded in different situations and at least one situation is partially or completely unknown (the trace recording situation). The speakers are not necessarily cooperative and may disguise their voices, with consequences on performance (Kajarekar et al., 2006). A speaker could also be ill, or under the influence of stress, alcohol or other factors. The social and linguistic environment of the unknown speaker is unknown by construction (so, for example, the influence of potential mother or second language should be taken into account by the forensic experts). The speech samples will most likely contain noise, may be very short, their content cannot be controlled (at least for the trace) and may not contain enough relevant information for comparative purposes (Ajili et al., 2015b). In their "*need for caution paper*" (Campbell et al., 2009), which inspired a large part of this paragraph, the authors said in the conclusion: "*Each of these variables, in addition to the known variability of speech in general, makes reliable discrimination of speakers a complicated and daunting task*".

This sentence remains in 2017 a nice synthesis of FVC's challenging aspects (for FVC in general and not only for ASpR-based FVC).

Intra-speaker variability Unlike DNA and fingerprints, Voice is far from constant and will change over the time (in short term and in long term), depending on several factors such as health and emotional state. The voice could also be altered voluntarily by the speaker (disguise or impersonation). This "within-speaker variability" can impact dramatically the FVC process. Despite its huge impact on ASpR systems (Kahn et al., 2010), this factor is still not well addressed in the different evaluation campaigns such as NIST framework (mainly due to the low number of available speech utterances per speaker.)

Limits of the performance evaluation If the desire to use ASpR in FVC is not novel, due to intrinsic interests of automatic processes in this field, this desire has increased significantly during the past years as the performance level reached by speaker recognition systems has become very attractive. The performance is measured thanks to international evaluation campaigns like NIST-SRE's ones (Greenberg et al., 2011, 2013). If the pros of such evaluation campaigns are well established, several research works emphasized the limits of the underlined evaluation protocols (Doddington et al., 1998; Doddington, 2012; Kahn et al., 2010; Ajili et al., 2015b). Moreover, the classical evaluation criterion and protocols used in ASpR are not designed for FVC. EER and DCF, quite often used as evaluation criterion, are based on hard score decisions and not on *LR*'s reliability. The protocols focus on global performance using a brute-force strategy and express the averaged behaviour of ASpR systems. In the same time, they ignore many sensitive cases which represent several distinct specific situations where the ASpR systems show a specific behaviour due, for example, to the recording conditions, the noises, the content of the recordings or the speakers themselves (and the evaluation databases are still missing a lot of variation factors).

A lack in content analysis of the audio recordings Human speech is a rich signal which embedded in its linguistic message information related to the gender, age, health, emotional state and identity of the speaker. In state-of-the-art ASpR systems (for example IVector(IV)-based ones) a recording is encoded by one low dimensional vector. The phonemic content of a recording is not used explicitly, as well as the presence or absence of different speaker-specific cues. However several research works like (Magrin-Chagnolleau et al., 1995; Besacier et al., 2000; Amino et al., 2006; Antal et Toderean, 2006) agree that speaker specific information is not equally distributed on the speech signal and particularly depends on the phoneme distribution. In (Ajili et al., 2015b,a, 2017a) the authors showed that homogeneity of the speaker-specific information between the two recordings of a voice comparison trial is also playing an important role and should not be ignored by the *LR* estimation process.

Considering that forensic practitioners are likely to use ASpR systems as a "black box", this is certainly a cause for concern. In order to ensure a reliable FVC, one should

start with analysing the ASpR system behaviour to understand the extent to which such dimensions of variability affect the estimation process of the LR. Several questions arise and need clarification:

- Which phonetic information is taken into account when performing ASpR-based FVC? Is this information the same for all the speakers?
- Does the information used to distinguish speakers depend on the pair of speakers involved in the comparison?
- When performing FVC, does the system have the same behaviour across different speakers? Is there any difference in terms of accuracy and reliability of the LRs corresponding to different speakers?

Answering these questions is mandatory to validate the use of LR-based ASpR systems or highlight their potential weaknesses and make explicit their limitations.

1.3 Major contributions of the thesis

As the title of the thesis "*Reliability of voice comparison for forensic applications* " suggests, this thesis mainly deals with two main points issued from the previous list:

Speaker specific information embedded in the speech signal If everybody agrees on the fact that voice signal is conveying information on the speaker, including speaker's identity, it is less easy to list the different cues which embed this aspect (this is true for both human perception and automatic systems). In this thesis, we do not wish to fully answer this question but we propose to use an ASpR system in order to investigate the links between phonological content and speaker discrimination abilities. In other words, we propose to analyse how an ASpR system takes into account the phonetic information of both voice recordings when performing forensic voice comparison and how this phonetic content affects the likelihood ratio (LR) accuracy. The aim of this study is to analyse whether certain classes of phonemes are bringing more speaker discriminative information than others and if these differences are stable among the speakers.

Homogeneity measure as indicator of LR accuracy and reliability In the second component of this thesis, we investigate an approach to predict the LR reliability. We define a "Homogeneity measure" (HM) able to measure the presence of speaker discriminant cues and the homogeneity of this information between the pair of voice records S_A-S_B. This measure is estimated only from the two in-interest voice records. Our motivation to define this measure is the following: It is obvious that the presence of speaker specific information inside S_A and S_B is mandatory for FVC accuracy, but we expect that it is not sufficient: examples tied with the same class of cues should be included in both speech recordings in order to be useful and thus ensure an accurate

and reliable LR. Therefore, we are expecting that lowest values of homogeneity are correlated with the lowest LRs accuracy, as well as the opposite behaviour for high values.

1.4 Organization of the manuscript

The remainder of this Thesis is organised as follows.

- Chapter 1 introduces the scientific context, the problematic and describes the motivation, outlines and contributions of this thesis.
- Chapter 2 introduces the topic of forensic science and particularly the forensic identification. This chapter details the evolution of forensic identification from traditional forensic sciences to the so-called paradigm shift. Further, it provides an overview of the acceptance of the Bayesian paradigm in forensic voice comparison.
- Chapter 3 discusses the difficulty of forensic voice comparison. Several source of variabilities are discussed such as recording conditions, mismatch in recording conditions, or intra-speaker variability.
- Chapter 4 presents the link between biometrics and forensic identification and explains why the ongoing paradigm shift in forensic sciences needs biometric methods to identify individuals. At the end of this chapter, we question in particular biometric aspects of speech evidence.
- Chapter 5 presents firstly an overview of an ASpR system. Then, it shows how ASpR system could be adapted in order to fit forensic context.
- Chapter 6 and 7 describe the speech databases, protocols and baseline speaker recognition systems used for the experiments presented in this Dissertation.
- Chapter 8 presents an exploratory investigation in order to analyse how an ASpR system behaves in different scenarios. This analysis is performed building upon the previous work of (Doddington et al., 1998; Kahn et al., 2010).
- Chapter 9 investigates the impact of phonetic content on the voice comparison process. We propose to analyse whether certain classes of phonemes are bringing more speaker discriminative information than others and if these differences are stable across the speakers.
- Chapter 10 is dedicated to investigate the amount of speaker specific information carried by each vowel. This study is performed based on formant parameters.
- Chapter 11 presents the homogeneity measure criteria. This measure estimates the amount of "speaker discriminant cues" and the homogeneity of this information between the pair of voice recordings.
- Chapter 12 is the final Chapter in which we draw the main conclusions of this Thesis, and we suggest possibilities for future studies.

Part I

State of the art

Chapter 2

Scientific evidence in courts: The case of Voice Evidence

True belief only becomes knowledge when backed by some kind of investigation and evidence.

–Karl Marx

Contents

Introduction

In this chapter, we introduce the topic of forensic science and particularly forensic identification. Indeed, the evolution of forensic identification from traditional forensic science to the so-called "*Bayesian Framework*" is detailed, while the resulting requirements are highlighted. Further, we focus on the case of voice evidence, how a piece of voice recording can be used in forensic identification. We provide therefore an overview of the position of the "Bayesian paradigm" in forensic voice comparison.

2.1 Justice and law: From evidence to source

2.1.1 Definitions and notions

According to (Bell, 2008), *the two words forensic and science each relate to the common theme of truth, either spoken or seen. The word forensic can be traced to the Latin forum or "in the public". The definition is roughly translated as "to speak the truth in public". In the modern world, this extends to speaking the truth in court, today's equivalent of the forum. Thus, the role of science is to help society define what is the fact; the role of forensic science is to help the legal system define it*.

The term "*forensic science*" is defined by the American Academy of Forensic Sciences as "*the study and practice of the application of science to the purposes of the law...*". *Forensic science is defined as the body of scientific knowledge and technical methods used to solve questions related to criminal, civil and administrative law* (Tistarelli et al., 2014). According to (Jackson et al., 2015), forensics means the use of scientific methods and techniques in the establishment of facts or evidence in the court of law .

One of the main branches of forensic science, namely criminalistics, is the profession and scientific discipline oriented to the recognition, identification, individualisation and evaluation of physical evidence by the application of natural science to law-science problems (Gialamas, 2000).

2.1.2 Evidence in forensic science

The cornerstone of forensic science since 1920s has been a maxim attributed to Edmund Locard *a pioneer in forensic science who became known as the "Sherlock Holmes of France". He formulated the basic principle of forensic science known as Locard's exchange principle* [1] "Every contact leaves a trace". According to this principle, whenever two objects come into contact, a mutual exchange of matter will take place between them (James et al., 2002). In other words, Locard speculated that when a perpetrator make contact with a victim, or an object in the crime scene, it results in an exchange of physical materials.

[1] *https://en.wikipedia.org/wiki/Edmond_Locard*

In Kirk's book Crime Investigation: Physical Evidence and the Police Laboratory (Kirk, 1953), Kirk articulates the principle as follows:

"[...] Wherever he steps, whatever he touches, whatever he leaves, even unconsciously, will serve as a silent witness against him. Not only his fingerprints or his footprints, but his hair, the fibers from his clothes, the glass he breaks, the tool marks he leaves, the paint he scratches, the blood or semen he deposits or collects. All of these and more, bear mute witness against him. This is evidence that does not forget. It is not confused by the excitement of the moment. It is not absent because human witnesses are. It is factual evidence. Physical evidence cannot perjure itself, it cannot be wholly absent. Only human failure to find it, study and understand it, can diminish its value" (Kirk, 1953)

According to Inman and Rudin, Locard believed that:

"[...] No one can commit a crime with the intensity that the criminal act requires without leaving numerous signs of it: either the offender has left signs at the scene of the crime, or on the other hand, has taken away with him - on his person or clothes- indications of where he has been or what he has done." (Inman et Rudin, 2000) (p.84)

The logic behind this principle is to establish a direct link between the perpetrator and the victim or between the perpetrator and a physical object in the crime scene in order to prove his guilt. This principle is at the foundation of crime scene investigation. Indeed, traces, signs or marks can be either left by the perpetrator or found on him. Forensic evidence can be found either as physical or digital. Physical traces can be made for example by fingers, ears or feet, while digital ones are mainly digital recordings typically from phone-tapping and/or security cameras. Figure 2.1 presents the Locard's evidence transfer theory.

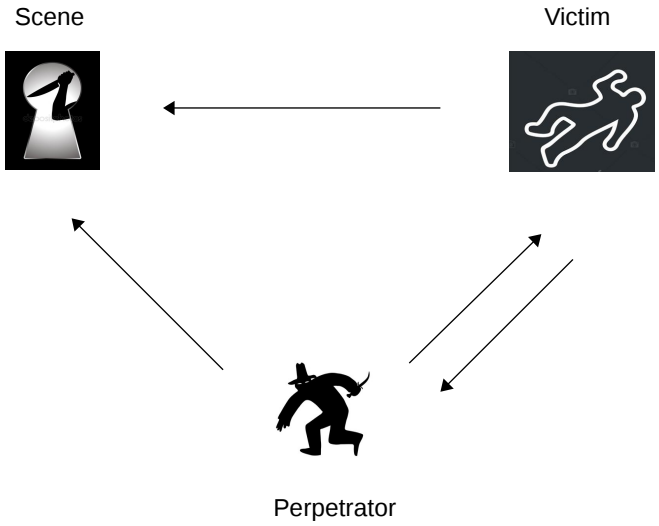

Figure 2.1: *Evidence transfer theory.*

Locard's theory reminds us the "adventures and memoirs of Sherlock Holmes": for

example, when Sherlock Holmes deduces the smoked cigarette by looking at the type of ash that it left, or discerns the part of London a person was from by the mud on his Jacket (Doyle, 2013). It is obvious that the presence of traces represents workable leads for detectives as they seek to identify and arrest potential suspects.

Establishing the identity of the criminal based on a piece of evidence is clearly an important issue in court. Indeed, it seems natural that both scientist and the trier-of-fact are interested to know if the trace of evidence originates from a particular suspect. Paul Kirk considers that "Criminalistics (also known as forensic science) is the science of individualization" (Inman et Rudin, 2000; Kirk, 1963). He states: « *The real aim of all forensic science is to establish individuality or to approach it as closely as the present state of science allows. Criminalistics is the science of individualization. The criminalist is not ultimately interested in the similarity of two objects but in their source* » (Kirk, 1963). According to this concept of individualization, every object has an individuality which is unique. Objects of the same morphology like fingerprints or DNA may seem similar but they are distinctly unique.

The forensic individualization sciences rest on a central assumption: two indistinguishable marks must have been produced by a single object. Traditional forensic scientists seek to link a piece of evidence to a single person or object "to the exclusion of all others in the world". They do so by leaning on the assumption of discernible uniqueness. According to this assumption marks produced by different people or objects are observably different. Thus, when a pair of marks is not observably different, the forensic expert concludes that the marks were left by the same person or object (Sallavaci, 2014).

2.1.3 Identity and individualization

In forensic sciences, the concept of identity is related to the identity of the source from which originates the trace. Establishing the identity of a source refers to an individualization process, which consists of determining if a particular individual entity is the source of a trace. According to (Meuwly, 2006), the term "identity" has a dual character which can give rise to ambiguity. In fact, Meuwly (Meuwly, 2006) and Kwan before (Kwan, 1977), recognized two situations which refer to two kinds of identities:

- Qualitative identity: when the source of the trace may be a class of individual entities from which this trace could originate. The operation in which a class is determined to be the source is called "*classification*" in science. Indeed, when only class characteristics are present in a piece of evidence, it is not possible to determine a single source for it. Therefore, one piece of evidence could have many possible sources.
- Numerical identity: when the source of the trace refers to a specific or particular individual entity from which this trace could originate. This operation is known by "*individualization*" in science. Indeed, when a piece of evidence shows several individualizing traits, the analyst may conclude that only one entity could be the source of the evidence. In this circumstance, the evidence has only one possible

source. Therefore, individualization is understood to mean the narrowing of possible sources of a forensic trace to a single one in the universe (Inman et Rudin, 2000).

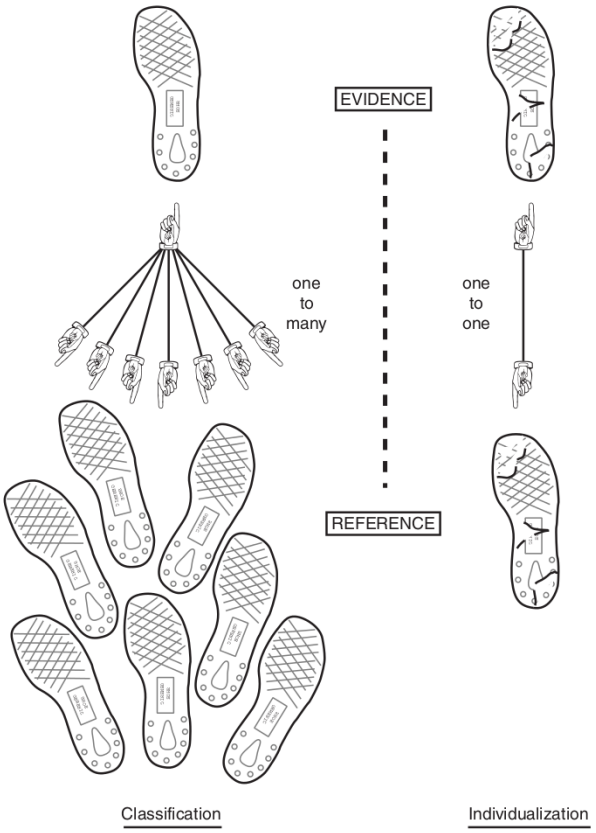

Figure 2.2: *Example of shoe-print evidence: One to many, one to one:Individualization and classification. Source (Inman et Rudin, 2000).*

The distinction between classification and individualization can be summarized as one-to-many versus one-to-one. An example of shoeprint evidence is presented in Figure 2.2.

In the source inference process (i.e the process of determining an identity based on a piece of evidence), the individualization process leads either to an absolute or a categorical identification. Indeed, three inferences are possible (Inman et Rudin, 2000):

1. The evidence originates from this source.
2. The evidence does not originate from this source.
3. The evidence can be classified but can not be individualized.

Meuwly and Champod (Meuwly, 2006) claim a real confusion surrounding the terms "identity", "identify" and "identification" in forensic science. Indeed, when a criminal is determined to be the source of a specific trace, we say that the criminal is individualised. It is not really "true" to say, the criminal is identified.

2.1.4 Criticisms on individualization

Popular television series promote strongly the notion of individualization in the viewer imagination as the forensic experts propose frequently confident decisions on, for example, whose hair was recovered from the crime scene or which gun fired the murderous bullet. But can forensic science really make such exact determinations? Can forensic scientists be sure that a particular hammer, to the exclusion of all other hammers in the world, produced the imprints observed on a victim's body?

Individualization could not be reached without the satisfaction of several requirements. As mentioned previously, we could not speak about individualization without evoking the uniqueness principle. It is claimed that uniqueness as well as permanence establish the validity of claims of individualization (Wertheim, 2001): " [..] *I say we all know and acknowledge biological uniqueness because biological uniqueness provides us with one of the foundational principles of our science; individuality.... The fact is, biological uniqueness allows us the liberty to identify persons through the comparisons we conduct. Of course, the other fundamental principle, permanence, also allows for identification, because if the detail examined in an comparison changed significantly, it wouldn't be much good for identification purposes!*".

Many forensic scientists and academic researchers consider the uniqueness principle as a fundamental premise whose validity is mandatory for forensic analyses. On the other side, many academic researchers and practitioners diverge in their understanding of the notion of "individualization", the claim to reduce a pool of potential sources of a forensic trace to a single source (Biedermann et al., 2016; Cole, 2014, 2009; Meuwly, 2006; Champod et al., 2001, 2016):

In (Meuwly, 2006), the authors claim that the concept of individualization is a complicated and daunting task as its main rule (establishing numerical identity) is based on "continuity" principle as there is no one to ensure unbroken continuity between the source and the piece of evidence since the creation of this piece. Meuwly et. al (Meuwly, 2006) and Kwan concluded that there is no way of knowing the numerical identity of the source.

In "*The Individualization Fallacy in Forensic Science Evidence*" (Saks et Koehler, 2008), Saks claims that the concept of individualization exists only in a metaphysical or rhetorical sense: " ***There is no scientific basis for the individualization claims in forensic sciences***". Saks also states that "*No basis exists in theory or data for the core contention that every distinct object leaves its own unique set of markers that can be identified by a skilled forensic scientist...Forensic scientists are not able to link a fingerprint, a hair, a handwriting sample, a tire-mark, a tool-mark, or any other evidentiary forensic item to its unique source, but they assert that ability every day in court.*" This is the classic reason that common law evidence doctrine required a heightened threshold for admission of expert testimony in court which will be detailed in subsection 2.2.1.

In his paper "*Forensics without uniqueness, conclusions without individualization: the new epistemology of forensic identification* (Cole, 2009)", Cole has questioned the fact of uniqueness considered one of the cornerstone of individualization and he adopts the

fact that individualization as -the perfect reduction of the potential sources of a forensic trace to one- is not possible or "not logically attainable". He claims that most forensic scholars who have sought to justify claims of individualization have appealed to the "**leap of faith**" argument first articulated by Stoney (Stoney, 1991). Stoney contended that reaching individualization through a subjective process, such as fingerprinting, was possible through a "leap of faith": the analyst becomes "subjectively certain that the patterns could not possibly be duplicated by chance". Many forensic scholars have echoed Stoney's claim (Inman et Rudin, 2000) (p. 139), (Champod et al., 2016) (p. 33). Thus, until recently, even those forensic scientists who recognized that individualization was logically unsupportable tended to go along with its reformulation as a "leap of faith". Several forensic and researchers have agreed that statements about the source of a trace are always **probabilistic** and not a categorical or deterministic (Stoney, 1991; Champod et al., 2001) (Inman et Rudin, 2000) (p. 148),(Champod et al., 2016) (p. 33). In (Sallavaci, 2014), the author claims that: *[...] leaning on the assumption of discernible uniqueness may well have served to perpetuate fundamental misunderstandings about the nature of the scientific evidence, which, contrary to the widespread belief, is essentially not of categorical or deterministic but of a probabilistic nature.*

This criticism of individualization was echoed by the 2009 U.S. National Research Council report (National Research Council, 2009), which called such claims unsupportable for any discipline except nuclear DNA profiling. The NRC Report adopts the denominations "classification" and "individualization". However, it goes on to note: *[...] With the exception of nuclear DNA analysis, however, no forensic method has been rigorously shown to have the capacity to consistently, and with a high degree of certainty, demonstrate a connection between evidence and a specific individual or source* (National Research Council, 2009) (p. 7). The report added: *Claims of "absolute" and "positive" identification should be replaced by more modest claims about the meaning and significance of a "match", (p. 142).*

Uniqueness assumption perpetuate fundamental misunderstandings

Although lacking theoretical or empirical foundations, the assumption of discernible uniqueness offers important practical benefits to the traditional forensic sciences. It enables forensic scientists to draw bold, definitive conclusions that can make or break cases. It excuses the forensic sciences from developing measures of object attributes, collecting population data on the frequencies of variations in those attributes, testing attribute independence, or calculating and explaining the probability that different objects share a common set of observable attributes. Without the discernible uniqueness assumption, far more scientific work would be needed, and criminalists would need to offer more tempered opinions in court (Saks et Koehler, 2005).

As a result, forensic sciences are moving toward a **new scientific paradigm**. According to (Saks et Koehler, 2005), three main forces are responsible for this paradigm shift:

- Emergence and outgrowth of DNA typing as a model for scientifically sound science. This model allows to discover many erroneous convictions.

- The considerable changes in the legal admissibility standards for expert testimony.
- Studies of error rates across the forensic sciences.

2.2 Paradigm shift in forensic sciences

In this section, the two major factors that forced the traditional forensic sciences to move toward a new scientific paradigm are reviewed.

2.2.1 Scientific evidence admissibility in courts

In the civil and criminal litigation systems, the need of a reliable scientific evidence is increasing. Indeed, the number of cases involving scientific evidence grows each year, focusing more attention on this need. Nevertheless, scientific evidences are not all admissible in court. There are often many requirements that must be considered before a piece of scientific evidence can be put forth in a courtroom as factual evidence. When dealing with the question of admissibility, we make reference to the American law, especially, "Frye test" and "Daubert v Merrell Dow Pharmaceuticals".

In this subsection, we start by giving a definition of the scientific evidence. Then, we briefly review the applicable standards controlling the admission of expert testimony in court for the example of the American civil and criminal litigation systems.

Scientific evidence

In law, a scientific evidence is based on a knowledge derived from scientific methods or techniques. This shall mean that the basis for the evidence has been, at least, hypothesized and tested and is generally accepted within the scientific community. For example, most forensic evidences like DNA matching, fingerprint are often considered as scientific evidence. Indeed, the methods used to develop these kinds of evidence are generally beyond the scope of judges and juries's knowledge and are therefore introduced as scientific evidence.

From Frye to Daubert

In 1923, the U.S. District Court for the District of Columbia announced the Frye test, which required a proponent of scientific evidence to establish that the expert witness's theory and method were generally accepted as reliable within the relevant scientific community. In Frye, the court gave a guideline for determining the admissibility of scientific examinations:

« Just when a scientific principle or discovery crosses the line between the experimental and demonstrable stages is difficult to define. Somewhere in this twilight zone the evidential force of the principle must be recognized, and while the courts will go a long way in admitting experimental testimony deduced from a well-recognized scientific principle or discovery, the thing from which the deduction is made must be sufficiently established to have gained general acceptance in the particular field in which it belongs. (Frye v.United States, 1923) »

In brief, to meet the Frye standard, scientific evidence presented to the court must be interpreted by the court as "generally accepted" by a meaningful segment of the associated scientific community. For that, two steps analysis are required:

1. Define the relevant scientific community.
2. Evaluate the testimony and publications to determine the existence of a general consensus. At its core, the purpose of the Frye test is to ensure that the scientific theory or discovery from which an expert derives an opinion is reliable.

In 1975, the Court reviewed the Frye test "in light of sharp divisions among the courts regarding the proper standard for the admission of expert testimony" and held that *Federal Rule of Evidence 702* ("FRE 702") superseded the Frye test. *FRE 702* included rules on expert testimony making it more flexible.

- The expert's scientific, technical, or other specialized knowledge will help the trier of fact to understand the evidence or to determine a fact in issue;
- The testimony is based on sufficient facts or data;
- The testimony is the product of reliable principles and methods; and
- The expert has reliably applied the principles and methods to the facts of the case.

Although criticized, the Frye test was adopted by many states and federal courts, remaining the dominant test until 1993, when the U.S. Supreme Court announced a new standard for determining the admissibility of scientific evidence in Daubert v.Merrell Dow Pharmaceuticals (Daubert, 1993). The Daubert ruling supersedes the Frye standard by setting several factors to be considered in determining the admissibility of scientific evidence. The Court retained « general acceptance » within the relevant scientific community as one factor, but it was no longer the exclusive test for determining admissibility. The criteria set out in the Daubert ruling, in order to determine whether a testimony is based on scientific theories, reasoning or methodology that is reliable and valid, are summarized in, (Loevinger, 1995), as follows:

- Whether the scientific theory or technique can be and has been tested to see if it can be falsified.
- Whether it has been subjected to peer review and publications.
- The known or potential error rate of the methodology.
- Whether the scientific theory or technique has attracted widespread acceptance within a relevant scientific community (the Frye "general acceptance" test).

- Whether standards exist and are maintained to control the operation of the technique.
- Whether the technique is based on facts or data of a type reasonably relied on by experts in the field.
- Whether the technique has a probative value that is not outweighed by the dangers of unfair prejudice, confusion of issues or misleading the jury.

The difference between the Frye and Daubert standards is that according to Frye, admissible evidence may be based on a method generally accepted in the scientific community, but according to the Daubert ruling, the inquiry is not limited to this but also depends on a demonstration, in an "objectively verifiable way", that the evidence is based on a reliable scientific method (Loevinger, 1995). The judge has a screening role between the testifying expert witnesses and the jury, to ensure the reliability and relevance of the expert testimony. For further information, refer to NRC 2009 report (National Research Council, 2009) and the excellent book by (Drozd et al., 2016) (section 1).

2.2.2 DNA typing as a scientific model

After all the criticisms on the traditional forensic science, the adoption of the principle of individualisation, as a categorical opinion of identity of sources, is no longer supported. The need of a logical and correct approach that sounds within the forensic scientific community seems mandatory. The real breakthrough came in the mid 1990s when a new scientific approach was developed for forensic DNA comparison, known as "*DNA typing*" or "*DNA profiling*" model. Indeed, DNA evidence was first presented in court in the 1980s and, after considerable debate, it was unanimously accepted in the 1990s as it meets the most court admissibility requirements demanding transparency in scientific evaluation of evidence and testability of systems and protocols (COUNCIL et al., 1992; Council et al., 1996).

In the DNA typing model, the evaluation of forensic evidence is no longer categorical or deterministic (hard match, i.e yes/no-response) but rather probabilistic (Buckleton et al., 2016). Indeed, the forensic expertise is expressed as a strength of evidence following the Bayesian paradigm: In a typical forensic scenario, a scientist is asked to evaluate the evidence in light of two competing hypotheses. Both hypotheses are mutually exclusive[2] and exhaustive[3] (Robertson et al., 2016).

- Hypothesis 1, H_1: The suspected source is the true source of the evidence.
- Hypothesis 2, H_2: The suspected source can not be (or is not) the true source of the evidence.

The result corresponds to a Likelihood ratio (LR) between the probability of the two hypotheses. The LR gives, therefore, an information about how much more likely the

[2] $H_2 \Rightarrow \overline{H_1}$ and $H_1 \Rightarrow \overline{H_2}$

[3] $P(H_1)+P(H_2)=P(\overline{H_1})+P(\overline{H_2})=1$

evidence would be under the first hypothesis compared with second one. The value of the LR could be interpreted as follows:

- when the LR is greater than one the evidence supports H_1 (increasingly for larger values).
- when the LR is less than one it supports H_2 (increasingly as the LR gets closer to zero).
- when the LR is equal to one then the evidence supports neither hypothesis and so is 'neutral'.

Bayesian approach an elegant solution for individualization

It is important to note that the Bayesian decision theory provides a rigorous framework through which the problem of individualization can be approached in a disciplined manner (Biedermann et al., 2016; Meuwly, 2006).

The emergence of DNA typing as a model for a scientifically defensible approach, has driven a revolution in forensic science: In an extremely short period of time, DNA typing has not just become the reference to be emulated but has also brought into question all other inference-of-the-source forensic disciplines, some of them with a long tradition of expert testimony in court, such as fingerprints or ballistics (Saks et Koehler, 2005). In (Saks et Koehler, 2005), authors point out that the DNA typing model echoes the requirements for admissibility of scientific evidence set out in the US Supreme Court ruling in Daubert (Daubert, 1993). First, the new paradigm is based on a logical scientific probabilistic approach (authors used « [...] *empirically grounded science* » (p. 892) and « [...] *data-based, probabilistic assessment* to describe the new approach » (p. 893)). Second, the limitations of the forensic comparison is quantified and reported via measurements of error rates. Saks and Koehler, therefore, recommend that other forensic comparison sciences emulate DNA comparison model. They state that " [...] *construct (other forensic discipline) databases of sample characteristics and use these databases to support a probabilistic approach*" (p. 893). Four years later, The 2009 release of the National Research Council (NRC) report to Congress on Strengthening Forensic Science in the United States (National Research Council, 2009) made a reiterated call for other of forensic sciences to be more "*scientific*", emulate DNA profile comparison, and conform to the Daubert requirements.

The idea of assessing the weight of the evidence using the likelihood ratio, explored at the beginning of the 20^{th} Century in the DNA typing model, made inroads into many other fields of forensic science (Lindley, 1977; Evett, 1983) only in the latter stages of the 20th century. It now dominates the literature as the logically and correct framework for interpreting forensic evidence(Aitken et Leese, 1995; Robertson et al., 2016; Evett et Weir, 1998).

2.3 Voice sample as a forensic evidence

In this section, we study the example of piece of voice recording as an evidence in the context of speaker comparison known as *Forensic Speaker Recognition* (FSR) or *Forensic Voice Comparison* (FVC).

2.3.1 Typical scenario of voice evidence

In court, to know the identity of a perpetrator based on his voice is quite often at issue: In a crime scene, (i) a victim may have heard but not seen the perpetrator and claim to identify him as a person whose voice is familiar or (ii) there may be one or many recordings, left by the criminal whose identity is unknown, to be compared with the voice of a suspect. "*Ear witness*" (The first case) is not in the scope of this thesis but it still raises several questions scientifically speaking. In the latter case, a voice expert is asked for an expertise based on a reliable scientific method.

A typical definition of forensic voice comparison is given by Geoffrey Stewart Morrison in the second Pan-American/Iberian Meeting on Acoustics:

« *Well, a typical scenario is that the police have an audio recording of an offender from a telephone intercept and another audio recording from an interview with a suspect. What the court wants to decide is whether the speaker on the two recordings is the same person or two different people. The task of the forensic scientist is to analyse the acoustic properties of the voices on the recordings and on the basis of that analysis present a weight-of-evidence statement to help the court to make its decision*[4]. »

2.3.2 Increasing demand for voice evidence expertise in court

Nowadays, forensic voice comparison is under the spotlight as there is an increasing demand for expertise in courts. It is becoming rare to see a law case in which there is no mention of the use of a smartphone or some other modern communication tool. This gives voice an overwhelming advantage over other pieces of evidence. In January 2017, an article [5] ("*Although voice recognition is often presented as evidence in legal cases, its scientific basis can be shaky*") was published in the "*Scientific American Magazine* [6]" by Michele Catanzaro and al, shows that hundreds of voice investigation cases per year are conducted in Italian and British courts. "*It's impossible to know how many voice investigations are conducted each year because no country keeps a register, but Italian and British experts estimate that in their respective countries there must be hundreds per year*".

The process of voice investigation is not limited to speaker comparison but usually involves several other tasks such as:

[4] *http://acoustics.org/pressroom/httpdocs/160th/morrison.html*
[5] https://www.scientificamerican.com/article/voice-analysis-should-be-used-with-caution-in-court/
[6] https://www.scientificamerican.com/

- Transcribing a recorded voice.
- Profiling a speaker based on dialect or language spoken.
- Verifying the authenticity of a recording.
- Putting the suspect's voice in a line-up of different voices.

In this thesis, we focus only on forensic voice comparison.

2.3.3 The myth of "voiceprint"

In 1962, Kersta (Kersta, 1962) introduced the first time the term "*Voiceprint identification*", referring to the speech spectrogram representation. Kersta believed that voiceprint (i.e speech spectrogram.) of a given individual is permanent and unique as fingerprint and thus could be used for speaker comparison purposes with a high degree of certainty. In 1970, Bolt et al. (Bolt et al., 1970) refuted these assertions and stated that the observable resemblance of speech spectrograms depends only partially and indirectly on the anatomical structure of the vocal tract. Indeed, "voiceprint" is mainly a speech visual representation which results from articulatory movements. There is no knowledge that this "voiceprint" will trace the speaker himself. Although the criticisms of "voiceprint" concept, Tosi et al (Tosi et Tosi, 1979) published a study focused on the ability of human to identify speakers, based on a visual comparison of speech spectrograms. The use of the expression "*voiceprint*" is a "*perversion of terminology*" as called by Bimbot and Chollet (Gibbon et al., 1997) but nevertheless it persists, even, quite remarkably, in reference books.

Today, several associations of forensic speaker recognition experts still remind us that speech spectrogram representations should not be used in their "best practices" or resolutions as it is not based on a logical and correct scientific approach. For example, in 2007, the *International Association for Forensic Phonetics and Acoustics* (IAFPA) voted a resolution considering that the spectrogram comparison approach (with a methodological reference to (Tosi et Tosi, 1979)) is "*without scientific foundation and it should not be used in a forensic casework*". Previously, the "*Groupe Francophone de la Communication Parlée*" (GFCP) in 1997 and then the "*Association Francophone de la Communication Parlée*" (AFCP[7]) drew attention to the scientific community to the impossibility of carrying out reliable forensic voice identification at that time. Their position has been outlined in several official statements (GFCP, 1999; AFCP, 2002), scientific publications (Boë et al., 1999; Boë, 2000) and several legal proceedings.

Even though "voiceprint" is still mentioned in some research studies, many researchers continue to reject the "*voiceprint*" concept and reaffirm the unscientific aspects of spectrogram reading. In his paper "*Forensic voice identification in France*" (Boë, 2000), Boë described the "*voiceprint*" history in detail, as well as several other examples of science misused in forensic speaker recognition, like the Micro-Surface "REVAO" tool

[7]The AFCP was initiated by the Groupe Francophone de la Communication Parlée (GFCP) of the Société Française d'Acoustique (SFA) in November 2001: $http://www.afcp-parole.org/$.

during the 1984 "Gregory" case [8]. He insisted that "*voiceprint*" does not yet exist in the sense in which one speaks today of fingerprints or genetic fingerprints. Bonastre et al. in (Bonastre et al., 2003) and later Campbell et al. (Campbell et al., 2009) considered that "*voiceprint*" is a fallacious term and stated that caution is needed if voice would be used in court.

2.3.4 Expressing conclusions in forensic voice comparison

Within the forensic speech science community, there is no consensus of how the forensic expertise should be expressed. Indeed, a number of frameworks are used for evaluating evidence and expressing expert conclusions across the world. These frameworks are surveyed in Gold and French (Gold et French, 2011) and may be grouped under the following headings:

- Binary decision; the expert is restricted to a two-way categorical decision: either the samples contain the voice of the same speaker or the samples contain voices of different speakers. Even though this approach has largely been rejected by the scientific community, it is still used by some experts surveyed in (Gold et French, 2011).
- Classical probability scales (CPS); The hard decision of the binary framework is resolved by the classical probability scales. Indeed, the expert expresses conclusions in terms of the gradient probability of the samples either they originate from the same speaker or they originate from different speakers. Typically, the assessment is a verbal and may use some terms such as "*likely*", "*highly likely*" or "probable", etc. For example, the two samples "*are likely*" to be from the same speakers. According to (Gold et French, 2011), this framework is the most commonly used framework for FVC (13 of 34 experts interviewed in this study use this framework) and is typically employed using auditory and acoustic analysis.
- UK position statement (UKPS); In the UKPS framework (French et Harrison, 2007), voice comparison consists of a two-stage evaluation: "*consistency* and *distinctiveness*". The first stage requires an assessment of the similarity between the suspect and the offender samples, "*whether the known and questioned samples are compatible, or consistent, with having been produced by the same speaker*". In sum, consistency refers to "the degree to which observable features are similar or different". Consistency is quantified on a three-point scale: consistent, not-consistent, or no-decision as shown in Figure 2.3. "*Not-consistent*" means that samples are spoken by different speakers. If the two samples are judged to be "*consistent*", the expert moves to the second stage, termed the distinctiveness judgement. This second stage is an assessment of the typicality of the shared features across the samples within the wider population as stated by Nolan (Nolan, 2001), the strength of evidence is dependent on "*whether the values found matching between two samples are vanishingly rare, or sporadic, or near universal in the general (relevant) population*".

[8] Gregory Villemin was a young boy murdered in 1984. This unresolved case involves several members of his family and is very famous in France.

Distinctiveness is classified using five-point scale ranging from "*not-distinctive*" to "*exceptionally-distinctive*". In brief, "*the expert is to report that the samples are consistent with having been produced by the same speaker, and provide the determined degree of distinctiveness as an indicator of how unusual it would be to find this consistency if the two samples were not spoken by the same speaker*" (Rose et al., 2009). In (Gold et French, 2011), this framework is used by 11 of 34 experts and is typically employed by experts using auditory and acoustic analysis.

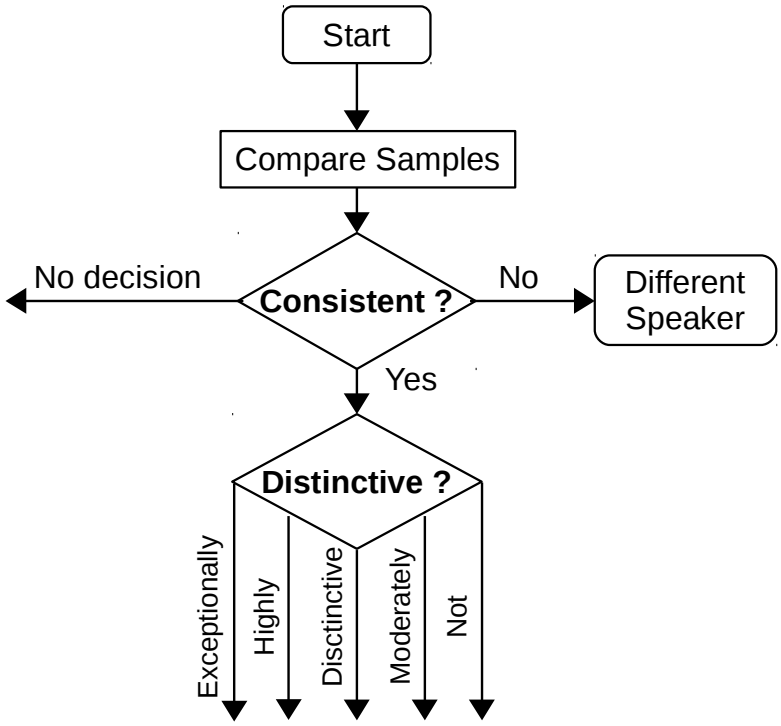

Figure 2.3: *Flow chart representation of the UK Framework (Rose et al., 2009).*

- Likelihood ratio (LR); Experts express their conclusions based on the Likelihood ratio, a numerical value which indicates the degree of support to the prosecution or the defence hypotheses as introduced in subsection 2.2.2. The LR overcomes logical shortcomings of the UKPS since it ensures that similarity is considered relative to typicality, rather than having two independent stages of analysis.

A study of satisfaction of each framework is reported in (Gold et French, 2011) using a Likert scale[9]. Results are summarized in Table 2.1.

[9] A Likert Scale was used to measure the level of satisfaction with a respondent's conclusion method. Likert ratings were averaged across respondents. The scale ranged from 1 (extremely dissatisfied) to 6 (extremely satisfied).

Table 2.1: *Satisfaction with conclusion framework expressed using the Likert Scale.*

Conclusion Framework	Mean Likert Rating
Numerical LR	5.00
UK Position Statement	4.27
Verbal LR	4.00
Classical Probability Scale	3.67
Binary Decision	3.50

Table 2.1 shows that the numerical likelihood ratio provides the best satisfaction for the experts for expressing conclusions, ranking in the top of the satisfaction scale followed by the UKPS and the verbal LR. The CPS and binary frameworks are at the end of the satisfaction scale. In this thesis, we focus only on the use of the Likelihood ratio framework in forensic voice comparison. For more details about the other approaches, their pros and cons, etc, readers are referred to (Hughes, 2014; Broeders, 2007; Rose et al., 2009).

2.3.5 Admissibility of the likelihood ratio in forensic voice comparison

In forensic voice comparison, there is a clear need for a common framework in order to evaluate voice evidence as it appears, in several crimes, that the only lead to proceed investigation is a piece of voice recording. This framework should satisfy the requirements of admissibility set out in the US Supreme Court ruling in Daubert (Daubert, 1993).

The emergence of DNA profile comparison in the 1990s as a scientific approach gradually (or widely) adopted in several forensic branches has pushed many academics and practitioners in forensic voice comparison to emulate (or recommend) this model.

A substantial forensic opinion argument made by Champod and Meuwly (Champod et Meuwly, 1998), initially at the "*Workshop on Speaker Recognition and its Commercial and Forensic Applications*" (RLA2C April 1998 Avignon, France) with a subsequent journal article published in 2000 (Champod et Meuwly, 2000), has had a great impact on the research community. This paper drew on the DNA model to make a lucid argument for its adoption in forensic voice comparison.

In 2003, Rose et al. (Rose, 2003; Rose et al., 2003) echoed Champod et Meuwly in using the LR framework for expressing the expert conclusions.

In their paper "*Emulating DNA: Rigorous Quantification of Evidential Weight in Transparent and Testable Forensic Speaker Recognition*", Gonzalez et al. (Gonzalez-Rodriguez et al., 2007) proposed a unified approach for forensic voice comparison oriented to fulfil the admissibility requirements within a framework which is transparent, testable, and understandable, both for scientists and fact-finders. This approach is founded based on the DNA model and recommends the use of the Likelihood ratio.

Morrison has joined Gonzalez et al. He believed that forensic voice comparison expertise should be expressed using the Likelihood ratio framework. In (Morrison, 2009a), Morrison states "*there is one other component of the new paradigm which I believe is implicit in Saks and Koehler's* (Saks et Koehler, 2005) *and the NRC report's* (National Research Council, 2009) *recommendation that other forensic comparison sciences emulate forensic DNA comparison: the adoption of the likelihood-ratio framework for the evaluation of evidence*".

Different other academics such as Aitken and Taroni (Aitken et Taroni, 2004), and Evett et al. (Evett et al., 2000), recommended in the NRC report (National Research Council, 2009) as providing "the essential building blocks for the proper assessment and communication of forensic findings"(p. 186), advocate the use of the likelihood-ratio framework.

In this section, we make a short review for one of the most accepted methods, the likelihood ratio-based approach and the full Bayesian approach for the interpretation of evidence.

Likelihood ratio approach for evidence evaluation

Forensic voice comparison (FVC) is based on the comparison of a recording of an unknown person (the evidence or trace) and a recording of a known suspect (the comparison piece). It aims to indicate whether the evidence supports the prosecution (the two speech excerpts are pronounced by the same speaker) or defence (the two speech excerpts are pronounced by two different speakers) hypotheses. The *LR* not only supports one of the hypothesis but also quantifies the strength of its support.

Let:

- H_p be a hypothesis or proposition advanced by the prosecution: the evidence and suspect recordings have the same origin.
- H_d be a hypothesis or alternative put forward by the defence.
- E signify the findings (referred to here as the evidence).

The *LR* is calculated using Equation 2.1.

$$LR = \frac{p(E \mid H_p)}{p(E \mid H_d)} \tag{2.1}$$

Another way of understanding the LR is that the numerator represents a numerical statement about the degree of similarity of the evidence with respect to the suspect samples and the denominator represents a numerical statement about the degree of typicality with respect to the relevant population (Champod et Meuwly, 2000; Drygajlo et al., 2015; Hughes, 2014; Ishihara et al., 2008; Kinoshita et al., 2014).

If the evidence is more likely to occur under the same-origin hypothesis than under the different-origin hypothesis then the value of the likelihood ratio will be greater than

one, and if the evidence is more likely to occur under the different-origin hypothesis than under the same-origin hypothesis then the value of the likelihood ratio will be less than one.

The responsibility of a forensic scientist is to provide the court with a LR and not a decision (Lucy, 2013; Morrison, 2009a), it is the responsibility of the trier of fact to make a decision using other sources of information about the case at hand. This decision is in itself a probabilistic statement, known as the posterior odds, and can be expressed mathematically as shown in Equation 2.2,

$$\underbrace{\frac{p(H_p \mid E)}{p(H_d \mid E)}}_{\text{posterior odds}} = \underbrace{\frac{p(E \mid H_p)}{p(E \mid H_d)}}_{LR} \times \underbrace{\frac{P(H_p)}{P(H_d)}}_{\text{prior odds}} \tag{2.2}$$

In Equation 2.2, The prior odds represent the view on the prosecution, H_p, and defence, H_d, hypotheses before the scientific evidence, E, is presented. This is something that is formed in the minds of the judge and jury. These odds are based on the non-scientific evidence and their assessment is the duty of judge and jury (for a dissenting view see Meester and Sjerps and the associated commentary (Meester et Sjerps, 2003; Triggs et Buckleton, 2004). The posterior odds could be seen as a revision of prior odds after the scientific evidence. The likelihood ratio informs us how to relate these two odds (prior and posterior odds) and how the opinion of trier of fact or jury are updated in a logical manner having heard the evidence. Figure 2.4 summarizes the main step from evidence to decision highlighting all the parties involved in this process.

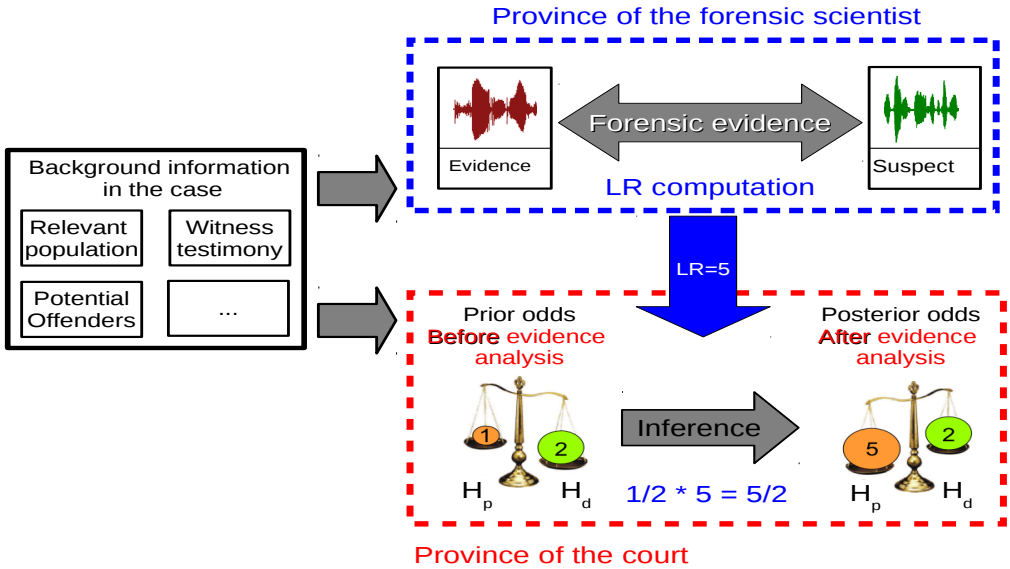

Figure 2.4: *Bayesian inferential framework in LR-based evidence analysis.*

It would be fair enough to leave the interpretation and decision procedure to the

court as it is the best body to undertake this interpretation as it disposes of several other information for this purpose. Furthermore, experts should only provide the court with the strength-of-the-evidence summarized by the LR value and should strictly avoid to form a view or make an influence to the value of prior odds. In (Buckleton et al., 2006), authors state "*Strictly it is not the remit of the scientist to form a view as to the value of the prior odds and most scientists would strictly avoid this.*"

Likelihood ratio framework: A logical approach for evidence interpretation

"The practical use of this philosophy typically leads scientists to report the strength-of-evidence in a form of a likelihood ratio. By doing this, the scientist reports the weight of the evidence without transgressing on those areas reserved for the judge and jury and thus being able to make an objective statement regarding the strength of the evidence. This is the reason why we have chosen the term 'logical' or 'likelihood ratio' to describe this approach." (Buckleton et al., 2006)

It is important to note that a verbal interpretation of the LR is sometimes used in the court. Even though this practice is put into question, it is still used in some countries. Table 2.2 represents an example of the interpretation of the LR value in New Zealand.

Table 2.2: *LR verbal interpretation in New Zealand.*

Likelihood Ratio	Verbal Equivalent
1	Inconclusive
1 to 10	Slightly supports the prosecution proposition
10 to 100	Moderately supports
100 to 1,000	Strongly supports
1000 to 1,000,000	Very strongly supports
Greater than 1,000,000	Extremely strongly supports

Issues with the likelihood ratio in forensic voice comparison

The Bayesian formalism becomes a cornerstone of forensic expertises and reports in several areas, including speech evidence. It provides a very elegant theoretical framework and places the expert (back) in his proper domain which is science and not judgement. However, implementing a theoretical framework to handle real world cases raises three main problems as highlighted by Bonastre et al in (Bonastre et al., 2015):

- **Estimation of** $P(E \mid H_d)$

In order to estimate the LR, one should calculate both $P(E \mid H_p)$ and $P(E \mid H_d)$. If, using a machine learning approach for example, it is possible to learn a class model for the hypothesis H_p using several samples of the suspect's voice, it is not trivial to train such a model for H_d. Indeed, three elements have to be evaluated in H_d in order to train such a model: the "concordant features", their "relative frequency" and the "relevant population" (Champod et Meuwly, 2000). This means that a forensic approach

claiming to comply with the Bayesian formalism, which is very often described as the only scientific formalism accepted for forensic evidence, should define these three elements explicitly. And the latter does not depend on the forensic expert, or at least not completely, since it is "dictated by the hypothesis proposed by the defence" (Champod et Meuwly, 2000). It means that the forensic expert referral should include a clear description of the expected relevant population. We should also remember that this hypothesis is not definitive and may evolve during the trial. In (Rose et al., 2009), Rose and Morrison highlight that the theoretical definition of the relevant population and practical issue of the collection of reference data are "real problems" for the numerical LR approach. They claimed "*We readily acknowledge that these are real problems: probably the most pressing at the moment. The first is theoretical and relates to the choice of the relevant population to sample, and the size of that sample; the second is practical and relates to the actual collection of data.*". French et al. (French et al., 2010) claim that the feasibility of the LR approach depends on a sufficient available data to estimate the distribution within the relevant population of all of the potential variables analysed in a given case "*it is unrealistic to see it as merely a matter of time and research before a rigorously and exclusively quantitative LR approach can be regarded as feasible*".

- **Background information**

Depending on the different legal systems, the forensic expert can have access to several pieces of background information concerning the current case, other than the piece of evidence E and the hypotheses to be evaluated. Therefore, the LR equation is very often formulated in order to include these background information in addition to the evidence, E, in the expert knowledge. As we have noted , the LR denominator is an estimation of the random probability in the "relevant population". It is understandable if the expert wishes to use as much information as possible in order to determine the H_d probability. Unfortunately, this is **in obvious contradiction** with the scientific position, which is **to be as little subjective as possible**. It may be useful here to remember the well-known double blind principle and why it is so important in medical research assessment.

- **Understandability of the LR by the court**

Champod believes that the LR is a useful tool "*for assisting scientists to assess the value of scientific evidence*", to "*clarify the respective roles of scientists and of members of the court*" and essentially "*to help jurists to interpret scientific evidence*" (Champod et Meuwly, 2000). Quite obviously, a forensic analysis has an interest only if judges, lawyers and jurors are able to understand the work done by the expert precisely, as well as the intrinsic nature of the scientific evidence presented. However, understanding probabilities in general and LR more specifically is not straightforward. Daniel Kahneman, the 2002 economics Nobel Prize (co)laureate, a specialist of judgement and decision-making and one of the two proposers of the prospect theory (Tversky et Kahneman, 1975), states in his 2011 book "Thinking, Fast and Slow" that "*Bayesian reasoning is not natural for humans*".

In (Thompson et al., 2013), the perception of LR by jurors is analysed. It appears that it is not easy for them to understand statistical evidence correctly. As highlighted by the authors, this is particularly true when forensic experts, prosecutors or lawyers

provide arguments that invite or encourage fallacious conclusions from statistical evidence, which is not uncommon in courts.

2.3.6 Likelihood ratio estimation

A diversity in approaches used for forensic voice comparison emerges from (Morrison et al., 2016; Gold et French, 2011). The approaches reported were: auditory, spectrographic or auditory-spectrographic, auditory-acoustic-phonetic, acoustic-phonetic, human-supervised automatic and fully automatic speaker recognition based approach. Acoustic-phonetic and fully automatic approaches are the most used solutions when *LR* framework is involved (Gold et French, 2011). In the following subsections, we review briefly both of them.

Acoustic-phonetic approach

The acoustic–phonetic approach consists mainly in making quantitative measurements of the acoustic properties in both voice recordings on comparable phonetic units. Indeed, similar phonetic units are extracted from the known and questioned speech samples, and various acoustic parameters measured from these segments are compared (Nolan, 1980; Nolan et Grigoras, 2005; Morrison, 2009b, 2011a; Zhang et al., 2011). The international survey on FVC practice published in 2011 (Gold et French, 2011) reports that 97% of the survey participants use formants, making formants one of the most favoured acoustic features in FVC casework. These formants correspond to the resonances of the vocal tract, and are considered as a good quality descriptors for vowels (Fant, 1970).

Automatic approach

Unlike the acoustic-phonetic approach which uses acoustic features extracted on a specific region of the signal, the automatic approach does not use explicitly the phonetic information: Acoustic features are extracted throughout the signal every a specific step, so-called frame.

The use of automatic approaches for forensic speaker recognition clearly offers important advantages in terms of objectivity and repeatability of the voice comparison measures, but also in terms of human time costs. The limited cost of automatic processes could also allow the expert to test several voices against the piece of evidence.

2.4 Miscarriage of justice

The term "*miscarriage of justice*" refers to a legal act or verdict that is clearly mistaken, unfair, or improper (Bedau et Radelet, 1987). Two errors could be seen in criminal

judicial system:

2.4.1 Wrongful conviction

Primarily, a miscarriage of justice is related to "Wrongful convictions". According to Duhaime's Law Dictionary a wrongful conviction is the conviction of a person accused of a crime which, in the result of subsequent investigation, proves erroneous.

Figure 2.5 shows statistics of wrongful conviction 's causes. It appears clearly that the primary causes of wrongful convictions is eyewitness misidentification and in the second place we found errors coming from forensic science. The two highlighted bars in "red" indicate causes related to forensic science.

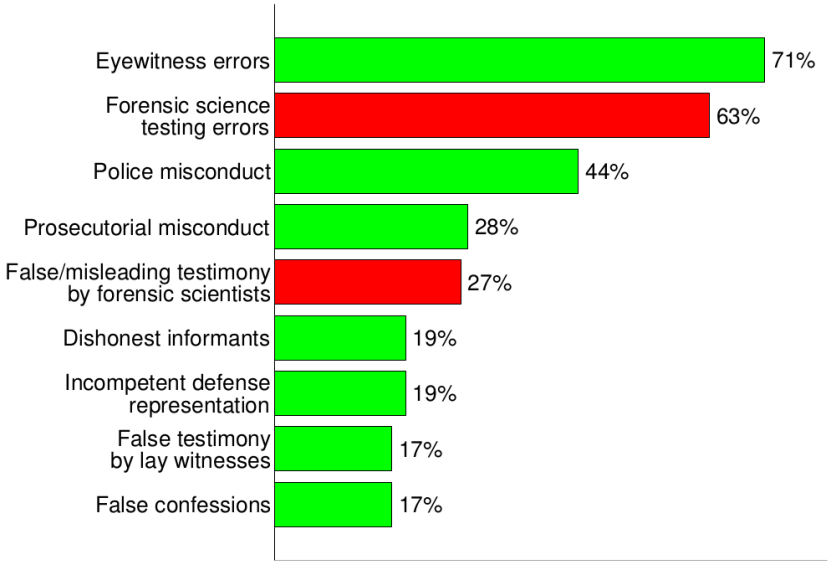

Figure 2.5: *Factors present in 85 wrongful convictions, based on the case analysis data from the Innocence Project. Source: Saks and Koehler (Saks et Koehler, 2005).*

In recent years, the emergence of forensic DNA analysis has allowed the criminal justice system's ability to clear many people wrongly convicted. The DNA testing has exposed a large number of cases in which innocent persons were wrongly convicted of crimes they did not commit. Thanks to the "innocence project" [10], the study of some cases in which suspects are announced guilty revealed numerous causes of errors. The "innocence project" has led to the exoneration of a significant number of innocents previously convicted (196 cases until now) through DNA testing.

The term of "miscarriage of justice", can also apply to errors in the other direction "errors of impunity", and to civil cases.

[10] http://www.innocenceproject.org/, founded in 1992 by Peter Neufeld and Barry Scheck at Cardozo School of Law.

2.4.2 Errors of impunity

"*Errors of impunity*" is a term used in Brian Forst's book "Errors of Justice" (Forst, 2004). It is defined as lapses that result in criminals either remaining at large or receiving sanctions that are below a socially optimal level. In this thesis, we refer to "*error of impunity*" as failing to find a culpable person guilty.

In the criminal judicial system, errors of impunity can be caused in much the same ways as wrongful conviction can. However, from an ethical point of view "wrongful conviction" is considered as the most outrageous in the miscarriage of justice. In this context, Voltaire said (Zadig 1747): "*It is better to risk saving a guilty person than to condemn an innocent one*".

2.4.3 Examples of miscarriage of justice

Guilty! This short word, in court, could lead to very heavy consequences on the defendant's life and thus announcing people guilt should be justified.

Figure 2.6: *From speech evidence to identity! Is it really possible?*

Voice as one of potential forensic evidence could play an important role in deciding verdict. But are speaker's voices always recognized correctly? Or concretely, is the forensic voice comparison always reliable?

Unfortunately, forensic voice comparison could lead to misleading results. For example, in (Hansen et Hasan, 2015), authors have mentioned an example about how to wrongly use automatic speaker recognition in a real forensic case. The example was seen during the recent U.S. legal case involving George Zimmerman, who was accused of shooting "*Trayvon Martin*[11]" during an argument. Another famous example could be mentioned too, the case of "Jerome Prieto", a man who served in jail 10 months because of a controversial police investigation that wrongly identified Prieto's voice in a phone call claiming credit for a car bombing. There are plenty of troubling examples of

[11] *https://en.wikipedia.org/wiki/Shooting_of_Trayvon_Martin*

dubious forensics in which innocents could be incriminated. "Michael Musmanno", an American jurist, said in this context:

> *"Can it happen? Can an innocent person be executed? Can it happen? It has happened!"*

Conclusion

In this chapter, we introduced the topic of forensic science, and particularly forensic identification. The evolution of forensic identification from traditional forensic science to the so-called "*paradigm shift*" was detailed, while the resulting requirements were highlighted. A particular focus was put on discussing the case of forensic evidence using a person's voice, with an overview of the position of the "*paradigm shift*" in forensic voice comparison.

Chapter 3

Voice comparison in courts: A challenging task

Contents

Introduction

For years, movies and television series have painted an unrealistic image about the "speech processing domain". For example, the movie "Clear and Present Danger" released in 1994, showed an expert who analysed a short speech recording and declared that the speaker is "Cuban, aged 35 to 45, educated in the [...] eastern United States". But is it really possible, from a brief recording, to extract all these information with high precision including speakers identities? Quite obviously most individuals do not

realize how difficult and complex the task really is. If they are to be asked whether it is possible to recognize speakers from their voice, they would readily (and quickly) answer "yes".

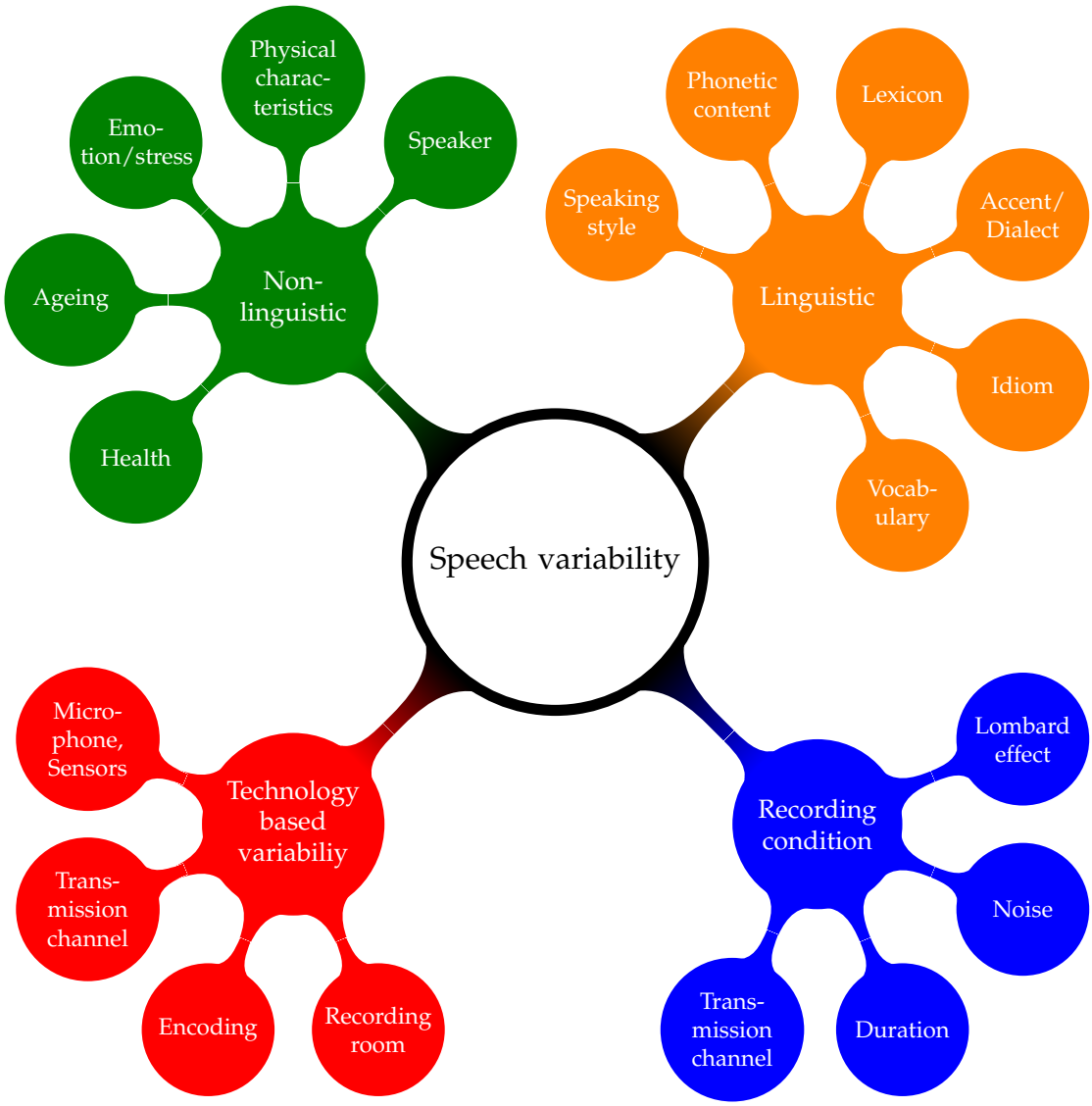

Figure 3.1: *Main sources of potential variability in voice comparison.*

In real world, using voice recording to recognize speakers is not so simple as promoted. Indeed, unlike other forms of forensic evidence such as DNA or fingerprint, speech is prone to a large degree of variability. Variability in the speech can be caused by numerous factors. In Figure 3.1, we classify the main source of variabilities on four broad classes[1] 1) linguistic based, 2) non-linguistic based, 3) technology based

[1]In this representation, we assume that all these factors are independent of each other, while in reality, it is not the case. For example, it is a little difficult for a person to speak angrily with normal rate and

and 4) factors linked to recording conditions. The first two factors (linguistic and non-linguistic) are related to the speech itself while the other factors are related to the technology.

This large variability and its diversity poses a real challenge in speaker recognition since any variability factor could decrease significantly the speaker recognition accuracy. Furthermore, the impact of variability is reinforced when there is a mismatch between the two voice recordings to be compared, test and enrolment samples. In commercial applications like banking application, some factors of variabilities could be controlled such as recording material, recording duration, speaker's distance to microphone, recording quality, etc. Moreover, in such application the speaker is cooperative. Speaker can be recorded many times and even further, the text to be produced can be fixed or pre-arranged, etc. Nevertheless, in forensic cases, the story is different. Forensic speaker recognition is still an extremely challenging field (Bonastre et al., 2003; Campbell et al., 2009; Bonastre et al., 2015) and reveals many problems which come either from the speech itself or from the way that the evidence and the suspect voice are collected...

In this chapter, we discuss some factors that make FVC a difficult task. First, we present the piece of evidence challenging conditions. Second, we briefly review the impact of mismatched conditions on the speaker recognition accuracy. At this level, only linguistic or technology based factors are discussed. Finally, we show the impact of within-speaker variability on FVC.

3.1 Forensic conditions

It is important to highlight the large difference between a voice comparison under "ideal" conditions[2] versus forensic voice comparison, where the real-world circumstances make the task much more complex. In a crime scene, the piece of evidence is collected under specific conditions. Indeed, a perpetrator is often acting under stress, sometimes yelling or speaking with a specific speech style. The forensic expert does not have a choice in defining the piece of voice conditions and should deal with the case at hand.

In forensic cases, the piece of evidence conditions are challenging:

- The piece of evidence is in most of the time very short: In (Künzel, 1994), Künzel estimated that 20% of the "voice recordings" analysed by the *German federal police* (BKA) contained less than 20 seconds of "*net*" speech (According to Künzel, "net" speech means "*what remains from the original signal after removing speech portions of dialogue partners and badly distributed portions*").
- The piece of evidence is most of the time degraded or of bad quality. In (Künzel, 1994), Künzel estimated that more than 95% of the cases treated by the BKA are

normal volume.

[2]In this thesis, ideal conditions refer to fully controlled conditions.

telephone-transmitted signals. This kind of transmission implies a multitude of degradation (detailed discussion could be found in (Moye, 1979) or in (Künzel, 2001)). Recent studies, dedicated to forensic applications, have addressed the influence of mobile phone communication on forensic voice comparison (Nair et al., 2016; Alzqhoul et al., 2012) or (Alzqhoul et al., 2016; Alzqhoul, 2015).

- Voice disguise: Perpetrators intend to hide their identities by mimicking another person or changing their voice. The *BKA* found that in 15% of the forensic cases (Künzel, 1994), criminal tried to change their voices in order to not be recognized.

3.2 Mismatched conditions

In forensic voice comparison casework, the criminal and suspect's voice are usually recorded in different situations. Indeed, the recordings could differ in telephone channel distortions, ambient noise in the recording environments, the recording devices, as well as their linguistic content or duration. For instance, the questioned speaker recording might be from a telephone conversation in a noisy environment whereas the suspected speaker recording is from a microphone-recorded police interview in a reverberant room. In many forensic cases, the expert does not have a choice in defining the recording conditions for the suspect and questioned recordings, as these recordings are provided by the police or the court, and additional recordings can not be made. A possible mismatch between the recording conditions remains a major challenge and therefore the impact of any mismatch on forensic speaker comparison should be addressed in order to assure the transparency and the reliability of the comparison process.

Several questions arise and have to be addressed: do the suspect and the questioned recordings have similar or equivalent conditions? If not, which is often the case, what is the impact of such a mismatch on FVC and how to handle it? To take this a step farther, another important question should be addressed concerning the mismatch between both suspect and criminal's voices from one hand and the recordings of the databases from the other hand. This question is so important as we know that the Bayesian paradigm relies heavily on the use of databases[3] to quantify the strength-of-evidence.

Many research works have dealt with various types of mismatch conditions. The following is a selected (short) list of them:

3.2.1 Transmission channel

Transmission channel mismatch between the recordings of suspect, offender and background data represents a frequent scenario in real forensic caseworks. Indeed, speech

[3]Here, databases refers to the background data used to evaluate the typicality factor in the likelihood ratio.

utterance could originate across a variety of telecommunication networks such as landline phone networks, mobile phone technologies, VoIP, etc. These transmission technologies do not have the same impact on the speech signal. Not only this, differences could also be seen using the same telecommunication technology. For example, within the mobile phone arena, the network providers use different technologies, such as *Global System for Mobile Communications* (GSM) and *Code Division Multiple Access* (CDMA). These technologies are different in the way in which they handle the speech signal, which in turn will lead to a significant mismatch between speech recordings.

Transmission channel variability is one among the challenging factors that poses a serious issue in forensic voice comparison (Zhang et al., 2013; BT Nair et al., 2014; Alzqhoul, 2015; Alexander, 2005; Alexander et al., 2005). In (BT Nair et al., 2014), authors investigated the impact of mismatch conditions associated with mobile phone speech recordings on forensic voice comparison (FVC). Table 3.1 reports the performance of FVC explained in term of C_{llr}[4] for matched and mismatched conditions.

Table 3.1: *FVC performance for matched and mismatched condition reported in term of c_{llr} (BT Nair et al., 2014).*

		Suspect speech from	
		GSM	**CDMA**
Offender speech from	**GSM**	0.216	0.544
	CDMA	0.597	0.249

From Table 3.1, authors showed that FVC accuracy degrades severely under mismatched conditions. Indeed, an average increase of C_{llr} values from 0.232 in matched conditions to 0.57 in mismatched conditions was observed (almost multiplied by a factor of 2.5).

In (Alzqhoul, 2015), authors investigated the impact of mismatch conditions between the test and background datasets on FVC. In this study, "Matched" conditions refers to the case when the background set is coded similarly to the development and testing sets. "Mismatched" conditions refers to the case when the background set uses no-coded speech, whereas the development and testing sets use coded speech. Table 3.2 presents a comparative of FVC performance between GSM and $CDMA$ in matched and mismatched condition at three SNR levels[5]: 9, 15, and 21 dB.

[4] C_{llr} is a measure of accuracy. Lower is the C_{llr}, better is the accuracy. C_{llr} is detailed in Chapter 5, Subsection 5.5.4.

[5] Three SNR levels are used in order to simulate the real-world conditions since the designers of mobile phone codecs typical test the performance of their products using three types of BN, namely car, street, and babble noise and at three SNR levels: 9, 15, and 21 dB.

Table 3.2: *FVC performance between the GSM and CDMA networks reported in term of C_{llr}. Babble noise is used. Adapted from (Alzqhoul, 2015).*

Network	**GSM**		**CDMA**	
Condition	**Matched**	**Mismatched**	**Matched**	**Mismatched**
BN at 9dB SNR	0.209	0.300	0.279	0.506
BN at 15dB SNR	0.216	0.263	0.249	0.358
BN at 21dB SNR	0.173	0.187	0.181	0.343
Uncoded speech	0.167			

From Table 3.2, authors showed that results of mismatched conditions were substantially worse than for matched conditions for both *GSM* and *CDMA* cases. A comparison with respect to the uncoded speech experiment shows an increase in the C_{llr} value by about 100% to 200% depending on the *SNR* level. The FVC accuracy under mismatched conditions are approximately 130% worse than matched conditions at an *SNR* value equals to 9 dB.

3.2.2 Duration

One of the issues with forensic speaker recognition is that the perpetrator and suspect speech data are often of different durations: In a typical forensic casework, the useful speech material in a piece of evidence is limited while suspect speech data could be of long duration especially when acquired from police interview. This situation requires the comparison of two utterances of different lengths. Speaker comparison systems are known to perform significantly worse when there is a duration mismatch between the voice recordings (Hasan et al., 2013). Several researchers have addressed the problem of duration mismatch (Scheffer et Lei, 2014; Ando et al., 2016; Hasan et al., 2013; Sarkar et al., 2012; Hong et al., 2015; Ben Kheder et al., 2016). In (Ando et al., 2016), authors presented the effect of duration mismatch between the two voice recordings to be compared, test and enrolment. Table 3.3 reports system performance explained in term of *EER*[6]. The test utterance varies between 1 second to 10 seconds while the enrolment utterance varies between 10 seconds to 60 seconds.

Table 3.3: *Effect of test/enrolment duration on system performance reported in term of EER (Ando et al., 2016).*

Enrolment	**Test**				
	1 sec	**2 sec**	**3 sec**	**5 sec**	**10 sec**
10 sec	5.99	1.47	0.96	0.42	0.17
60 sec	10.05	3.07	2.27	0.53	0.00

Based on the results reported in Table 3.3, authors showed that when the mismatch

[6] *EER* is a measure of error. Lower is the *EER*, better is the performance. *EER* is detailed in Chapter 5, Subsection 5.4.

between the two speech recordings is high, the *EER* is large. For example, when the test utterance is equal to 1 second, the *EER* is equal to 10.05% when the enrolment is 60 seconds whereas it is equal to 5.99% when the enrolment is equal to 10 seconds.

In (Ben Kheder et al., 2016), authors investigated the impact of duration mismatch between training and testing sets. Speech durations of both sets are uniformly sampled ranging from 5 seconds to 30 seconds. Table 3.4 summarizes the effect of train/test durations in various mismatched conditions.

Table 3.4: *Effect of the train/test duration on the system performance reported in terms of EER (Ben Kheder et al., 2016). "**Full**" corresponds to the case where the speech segment length are more than 30 seconds. "**Red**" color refers to the highest EER value while the "**green**" color refers to the lowest EER obtained for a specific "Test" condition.*

		Train speech duration					
		Full	**30s**	**20s**	**15s**	**10s**	**5s**
Test speech duration	**Full**	***1.59***	2.05	2.49	2.73	3.18	4.56
	30s	3.59	3.18	***2.96***	3.18	3.87	5.21
	20s	5.26	4.32	***3.87***	3.87	4.78	5.69
	15s	7.28	5.92	5.89	***5.50***	5.72	6.54
	10s	11.84	8.65	7.99	***7.28***	7.75	8.43
	5s	21.83	17.31	15.91	15.26	13.62	***13.21***

From Table 3.4, authors showed that system performance decreases dramatically when there is a mismatched duration between the training and testing sets.

3.2.3 Material and technical conditions

In forensic cases, recording technical conditions could be another salient source of mismatch. Indeed, speech samples are often recorded with various recording devices, the distance to the microphone is varying, etc. The mismatched conditions between the criminal and suspect's voice recordings regarding these factors are known problems in forensic speaker recognition (Enzinger et Morrison, 2015; Alexander et al., 2004). During his Speaker Odyssey 2014 keynote talk (Campbell, 2014), Campbell presented some factors of variability with a potentially strong impact for forensic speaker recognition notably the recording device and the speaker's distance to the microphone. Figure 3.2 shows the *EER* variations depending on the recording device.

Figure 3.2 shows a wide accuracy gap depending on the recording device used and this gap widens significantly when different devices are used for the two recordings. Figure 3.2 shows also that "*Mic-Tel*" condition presents the high error rate. This could be explained by the large quality mismatch between the two speech recordings to be compared.

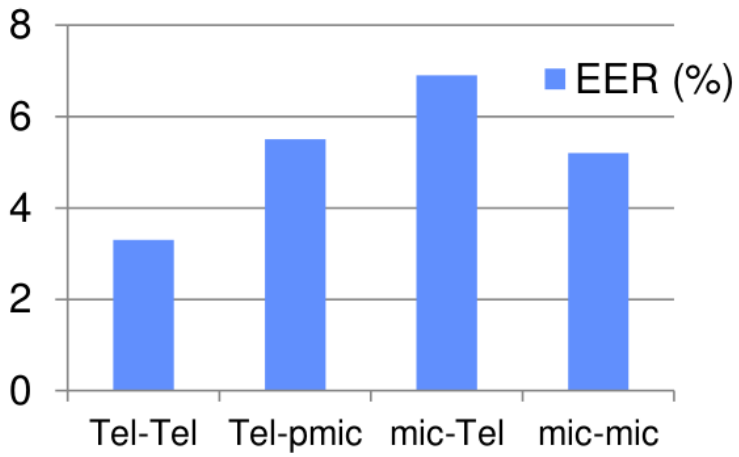

Figure 3.2: *EER variations depending on the recording device (Campbell, 2014).*

3.2.4 Environmental noise

Speech utterances recorded in real environments often contain different types of environmental noises such as traffic noise, car noise, music etc. Moreover, speech signal could be affected by noise at different levels (i.e different SNR levels). Environmental noise mismatch among training and testing datasets is known to degrade outstandingly the speaker recognition accuracy (Ortega-García et González-Rodríguez, 1996; Sadjadi et Hansen, 2010; Mandasari et al., 2012; Ribas et al., 2015; Mak et al., 2016; Li et al., 2017). In (Mak et al., 2016), authors investigated the impact of environmental noise mismatch between training and testing sets on the speaker recognition accuracy. They used three sets, χ_1, χ_2 and χ_3 with different SNR levels. χ_1 is a clean set while χ_2 and χ_3 are corrupted with babble noise at 6 dB and 15 dB respectively. Table 3.5 reports the results for matched and mismatched conditions.

Table 3.5: *Speaker identification accuracy under matched (diagonal) and mismatched (off diagonal) training and test conditions. χ_1, χ_2 and χ_3 refer to clean, 15 dB and 6 dB conditions (Mak et al., 2016).*

		Test data		
		χ_1 (clean)	**χ_2(15 dB)**	**χ_3(6 dB)**
Training Data	**χ_1 (clean)**	95.6%	83.7%	55.2%
	χ_2(15 dB)	93.9%	93.9%	83.7%
	χ_3(6 dB)	90.1%	93.3%	88.5%

Table 3.5 shows that matched environmental noise between test and training data (green) presents the highest speaker recognition accuracy. Moreover, results in the first and third columns suggest that speaker recognition accuracy gradually decreases when the SNR of the training data progressively deviates from that of the test data (i.e when the mismatch is reinforced).

3.2.5 Linguistic content

In forensic voice comparison, speech linguistic information such as dialect, accent, grammar, idiom, etc, could be a real source of mismatch. For example, considering phonetic content, both voice recordings could have completely different phonetic content which could affect the comparison accuracy. Phonetic mismatch problem has been attacked with phonetically-motivated purposes in many research works (Scheffer et Lei, 2014; Chaudhari et al., 2003; Hébert et Heck, 2003). Spoken language could also be another source of mismatch. The impact of language mismatch on speaker recognition performance was addressed in many research studies (Auckenthaler et al., 2001; Ma et Meng, 2004; Ma et al., 2007; Misra et Hansen, 2014).

Table 3.6: *Results explained in terms of EER in the language matched and mismatched conditions. "EN-1" refers to English telephone data. "EN-2" refers to multilingual telephone and microphone data. "ML-1" refers to multilingual telephone and microphone data. "ML-2" refers to telephone, microphone noisy data. "UBM", "TV" and "PLDA" correspond to the models used by the ASpR system (Misra et Hansen, 2014).*

Training models			**Matched**	**Mismatched**
UBM	**TV**	**PLDA**	**EER (%)**	
EN-1	EN-2	EN-2	1.745	4.395
EN-1	EN-2	ML-2	0.868	1.662
ML-1	EN-2	EN-2	1.495	3.602
ML-1	EN-2	ML-2	1.188	2.291
EN-1	ML-2	EN-2	2.178	6.411
EN-1	ML-2	ML-2	0.485	2.288
ML-1	ML-2	EN-2	1.869	5.11
ML-1	ML-2	ML-2	0.526	2.781

In (Misra et Hansen, 2014), Misra and Hansen studied the impact of language mismatch on the speaker recognition performance. In this study, the enrolment segments correspond to English while the test set is divided in two subsets: The first one contains only English as the spoken language (matched condition) and the second one contains all the languages other than English (mismatched condition). In order to ensure that the dominant mismatch is the language, long test and enrolment segments issued from telephone conversations are selected. The results obtained in matched and mismatched conditions are summarized in Table 3.6.

From the results in Tables 3.6, authors observed that when only English data is used for training (i.e, UBM, TV space and $PLDA$ model), the EER in the matched and mismatched conditions is 1.745% and 4.395%, respectively. This shows the severe degradation caused by language mismatch alone, dropping the errors by a factor of 2.5.

If changes in the acoustic environment, duration, linguistic content and so on are considered as recording conditions mismatch, variations of the speaker him/herself (state of health, mood, ageing) represent other undesirable factors for FVC.

3.3 Within-speaker variability

In daily life, we are frequently identifying, fairly successfully, familiar speakers without seeing them. This human ability is possible thanks to the variation between speaker voices, usually known as "*between-speaker or inter-speaker variability*". "*Although it is a general assumption, without any scientific basis, that different speakers have different voices*" (Rose, 2003), the voice of a speaker is far from constant and will always vary. It is a phonetic truism that no one can say the same word in exactly the same way two times. This phenomenon is referred as "*within-speaker variability, style shifting or intra-speaker variability*" (Eckert et Rickford, 2001). In this context, Doddington stated: "*The problem is that people sometimes are just not themselves(!)*"(Doddington, 1985).

Within-speaker variability remains one among the difficult challenges that forensic voice comparison faces (Ajili et al., 2016; Kahn et al., 2010; Karlsson et al., 1998; Stevens, 1971). High intra-speaker variability could lead to erroneous or misleading voice comparison. Indeed, this variability could make voices of different speakers closer.

The main reason which makes within-speaker variability difficult to handle is the lack of control over speech variation factors. (Stevens, 1971) provides an introduction to the acoustic theory of speech production and describes the various factors that cause variation, with special emphasis of physiological causes of variability. In the following, some selected factors are discussed because they would mostly appear in a typical forensic scenario.

- Speaking rate and style; a criminal could speak fast or slow, spontaneous or interview context, etc.
- Vocal effort; a criminal could speak loudly or whispering, etc.
- Ageing; suspect could be interviewed long time after the crime act.
- Emotional state; The crime act does influence the criminal emotion state. Indeed, a criminal is often acting under stress, fear and other different emotion states. On the other hand, suspect voice is collected in another condition, generally from police interview.

3.3.1 Speaking rate and style

Speaking rate, expressed in phonemes or syllables per second, is considered as an important factor of intra-speaker variability. When speaking fast for example, the acoustic realization of phones and their timing are strongly affected due in part to the limitations of the articulatory machinery. The effect of speaking rate was addressed earlier in many research works (Gay, 1978, 1968). In speaker comparison, if a small difference in speaking rate between the two voice recordings will not be a problem, a substantial mismatch may lead to serious performance degradation as observed in (Rozi et al., 2016). Authors showed that when there is a mismatch on speaking rates (fast and slow) between the two speech recordings performance decreases drastically. Table 3.7 illustrates the gap

of performance between matched and mismatched conditions for "normal" and "slow" speech.

Table 3.7: *Equal error rate (EER%) of speaker verification system with training and testing speech in varied speaking style (Rozi et al., 2016). "**Normal**" refers to normal speech while "**Slow**" refers to slow speech.*

Condition	EER
Normal-Normal	4.00%
Normal-Slow	13.09%

Changes in speaking rate could be considered as a part of speaking style. Speaking style variations have a huge effect on speaker recognition, especially, when two voice recordings corresponding to two different speaking styles are compared (Park et al., 2016; Karlsson et al., 2000; Shriberg et al., 2008; Wang et Xu, 2011). In (Shriberg et al., 2008), authors have performed a study using , "read[7]", "Spontaneous", "conversational", "interview", as variation of speaking style. Table 3.8 reports the system performance when enrolment and test speech samples are under various speaking styles.

Table 3.8: *System performance reported in term of Equal error rate (EER%) when "Enrolment" and "testing" speech are under varied speaking style. "Green" refers to the matched conditions (Shriberg et al., 2008).*

		Test from		
		Inter	**Conv**	**Read**
Enrolment from	**Inter**	8.33	11.88	10.83
	Conv	7.50	8.33	8.33
	Read	12.56	16.79	10.00

Authors showed that mismatched conditions involving "read" and "conversational" speech show higher error rates than matched conditions.

Many studies addressed the speaking style mismatch impact on speaker recognition (Grimaldi et Cummins, 2009; Wang et Xu, 2011) and agreed that speaking style mismatch results in performance degradation in speaker recognition application.

3.3.2 Vocal effort

Variation in vocal effort represents one of the most challenging problems in speaker recognition. Changes in speaker vocal effort result in a fundamental change in speech production (Zhang et Hansen, 2007; Shriberg et al., 2008; Jessen et al., 2007). In (Zhang et Hansen, 2007), Zhang and Hansen studied the impact of five speech modes ("whispered", "soft", "neutral", "loud" and "shouted") on speaker recognition accuracy. In

[7] "read" is more than a speaking style as it implies different brain mechanisms than conversational speech.

this study, "*the whispered speech is defined as the lowest vocal mode of speech with limited vocal cord vibration. On the other hand, shouted speech is referred to as the highest vocal mode of speech, which requires the most dramatic change in vocal excitation*". Authors showed that a mismatch in speech mode between two voice recordings can seriously impact speaker recognition performance: an average reduction of accuracy from 97.26% in matched to 54.02% in mismatched modes. More details about matched and mismatched condition are summarized in Table 3.9.

Table 3.9: *Accuracy rate (%) of the speaker recognition system for Enrolment and test data under five speech modes. Source (Zhang et Hansen, 2007).*

		Test				
		Whispered	**Soft**	**Neutral**	**Loud**	**Shouted**
Enrolment	**Whispered**	**94.6**	33.3	30.4	23.3	17.9
	Soft	57.9	**97.5**	86.3	61.7	41.7
	Neutral	46.7	86.7	**98.8**	86.3	56.3
	Loud	39.2	66.7	92.1	**98.3**	64.2
	Shouted	27.1	40.4	53.8	68.3	**97.1**

3.3.3 Emotional state

Emotion is an intrinsic nature of human beings. It is recognized that a speaker mood change, for example, has a considerable impact on his speech. Compared with other intra-speaker variations such as the speaking rate, emotions tend to cause more substantial variations on speech properties, such as harmonic forms, formant structures and the entire temporal-spectral patterns. Hence, emotion directly affects the basis of speaker voice comparison process.

Table 3.10: *Equal error rate (EER%) of speaker verification system with enrolment and testing speech in varied emotions. Source (Wu et al., 2006).*

		Test				
		Neutral	**Anger**	**Fear**	**Happiness**	**Sadness**
Enrolment	**Neutral**	4.48	17.93	18.62	17.24	12.59
	Anger	17.76	17.24	20.17	17.24	21.38
	Fear	20.52	18.28	17.59	15.00	22.93
	Happiness	14.48	16.55	12.07	11.21	19.83
	Sadness	4.48	19.83	16.55	19.66	6.90

Several research works have been conducted to address the emotion variations. They can be divided into two main categories:

- First category; Research works are dedicated to the study of various emotion-related acoustic factors such as prosody and voice quality (Zetterholm, 1998), pitch (Wu et al., 2005; Pereira et Watson, 1998), duration and sound intensity (Pereira et Watson, 1998).

- Second category; Research studies tends to show the impact of mismatched emotions, between both voice recordings to be compared, on the comparison accuracy and propose several emotion-compensation methods to deal with this mismatch (Wu et al., 2006; Bao et al., 2007; Scherer et al., 2000). In (Wu et al., 2006), the authors had studied the impact of speaker emotional state on voice comparison accuracy. This study shows that comparing speech utterances with various emotions increases significantly the error rates. Indeed, an average increase of error rate from 11.48% in matched to 17.15% in mismatched emotion is observed as shown in Table 3.10. According to (Wu et al., 2006), these results could be explained by two reasons: (i) mismatched emotions between the two voice comparison to be compared, (ii) the articulating styles of certain emotions create intense intra-speaker vocal variability.

3.3.4 Non-contemporaneous speech and ageing effect

In non-contemporaneous speech there is a time delay between the date of the recording of the questioned speaker and the one of the suspected speaker. If the time delay is at the order of several years, we speak about ageing effect whereas scenarios in which "*the time delay is at the order of just a few days, weeks or months are commonly referred to as session-mismatch*" (Drygajlo et al., 2016). In the following paragraphs, we briefly review the impact of short-term non-contemporaneous speech and ageing on speaker recognition performance.

Short-term non-contemporaneous speech

Short-term non-contemporaneousness is common in almost all FSR caseworks (Drygajlo et al., 2016) as "*logically, no evidential and suspect recordings are made contemporaneously*" (Rose, 2003). In (Campbell et al., 2009), authors showed a striking non-contemporaneous speech effect detected by NIST after the SRE 2005 evaluation: performance decreased significantly when only a few weeks separated the two recordings of a voice comparison trial. Figure 3.3 shows the impact of the elapsed time between recording the enrolment speech and the test speech[8].

The *EER* moves from less than 5% when only five days or fewer separate the two voice recordings to more than 9% when 30 days or more separate the two speech recordings.

Ageing effect

Other than speaking rate and emotion, age is also a source of intrinsic variation. Alongside regular variability, the process of ageing leads to change the voice over the long-term. Noticeable changes to the voice appear mainly in three periods of life: childhood,

[8]The figure corresponds to the det curve explained in Chapter 5, Subsection 5.4

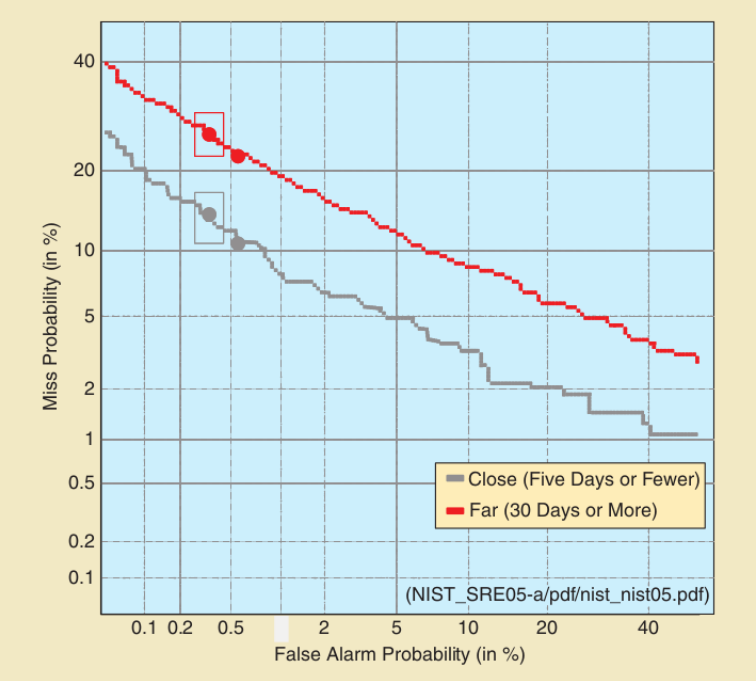

Figure 3.3: *Effect of time between enrollment and test recordings, NIST-SRE '05. source (Campbell et al., 2009).*

adolescence and old age (Markova et al., 2016; Reubold et al., 2010). Human articulatory system (including velum, pharynx, larynx, vocal fold, etc.) is affected by ageing:

- Vocal tract and related organs are subject to gradual deformation (Lindblom et Sundberg, 1971; Linville et Rens, 2001).
- Muscle strength of the tongue declines with ageing (Rother et al., 2002). This has a noticeable effect on the production of many phonemes which depend on tongue hump position and height. The effect of ageing on speech features and phoneme is well addressed in (Das et al., 2013).
- Pharynx cavity and larynx tube are also affected by ageing effect (Xue et Hao, 2003).

These physiological changes tend to cause a significant variation on speech parameters. The fundamental frequency (F0), formant frequencies (F1, F2, F3, F4,...) (Reubold et al., 2010), jitter, shimmer, voice onset time, are some of the speech properties which are the most affected by ageing. Moroever, several degradation of voice quality appear with ageing. Indeed, the speaking rate reduces with ageing due to cognitive decline (Ulatowska, 1985).

In a typical forensic case, age is considered as a major cause of variability as a piece of evidence can be years or even decades old. The impact of ageing in speaker comparison is well addressed in several research works and has been well documented (Kelly et al., 2012; Kelly et Hansen, 2015; Kelly et Harte, 2011; Kelly et al., 2014; Matveev, 2013).

In (Kelly et al., 2014), authors observed that the voice comparison accuracy degrades progressively as the absolute age difference between voice samples increases. The error rate is multiplied by 7 when using samples where ageing difference is about 0-1 years compared with experiment using samples of ageing difference is 51-60 years as illustrated in Figure 3.4.

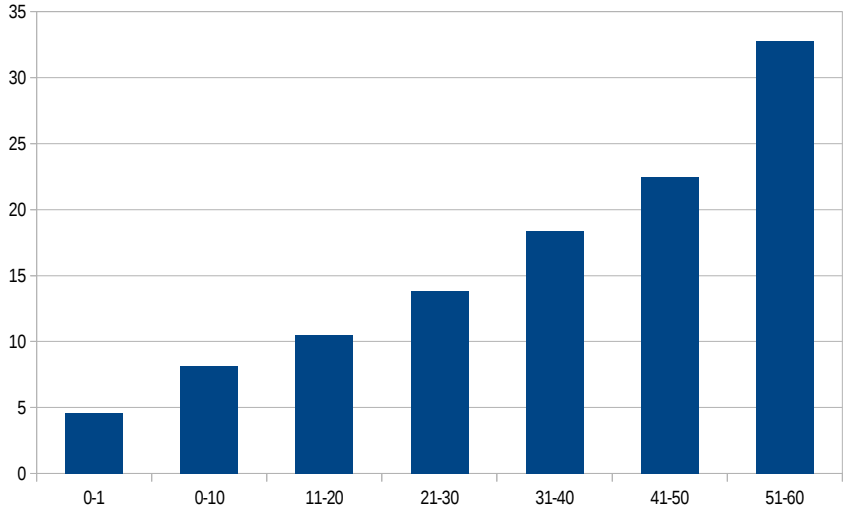

Figure 3.4: *Accuracy variations depending on absolute age difference range (years) adapted from (Kelly et al., 2014).*

In this study, authors suggested that long-term ageing variability is distinct from everyday intersession variability, and therefore an age compensation approach is needed to deal with age-related variations.

3.3.5 Speaker factor

After evoking the intrinsic variation factors, it is therefore necessary to come to "*speaker factor*" question. "*Speaker factor*" is a denomination that reflects mainly intra-speaker variability and differences between the speakers according to this variability. "Speaker factor" was earlier addressed in the famous article (Doddington et al., 1998) where authors characterized the different speakers in terms of their error tendencies using an automatic speaker recognition and proposed 4 "speaker profiles", "sheep", "goats", "lambs" and "wolves":

- Speakers for which the system has a "normal" behaviour are denoted as "sheep".
- Speakers who cause a proportionately high number of false rejection errors are called "goats".
- Speakers who tend to cause false acceptance errors because they are accepting too much impostors are "lambs".

- Speakers who tend to cause false acceptance errors as the impostor speaker are called "wolves".

This concept is known as "Doddington zoo menagerie".

Revisiting the avenue opened by Doddington, (Kahn et al., 2010) demonstrated that this notion of "speaker profile" is in fact a simplified view of a more general problem: speaker recognition systems model speech files and not, or not only, the speech or the voice of a given speaker. In order to demonstrate this assumption, the authors built an experimental setup using the NIST 2008 evaluation database. The experiment is composed of voice comparison trials, represented by a couple of speech signals (X^i, Y_k). The right value, Y_k, is fixed and simply one of the K speech extracts from recording set Y. The left value, X^i, is the in-interest factor. X^i is the recording of speaker S^i, taken from a subset of recordings X^i_j, pronounced by S^i. For each S^i speaker, voice comparison trials (X^i, Y_k) with k varying from 1 to K are carried out using each available speech signal X^i_j, j varying from 1 to J. For each S^i speaker, the speech extract which allowed the speaker recognition system to make the least errors is labelled with a "best" label. Respectively, the speech extract showing the maximum number of errors is labelled with a "worst" label. Figure 3.5 plots the performance of the system when the recordings selected for the X^i parts of the voice comparisons are the "best" ones or the "worst" ones.

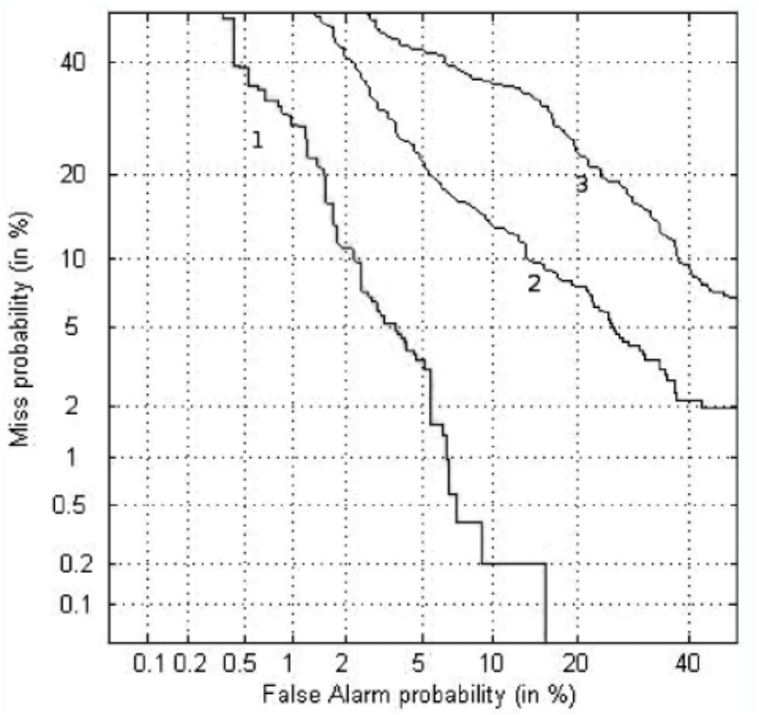

Figure 3.5: *DET performance curves of a speaker recognition system using (1) the "best" speech extracts, (3) the "worst" speech extracts and (2) randomly selected speech extracts. (Kahn et al., 2010).*

The *EER* moves from less than 5% for the recordings with the "best" labels to

more than 20% for the recordings with the "worst" labels. It is important to emphasize that the only difference between the "best" condition and the "worst" condition is the speech sample selected to represent a given speaker. Clearly, the speaker recognition system gives a great importance to the speech extract itself. In forensic voice comparison, it means that the choice of the speech material used as comparison has an important effect on the voice comparison result itself.

Conclusion

In this chapter, we discussed the complexity of forensic voice comparison process. We highlighted this complexity at three levels: 1) Forensic and challenging conditions of the piece of evidence, 2) Mismatched conditions between the questioned and the suspect's recordings. Indeed, unlike other forms of forensic evidence such as DNA, speech is inherently variable and depends on many factor of variabilities. 3) Within-speaker variability. We showed that speech variability's factors, either intrinsic or extrinsic, make the FVC a daunting task. At the end of this chapter,the reliability of FVC appears as questionable point.

Chapter 4

Voice and "biometrics"

Contents

Introduction

In practice, it is possible to verify an individual identity through three different methods: something that he knows, like a password or a code PIN, something that he has, like a key or a badge, or something that he is, biometric characteristics. Using biometric methods to establish the identity of an individual has two main advantages over the other ones:

- Simple and secure because one does not have to remember the password or to be afraid of losing it.
- Reliable since it is based on characteristics linked to the individual itself.

In this chapter, we highlight the importance of biometric technology in forensic identification. First, we provide an overview of the biometric technologies. Second, we show how biometric technology constitutes an advantage for forensic identification sciences. Third, we address the question of voice and its "biometric aspect".

4.1 Overview of biometrics

4.1.1 Definitions and notions

The term Biometry is composed of two words -Bio (Greek word for Life) and Metrics (Measurements). Biometry is *"the analysis of biological data using mathematical and statistical methods*[1]". Biometry should not be confused with biometrics: The latter is more recent and corresponds to the identification of a person using biometry. According to (Jain et al., 2007), biometrics is the science of establishing the identity of an individual based on his physical or behavioural attributes. In (Kay, 2005), Kay defines biometrics as *''the verification of a person's identity by means of a physical trait or behavioural characteristic that can not easily be changed, such as a fingerprint*". In brief, biometrics is related to measure intrinsically the individual's bodily characteristics for identity inference purpose. Biometric properties of a person **are assumed** not to change. There is a strong link between a person and its biometric traits because biometric traits are inherent to an individual.

Researchers have distinguished between physical and behavioural attributes:

- Physical attributes such as fingerprints, color of iris, hand geometry, etc.
- Behavioural attributes such as voice, handwriting, signature, gait, or the way of typing keys of computer keyboard, etc.

Figure 4.1 presents an example of potential biometric traits.

4.1.2 How to choose a biometric trait?

Any physical and/or behavioural characteristic of an individual can be used as a biometric characteristic as long as it satisfies some requirements (Maltoni et al., 2009; Marzotti et Nardini, 2006; Jain et al., 2004). Jain et al (Maltoni et al., 2009; Jain et al., 2004) have identified seven factors that determine the suitability of a biometric characteristic:

1. Universality: each person should have a biometric characteristic.
2. Uniqueness (Distinctiveness is often used to avoid the "uniqueness" term): Each human being is unique in terms of characteristics, which make him or her different from all others.
3. Permanence: The biometric characteristic do not change over time.

[1] https://www.thefreedictionary.com/

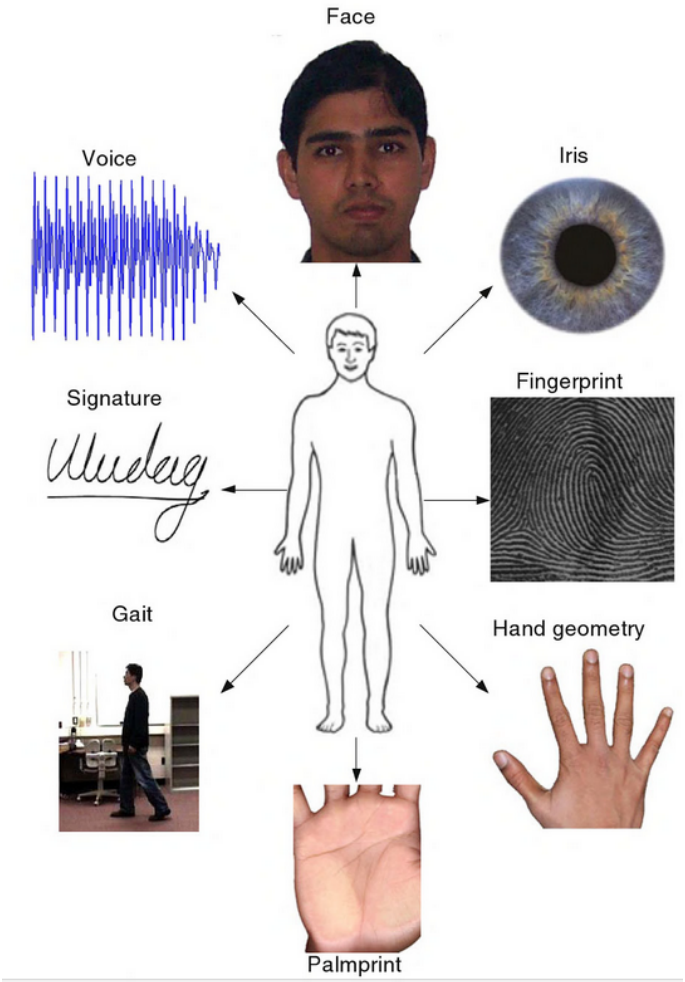

Figure 4.1: *Example of biometric traits that can be used for authenticating an individual (Jain et al., 2011).*

4. Collectability: The biometric characteristic can be measured in a quantitative way.

   In real life applications, three additional factors should also be considered such as:

5. Performance, which refers to the achievable recognition accuracy and speed, the resources required to achieve the desired recognition accuracy and speed.

6. Acceptability, in which extent each person will accept a particular biometric technology.

7. Circumvention, reflects the extent in which one could easily fool the system.

If all these factors play a major role in determining the "interest" of a biometric characteristic, the distinctiveness and permanence factors constitute the fundamental premise of biometric recognition. Indeed, a biometric trait should be permanent and

should retain its discriminatory power over the lifetime of an individual.

It is difficult to find a biometric characteristic that effectively satisfies all these requirements and therefore could fit with all biometric applications. Every biometric trait has its strengths and weaknesses. The relevance of a specific biometric characteristic to an application is established depending upon the nature and requirements of the application, and the properties of the biometric characteristic.

According to (Maltoni et al., 2009): "*A practical biometric system should have acceptable recognition accuracy and speed with reasonable resource requirements, harmless to the users, accepted by the intended population, and sufficiently robust to various fraudulent methods.*"

4.1.3 A Schematic view of a biometric system

Despite the variety of the biometric traits, they all share the same biometric processing chain. It consists of 3 phases as shown in Figure 4.2: feature extraction, feature modelling and finally comparison called also scoring (Li et Jain, 2015; Drygajlo, 2012).

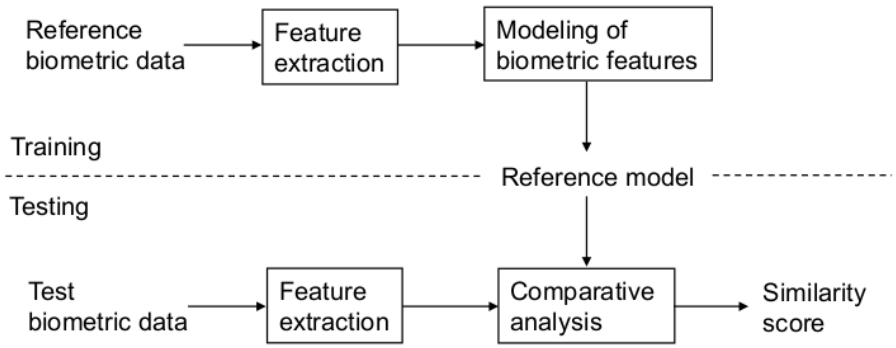

Figure 4.2: *Generic processing chain for biometric recognition (Drygajlo, 2012).*

4.1.4 Biometric tasks

Based on a particular biometric trait for example a person's voice, two tasks could be performed: Verification and identification.

- Verification: Generally, an unknown speaker claims an identity, and the task is to verify whether his claim is correct. This essentially comes down to comparing two speech samples and deciding if they are spoken by the same speakers. In simple term, verification is a *one-to-one match*, (1 against 1).
- Identification: The identification task is different from the verification one. Here, the task is to identify an unknown speaker from a set of known speakers. In other words, the goal is to find the speaker who sounds closest to the speech stemming from an unknown speaker. In simple term, identification is a *one-to-many* matching, (1 against N, N is the number of known speakers in the identification set).

When all speakers within a given set are known, it is called a closed or in-set scenario. Alternatively, if the potential input test subject could also be from outside the predefined known speaker group, this becomes an open-set scenario.

Throughout this thesis, the generic term recognition will be used when we do not make a distinction between the verification and identification functionalities.

These two tasks could be of utmost importance in forensic investigation and evaluation.

4.2 Biometrics and forensic sciences

If we go back to the definition of forensic science given in chapter 2, ("*Forensic science refers to the applications of scientific principles and technical methods to an investigation in relation to a crime in order to determine the identity of its perpetrator*"), it is thus logical that biometric science constitutes a fertile ground for the use of distinctive physiological or behavioural traits to identify protagonists involved in a crime. Nowadays, biometric traits have the possibility to become an irreplaceable part of any identification task. Indeed, digital evidence (linked generally to behavioural traits such as facial imaging, voice, etc.) as well as physical one (physiological traits) are increasingly getting easier to collect from the crime scene. By exploiting biometric technologies, capturing the identity of a perpetrator from a biometric trait left at the crime scene will be possible (in the sense of more reliability). In the same context, Meuwly and al. (Tistarelli et al., 2014) state: "*there is an almost endless list of open problems and processes in Forensics which may benefit from the introduction of tailored Biometric technologies. Joining the two disciplines, on a proper scientific ground, may only result in the success for both fields, as well as a tangible benefit for the society*".

It is important to note that using biometric traits to identify people is not new and dates back to the 19^{th} century! Forensics was one of the earliest application of biometrics.

4.2.1 Historical review

Biometrics has historically found its natural mate in Forensics. The use of biometric characteristics in order to identify an individual in forensic science dates back to the 1870s, more precisely to Alphonse Bertillon measurement system known as "*anthropometry*" or also popularly known as "*Bertillonage*" (Rhodes, 1956). Bertillon was the first who tried to find a scientifically way to identify criminals by taking body measurements such as skull diameter, arm, foot length, etc. These measures were used in the USA to identify prisoners until the 1920s.

Soon after Bertillon's system, in the 1880s, an identification technique taking advantage of fingerprint was proposed following the work by Henry Faulds, William Her-

schel and Sir Francis Galton (Kirk, 1963). Fingerprint had replaced Bertillon's system due to their ease of taking, classification and retrieval.

At the beginning of 1960s, the development of digital signal processing techniques allowed to work immediately in automating human identification. Speaker (Pruzansky, 1963; Li et al., 1966; Luck, 1969; Atal, 1976; Rosenberg, 1976) and fingerprint recognition (Trauring, 1963) were among the first systems to be explored. The need to apply these emerging technologies to high-security access control, financial transactions and so forth, was firstly enunciated in the early 1960s (Trauring, 1963).

The 1970s saw development and deployment of hand geometry systems (Zunkel, 1996). Retinal verification systems (Crane et Ostrem, 1983; Hill, 1996) came in the 1980s. The first behavioural biometric used in forensic science is handwriting (Alfred Dreyfus case (Taroni et al., 1998)) and signature whose usage goes back a few centuries (Huber et Headrick, 1999).

Iris recognition systems were developed in the 1990s.

4.2.2 Biometric Evidence for Forensic Evaluation and Investigation

The ongoing paradigm shift in forensic sciences (Saks et Koehler, 2005; National Research Council, 2009; Morrison, 2009a) needs biometric methods in order to identify individuals based on their physiological and behavioural characteristics, as a common practice (Drygajlo, 2012; Jain et al., 2005; Dessimoz et Champod, 2008; Jain et al., 2007). According to (Jain et al., 2016), one of the main reasons of the intersection (or what Champod called "linkage") between forensic sciences and biometrics is the development of evidence based on DNA profiles which have been put in place for around two decades. The DNA analysis deals with the comparison of the quantifiable properties of samples of known and questioned origin based on biological material (i.e Nuclear DNA contains genetic instructions to encode the different **biological** functions: It is therefore a direct reading of body traces.). DNA has become the new "*golden standard*" in forensic science. Moreover, the evolving requirements for the admissibility of forensic evidence following the Daubert rules constitutes another driving force for this linkage. In (Jain et al., 2016), Jain et al. state: "*[...] bolstering the scientific basis for biometric methods used in forensic investigations. In particular, it will be the first step in assuaging criticism levelled by the 2009 National Academy of Sciences' report, Strengthening Forensic Science in the United States: A Path Forward, which concluded that claims about the evidential value of forensic data are not supported by rigorous scientific study.*"

Biometrics can be used in forensic in two different ways:

- as a tool to assist an investigation in order to identify potential suspects.
- as a tool for evidence evaluation in a court of law.

In brief, "*biometric systems in forensic science today aim at (1) filtering potential candidates (i.e suspects) and (2) putting forward candidates for further 1-to-1 verification by a forensic expert (Dessimoz et Champod, 2008)*".

It is worth to note that these cases have very different requirements. Indeed, the speed and the accuracy of the biometric process constitutes the key requirements for the first case. Whereas, the primary requirement in the second scenario is a convincing presentation of biometric evidence with strong scientific basis. This in turn involves obtaining a reliable estimate of the individuality of a biometric trait (Jain et al., 2016).

For more information, the readers could be referred to the recent survey for biometrics research carried out by Jain et al (Jain et al., 2016).

4.3 Voice for human identification

The development of mobile phone networks and more recently VoIP confirms that "*a person's voice is the most accessible biometric trait*" (Gupta et Chatterjee, 2012). To collect a speech sample is an easy task and does not need any extra acquisition device or transmission system. This fact gives voice an overwhelming advantage over other biometric traits. However, the voice is not only related with speaker characteristics, but also with many environmental and sociolinguistic variables, as voice is the result of an extremely complex process. In the next subsections, we list the main speaker specific information. Then, we describe briefly the main factors (speaker and other than speaker) which influence the speech production.

4.3.1 Speaker specific information

Different studies agree that speaker specific information is not equally distributed over the spectral domain and particularly depends on the phoneme distribution (Magrin-Chagnolleau et al., 1995; Besacier et al., 2000; Amino et al., 2006; Antal et Toderean, 2006). In the following paragraphs, we discuss briefly how speaker-specific information are distributed over the phonetic content and the spectral domain.

Phonetic content

Several earlier studies have analysed the speaker-discriminant properties of individual phonemes or phoneme classes (Wolf, 1972; Sambur, 1975; Eatock et Mason, 1994). The authors agreed that vowels and nasals provide the best discrimination between speakers. (Hofker, 1977) presents a ranking of 24 isolated German phonemes, which indicates nasals as providing the best ASpR performance, with the voiced alveolar fricative /z/ and the voiced uvular fricative /ʁ/ also performing fairly well. In (Kashyap, 1976), /s/, /t/ and /b/ are found to perform worse than vowels and nasals. (Wolf, 1972; Sambur, 1975) strongly promote the nasals and vowels as best performers. The influence of the phonemic content on ASpR performance was also evaluated in (Magrin-Chagnolleau et al., 1995) in which authors suggest that glides and liquids together, vowels -and more particularly nasal vowels- and nasal consonants contain more speaker-specific information than phonemically balanced speech utterances. According to (Amino et al., 2006,

2012; Antal et Toderean, 2006; Eatock et Mason, 1994), nasals and vowels are conveying speaker specific information and nasal vowels are more discriminant than oral vowels.

Spectral domain

Different studies (Besacier et al., 2000; Gallardo et al., 2014b) showed that some frequency sub-bands seem to be more relevant to characterize speakers than some others. Indeed, fricatives and nasals -known as speaker specific information and can even be more useful than vowels for speaker discrimination (Schindler et Draxler, 2013)- exhibit spectral peaks at high frequencies, ranged from 4 to 7 kHz depending on the particular phoneme (Jongman et al., 2000). Moreover, Nasals present a high speaker discrimination power in low and mid-high frequencies (Hyon et al., 2012) due to physiological characteristics of speaker. Other consonants contain speaker specific information in the upper part of the frequency spectrum, above 6kHz (Hyon et al., 2012). On the other hand, oral vowels show an important ability for speaker discrimination (Sambur, 1975). A recent research work (Gallardo et al., 2014a) shows that speaker specific-information conveyed in the band 4-8 kHz provides a performance similar to that obtained with the band 0-4 kHz. This result confirms that speaker specific information are distributed over the entire spectral domain with different extents.

4.3.2 Factors which influence speech production

Speech production is a complex process which is influenced by many factors at different levels. Indeed, a single speech sample represents several kinds of information, both linguistic and speaker specific attributes, encapsulated together. Speaker attributes certainly include cues linked to the physical properties of the speaker's vocal organs. In addition, many non-physiological speaker specific characteristics leave their imprint on the acoustic signal such as information about the speaker's emotional state. Moreover, speech is constrained by diastratic, diatopic, and diaphasic variations. These factors include without limitation:

- Sociolinguistic factors: concern the difference of language between social groups (age, sex, level of education, status, ethnicity, dialectal differences...) (Labov, 1972).
- Geolinguistics is concerned with the spatial distribution of linguistic phenomena.
- Pragmatics studies how speech production depends not only on phonology, lexicon and syntax but also on the inferred intent of the speaker and the context of the utterance (Austin, 1970). Speech also conveys the emotional and psychological states of the speaker (Scherer, 1986).
- The physiological aspects are mainly related to "the physical shape of the vocal tract" including the pharynx, larynx, nasal cavities, etc and depends also on the mouth, tongue and lungs. The different structures of all those elements in any human being are "invariant" (Campbell, 1997).

- The behavioural aspects are related to all different aspects like movements and manners that may influence the speech. They are changing over the time (Campbell, 1997).

It is thus logical that when hearing speech, one imagines a speaker, and assigns sex, age, geographical and social origin, and even some personality traits. In the presence of all these factors of variability that are difficult to deal with, a speech sample will therefore embed (encode) a "*degraded version*" of speaker specific information.

4.4 Is voice a biometric?

As seen previously, the term biometrics usually refers "*to identifying an individual based on his or her distinguishing characteristics*" (Bolle et al., 2003). Particularly, voice "biometrics" aim to identify idiosyncratic features in the speech signal, produced at a given time and in specific conditions, for person recognition purpose. As mentioned previously, the uniqueness and permanence factors play a major role in determining the "accuracy" of a biometric trait and thereby the biometric system accuracy. The analysis of these two properties for voice reveals a core challenge:

- A speech signal is not a direct reading of body traces like fingerprints or DNA and includes a high number of variability factors related to the condition of the recording or even to the speaker itself (These factor of variabilities are detailed in Chapter 3). Using voice for speaker recognition relies heavily on behavioural variables: it looks more at the way of speaking than to the physical properties of the speaker's body. In forensic science, the principle of the extension of the properties of the source to the trace is assumed without question. This principle is particularly problematic for voice (Doddington, 1985; Meuwly, 2006).
- Distinctiveness (i.e uniqueness) of speaker's voice does not reach a sufficient level of "verisimilitude" (Meuwly, 2006). The uniqueness of properties of the human voice is questionable, as no particular set of speaker idiosyncratic property is present or detected in speech (Meuwly, 2006). Indeed, the biometric task is essentially a statistical-probabilistic process based on speaker's models. The information used to learn a speaker model depends on the speech organs, the language used and many other variables as detailed in subsection 4.3.2. In brief, the uniqueness of the speaker attributes in the speech is assumed without debate, both in forensic and commercial biometric application (Doddington, 1985). In this context, Meuwly (Meuwly, 2006) stated: "*the hypothesis that no two humans speak exactly alike is plausible, but to date no large-scale demonstration of the extent of idiosyncrasy in a homogeneous community of speakers has been adduced in support of this hypothesis (Nolan, 1991)*".

The considerations mentioned supra are true for the majority of behavioural biometric traces (face, voice, gait, handwriting,..). Several research studies confirmed the limitations of human beings and machines in their capacity for person individualisation using speaker recognition or face recognition or the two combined (Clifford, 1980;

Boves, 1998). This situation is much more complex in forensic science, where the piece of evidence is normally of bad conditions.

Conclusion

In this chapter, we highlighted the importance of biometric in forensic identification and we questioned the voice's biometric aspect. We showed that voice is different to other biometric traits such as DNA or fingerprint in two main points: Uniqueness and permanence. These characteristics do not mean that voice can not be used or are not suitable for speaker individualisation in forensic science, but it clearly lays down limits on the reliability that can be expected.

Chapter 5

Automatic speaker recognition for forensic voice comparison

«Judges and lawyers usually react to science with all the enthusiasm of a child about to get a tetanus shot. They know it is painful and believe it is necessary, but have not the foggiest idea how or why it works.»

Black et al.: "Science and the Law After Daubert" Texas Law Review 1994

Contents

Introduction

An automatic speaker recognition system could be seen as a sequential system. It is composed of three basic components as shown in Figure 5.1: First, acoustic feature parameters are extracted from the enrolment audio recording. These features are supposed to capture, among others, the idiosyncratic characteristics of a speaker. The next step consists to build/train a model that summarizes speaker-specific information using features obtained from the enrolment sample. This step is called "modelling". Finally, for an unknown test segment, the same features are extracted, and they are compared against the model of the enrolment/claimed speaker. This comparison provides a score, a scalar value, which indicates whether both voice recordings are pronounced by the same speaker (i.e the same source) or not. The principle, then, is easy: if this score is higher (or lower) than a fixed threshold then the system accepts (or rejects) the speaker.

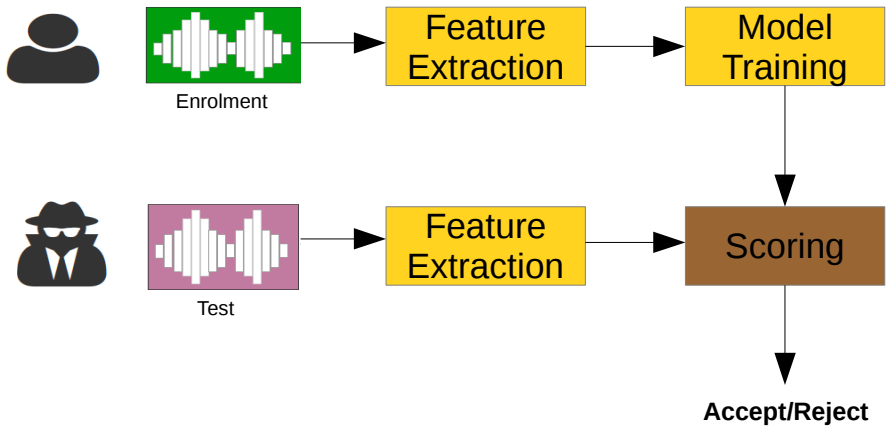

Figure 5.1: *Schematic view of main component of automatic speaker recognition.*

5.1 Feature extraction and preprocessing

Speech parametrization consists in transforming the speech signal into a set of feature vectors in which speaker idiosyncratic properties are emphasized. This transformation aims to obtain a new representation more compact and more suitable for speaker modelling. Several speech parametrization appear in the speaker recognition state-of-the-art. In the following, we review some of them.

Short-term spectral parameters

The most used parametrizations relies on a cepstral representation of speech. This process consist of extracting feature vectors on a local part of the signal by the application of a temporal sliding window. This window is applied from the beginning of the signal,

then moved further and so on, until the end of the signal is reached. Each application of the window to a portion of the speech signal provides a feature vector as presented in Figure 5.2. The length of the window is generally fixed between 20 milliseconds and 30 milliseconds. These values represent the average duration which ensure the pseudo-stationarity assumption to be true. On the other hand, the step value is fixed in order to have an overlap between two consecutive windows, 10 milliseconds is commonly used.

In order to capture the temporal variation, these feature vectors are usually appended with their time derivative coefficients (deltas Δ and double-deltas Δ^2). Log-energy is also provided in the final feature vector. For more details, interested readers are referred to (Davis et Mermelstein, 1980)

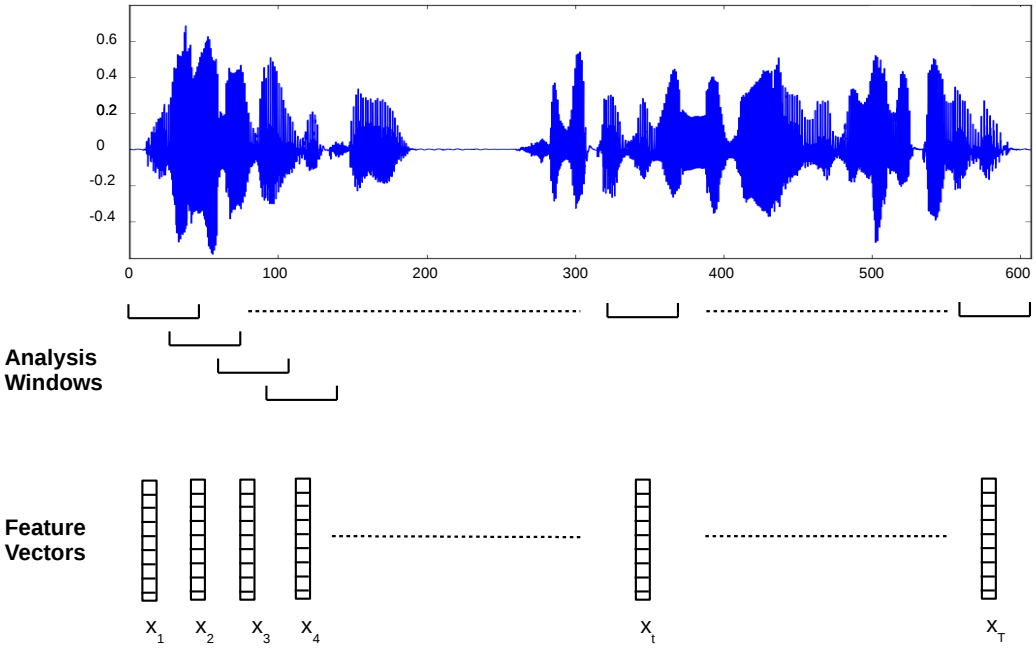

Figure 5.2: *Short-term analysis and parametrization of a speech signal.*

The output of this stage is a sequence of feature vectors representing a speech segment S={ x_1,...,x_T }, where x_t is a feature vector resulted of the application of the window, t (t $\in$[1, 2,...,T]).

MFCCs are considered as a descriptor of the short-term power spectrum. Linear frequency cepstral coefficients (LFCCs), Linear predictive cepstral coefficient (LPCC) or perceptual linear prediction (PLP) features are also used as a descriptor of the short-term power spectrum.

In this Thesis, short-term spectral features, LFCCs specifically, have been used exclusively.

Formant parameters

One of the main analysis performed by forensic speaker comparison experts has traditionally been formant analysis (Nolan, 1980; LaRiviere, 1975; Gold et French, 2011; Rose, 2003, 2006). Formants parameters are defined as the resonant frequencies of the vocal tract. Formants are connected to the anatomy and physiology structure of the supralaryngeal articulators (Stevens, 1971). Multiple formants could be extracted from the speech signal. The first three formant are the most used in speaker recognition.

Prosodic and high-level parameters

Prosodic and high-level features have been studied in speaker verification context (Shriberg et al., 2005; Siddiq et al., 2012; Dehak et al., 2007; Kockmann et al., 2011).

Unlike short term features, prosodic features are extracted from longer segments (syllables or word-like units) in the range of 100ms. These features characterise information such as rhythm, energy, pitch and speaking rate (Reynolds et al., 2003a; Shriberg et al., 2005). These features reflect information related to both physical (gender, age...) and behavioural characteristics such as speaking style.

High-level features need much more speech material in order to be extracted. These features capture phonetic-level information, such as speaker's accent, and word-level information such as semantics and idiolect (Reynolds et al., 2003a; Shriberg, 2007).

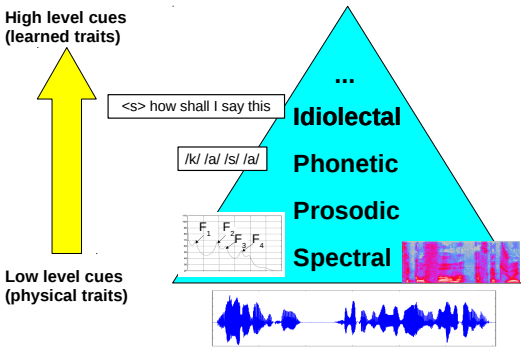

Figure 5.3: *Identity level in the speech signal. Adapted from (Reynolds et al., 2003b)*

In the state-of-the-art, features derived from the magnitude spectrum of the speech signal are the most successful. These features are known to convey information on low-level cues that are related to the acoustic aspects of speech as well as the speaker glottal and vocal-tract characteristics (i.e physical characteristics) while "high-level" characteristics convey behavioural information (Figure 5.3).

Ideal feature for speaker discrimination

Whatever the used feature parameters, an ideal feature would satisfy some requirement to ensure a high speaker discrimination accuracy. Ideal features would (Wolf, 1972; Nolan, 1980):

1. have large between-speaker variability and small within-speaker variability.
2. be robust against noise and distortion.
3. occur frequently and naturally in speech.
4. be easy to measure from speech signal.
5. be difficult to mimic.
6. not be affected by the speaker's health or long-term variations in voice.
7. be manageable in size, avoiding the so-called curse of dimensionality (Jain et al., 2000).

In practice, it is difficult that a single feature would satisfy all these requirements. As mentioned in (Kinnunen et Li, 2010), there is no globally "best" feature, and a trade-off must be made between speaker discrimination, robustness and practicality.

Voice Activity Detection

Voice Activity Detection (VAD) aims to identify speech portions from the signal while portions corresponding to silence or background noise are removed or not used in the recognition process.

Detecting speech segments becomes challenging especially when degraded acoustic conditions are considered. Various VAD techniques are used such as: periodicity measure (Tucker, 1992), zero-crossing rate (Junqua et al., 1991), pitch (Li et al., 2002), spectrum analysis (Marzinzik et Kollmeier, 2002), higher order statistics in the LPC residual domain (Nemer et al., 2001), combinations of different features (Tanyer et Ozer, 2000), but the most employed VAD approach is energy-based methods following this principle: Frames with high energy correspond to speech, frame with low energy correspond to silence and background noise. This technique poses a real problem: Some phonemes especially consonants which are low-energy, will not be considered as speech and therefore will be discarded. Thus, a part of speaker specific information will be lost and not used for the speaker recognition process.

Feature normalisation

Once features have been extracted and speech portions have been identified, different kind of normalisation could be applied on the feature vectors. The aims of these feature-level normalisation approaches is mainly to attenuate the channel and noise effects.

One of the commonly used techniques in automatic speaker recognition systems is

cepstral mean subtraction (CMS) (Furui, 1981). This operation consists in centring the data: the cepstral mean vector is subtracted from each cepstral vector. Also, variance normalisation could be applied. This normalisation consists in normalising to one the variance. The normalisation involving the CMS and the variance normalisation is the most used and it is denoted as *Cepstral Mean and Variance Normalization* (CMVN).

Other feature-level normalization approaches have been investigated in the state of-the-art including RASTA filtering (Hermansky et Morgan, 1994), feature warping (Pelecanos et Sridharan, 2001) and feature mapping (Reynolds, 2003).

In this thesis, cepstral mean and variance normalization (CMVN) is adopted.

5.2 Speaker modelling approaches

Speaker modelling refers to the process of describing the feature parameters in an effective way in order to generate a "speaker representation". Indeed, this representation must be general enough to describe the typical feature space of a speaker and also discriminative enough to distinguish between the feature spaces of different speakers. In the state-of-the-art different modelling approaches have been employed. The approaches are divided in two main categories: template models and stochastic models also known by non-parametric and parametric models (Campbell, 1997). *Vector Quantisation* (VQ) approach developed in the 1980s (Soong et al., 1985) and *Dynamic Time Warping* (DTW) (Furui, 1981) are good examples of template models. On the other hand, *Gaussian Mixture Model* (GMM) (Reynolds et al., 2000; Reynolds, 1995) and *Hidden Markov Model* (HMM) are the most popular stochastic models. The state-of-the-art has advanced significantly since the early days of GMM which have been the most dominant modelling approach in speaker recognition. Other approaches appeared such as factor analysis methods (Kenny et Dumouchel, 2004; Kenny et al., 2007) and I-vector approach (Dehak et al., 2011) which became the state-of-the-art in speaker recognition. Recently, there have been some attempts to use "*deep neural networks*" (DNN) approaches for speaker modelling (Variani et al., 2014; Rouvier et al., 2015; Matějka et al., 2016).

In this subsection, we review some of the popular models used in speaker recognition: We describe briefly the evolution from the GMM-UBM to I-vector approach.

5.2.1 GMM-UBM approach

In order to better understand the GMM-UBM approach, we first present the likelihood-ratio test for which it was intended. Then, we briefly describe the Gaussian Mixture Model and the UBM. Finally, we present the adaptation of a speaker model.

Likelihood Ratio Test

Given a speech utterance, X, and a speaker identity, S, the task of speaker verification can be stated as a test between two competing hypotheses:

- H_0 : X comes from speaker S.
- H_1 : X does not come from speaker S.

The similarity score in this approach corresponds to a likelihood ratio (LR) given by:

$$LR = \frac{p(X|H_0)}{p(X|H_1)} \begin{cases} \geq \tau, & H_0 \text{ is retained.} \\ < \tau, & H_0 \text{ is rejected.} \end{cases} \tag{5.1}$$

where $p(X|H_i)$, $i = 0, 1$, is the probability density function for the hypothesis H_i evaluated for the observed speech sample X, also referred to as the likelihood of the hypothesis H_i given the speech segment. The decision threshold for accepting or rejecting H_0 is τ.

In order to perform speaker verification, the two likelihoods, $p(X|H_0)$ and $p(X|H_1)$ should be estimated. In GMM-UBM approach a hypothesis is modelled by a GMM (Reynolds et al., 2000).

Gaussian Mixture Models (GMM)

A GMM is a weighted linear combination of a finite unimodal multivariate Gaussian densities, $p_i(x)$, each parametrized by a $D \times 1$ mean vector, μ_i and a $D \times D$ covariance matrix, Σ_i:

$$p_i(x) = \frac{1}{(2\pi)^{\frac{D}{2}}|\Sigma_i|^{\frac{1}{2}}} exp\{-\frac{1}{2}(x-\mu_i)^t)(\Sigma_i^{-1})(x-\mu_i)\} \tag{5.2}$$

Let $\lambda = \{w_i, \mu_i, \Sigma_i\}_{i=1}^{i=M}$ be a GMM of M Gaussians and x a D-dimensional feature vector. The mixture density used for the likelihood function is defined as:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i p_i(x) \tag{5.3}$$

Where w_i is the weight of the i^{th} Gaussian represented by the pdf p_i. The mixture weights, w_i, satisfy the constraint $\sum_{i=1}^{M} w_i = 1$.

The GMM parameters are tuned using *Expectation Maximization* (EM) algorithms (Bilmes et al., 1998) that iteratively increases the likelihood of the data given the model using the "*Maximum Likelihood*" criterion.

Let λ_S be a model of a target speaker S, and X a speech utterance of an unknown speaker. λ_S is used to represent the hypothesis H_0 in the likelihood ratio. Thus, the numerator of the LR (Equation 5.1) becomes $p(X|\lambda_S)$.

Universal Background Model (UBM)

Besides the claimed speaker model λ_S, an alternate speaker model $\lambda_{\overline{S}}$ is needed. This model is known as *Universal Background Model* (UBM). Concretely, the UBM is a large GMM learned to represent the "*speaker-independent distribution of features*" using a large amount of data from a pool of background speakers (Reynolds et al., 2000). The UBM should have a good representativity of the whole feature space in order to be able to describe a typical speaker-feature-space. $\lambda_{\overline{S}}$ is used to represent the alternative hypothesis in the likelihood ratio, $p(X|\lambda_{\overline{S}})$.

Adaptation of Speaker Model

In the GMM-UBM approach, the speaker model, λ_S, is derived by adapting the parameters of the UBM using the speaker's training speech and a form of Bayesian adaptation. The most used adaptation techniques is the "*Maximum A Posteriori*" (MAP) adaptation (Gauvain et Lee, 1994). Figure 5.4 presents an illustration of MAP adaptation procedure for a 2-dimensional feature and 4-mixture UBM case.

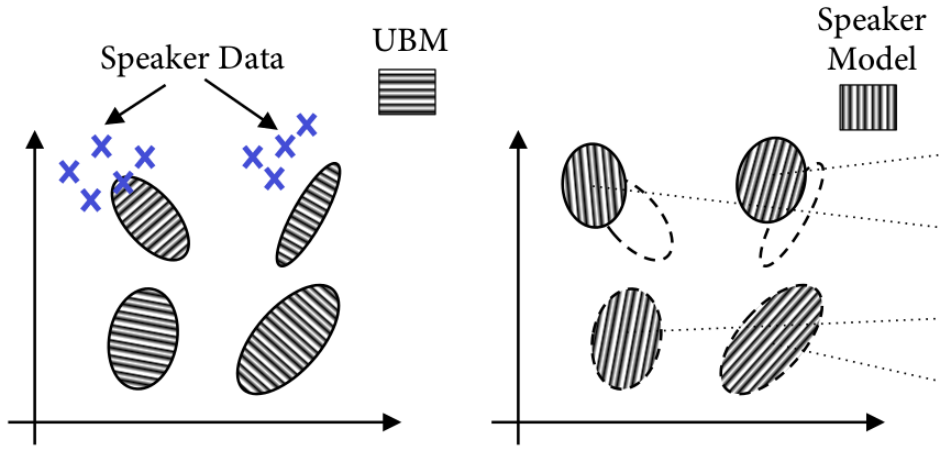

Figure 5.4: *MAP adaptation procedure is illustrated for a 4 mixtures UBM. Source (Reynolds et al., 2000).*

In the GMM-UBM paradigm, estimating a speaker's model based on an enrolment recording by a classical MAP produces a model adapted not only to the speaker but also to the enrolment recording conditions itself. Using this GMM to recognize the speaker under different recording conditions is therefore problematic.

5.2.2 The GMM supervectors

A noticeable advancement of the GMM-UBM framework has started with the introduction of supervector approaches. The term supervector was introduced for speaker recognition in (Campbell et al., 2006). The GMM supervector is a high-dimensional vector obtained by stacking the GMM mean vectors into a single big vector. GMM supervectors offer the advantage to obtain a fixed dimensional representation of a length variable speech recording. This representation has proven its usefulness for designing channel compensation methods. Several modelling techniques were proposed to operate on supervector space. *Factor Analysis* (FA) and *Support Vector Machines* (SVMs) with *Nuisance Attribute Projection* (NAP) are considered the most popular in the state-of-the-art. In this thesis, only the FA model is reviewed.

5.2.3 Factor analysis approach for GMM supervector

The variation in recording conditions of speech signals is one of the major problems that face automatic speaker recognition. Indeed, a speech recording is a mixture of different kind of information due to the speaker characteristics, as well as the environment and the channel variabilities. Thereby, to build a reliable speaker model, one should be based only on the speaker specific information. Factor Analysis (FA) technique proposed in (Kenny et Dumouchel, 2004) is one of the solutions that has proven its usefulness. This technique is based on the assumption that speech recordings depend on two main factors: the speaker itself and the channel information. In this context, the channel represents all information which can vary in the speech recording and which is not due to the speaker such as the recording conditions or the language. In consequence, it would be possible to put the inter-speaker variability into one sub-space and constrain the channel variability into a different low-dimensional sub-space. This assumes that these two variabilities are of additive nature (in the supervector space).

Many variants of factor analysis methods have been employed in the state-of-the-art such as Classical MAP adaptation, EigenVoice adaptation (Kenny et al., 2003), Eigen-Channel adaptation (Kenny et al., 2003) and Joint Factor Analysis (JFA).

JFA model assumes that both speaker and channel variability lie in a lower dimensional subspace of the GMM supervector space: the speaker space defined by the Eigenvoice matrix V and the channel space represented by the Eigenchannel matrix U. JFA decomposition for speaker s and session h, is given by Equation 5.4.

$$m_{s,h} = m_0 + Ux_h + Vy_s + Dz_{s,h} \tag{5.4}$$

where,

- $m_{s,h}$ is the speaker- and session dependent GMM supervector.
- m_0 is the UBM supervector.
- Ux_h; U is the low rank matrix that spans the channel subspace, x_h is the projection of the supervector on this subspace.
- Vy_s; V is the low rank matrix that spans the Speaker subspace, y_s is the projection of the supervector on this subspace.
- $Dz_{s,h}$; $z_{s,h}$ is the residual term of speaker and channel information.

5.2.4 Total variability modelling: i-vector paradigm

FA model evolved into a simplified model, the total variability model denoted the "i-vector" approach (Dehak et al., 2011). This approach defines only a single space which encodes all the speech variability. This new subspace does not make distinction between the speaker and channel like JFA. Given an utterance, the new speaker- and channel-dependent GMM supervector defined by Equation 5.4 is rewritten as follows:

$$m_{s,h} = m_0 + Tw_{s,h} \tag{5.5}$$

where,

- $m_{s,h}$ is the speaker model (GMM supervector).
- m_0 is the UBM supervector.
- T is the total variability matrix of speaker and session, a rectangular matrix of low rank.
- $w_{s,h}$ is the identity vector "i-vector", a random vector assumed to have a standard normal distribution $\mathcal{N}(0,I)$.

The low dimensional nature of this representation is very appealing and has opened the door for new ways to explore one of the key problems in speaker recognition, that is, how to deal with session variability.

5.3 I-vector preprocessing and scoring

Once i-vectors are extracted, preprocessing normalisation techniques are applied in order to attenuate session effects and to prepare the data for scoring. In this subsection, we briefly review some normalisation and channel compensation techniques. Then, we present some scoring methods in the i-vector space.

5.3.1 Normalisation techniques

Various normalization techniques appeared to be effective for system performance, such as Length normalisation (Garcia-Romero et Espy-Wilson, 2011; Bousquet et al., 2011), Eigen Factor Radial (EFR) (Bousquet et al., 2011), Spherical normalisation[1] (sphNorm) (Bousquet et al., 2012).

Length normalisation

Length normalisation technique is proposed (Garcia-Romero et Espy-Wilson, 2011; Bousquet et al., 2011). This method deals with the non-Gaussian behaviour of i-vector by performing a simple length normalisation (i.e scale the length of each i-vector to unit length).

Eigen Factor Radial (EFR)

The *Eigen Factor Radial* (EFR) normalization is presented in (Bousquet et al., 2011). This method iteratively modifies the distribution of i-vectors such that it becomes standard normal and the i-vectors have a unitary norm. Given a development set χ of i-vectors, of mean μ and total covariance matrix Σ, the algorithm of i-vector normalisation is modified according to:

Algorithm 1 Algorithm for i-vector transformation using EFR method.

for i=1 **to** number of iterations **do**
1- Compute mean, μ_i, and total covariance matrix, Σ, of χ
2- For each w of χ: $w \leftarrow \frac{\Sigma_i^{\frac{-1}{2}}(w-\mu_i)}{|\Sigma_i^{\frac{-1}{2}}(w-\mu_i)|}$

5.3.2 Channel compensation

Different channel compensation methods could be applied before scoring in order to get better performance. The most popular techniques are *Linear Discriminant Analysis* (LDA), *Nuisance Attribute Projection* (NAP) (Solomonoff et al., 2005) and *Within-Class Covariance Normalization* WCCN (Hatch et al., 2006).

Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a technique for dimensionality reduction that projects the data onto a subspace which satisfies the requirement of maximizing be-

[1]Spherical Nuisance Normalization is a variant of EFR. Unlike EFR, the total covariance matrix Σ is replaced by the within class covariance matrix of the development set.

tween class variance and minimizing within class variance. The LDA optimisation problem, looking for a basis of this subspace, can be defined according to the following ratio:

$$J(v) = \frac{v^t \mathbf{B} v}{v^t \mathbf{W} v} \tag{5.6}$$

B and W are respectively the between- and within- speaker covariance matrices.

$$\mathbf{B} = \sum_{s=1}^{S} \frac{n_s}{n} (\overline{w_s} - \mu)(\overline{w_s} - \mu)^t \tag{5.7}$$

$$\mathbf{W} = \sum_{s=1}^{S} \frac{n_s}{n} W_s = \frac{1}{n} \sum_{s=1}^{S} \sum_{i=1}^{n_s} (w_i^s - \overline{w_s})(w_i^s - \overline{w_s})^t \tag{5.8}$$

Where,

- W_s is the covariance matrix of speaker s,
- n_s is the number of utterances for speaker s and n is the total number of utterances.
- w_i^s are the i-vectors of speaker s from different sessions.
- μ is represents the overall mean of the training dataset.

The optimal subspace is comprised by the first eigenvectors (those with highest eigenvalues) of $W^{-1}B$.

Nuisance Attribute Projection (NAP)

The NAP algorithm was originally proposed in (Solomonoff et al., 2005) to deal with channel variability in the supervector space. This technique estimates session variability as a subspace of intermediate rank obtained using principal axes (eigenvectors having the largest eigenvalues) of the within-class covariance matrix, and projects the supervector into the orthogonal complementary subspace, assumed to be the speaker space.

The projection matrix, P, is given by the following formula:

$$P = I - M_k M_k^T \tag{5.9}$$

M_k is a rectangular matrix of low rank whose columns are the k principal eigenvectors of the within-class covariance matrix.

The NAP transformation of a given vector w is:

$$NAP(w) = Pw. \tag{5.10}$$

This approach is adapted to operate on the i-vector space by Bousquet et al. in (Bousquet et al., 2011).

Within-Class Covariance Normalization (WCCN)

This technique was proposed by Hatch (Hatch et al., 2006) in the context of SVM-based speaker recognition. The $WCCN$ projection aims at minimizing the expected error rate of false acceptances and false rejections during SVM training. The $WCCN$ projection is performed as:

$$WCCN(w) = B^T w \tag{5.11}$$

Where B is computed through the Cholesky factorization of the inverse of the within-class covariance matrix, W^{-1} such that $W^{-1} = BB^T$.

In contrast to LDA and NAP, the $WCCN$ projection preserves the directions of the feature space.

5.3.3 Scoring methods

Various scoring techniques are employed in the i-vector space. In the following, we review some of them.

Cosine Distance Scoring

The cosine similarity measure-based scoring (CDS) was proposed for speaker verification. In this measure, the similarity (or also the match) score between two i-vectors w_1 and w_2 is computed as follow.

$$score(w_1, w_2) = \frac{w_1 \cdot w_2}{||w_1|| \times ||w_2||} \tag{5.12}$$

Mahalanobis Scoring

Unlike the cosine distance scoring, Mahalanobis scoring needs a training phase based on a training set of i-vectors belonging to k speakers. These i-vectors can be classified according to the known class of the speaker. Given a new observation w, the goal of a statistical classifier is to identify to which class it belongs. If we assume homoscedasticity (equality of class covariances) and Gaussian conditional density models, the most likely class can be obtained by the Bayes optimal solution. An i-vector w is assigned to the speaker s that minimizes:

$$(w - \overline{w_s})^t W^{-1} (w - \overline{w_s}) \tag{5.13}$$

Where,

$\overline{w_s}$ is the mean of i-vectors belonging to class s,

W is the within-class covariance matrix given by Equation 5.14.

$$W = \sum_{s=1}^{S} \frac{n_s}{n} W_s = \frac{1}{n} \sum_{s=1}^{S} \sum_{i=1}^{n_s} (w_i^s - \overline{w_s})(w_i^s - \overline{w_s})^t \tag{5.14}$$

Where,

- W_s is the covariance matrix of speaker s,
- n_s is the number of utterances for speaker s and n is the total number of utterances.
- w_i^s are the i-vectors of speaker s from different sessions.

The Mahlanobis score between two i-vectors w_1 and w_2 is given by Equation 5.15.

$$score(w_1, w_2) = -(w_1 - w_2)^t W^{-1} (w1 - w_2) \tag{5.15}$$

Probabilistic Linear Discriminant Analysis

Probabilistic Linear Discriminant Analysis (PLDA) was first used for session variability compensation for facial recognition (Prince et Elder, 2007). It essentially follows the same modelling assumptions as *JFA* but performing at the i-vector level (an i-vector vector contains class-dependent and session-dependent variabilities, both lying in lower-dimensional subspaces). A p-dimensional i-vector w extracted from a speech recording s is decomposed as shown in Equation 5.16.

$$w = \underbrace{\mu + \Phi y_s}_{\text{speaker component}} + \underbrace{\Gamma z + \epsilon}_{\text{Noise component}} \tag{5.16}$$

Where,

- Φ is a rectangular matrix, with r_{voices} columns ($r_{voices} < p$) providing a basis for a speaker subspace, usually called "eigenvoices".
- Γ is a rectangular matrix, with $r_{channels}$ column providing a basis for a channel subspace, usually called "eigenchannels".
- Standard normal priors are assumed for y_s and z . Lastly, the residual term ϵ is assumed to be Gaussian with zero mean and diagonal covariance Σ.

It is important to note that in speaker recognition, the channel term is ignored. Thus, the PLDA model is rewritten as follow (Kenny, 2010; Garcia-Romero et Espy-Wilson, 2011):

$$w = \underbrace{\mu + \Phi y_s}_{\text{speaker component}} + \underbrace{\epsilon}_{\text{Noise component}} \tag{5.17}$$

After estimation of the PLDA meta-parameters, the speaker verification score given two i-vectors w_1 and w_2 is the likelihood-ratio described by Equation 5.18, where the

hypothesis H_0 states that inputs w_1 and w_2 are from the same speaker and the hypothesis H_1 states they are from different speakers.

$$score(w_1, w_2) = log\left(\frac{(w_1, w_2)|H_0}{(w_1, w_2)|H_1}\right) \tag{5.18}$$

Two-covariance scoring

The two covariance scoring is proposed in (Brümmer et De Villiers, 2010) and could be seen as a particular case of $PLDA$ (Burget et al., 2011). It consists of a simple linear-Gaussian generative model where an i-vector w extracted from a speech recording is decomposed as shown in Equation 5.19.

$$w = y_s + \epsilon \tag{5.19}$$

The speaker model, y_s, is a vector of the same dimensionality as the i-vector, w, while ϵ is Gaussian noise.

The two-covariance scoring uses between-speaker covariance matrix B and within-speaker covariance matrix W. The speaker verification score given two i-vectors, w_1 and w_2 is given by 5.20[2].

$$score(w_1, w_2) = -\frac{1}{2}\{w_1^t \Psi w_1 + w_2^t \Psi w_2 + 2w_1^t \Phi w_2\} \tag{5.20}$$

Where,

$$\Psi = W^{-1}\{\left(B^{-1} + 2W^{-1}\right)^{-1} - \left(B^{-1} + W^{-1}\right)\}W^{-1} \tag{5.21}$$

$$\Phi = W^{-1}\left(B^{-1} + 2W^{-1}\right)^{-1}W^{-1} \tag{5.22}$$

5.4 Performance evaluation

In order to evaluate the performance of speaker recognition systems, a large number of target and non-target trials (also known as matched and non-matched comparisons or genuine and impostor trials) are required. Target trial is considered where both voice recordings are pronounced by the same speaker (same identity), and non-target trial where voice recordings are pronounced by two different speakers (two different identities). An ASpR system computes a score for each trial and compares the score

[2]This formula assumes that the i-vector distribution is normalized, $\mathcal{N}(0, I)$.

against a fixed threshold τ to effect a **binary** accept/reject decision. This process results in one of four possible outcomes as shown in Table 5.1.

Table 5.1: *Possible scenarios for trial decision.*

	Accept	Reject
Matched pair	√	×
Non-matched pair	×	√

Two types of errors may result:

- For matched pair, the error when ASpR rejects a client is called *False Reject* (FR) or miss.
- For non-matched pair, the error when ASpR accepts an impostor is called *False Acceptance* (FA).

The *False-Acceptance-Rate* (FAR) is the count of *FA*, normalized by the number of non-target trials as shown in Equation 5.23.

$$\text{False Acceptance Rate}(FAR) = \frac{\text{Number of FA errors}}{\text{Number of non-matched pairs}}. \tag{5.23}$$

The *False Reject Rate* (FRR) is the miss count, normalized by the number of target trials as shown in Equation 5.24.

$$\text{False Reject Rate}(FRR) = \frac{\text{Number of FR errors}}{\text{Number of matched pairs}}. \tag{5.24}$$

The pair (*FAR*, *FRR*) can be considered as the evaluation outcome. In practice, the system parameters are optimised to minimise the balance between *FAR* and *FRR* which could be measured for example by the *Equal Error Rate* (EER) or the minimum *Decision Cost Function* minDCF.

Equal Error Rate

EER is defined as the *FAR* and *FRR* values becoming equal. The *EER* is a commonly used measure which gives, in one number, an information about system's accuracy. A graphical illustration is shown in Figure 5.5.

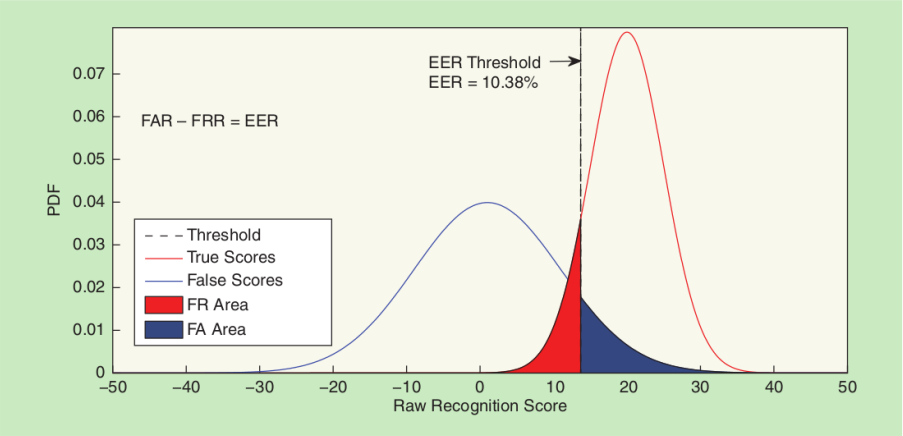

Figure 5.5: *An illustration of target and non-target score distributions and the decision threshold. Areas under the curves with blue and red colors represent FAR and FRR errors, respectively (Hansen et Hasan, 2015).*

Detection Cost Function (DCF)

The *Detection Cost Function* (DCF) introduces numerical costs/penalties for the two types of errors (*FAR* and *FRR*). The a priori probability of encountering a target speaker, P_{Target}, is also provided. The *DCF* is computed as a function of the decision threshold as shown in Equation 5.25:

$$DCF(\tau) = C_{MISS} \times FRR(\tau) \times P_{Target} + C_{FA} \times FAR(\tau) \times (1 - P_{TARGET}) \tag{5.25}$$

Where,

- C_{MISS} is the cost of a FR error.
- C_{FA} is the cost of a FA error.
- P_{Target} is the a priori probability of observing a target speaker.
- FRR is the false rejection rate.
- FAR is the false acceptance rate.

Two performance measures, derived from the DCF, could be calculated:

1. The actual DCF denoted *ActDCF*.
2. The minimum DCF usually denoted *"minDCF"*. This measure correspond to the minimum value of *DCF* obtained by varying the threshold τ.

Several graphical tools could be used in order to visualize system performance like *Detection Error Trade-off* (DET) curves, Tippett plots, *Applied Probability of Error* (APE) plot and other kind of plots.

Detection Error Trade-off (DET) curves

The performance of a biometric system is graphically represented by the popular *Receiver Operating Characteristic* (ROC) curve. This curve represents *FRR* in function of *FAR* for different threshold values. The DET-curve (see (Martin et al., 1997)) is a re-parameterization of the receiver-operating-curve (ROC). An example of the det-curve is presented in Figure 5.6.

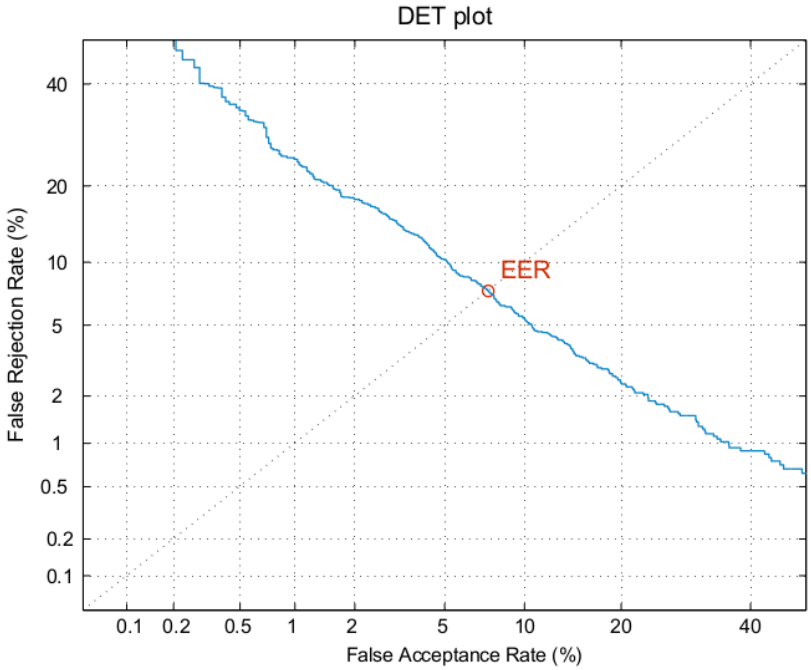

Figure 5.6: *An example of a DET curve and the EER point.*

5.5 Automatic Speaker recognition in Forensic context

Forensic Automatic Speaker Recognition (FASpR) is a term used when automatic speaker recognition methods are adapted to forensic applications (Drygajlo, 2007). In this case, automatic speaker recognition is used to identify whether an unknown voice of a questioned recording (trace) is pronounced by a suspected speaker. Following the same steps as a typical biometric system, a similarity score quantifying the degree of similarity between speaker-dependent features extracted from the trace and speaker-dependent features extracted from recorded speech of a suspect, is calculated (Drygajlo, 2015; Drygajlo et al., 2003; Drygajlo, 2007). Then, a normalisation method is applied to map the calculated score to a LR. The FVC process is summarized in Figure 5.7.

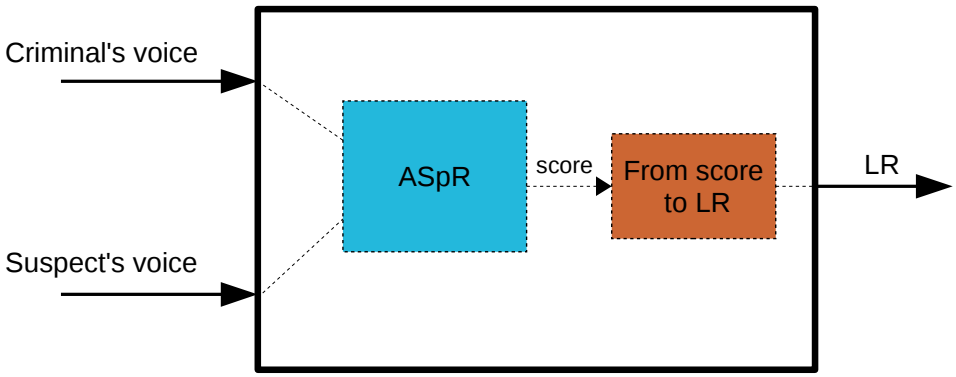

Figure 5.7: *A schematic view of FVC process using ASpR system.*

In this subsection, we briefly discuss three main points:

- why automatic speaker recognition can be considered well suited in forensic context.
- how ASpR provides a coherent way to quantify and present a speech recording as an evidence, as well as the assessment of its strength in the Bayesian interpretation framework.
- how to evaluate the validity and reliability of the method.

5.5.1 Advantages of using automatic speaker recognition in forensic context

According to (Ramos, 2007), an ASpR has several characteristics that make it well adapted to the "*paradigm shift*" requirements and particularly satisfies Daubert criterion. Indeed, ASpR system ensures:

- Transparency: For a transparent framework, two main conditions should be satisfied: (i) The framework is based on scientific disciplines and (ii) accepted widely in the scientific community.
- Testability: Collection and clear definition of materials and corpus. Second, clear and well-defined forensic testing protocols. These two requirements guarantee the accuracy of a forensic *ASpR* technique to be determined in controlled and transparent conditions for the court.
- Accuracy: *ASpR* has reached impressive low error rates.
- Common procedures: score-based architecture, common database. The use of common methods and procedures in order to report forensic expertise is a key

issue for the admissibility as stated by Daubert.

5.5.2 From score to Likelihood ratio: Similarity/Typicality approach

The first application of the LR computation methodology to ASpR was proposed by Meuwly (Meuwly, 2001; Meuwly et Drygajlo, 2001) for a GMM-based system. In this approach, the outputted scores of an ASpR system are used in order to model both the pdfs of the two hypotheses used in the Bayesian framework. These hypotheses correspond respectively to the numerator and the denominator of the likelihood ratio, $p(E \mid H_p)$ and $p(E \mid H_d)$. Generally, the methodology proposed by Meuwly needs three databases for the calculation and the interpretation of the evidence.

- Relevant population database (P).
- Control speech database (C).
- Reference database (R).

The between-source pdf, $p(E \mid H_d)$, models scores assuming that H_d is true. The non-target scores are obtained comparing the questioned speech under analysis with the relevant population of individuals (i.e database P).

The within-source distribution, $p(E \mid H_p)$, models scores assuming that H_p is true. The target scores are obtained comparing different utterances from the control speech material.

For more details, readers may refer to (Meuwly, 2001; Gonzalez-Rodriguez et al., 2006; Drygajlo et al., 2003). The LR estimation process is shown in Figure 5.8.

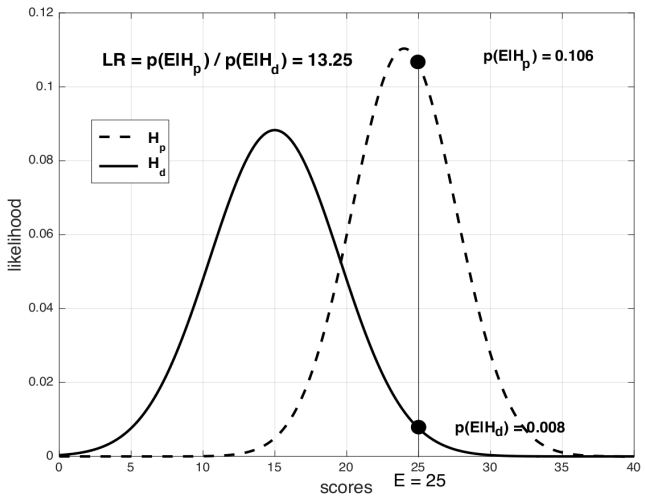

Figure 5.8: *Example of LR computation given a value of the evidence of 25. The probability density functions (pdfs) of the within-source,$p(E \mid H_p)$, and between-sources, $p(E \mid H_d)$), similarity scores are also presented. LR value (13.25) goes in favour of the prosecution. Adapted from (Drygajlo et Haraksim, 2017).*

5.5.3 From score to Likelihood ratio: Calibration approach

For proper *LR* interpretation, a score normalization known as "*calibration*" is applied. The latter is a process of transforming the *ASpR* outputted raw scores into *LRs*. Several kinds of calibration methods appeared in the state-of-the-art (Brummer et van Leeuwen, 2006; Brummer, 2007; Gonzalez-Rodriguez et al., 2006; Gonzalez-Rodriguez et Ramos, 2007; Gonzalez-Rodriguez et al., 2007; Kinoshita et al., 2014; Nautsch et al., 2016). Particularly, linear calibration has proven to be working well in many NIST SREs (Brümmer et al., 2007). Linear calibration is an affine transformation to map *ASpR* "raw" scores, denoted s^{raw}, to calibrated score *LR* using two parameters, offset w_0 and scaling w_1 parameter. The transformation of raw scores is shown in Equation 5.26.

$$LR = w_0 + w_1 s^{raw} \tag{5.26}$$

The calibration parameters (the offset and the linear scaling parameters) are estimated using logistic regression based on training scores sourced from a development database.

The best calibration is performed based on *Pool Adjacent Violator* (PAV) algorithm (Ahuja et Orlin, 2001; Zadrozny et Elkan, 2002).

For more details, readers may be referred to the recent published paper "*From Biometric Scores to Forensic Likelihood Ratios*" (Ramos et al., 2017).

5.5.4 Likelihood ratio accuracy

This subsection discusses numerical and graphical tools for assessing FVC accuracy. The main evaluation measure discussed here is the *Log-likelihood-ratio Cost* denoted (C_{llr}). APE curve is also presented for graphical assessment.

Log-likelihood-ratio Cost

The *Log-likelihood-ratio cost*, C_{llr}, (Brümmer et du Preez, 2006; Morrison, 2009a; Castro, 2007; Gonzalez-Rodriguez et Ramos, 2007) is largely used in the forensic arena as it wishes to evaluate the *LR* and is not based on hard decisions like, for example, *equal error rate* (EER). C_{llr} has the meaning of a cost or a loss: lower is the C_{llr}, better is the performance. C_{llr} could be calculated as follows:

$$C_{llr} = \frac{1}{2N_{tar}} \sum_{LR \in \chi_{tar}} \log_2 \left(1 + \frac{1}{LR}\right) + \frac{1}{2N_{non}} \sum_{LR \in \chi_{non}} \log_2 \left(1 + LR\right) \tag{5.27}$$

Where χ_{tar} and χ_{non} are the same-speaker and different-speaker comparisons sets, of cardinality N_{tar} and N_{non}.

Figure 5.9 provides a plot of C_{llr} computation function for both same- and different-origin pair. It shows that large positive likelihood-ratio values which correctly support the same-origin hypothesis correspond to very low C_{llr} values. On the other hand, negative log-likelihood-ratio which contrary-to-fact support the different-origin hypothesis are assigned with high C_{llr} component values.

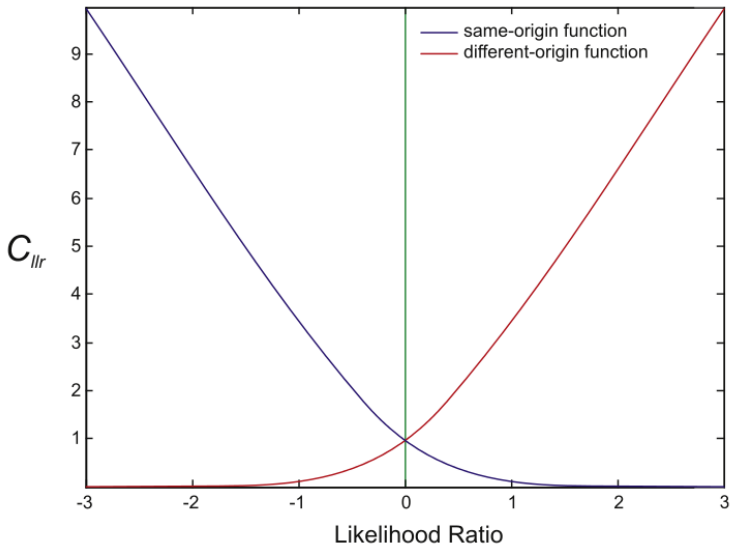

Figure 5.9: *Components functions of C_{llr} used to calculate penalty values assigned to likelihood ratios from same-speaker comparisons (curve rising to the left) and different-speakers comparisons (curve rising to the right) (Morrison, 2011a).*

The minimum value of C_{llr} is denoted C_{llr}^{min}. As mentioned previously, this value is obtained using *PAV* algorithm for calibration.

C_{llr} involves calibration loss while C_{llr}^{min} contains only discrimination loss. We can judge the quality of the calibration (i.e., the mapping from score to log-likelihood-ratio which is actually present in the detector) by calculating C_{llr}^{cal} given by the following Equation:

$$C_{llr}^{cal} = C_{llr} - C_{llr}^{min}. \tag{5.28}$$

A system is deemed well-calibrated when it has a low calibration cost and is, therefore, able to provide more reliable likelihood ratio values.

APE curve

The APE-curve (Brümmer et du Preez, 2006) is proposed as a way of measuring the probability of error of the LR values computed by the forensic system over a wide range of applications (different costs and priors). The APE curve plots the total probability of error with respect to the prior probabilities. Figure 5.10 is an example of APE curve.

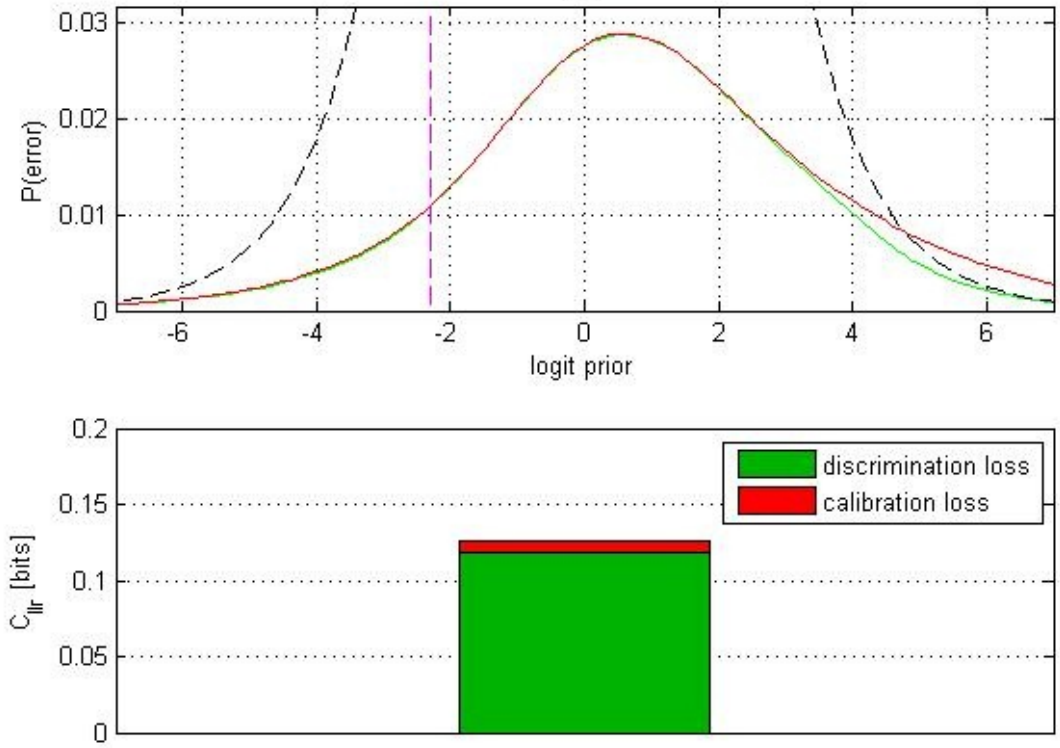

Figure 5.10: *Example of APE and C_{llr} plots. (i) The magenta dashed vertical line to indicate where the traditional NIST operating point is (at -2.29). (ii) The red and green error-rates at -2.29 are scaled versions of the traditional CDET and 'min CDET' values. The max of the green curve is also the EER. (iii) GREEN: (Minimum) total error-rate over the whole range of applications. This is the performance that the system under evaluation could have obtained with a perfect (for this data) score-to-llr calibration. (iv) RED: This is the area between the red and the green APE-curves and is the measure of how well the score to log-likelihood-ratio mapping is 'calibrated'.*

Conclusion

In this chapter, we provided an overview of the different components of the automatic speaker recognition (ASpR) system including feature extraction, speaker modelling and scoring methods. Then, we showed how ASpR systems could be adapted in order to perform forensic voice comparison.

Part II

Methodology and resources

Chapter 6

Databases and protocols

Contents

Introduction

Speech databases constitute an important factor for automatic speaker recognition. The NIST-SRE evaluation campaigns organized by the *National Institute of Standards and Technology* (NIST) (Martin et Przybocki, 2001; NIST, 2010) paved the way for an outstanding evolution of ASpR systems. Indeed, since the first SRE campaign organized in 1996, ASpR systems have achieved significant progress and have reached remarkably low error rate ($\approx$ 1%). Many other speech corpora such as TIMIT, Aurora, CHAINS, YOHO, CSLU and Switchboard, are widely used for speaker recognition. Nevertheless, in order to better suit speaker comparison task, speech databases should satisfy some minimal requirements:

- Speech databases should have a large number of speakers for both background model training and testing stage.

- Speech databases should provide multi-session voice samples for each speaker.

If the first point is well addressed in the majority of the quoted databases, the second point is still not taken into account seriously. In this chapter, we review briefly the NIST SRE framework. We describe the NIST SRE'08, a database used in this thesis. Second, we present Fabiole a new database dedicated to study the "speaker factor".

6.1 NIST SRE framework

6.1.1 Motivation and challenge

The NIST SRE evaluation campaigns [1] started in 1996. The main goal of the NIST-SRE is to provide a common framework in order to evaluate the system's accuracy and the developed approaches in the field of speaker recognition. Through these evaluations, researchers have been able to compare different approaches using common experimental protocols and common large datasets. This had facilitated much of the progress made in the two past decades. Indeed, these yearly evaluations widely spread the highest-performing speaker recognition techniques, allow to exchange ideas and show the most promising research directions.

The continuous success of NIST SRE is confirmed by the growing number of participating sites in the follow-up workshop after each evaluation and by the the number of NIST-SRE-related scientific publications in major conferences and journals.

6.1.2 NIST SRE'2008 database

The NIST SRE'2008 database is a part of an ongoing series of databases provided by NIST. It was developed by the *Linguistic Data Consortium* (LDC) and NIST. NIST SRE'2008 evaluation was distinguished from prior evaluations, in particular those in 2005 and 2006, by including not only conversational telephone speech data but also speech data recorded over a microphone channel involving an interview scenario (Martin et Greenberg, 2009).

Speech data description

NIST SRE'2008 database contains 942 hours of multilingual telephone speech and English interview speech:

- The telephone speech is predominantly English, but the corpus also includes other languages. All interview segments are in English.
- Telephone speech represents approximately 368 hours, whereas microphone speech represents the other 574 hours.

[1] NIST evaluations are mainly funded by the U.S. Department of defence.

- The telephone speech segments include two-channel excerpts and summed-channel excerpts of approximately 10 seconds and 5 minutes. On the other hand, the microphone excerpts are either 3 or 8 minutes in length.

These data are collected from a large pool of speakers recorded across numerous microphones and in different communicative situations, sometime in multiple languages. Speakers were native English and bilingual English. Table 6.1 summarizes information of the telephone data related to the spoken language, the number of sessions per language as well as the number of speaker for each language.

Table 6.1: *Speaker in the NIST SRE'2008 database. Source (Juliette, 2011)*

Language	Number of sessions	Number of speakers	Average of sessions per speaker
Arabic	3	1	3
Bengali	8	5	1
Chinese Min Nan	6	1	6
Chinese Min Nan and Mandarin	1	1	1
Mandarin	12	10	1
Mandarin and chinese cantonese	1	1	1
English (native and non-native)	554	164	3
Persian	3	3	1
Hindi	48	24	2
Hindi. Indian English	7	7	1
Italian	6	4	1
Japanese	19	7	2
Korean	12	6	2
Russian	23	17	1
Tagalog	4	2	2
Thai	41	15	2
Uzbek	3	2	1
Vietnamese	29	12	2
Wu Chinese of Shanghai	6	2	3
Chinese Cantonese	30	13	2

Evaluation protocol

The NIST SRE-2008 evaluation protocol includes six enrolment data conditions (10-sec, short2, 3conv, 8conv, long, 3summed) and four test conditions (10-sec, short3, long, summed). For more details, readers could be referred to "*The NIST Year 2008 Speaker Recognition Evaluation Plan*" (Martin et Greenberg, 2009). The combination Short2-Short3 represents the "*core*" evaluation or the common evaluation condition. This condition offers about 2.5 minutes of data each for enrolment and testing and has eight different subsets for scoring purposes (Martin et Greenberg, 2009). Table 6.2 presents the number of target and non-target trials in each condition. A brief description for each condition is also provided.

Table 6.2: *Number of target and non-target comparisons in the common condition short2-short3 in NIST SRE 2008.*

	#target comparisons	#non-target comparisons	Description: All trials involving
det1	4901	9504	only interview speech in training and test.
det2	248	483	interview speech from the same microphone type in training and test.
det3	4653	9021	interview speech from different microphones types in training and test.
det4	439	4609	interview training speech and telephone test speech.
det5	640	3406	telephone training speech and non-interview microphone test speech.
det6	874	11637	only telephone speech in training and test.
det7	439	6176	only English language telephone speech in training and test and finally.
det8	228	3028	only English language telephone speech spoken by a native U.S. English speaker in training and test.

In this thesis, we focus on the first condition (det1) as it provides the highest number of trials.

6.2 Fabiole database

"FABIOLE" [2] is a speech database created inside the ANR-12-BS03-0011 FABIOLE project [3]. The main goal of this database is to investigate the reliability of ASpR-based FVC. FABIOLE is primarily designed to allow studies on intra-speaker variability as it is a major issue in scientific research generally and in forensic voice comparison particularly. In the following subsections, we present first the motivations and challenges behind the creation of this database. Then, we present a detailed description of Fabiole.

6.2.1 Motivation and challenge

Even though NIST SRE is the most commonly used framework for the evaluation, it does not allow to have a robust study on the intra-speaker variability effect: For example NIST'SRE 2008 contains a large number of speakers with only 3 speech recordings in average per speaker. Moreover, some forensic databases (Ramos et al., 2008; van der Vloed et al., 2014; Morrison et al., 2015), -that logically have to pay attention to all the sensitive factors- do not address properly the intra-speaker variability effect for the same reason (low utterance number per speaker). For example, a recent forensic database (Morrison et al., 2015) proposes a large number of speakers (301 male and 231 female) but every speaker disposes a limited number of speech recordings: only 5 male and 19 female have been recorded more than 3 times.

As discussed in Subsection 3.3, intra-speaker variability is a major challenge and should not be neglected anymore especially for FVC. In order to study this factor, the availability of adequate speech databases is crucial. This database should differ from previous corpora in three main points:

[2] Public corpus easily accessible to the scientific community.
[3] French project funded by the National Research Agency, ANR.

- The number of voice recordings per speaker should be high enough to allow a robust study of the intra-speaker variability effect.
- Variation related to speech conditions should be controlled as much as possible in order to focus mainly on the difference related to the speaker itself rather than the other factor of variabilities.
- Presence of speaker in different communicative situations to introduce a significant intra-speaker variability.

6.2.2 Speech conditions

In the Fabiole database, the factors responsible of speech variabilities not linked to the speaker itself are controlled as much as possible:

- First, channel variability is reduced as all the excerpts come from French radio or television shows.
- Second, for all speech samples, the quality of recordings is high in order to decrease noise effects as our main interest is the intra-speaker variability.
- Then, all the speech samples turn with a minimum duration of 30 seconds of speech (short utterance is no longer a matter).
- Finally, due to financial constraints, Fabiole contains recordings from only 30 male speakers.

Fabiole database takes advantage of some French corpora which contain speech materials *close* [4] to Fabiole excerpts. Indeed, these corpus could be used as training data to build state-of-art models. In the literature, the following databases are available:

- ESTER 1 (Galliano et al., 2006): About 100 hours of transcribed data make up the corpus, recorded between 1998 and 2004 from six French speaking radio stations: France Inter, France Info, RFI, RTM, France Culture and Radio Classique. Shows last from 10 minutes up to 60 minutes. They consist mostly of prepared speech such as news reports, and a little conversational speech (such as interviews).
- ESTER 2 (Galliano et al., 2009) comes to supplement ESTER 1 corpus with about 100 hours of transcribed broadcast news recorded from 1998 to 2004.
- REPERE (Kahn et al., 2012; Galibert et Kahn, 2013) contains currently 60 hours of video with multi-modal annotations. The aim in this evaluation is to develop a system able to answer the following questions: Who is speaking? Who is present in the video? What names are cited? What names are displayed? The challenge is to combine the various information coming from the speech and the images.

[4] Close in term of acoustic conditions. Firstly, FABIOLE and these corpus contain excerpts from the same television shows and radio. Second, they contain the same kind of speech, political speech, news report...

- ETAPE corpus (Gravier et al., 2012) consists of 30 hours of TV and radio broadcasts, selected to cover a wide variety of topics and speaking styles, emphasizing spontaneous speech and multiple speaker areas.

FABIOLE contains the same type of records as those contained in these databases (details are given in the following subsection). The list of speakers of REPERE, ETAPE and ESTER 1&2 that must be excluded to avoid biasing the system is provided in the FABIOLE package.

6.2.3 Shows and sampling

Fabiole speech recordings are sampled to 16 kHz using 16 bit wave files. All excerpts are collected from 10 different sources including speech recordings from television and radio sequences of 10 programs recently recorded (from 2013 to 2014). Excerpts come from:

- Debates such as “Ca vous regarde - Le débat” (*cvgddebat*), “Entre les Lignes” (*entreligne*), “Le masque et la plume” (*MsqPlum*).
- Chronicles such as “Service public" (*Spublic*), “Comme on nous parle" (*ComParle*).
- Interviews such as “Un Temps de Pauchon" (*temPauch*).
- Parliamentary discussions such as “Top Questions” (*topquestions*).
- News such as “BFM Story" (*bfmstory*), “LCP Info” (*parlinfo*), “Ca vous regarde-l’Info (*cvgdinfo*)”.

Figure 6.1 presents the amount of data per broadcaster. We can see that excerpts come from 10 different shows with high difference of contribution. For example, *Comme on nous parle* represents 22.5% of Fabiole materials while *temps de Pauchon* represents only 2.2% of the database.

6.2.4 Speaker information

Regarding inter-speaker variability, it is of course desirable to include many speakers to achieve a good sampling of a speaker population. At the same time, it is often desirable to get a good sampling of individual speakers over multiple sessions to study intra-speaker variability. With only limited resources for database creation, a trade-off between the number of speakers, number of sessions per speaker is applied.

FABIOLE database contains 130 male French native speakers divided into two sets:

- Set *T*: 30 target speakers with 100 speech recordings each. Hence, each speaker can be associated with a large number of targets trials, which is a clear advantage compared to various other databases in which the number of target trials per speaker is very low.

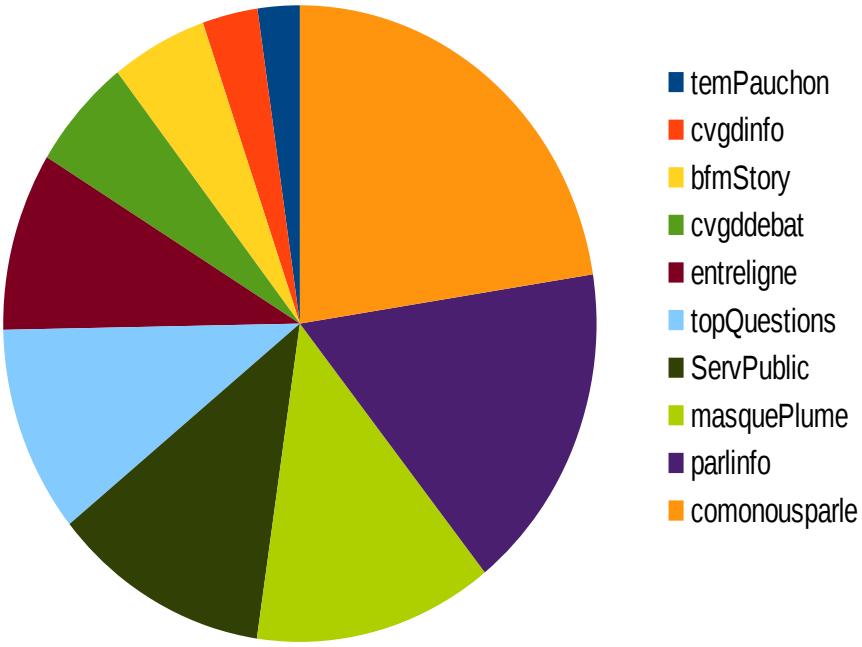

Figure 6.1: *Proportion of shows contribution in FABIOLE database ranged from the lowest (Blue) to the highest (Orange) contribution.*

- Set *I*: 100 "impostors" who everyone has one speech recording (one session). These test files are used essentially to create non-targets trials. It allows to associate a given impostor recording with all the *T* speakers, removing one of the frequent bias in NIST-based experiments.

Table 6.3 presents the amount of data per broadcaster for every speaker. Table 6.3 shows that some speakers appear only in one show such as *Arnaud Ardoin* and *Michel Ciment* whereas others appear in more than one show such as *Manuels Valls*: 13 speakers were recorded from 1 kind of show, 10 speakers from 2 kinds of show and 7 speakers from 3-5 kinds of show.

Table 6.3: *Amount of data per broadcaster for every speaker.*

Spk vs show	**cvgdinfo**	**cvgddebat**	**entreligne**	**parlinfo**	**topquestions**	**bfmstory**	**temPauch**	**MsqPlum**	**ComParle**	**Spublic**
“1”	0	0	0	0	0	0	0	0	1:04:15	0
“2”	0	0	0:58:00.13	0:13:47.99	0	0	0	0	0	0
“3”	0	0	0	0	0	0	0	0	1:26:08.53	0
“4”	0	0	1:05:48.02	0	0	0:1:23.94	0	0	0	0
“5”	0	2:13:41	0	0	0	0	0	0	0	0
“6”	0	0	0	1:18:15.50	0	0	0	0	0	0
“7”	0:9:01.94	0	0	1:21:18.63	0	0	0	0	0	0
“8”	0	0	0	0:0:47.90	0	0	0	0	1:17:23.49	0
“9”	0	0	0	0	0	0	0	0	1:08:09.38	0
“10”	0:0:38.80	0	0	1:02:15.73	0	0	0	0	0	0
“11”	0	0	0	0	0	0	0	1:11:31.54	0	0
“12”	0	0	0	0	0	0	0	1:58:45.01	0	0
“13”	0	0	0	0	0	1:13:24.20	0	0	0	0
“14”	0	0	0	0	0	0	0	0	0	1:28:54.79
“15”	0	0	0	0	0	0	0	0	0	2:58:14
“16”	0	0	0	0	0	0:5:18.27	0	0	1:01:39.57	0
“17”	0	0	0:59:40.36	0:3:47.07	0	0	0	0	0	0
“18”	0	0	0	0	0	0	0	0:54:40.01	0:11:36.74	0
“19”	0	0	0	0	0	0	0	1:09:27.19	0	0
“20”	0	0	0	0	0	0:0:31.18	0	0	1:02:17.13	0
“21”	0	0	0	0	0	0	0	0	1:07:10.21	0
“22”	0:2:35.97	0	0:1:21.95	0:2:18.82	1:49:02.31	0:4:12.24	0	0	0	0
“23”	0	0	0:33:39.01	0:4:09.54	0	0:29:18.56	0	0	0	0
“24”	0:9:28.61	0:1:52.86	0	0:48:11.11	0	0	0	0	0	0
“25”	0:20:16.74	0:2:20.55	0	0:34:34.68	0	0	0	0	0	0
“26”	0	0	0	0	0	0	0:53:50.50	0	0	0
“27”	0	0	0	0:15:26.83	0	0	0	0	0:43:13.07	0
“28”	0:17:19.52	0:0:34.13	0	0:39:32.99	0	0	0	0	0	0
“29”	0:12:05.82	0	0	0:36:18.00	0:3:45.29	0:5:39.76	0	0	0	0
“30”	0:1:28.30	0	0	0:1:49.59	0:52:47.12	0	0	0	0	0
Speakers of set I	0	0	0:1:06.40	0:4:21.74	1:29:44.85	0:7:15.74	0:1:36.96	0:1:33.17	0:4:10.96	0:19:43

As our main interest are speakers of set T, we present in Table 6.4 detailed information related to their age and profession while for speakers of set I the same information is given globally. Table 6.4 shows that the majority of target speakers are 40 to 60 years old and the most of them are interviewers (10 speakers), chroniclers(7 speakers) and debaters(7 speakers).

Note that for some speakers, age information is missing.

Table 6.4: *Speaker's ages and professions.*

Spk vs age	**20-30**	**30-40**	**40-50**	**50-60**	**60-70**	**>70**	**Profession**
"1"	0	0	0	1	0	0	chronicler
"2"	0	0	0	1	0	0	chronicler
"3"	0	0	1	0	0	0	chronicler
"4"	0	0	0	1	0	0	interviewer
"5"	0	0	1	0	0	0	interviewer
"6"	0	0	0	0	0	1	scientist
"7"	0	0	1	0	0	0	interviewer
"8"	0	0	0	1	0	0	chronicler
"9"	0	0	0	0	1	0	debater
"10"	0	0	0	0	0	1	interviewer
"11"	0	0	0	0	0	1	chronicler
"12"	0	0	0	1	0	0	interviewer
"13"	0	0	1	0	0	0	interviewer
"14"	0	0	0	0	1	0	interviewer
"15"	0	0	1	0	0	0	interviewer
"16"	-	-	-	-	-	-	debater
"17"	0	0	1	0	0	0	debater
"18"	0	0	0	0	1	0	debater
"19"	0	0	0	1	0	0	debater
"20"	0	0	0	1	0	0	debater
"21"	0	0	1	0	0	0	chronicler
"22"	0	0	0	1	0	0	politician
"23"	0	0	0	1	0	0	chronicler
"24"	0	1	0	0	0	0	interviewer
"25"	0	0	1	0	0	0	interviewer
"26"	0	0	0	1	0	0	journalist
"27"	0	1	0	0	0	0	debater
"28"	0	1	0	0	0	0	journalist
"29"	0	0	1	0	0	0	politician
"30"	0	0	0	0	1	0	politician
Speakers of set I	1	9	30	27	30	2	all profession

6.2.5 Orthographic transcription

Only 10% of Fabiole data are manually transcribed. The remaining data has been automatically transcribed thanks to Speeral automatic transcription system (Linares et al., 2007). This system was used to transcribe REPERE development set (which contains speech recordings close to FABIOLE excerpts) with an overall Word Error Rate of 29% (Bigot et al., 2013).

6.2.6 Experimental protocols

Main protocol

This protocol relies only on the data of set T. In this context, FABIOLE proposes more than $150,000$ matched pairs (target trials) and more than $4.5M$ non-matched pairs (non-target trials). The trials are divided into 30 subsets, one for each T speaker. For one subset, the voice comparison pairs are composed with one recording pronounced by the corresponding T speaker. It gives for a given subset 294950 pairs of recordings distributed as follows:

- 4950 same-speaker pairs. These target pairs are obtained using all the combinations of the 100 recordings available for the corresponding T speaker (C^2_{100} target pairs).
- $290k$ different-speakers pairs. Non-targets pairs are obtained by pairing each of the target speaker's recording (100 are available) with each of the recordings of the 29 remaining speakers, forming consequently $(100 \times 100 \times 29 = 290k)$ non-targets pairs.

Secondary protocol

This protocol relies on the data of both set T and set I. Indeed, this protocol takes advantage of set I in order to propose non-target pairs based on a same impostor set: The set I offers the possibility to separate "*impostors*" from the target speakers.

This protocol propose the same number of matched pairs (target trials) as the main protocol while it proposes about $300k$ non-matched pairs (non-target trials). The non-target trials could be divided into 30 subsets, one for each T speaker. For one subset, the voice comparison pairs are composed with at least one recording pronounced by the corresponding T speaker and one recording of an impostor speaker. Consequently, each target speaker disposes $10k$ non-target trials. In brief, in this protocol, a speaker subset corresponds to $4.95k$ of target trials and $10k$ of non-target trials.

Conclusion

In this chapter, we described the databases and protocols used throughout this Dissertation. First, NIST SRE framework was described with a focus on NIST'2008 database, one of the database used to generate a part of our results. Second, we presented Fabiole, a new corpus created within FABIOLE project[5]. This corpus was mostly used for generating results in this Thesis. The adopted evaluation protocol in our experiments was also introduced.

[5] A project funded by the French Research Agency (ANR).

Chapter 7

Baseline systems

Contents

Introduction

This chapter is dedicated to provide information about the configurations of the systems used in this Thesis. Two PLDA I-vector systems are employed: The first system is used in order to perform experiments on NIST SRE whereas the second one is used to carry out experiments on Fabiole.

7.1 NIST SRE I-vector system

7.1.1 Feature processing

Feature extraction and normalization

The acoustic features are composed of 19 LFCC coefficients (plus energy), using 24 filter banks ranging from 300 Hz to 3400 Hz. Frame length and the step size are fixed respectively to 20 ms and 10 ms. These features are augmented with their 19 first (Δ) and 11 second ($\Delta\Delta$) order derivatives giving a total of 50-dimensional features. At the end of this step, a file-based $CMVN$ normalisation is applied on each utterance producing a 0-mean and 1-variance feature distribution.

Voice activity detection

The VAD system used in this Thesis is described in (Larcher et al., 2013). It is based on the log-energy distribution of frames. First, the log-energies of each frame of an utterance are computed. Then, using the EM algorithm, the distribution of log-energy coefficients is estimated using a GMM with 3 components. Frames which correspond to the highest mean Gaussian (high energy frames) are then used as speech frames while low energy frames, corresponding mainly to silence and noise, are discarded. A threshold is computed to determine the decision making boundary between speech and non-speech classes as defined in 7.1.

$$\tau = \mu_i - \alpha \times \sigma_i \tag{7.1}$$

Where μ_i and σ_i are respectively the mean and the standard deviation of the Gaussian corresponding to high-energy frames and α is a value controlling the selectivity. Increasing the value of the coefficient α allows to take more high-energy frames into account. Finally, the selection of frames is smoothed using a morphological window (Larcher et al., 2013).

In this system, we used α=0 during the voice activity detection process. In NIST evaluation (2008) and under "short2- short3" condition, the rate of frame selection is between 30% to 70% of the original speech.

7.1.2 Modelling

A gender-dependent 512 diagonal component UBM and Total Variability matrix (TV) of low rank, 400, are estimated using NIST SRE 2004, 2005, 2006 and Switchboard data. These models are trained using 15,660 speech utterances corresponding to 1147 speakers. This system is developed using the ALIZE/SpkDet open-source toolkit (Bonastre et al., 2005, 2008; Larcher et al., 2013). A detailed description for the estimation of the total variability matrix and the i-vector extraction are presented in (Matrouf et al., 2007).

7.1.3 I-vector processing and scoring

A PLDA model (Prince et Elder, 2007) is applied for scoring. The eigenvoice rank is fixed to 100 and the eigenchannel matrix is 400 (full-rank). PLDA is preceded by 2 iterations of LW-normalization (spherical nuisance normalization (Bousquet et al., 2012)). The PLDA model is trained using Switchboard II as well as *SRE* telephone data from 2004, 2005, and 2006. Between-class and within-class covariance matrices are estimated from sample covariance matrices, using the same data.

7.2 Fabiole I-vector system

7.2.1 Feature processing

Feature extraction and normalisation

The same feature extraction and normalisation processes as the first system are employed.

Voice activity detection

Speech portions are detected using automatic text-constrained phone alignment. It is important to note that unlike the first system, this VAD covers all the speech portions and not only the speech portions with high energy.

FABIOLE database has been automatically transcribed using Speeral, an automatic transcription system (Linares et al., 2007) whereas training data such as ESTER 1&2, REPERE and ETAPE were manually transcribed.

7.2.2 Modelling

The UBM has 512 components. The UBM and the total variability matrix, TV, are trained on ESTER 1&2, REPERE and ETAPE databases on male speakers who do not appear in FABIOLE database. They are estimated using $7,690$ sessions from $2,906$ speakers whereas the inter-session matrix W is estimated on a subset (selected by keeping only the speakers who have at least two sessions) of $3,410$ sessions from 617 speakers. The dimension of the I-Vectors in the total variability space is 400.

7.2.3 I-vector processing and scoring

A PLDA model is used for scoring. We have set a 250-dimensional subspace for the PLDA eigenvoice and a 0-dimensional subspace for eigenchannel components. PLDA is preceded by 2 EFR iterations and I-vector length normalization (Garcia-Romero et

Espy-Wilson, 2011) since it has been widely shown to provide performance improvements for speaker recognition. The PLDA model is trained using data from ESTER 1&2, REPERE and ETAPE.

7.2.4 Calibration

In this Thesis, we use an affine calibration transformation (Brümmer et al., 2007) estimated using all the trial subsets (pooled condition) using FoCal Toolkit (Brummer, 2007). The parameters of this transformation are obtained using logistic regression (Pigeon et al., 2000; Brummer, 2007). As the transformation applied to the output score is affine, the application of logistic regression calibration to a system will not change its discrimination performance.

Conclusion

In this chapter, we described briefly the configuration of the two systems used in this Thesis. The configuration details are provided at the different system levels notably feature extraction, modelling and I-vector processing and scoring.

Part III

Contributions

Chapter 8

"The devil lies in the details": A deep look inside accuracy

Contents

Introduction

Even though ASpR systems have witnessed significant evolution and have reached low error rate over the two last decades, ASpR performance evaluation is still one of the main weaknesses of these systems. Indeed, the protocols, established in the evaluation campaigns focus on global performance, they use a brute-force strategy and take into account only the averaged behaviour of ASpR systems. At the same time, they do not put enough attention on many sensitive cases where the ASpR systems show a specific behaviour due, for example, to the recording conditions, the noises, the content of the recordings or the speakers themselves. In this chapter, we will deeply investigate the FVC accuracy depending on the speaker's behaviour using an ASpR system as a gauge.

This investigation will take advantage of our Fabiole database to highlight the intra-speaker variability effects and the inter-speaker's ones. This analysis is carried at two levels:

- The first level is dedicated to highlight the differences in performance between the 30 Fabiole speakers. If the system has the same behaviour across all the speakers or there are some speakers who are well recognized by the system while others are not? This investigation will also look separately at the intra-speaker variability effect and inter-speakers ones.
- At the second level, we push our analysis a step farther. This time the study is not limited to the speakers separately but is focused on speaker pairs.

This chapter is organized as follows: Section 1 presents the methodology we adopted in this investigation. Section 2 presents the performance of the PLDA i-vector system following the traditional protocols showed during the large evaluation campaigns (like the NIST's ones). Section 3 shows the difference in system behaviour across the speakers taken separately (first level). Section 4 shows the difference in system behaviour when a pair of speaker is compared (second level).

8.1 Methodology

In this section, we start by defining criteria to quantify intra-speaker variability and speaker discrimination power. Then, we propose the total cross entropy as a distance between a pair of speakers.

8.1.1 Criteria to quantify intra-speaker variability and speaker discrimination power

In order to quantify the intra-speaker variability effect and the speaker discrimination power, we propose two measures derived on C_{llr} given by the following formula (detailed in section 5.5.4).

$$C_{llr} = \frac{1}{2N_{tar}} \sum_{LR \in \chi_{tar}} \log_2 \left(1 + \frac{1}{LR}\right) + \frac{1}{2N_{non}} \sum_{LR \in \chi_{non}} \log_2 \left(1 + LR\right)$$

As shown in Equation 8.1, C_{llr} can be decomposed into the sum of two components:

$$C_{llr} = \underbrace{\frac{1}{2N_{tar}} \sum_{LR \in \chi_{tar}} \log_2 \left(1 + \frac{1}{LR}\right)}_{C_{llr}^{TAR}} + \underbrace{\frac{1}{2N_{non}} \sum_{LR \in \chi_{non}} \log_2 \left(1 + LR\right)}_{C_{llr}^{NON}} \tag{8.1}$$

- C_{llr}^{TAR}, which is the average information loss related to target trials.

- C_{llr}^{NON}, which is the average information loss related to non-target trials.

In FVC applications, the first component will give an idea of the risk of "impunity" and the second component will express the risk of "wrongful conviction" (convicting a person for a crime that he did not commit.).

Within our experimental conditions, we make the hypothesis that C_{llr}^{TAR} reflects mainly *intra-speaker variability*. Indeed, we base this statement on the fact that in the Fabiole context several factors of variability are controlled. Respectively, we assume that C_{llr}^{NON} is mainly linked to the *speaker discrimination power*. Let S_A and S_B be two speakers and χ_{AB} a trial dataset corresponding to non-target pairs involving these speakers {non-targets $= S_A S_B$}. The C_{llr} of this trial dataset given in Equation 8.2 can be seen as a degree of confusion between the two speakers, S_A and S_B. Indeed:

- A low value of C_{llr}^{NON} indicates that the two speakers are fairly discriminated.
- A high value of C_{llr}^{NON} indicates that the two speakers are confused.

$$C_{llr}^{NON} = \frac{1}{2N_{non}} \sum_{LR \in \chi_{non}} \log_2 (1 + LR) \tag{8.2}$$

8.1.2 Total cross entropy as a distance between two speakers

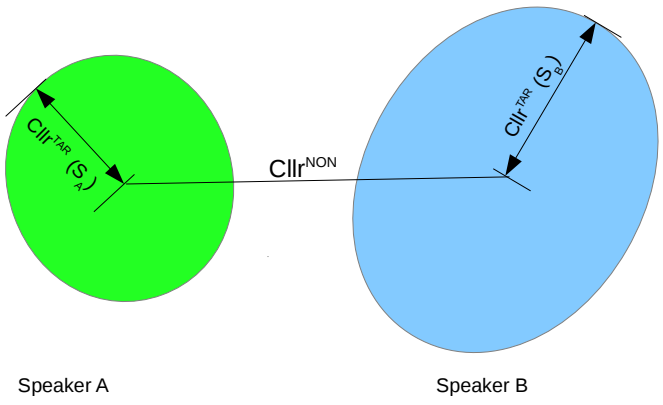

Figure 8.1: *"Distance" between speaker* S_A *and* S_B. *The distance between two different speakers depends on three factors: The strength of intra-speaker variability of speaker* S_A, $C_{llr}^{TAR}(S_A)$. *The strength of intra-speaker variability of speaker* S_B, $C_{llr}^{TAR}(S_B)$ *and the discrimination power of the two speakers,* S_A *and* S_B, $C_{llr}^{NON}(S_A - S_B)$.

To define a distance between two speakers is not a straightforward task. Let S_A,S_B two speakers and χ_{AB} a trial dataset corresponding to target and non-target trials involving these two speakers:

χ_{AB}={targets $= S_A S_A, S_B S_B$; non-targets $= S_A S_B$}.

The C_{llr} of this trial dataset can be seen as a similarity measurement between the two speakers, S_A and S_B (see Figure 8.1), but it can not be used as a distance, since an ideal value $C_{llr} = 0$ is achieved when S_A and S_B are perfectly discriminated (S_A,S_B are "far" from each other).

A distance can be stated by using the *Total Cross Entropy*, (TCE), which is the probability of correctly classifying the trials [1]. Given the dataset of target and non-target trials χ_{tar}, χ_{non} with cardinalities N_{tar}, N_{non} respectively, TCE is defined as follows (Brümmer et du Preez, 2006):

$$TCE = \left(\prod_{t \in \chi_{non}} P\left(H_{non}|t\right) \right)^{\frac{1}{N_{non}}} \left(\prod_{t \in \chi_{tar}} P\left(H_{tar}|t\right) \right)^{\frac{1}{N_{tar}}} \tag{8.3}$$

Given the LLR of a trial t, $s(t) = \log \frac{P\left(H_{tar}|t\right)}{P\left(H_{non}|t\right)} + log\left(\frac{P(H_{tar})}{P(H_{non})}\right)$, and the formula of C_{llr}:

$$C_{llr} = \frac{1}{2\log\left(2\right)} \left(\frac{1}{N_{non}} \sum_{t \in \chi_{non}} \log\left(1 + e^{s(t)}\right) + \frac{1}{N_{tar}} \sum_{t \in \chi_{tar}} \log\left(1 + e^{-s(t)}\right) \right) \tag{8.4}$$

a straightforward computation shows that:

$$C_{llr} = -\frac{1}{2\log\left(2\right)} \log TCE \tag{8.5}$$

Ideally, if all the target scores are equal to $+\infty$ and all the non-target scores are equal to $-\infty$ then $TCE = 1$ and $C_{llr} = 0$.

Given two speakers S_A and S_B, we define a clustering distance between S_A and S_B based on $Cllr$ as follow:

$$\begin{aligned} d\left(S_A, S_B\right) &= TCE \\ &= \exp\left(-2\log\left(2\right) C_{llr}\right) \end{aligned} \tag{8.6}$$

Note that:

- $d\left(S_A, S_B\right) \in [0, 1]$
- $d\left(S_A, S_A\right) = 0$ arbitrarily ($C_{llr} = +\infty$, impossible to separate one speaker from himself).

[1] Sometimes defined as the *log*-probability.

8.1.3 Statistical significance evaluation

In this subsection, we present the statistical methods used to study the significance of our results. We selected *"ANalysis Of VAriance"* (ANOVA) one of the most widely used statistical hypothesis tests. A difference in term of C_{llr} is considered significant if the obtained p-value is below an arbitrary threshold, classically set to 0.05. The p-value is a number between 0 and 1 and interpreted in the following way:

- A small p-value (typically < 0.05) indicates strong evidence against the "null hypothesis", so you reject the "null hypothesis[2]".
- A large p-value (> 0.05) indicates weak evidence against the "null hypothesis", so you fail to reject the "null hypothesis".
- p-values very close to the cut-off (0.05) are considered to be marginal (could go either way).

In general, a very low value indicates a significant test. A common interpretation of the significance of a test is summarized in Table 8.1.

Table 8.1: *Correspondence between significance level and p-value. The number of (*) represents the significance level. (n.s) refers to non-significant test.*

p-value	significance
≤ 0.001	***
≤ 0.01	**
≤ 0.05	*
> 0.05	n.s

In order to study the size of an effect, several standardized measures have been proposed. An effect size is a quantitative measure designed to quantify the degree of association between an effect (e.g., a main effect, an interaction, a linear contrast) and the dependent variable (Lakens, 2013; Fritz et al., 2012). The value of the measure of association is squared and it can be interpreted as the proportion of variance in the dependent variable that is attributable to each effect. Eta squared η^2 (Levine et Hullett, 2002), one among these measures, is the proportion of the total variance that is attributed to an effect. It is calculated as the ratio of the effect variance (SS_{effect}) to the total variance (SS_{total}). As shown in Equation 8.7, η^2 can be interpreted as the ratio of variance explained by the factor of interest.

$$\eta^2 = \frac{SS_{effect}}{SS_{total}} \tag{8.7}$$

[2]The null hypothesis for ANOVA is that the mean (average value of the dependent variable) is the same for all groups. The alternative hypothesis is that the average is not the same for all groups. The null hypothesis here should not be confused with the likelihood ratio defence hypothesis called sometimes the null hypothesis.

A larger value of Eta-squared, η^2, always indicates a stronger effect. A commonly used interpretation, mentioned in (Cohen, 1977, 1988) (pp. 283−287), is to refer to effect sizes as:

- Small when η^2 ≈1%.
- Medium when η^2 ≈6%.
- Large when η^2 ≈14% or higher.

8.2 Global performance of Fabiole I-vector system

In this thesis, ASpR systems are assessed essentially using C_{llr} as measure of accuracy as well as APE plots for intuitive interpretation. Other measures of accuracy such as EER and $minDCF$ are given for comparative purpose.

Figure 8.2 presents the target and non-target score distributions corresponding to Fabiole main protocol. We remind that this protocol uses approximately 160k target trials and 4.5 millions non-target trials.

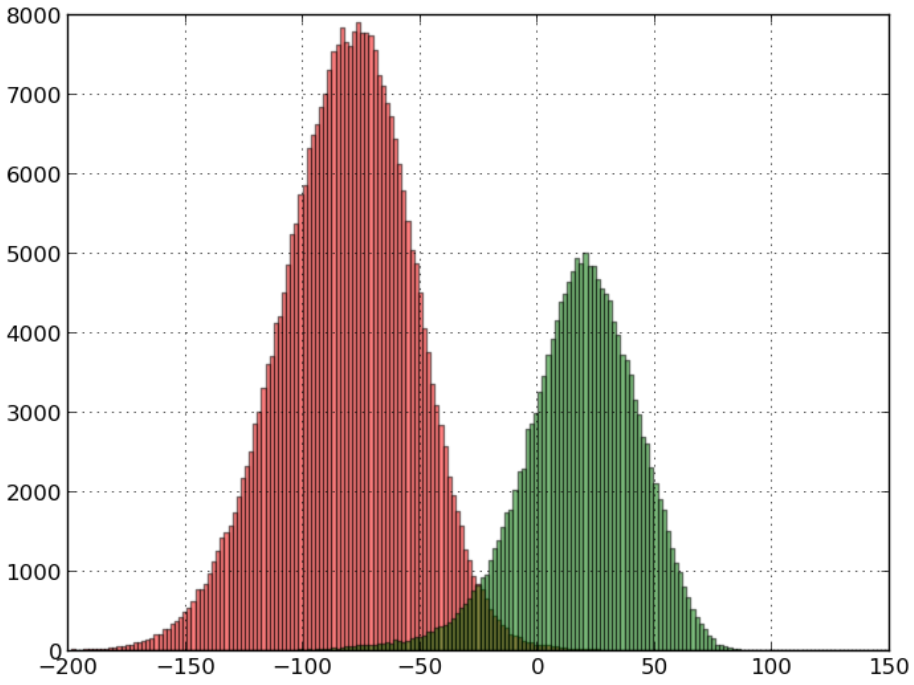

Figure 8.2: *Target (green) and non-target (red) score distributions for the pooled condition (all the comparison tests taken together).*

Figure 8.3 presents the Det plot. The global EER is equal to 2.88% (intersection of det curve with the bisector) and the minimum DCF is equal to 0.0276 (C_{Miss}=1, C_{FA}=1, P_{target}=0.5). It is important to notice that we obtain comparable performance to the state-of-the-art systems (Saedi et al., 2013; Bousquet et al., 2014). Indeed, across NIST SRE evaluation, SRE 2008, 2010 or 2012, best systems obtain an EER between 1% and 5%.

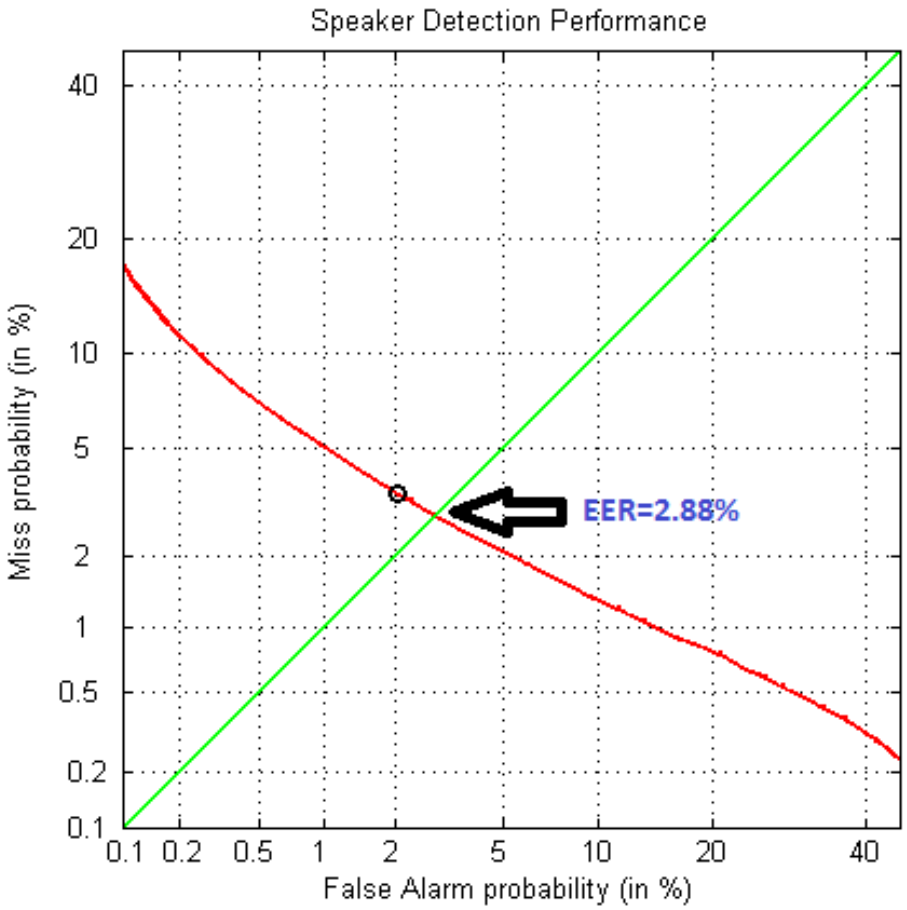

Figure 8.3: *DET plot of the pooled condition (All trials put together). Black circle represents the pair (FA, FR) corresponding to the minimum DCF.*

Figure 8.4, shows the APE and C_{llr} plots for the system using Fabiole main protocol. The global C_{llr} (respectively C_{llr}^{min}), computed using all the trial subsets put together, is equal to 0.12631 *bits* (respectively 0.11779 bits). From Figure 8.4, it can be seen that the LRs, estimated using i-vector PLDA system, show a good calibration. This is indicated by a fairly small calibration loss, 8.52×10^{-3} bits.

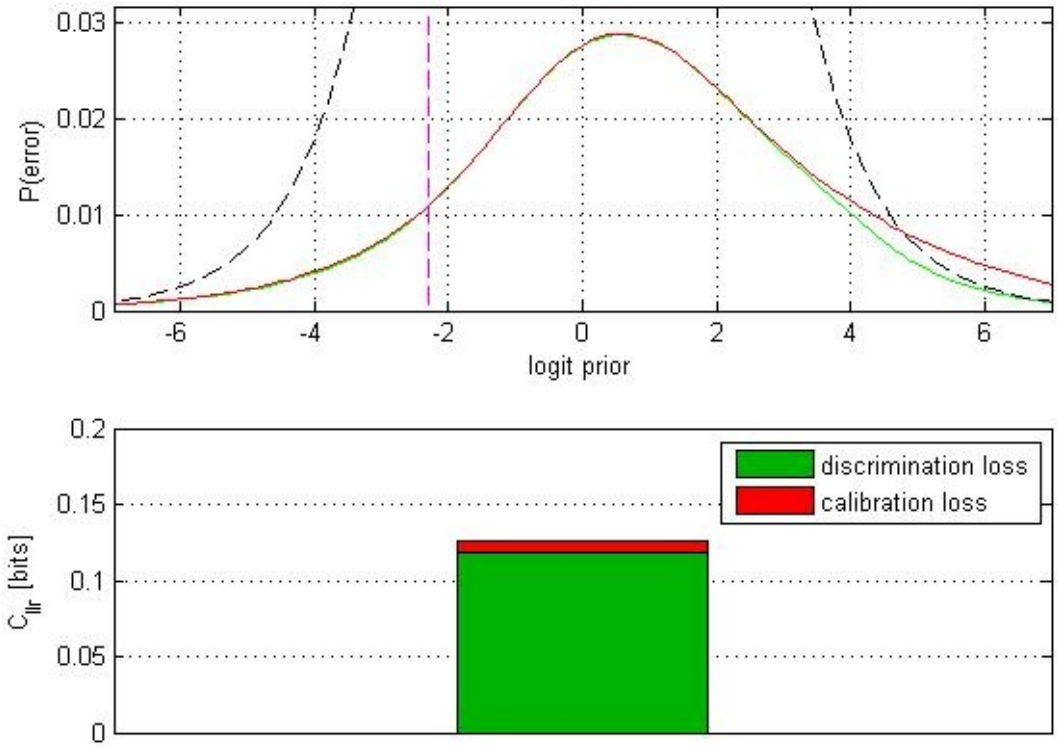

Figure 8.4: *APE and C_{llr} plots for the pooled conditions (all trials put together: Fabiole main protocol). It is important to bear in mind when reading the APE curve that: (i) The magenta dashed vertical line to indicate where the traditional NIST operating point is (at -2.29). (ii) The red and green error-rates at -2.29 are scaled versions of the traditional CDET and 'min CDET' values. The max of the green curve is also the EER. (iii) GREEN: (Minimum) total error-rate over the whole range of applications. This is the performance that the system under evaluation could have obtained with a perfect (for this data) score-to-llr calibration. (iv) RED: This is the area between the red and the green APE-curves and is the measure of how well the score to log-likelihood-ratio mapping is 'calibrated'.*

The low C_{llr}^{min} and EER values indicate that the system has a fairly good discrimination power.

8.3 Inter-speaker differences in genuine and impostor likelihood ratios

The global representation of the system performance, presented in the previous subsection, hides the impact of the inter-speaker differences on performance. In order to highlight the variations linked to the speaker, in Figure 8.5 we present C_{llr} estimated

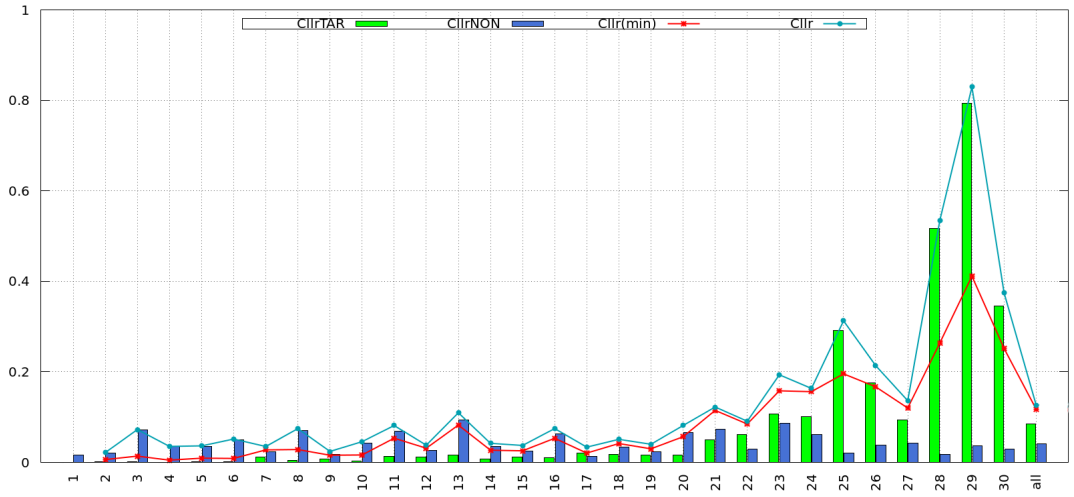

Figure 8.5: C_{llr}, C_{llr}^{min}, C_{llr}^{TAR}, C_{llr}^{NON} *per speaker and for "all" (data from all the speakers are pooled together).*

individually for each T speaker (the results are presented following the same ranking as (Ajili et al., 2016a), which was based on general C_{llr} performance). In this figure, C_{llr} is divided into two components, C_{llr}^{TAR} and C_{llr}^{NON}, in order to quantify separately the information loss relative to target and non-target trials.

Figure 8.5 shows a high difference on performance depending on the speakers. Indeed, when the global C_{llr} is equal to 0.12631 bits, we observe that more than half of the speakers obtain low C_{llr} values (lower than 0.05 bits) while about 10% of the speakers present significantly higher costs (higher than 0.4 bits) compared to the average cost.

Figure 8.5 also shows that information loss related to non-target trials (measured by C_{llr}^{NON}) presents a quite small variation between speakers while there is a huge variation of the information loss related to target trials (measured by C_{llr}^{TAR}). The information loss coming from target trials is mainly responsible for the high reported costs obtained for some speakers. This result is clear if we look at speaker 28, 29 or 30 where C_{llr}^{TAR} explains more than 90% from the total cost.

In order to illustrate the large difference of speaker behaviours, we present in Figure 8.6 examples of speakers presenting the best and the worst c_{llr}, in a form corresponding to Figure 8.2. Here, speakers 1 and 29 from Figure 8.5 are selected. This figure shows that non-target score distributions corresponding to the "best" and the "worst" speakers are quite similar while the main difference is observed on the target distributions. Indeed, the "best" speaker presents a narrow target distribution while the "worst" speaker shows a wider one. This finding suggests that the latter speaker is accepting a high intra-speaker variability.

This experiment shows that even if the trial subsets are mainly similar regarding their recording conditions (number of recordings, duration, signal quality, channel variability, etc.), a large variability in terms of performance is present which means that

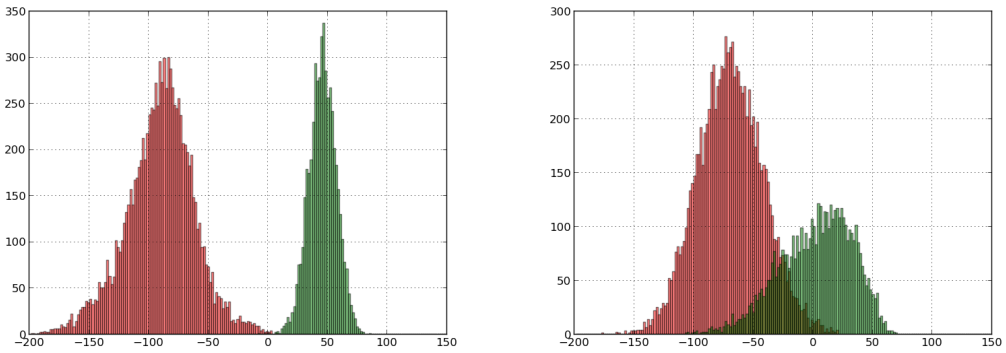

Figure 8.6: *Examples of target (green) and non-target (red) score distributions for the "best" speaker (left) and an the "worst" speaker (right) in term of performance.*

speakers do not behave the same way.

An analysis of variance (ANOVA) with the speaker as fixed factor and C_{llr} calculated for each trial, as the dependent variable, shows that the differences observed between speakers in C_{llr} are significant with a p-value <0.001. The speaker factor explains about 32.4% and 6% of the variance of, respectively, C_{llr}^{TAR} and C_{llr}^{NON} values. Results are summarized in Table 8.2. The effect of the speaker is large on target trials and it is medium for non-target trials.

Table 8.2: *Effect size of "speaker" on both target (TAR) and non-target (NON) comparisons explained in terms of, Eta-square η^2. (*) represents the significance level. "bold", "italic" represent respectively high and medium effect.*

	Speaker	**p-value**
TAR	**32.4**	***
NON	*6.0*	***

8.4 Speaker discrimination

In the previous subsection, we showed that non-target information loss presents a small variation between speakers compared to target information loss. For a deep investigation, this section is dedicated to study the speaker difference in performance when only a pair of speakers is involved.

8.4.1 Information loss relative to non-target trials

The C_{llr}^{NON} computed using all the non-target trial subsets put together is equal to 0.04bit. Averaging the results on all the speaker pairs could hide large differences be-

tween the speakers or the speaker pairs. In the following paragraphs, we investigate the variation of C_{llr}^{NON} between speakers and between speaker pairs.

- **One-to-many speaker comparison**

Let $\{spk_i\}_{i=1..n}$ be the set of Fabiole speakers. Each target speaker, spk_i, is compared with the rest of the speakers, $\{spk_j\}_{j\neq i}$, and the averaged information loss value is retained as described in Figure 8.7.

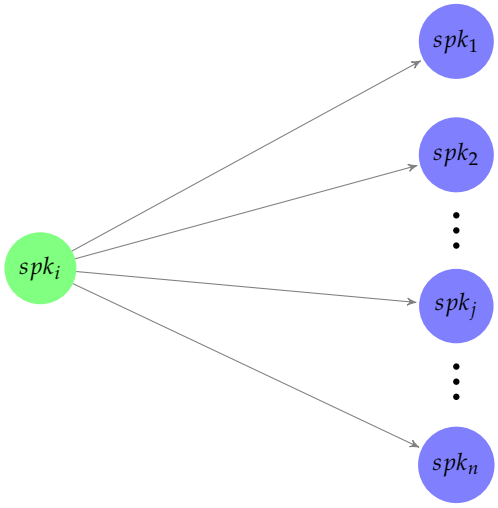

Figure 8.7: *One-to-many comparison: the speaker, spk_i is compared with $\{spk_j\}$ where $i \neq j$. The average information loss value, $\frac{1}{(n-1)}\sum_{j\neq i} C_{llr}^{NON}(spk_i, spk_j)$, is retained. C_{llr}^{NON} is computed using 290k trials.*

Figure 8.8 presents C_{llr}^{NON} estimated individually for each T speaker (the results are presented following the same ranking as (Ajili et al., 2016a)). Results show that C_{llr}^{NON} averaged per speaker present a significant variation across speakers: The lowest C_{llr}^{NON} value, $0.013bits$, is seen for speaker 17 while the highest value, $0.093bits$, is seen for speaker 13 (almost multiplied by 7).

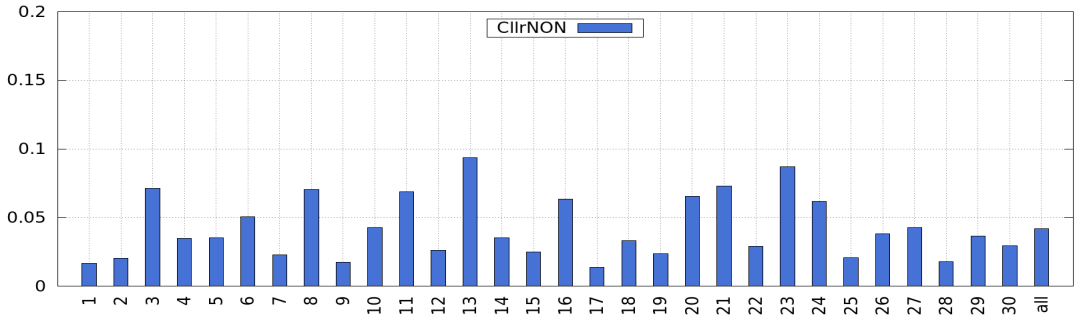

Figure 8.8: *Bar-plot of C_{llr}^{NON} per speaker and for "all" (all speakers pooled together) extracted from 8.5.*

- **One-to-one speaker comparison**

In this step, a target speaker, spk_i is compared with only one speaker, spk_j where $i \neq j$, as shown in Figure 8.9. Here, the information loss is investigated at the speaker pair level.

Figure 8.9: *One-to-one comparison: the speaker, spk_i is compared with $\{spk_j\}$ where $i \neq j$. The information loss value, $C_{llr}^{NON}(spk_i, spk_j)$, is retained. C_{llr}^{NON} is computed using 10k trials.*

The C_{llr}^{NON} is computed for all the pairs, C_{30}^{2}=435 pairs. Figure 8.10 presents the histogram of the resulting C_{llr}^{NON} values. Table 8.3 summarizes the distribution of the 435 speaker pairs according to their C_{llr}^{NON} values.

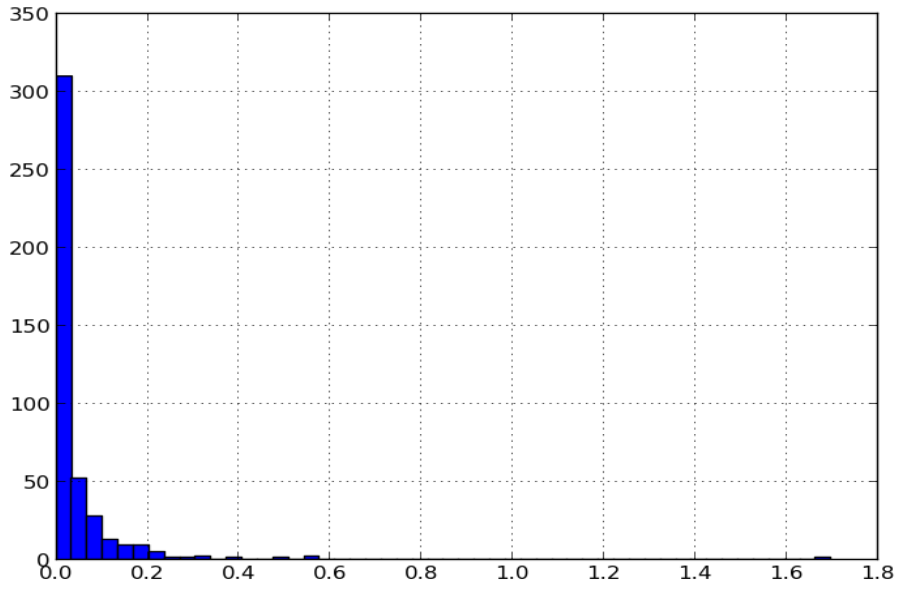

Figure 8.10: *Histogram of C_{llr}^{NON} of the 435 speaker pairs. Each C_{llr}^{NON} value is computed using 10k trials.*

Table 8.3: *Distribution of the 435 speaker pairs according to their C_{llr}^{NON} values.*

C_{llr}^{NON}	C_{llr}^{NON} <0.1	$0.1 < C_{llr}^{NON} < 0.2$	$0.2 < C_{llr}^{NON} < 0.3$	$0.3 < C_{llr}^{NON}$
Rate (%)	89.42	7.12	1.83	1.60

Table 8.3 and Figure 8.10 show that the vast majority of pairs (about 90%) presents a very low C_{llr}^{NON} (< 0.01). However, there exist few pairs who present a quite high C_{llr}^{NON} which can reach **1.67bits**.

In order to highlight this result, we present in Figure 8.11 the speaker confusion matrix, M_c. Each value of this matrix, $m_{i,j}$, corresponds to the C_{llr}^{NON} value calculated for the pair of speakers, spk_i and spk_j. This representation offers the possibility to visualize the "*magnitude*" of information loss of one speaker with regard to the other speakers.

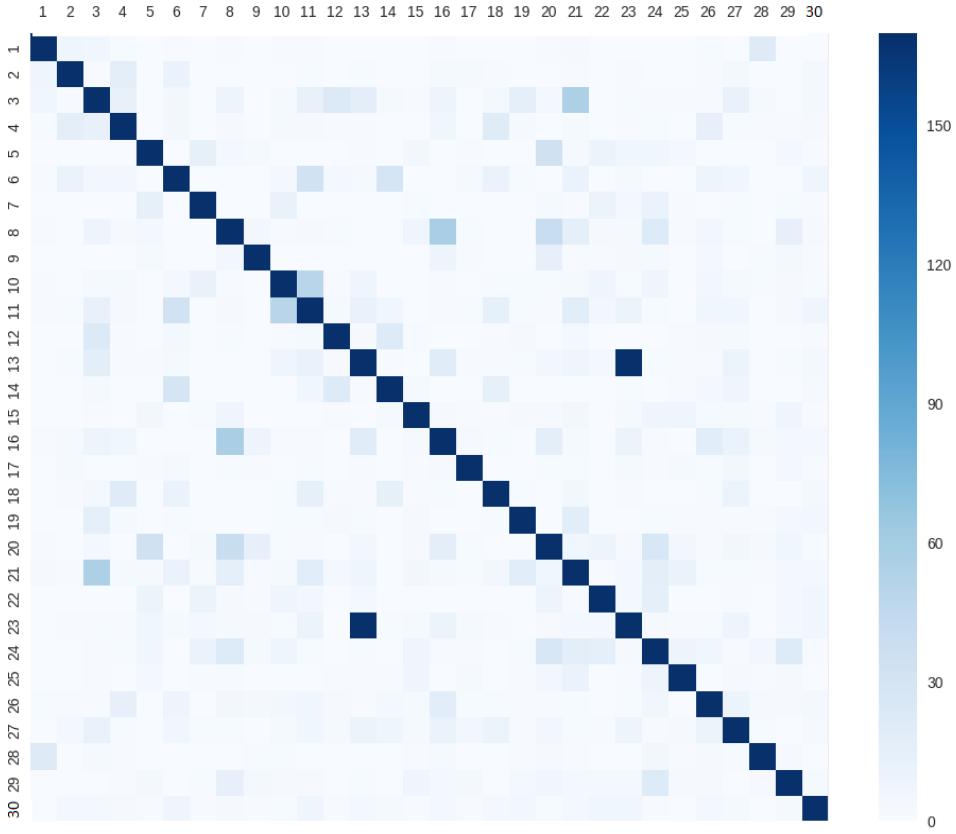

Figure 8.11: C_{llr}^{NON} *confusion matrix for the 30 speakers. The diagonal value is fixed to the maximum* C_{llr}^{NON} *value obtained across the speaker pairs.*

Figure 8.11 shows that each speaker is confused with the remaining speakers with different degrees. Indeed, each speaker is mainly confused with one or two speakers while being well discriminated with the remaining speakers. For instance:

- Speaker 13 appears to be highly confused with speaker 23 and at the same time has fairly low C_{llr}^{NON} across all the remaining speakers.
- Speaker 11 is highly confused with both speaker 10 and 6 (with lesser degree of confusion for speaker 6).

- Speaker 20 is mainly confused with speakers 8, 5 and 24 respectively.

8.4.2 Speaker clustering

In this subsection, we use firstly t-sne[3] algorithm (Maaten et Hinton, 2008) to visualize the speaker in the i-vector space. Figure 8.12 shows a 2-dimensional visualization of the speaker space.

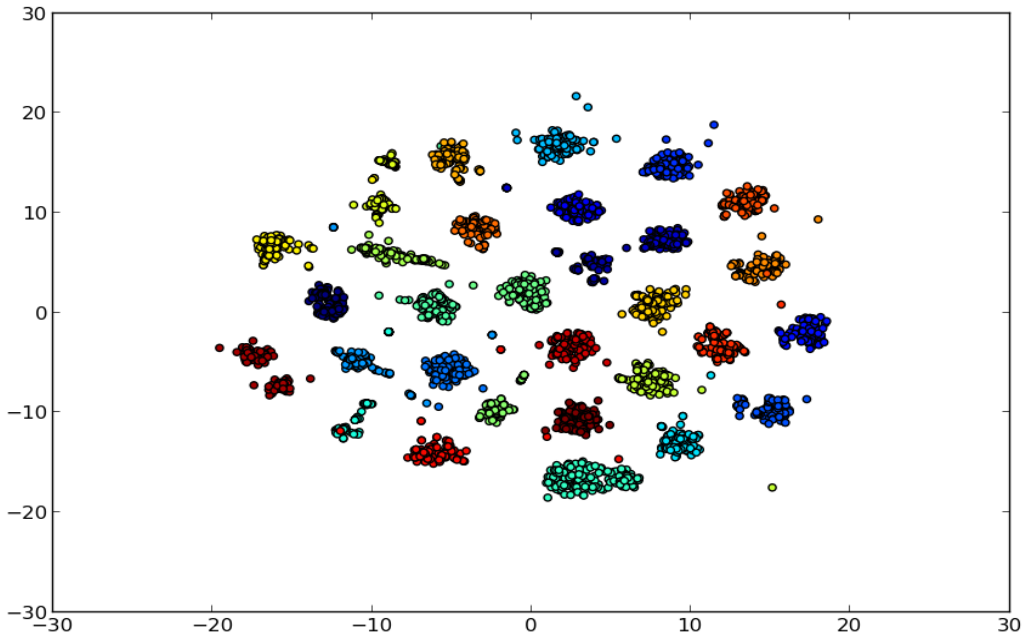

Figure 8.12: *A 2-dimensional visualization of speakers in the i-vector space. Each color represents a specific speaker.*

Figure 8.12 shows that i-vectors from the same speaker are usually close in the i-vector space, however, different speakers are quite tightly packed in this space.

In the following, we use the TCE distance in order to cluster the speaker according to their "*relative positions*". We apply a hierarchical clustering to the speaker dataset, using Ward's criterion (Ward Jr, 1963) which minimizes the loss of between-class variance at each step. Figure 8.13 shows the resulting clustering tree and Figure 8.14 displays the rate of between-class variance for each q-class set.

[3]T-sne is a popular tool in the machine learning community, and tends to cluster similar data points close together.

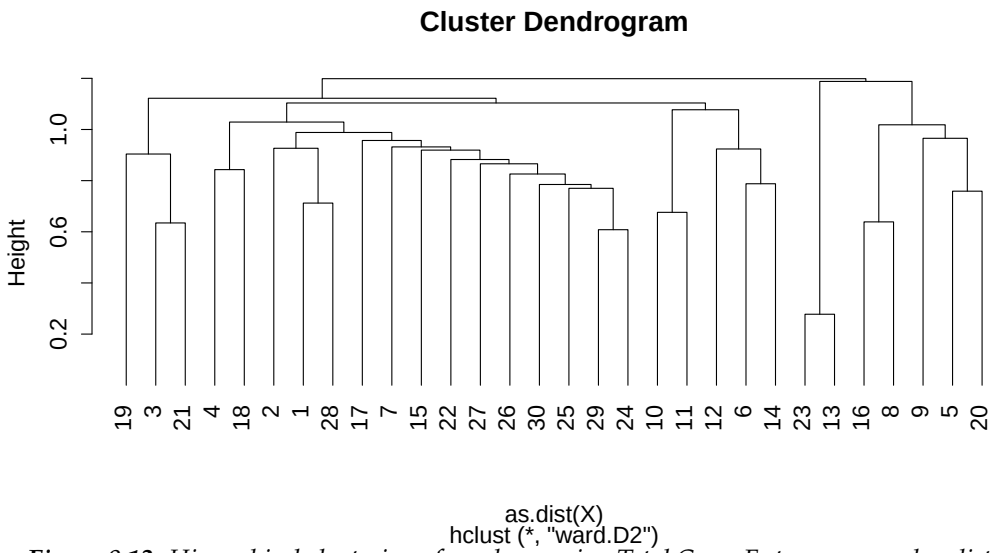

Figure 8.13: *Hierarchical clustering of speakers, using Total Cross Entropy as speaker distance.*

Figure 8.13 and 8.14 show that there is no clear/possible partition of the speaker set. Indeed, except some speakers who are close to each others such as $\text{spk}_{13} - \text{spk}_{23}$, the majority of the other speakers are well discriminated.

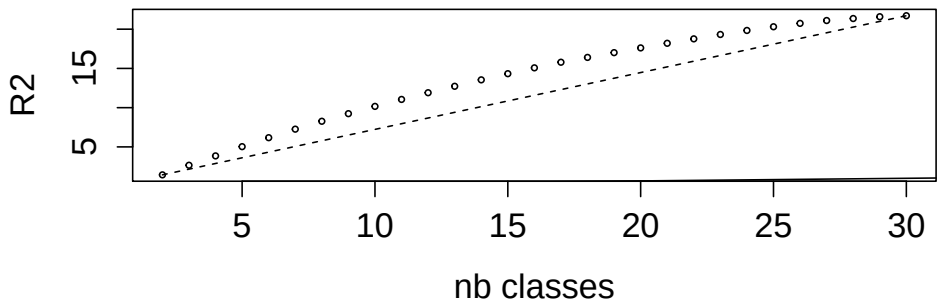

Figure 8.14: *Estimating the clustering ability of a dataset by using the area under curve (AUC). The AUC (Area Under Curve), also referred to as Gini, is the area between the curve and the bisector, equal to* 0 *if the set of observations is not classifiable (all the speakers are well discriminated) and maximal, equal to* 0.5*, if all speakers coincide.*

Conclusion

In this chapter, we overpassed the limits of the traditional evaluation protocols which look only at the global performance of an ASpR system. We decided to investigate deeper the variation of performance in a specific situation.

In a first step, we showed the importance of the "speaker factor": When the global C_{llr} (computed using all the trial subsets put together) is equal to 0.12631 bits, we observed that more than half of the speakers obtain low C_{llr} values (lower than 0.05 bits) and about 10% of the speakers present significantly higher costs (higher than 0.4*bits*) compared to the average cost.

In a second step, we pushed our analysis deeper by looking separately at the information loss relative to target and non-target LRs. These two losses were quantified based on C_{llr}^{TAR} and C_{llr}^{NON} respectively. We showed that LR target trials bring in general about two third of C_{llr} loss (0.67 vs 0.33). This proportion is significantly higher (up to 0.94 vs 0.06) for the speakers who present the largest C_{llr} loss contribution. On the other hand, we showed that the information loss for non-target LRs presents a quite small variation regarding speakers. The huge inter-speaker variation of the information loss related to target trials (measured by C_{llr}^{TAR}) suggests strongly the presence of a high intra-speaker variability effect in FVC. This factor should be taken into account in reliability evaluation of the LR.

Finally, we explored deeply the differences in performance by focusing on speaker pairs. We have shown that in our database the vast majority of pairs (more than 90%) presents a very low C_{llr}^{NON} ($<$0.01) while few pairs present a quite high C_{llr}^{NON}.

The main conclusion of this chapter is that averaging the system behaviour over a high number of trials hides potentially many important details. For a better understanding of an ASpR system, it is mandatory to explore the behaviour of the system at as-detailed-as-possible scale (*The devil lies in the details*).

Chapter 9

Phonological content analysis

Contents

Introduction

In state-of-the-art ASpR systems, like i-vector systems, a recording is encoded by a unique low dimensional vector. Such a system does not work on explicit information linked to the phonemic content or on speech specific cues: All the information is embedded in the low dimensional vector. However, several studies (Magrin-Chagnolleau et al., 1995; Besacier et al., 2000; Amino et al., 2006; Antal et Toderean, 2006) agree that

speaker specific information is not equally distributed inside the speech signal and particularly depends on the phoneme distribution. Despite its apparent richness, the above literature review reveals different lacks. First, the majority of the quoted research works are dedicated to ASpR and do not take into account the specific context of FVC. Second, they do not take into account enough intra-speaker variability, mainly due to a lack in terms of databases (a small number of available recordings per speaker).

This chapter investigates the impact of phonetic content on voice comparison process. We propose to analyse whether certain classes of phonemes are bringing more speaker discrimination information than others and if these differences are stable across the speakers. We wish also to investigate the impact of phonemic categories on both intra- and inter-speaker variability.

This chapter is organized as follows: Section 9.1 presents the methodology and protocols we adopted in this investigation. Section 9.2 presents the study of the impact of phonemic categories on FVC. Section 9.3 studies the impact of the phonemic categories at the speaker pairs level. In Section 9.4, we redo the same analysis, as in Section 9.2, in the wideband context.

9.1 Methodology and protocols

In this subsection, we propose to classify the speech content into phonemic classes. Then, we describe the adopted protocol used to investigate the impact of each class on FVC.

9.1.1 Phonemic categories

To conduct our work, we propose to use phoneme classes in place of individual phonemes. Working on phoneme classes presents two main advantages in the context of our study. First, a phoneme transcription/alignment process is mandatory and always accept imprecisions and errors. If the classification is well chosen, the use of phoneme classes will allow to reduce the effect of these errors. Second, the speech extracts involved in FVC trials are usually of a relatively short duration. To work at phoneme level presents a risk of piecemeal or inconsistent results, due to insufficient amount of speech material for some phonemes. Working with a short set of phoneme classes will allow to overcome this risk. In this work, we propose to classify the speech content into 6 phoneme categories based on phonological features. The phoneme classification is described below:

- Oral vowels (OV) which includes /i/, /y/, /u/, /e/, /ø/, /o/,/ɛ/, /œ/,/ɔ/, /a/.
- Nasal vowels (NV) which includes /ɑ̃/, /ɔ̃/, /œ̃/, /ɛ̃/.
- Nasal consonants (NC) which includes /m/, /n/.

- Plosive (P) which includes /p/, /t/, /k/, /b/, /d/, /g/.
- Fricatives (F) which includes /f/, /s/, /ʃ/, /v/, /z/, /ʒ/.
- Liquids (L) which includes /l/, /ʁ/.

9.1.2 Phoneme filtering protocol

In order to study the influence of a specific phonemic class, we use a knock-out strategy: the in-interest information is withdrawn from the trials and the amount of performance loss indicates the influence of the corresponding speech material. So, we perform several experiments where the speech material corresponding to a given class is removed from the two speech recordings of each trial. This condition is denoted here "**Specific**". Since the amount of speech material is largely unbalanced (for example, in our experiments, nasal consonants represent 6% of the speech material and oral vowels 36%), in order to avoid a potential bias, we create a control condition denoted "**Random**", where the corresponding amount of speech material is randomly withdrawn. More precisely, for each speech signal, when a certain percentage of speech frames is withdrawn for the "**Specific**" condition, the same percentage of frames is randomly withdrawn for the "**Random**" condition. This process is repeated 20 times (giving 20 times more trials in "**Random**" condition than in "**Specific**" one).

The impact of a specific phonemic class is quantified by estimating the relative C_{llr}^{R} given by Equation 9.1.

$$C_{llr}^{R} = \frac{C_{llr}^{random} - C_{llr}^{specific}}{C_{llr}^{random}} \times 100\% \tag{9.1}$$

A positive value of C_{llr}^{R} indicates that the speech material related to the corresponding phonemic class brings a larger part of the speaker-discriminant loss than averaged speech material. A negative value says the opposite: the corresponding phonemic class reduces the discrimination loss compared to averaged phonemic content. An illustration of the phoneme filtering protocol is given in Figure 9.1.

9.2 Phonemic content impact in FVC

In this section, the study of the phonemic category impact on FVC is carried at three levels: The first level focuses on both target and non-target trials. The second and the third levels are dedicated to study the phonemic content impact separately on target and non-target trials. It is important to remind that we discuss here results obtained using an ASpR system as a measurement instrument. Therefore, it is not possible to discriminate between the intrinsic characteristics of a cue and the way that this cue is taken into account by an ASpR system.

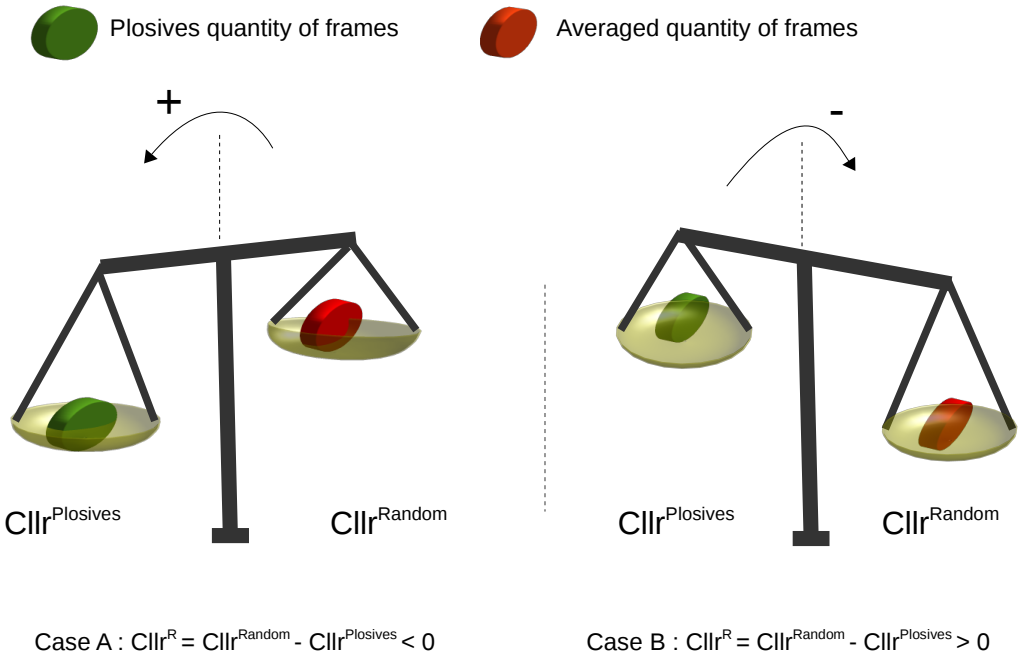

Figure 9.1: *Phoneme filtering protocol illustration when the specific class is plosives. Two cases, A and B, are presented. Case A when the withdrawal of plosives shows a higher information loss than an averaged phonemic content: plosives are more speaker-specific information than an averaged phonemic content. Case B when the withdrawal of plosives shows a lower information loss than an averaged phonemic content: plosives are less speaker-specific information than an averaged phonemic content. The same reasoning could be done in which C_{llr}^{Random} is considered as the baseline information loss. "+" indicates that plosives have a positive role in FVC while "-" indicates that plosives have a negative role in FVC.*

9.2.1 Global effect

Table 9.1 shows the impact of the 6 phonemic categories on C_{llr} for "Specific" and "Random" conditions (C_{llr}^{min} results are also provided for comparison purposes). It gives also the amount of speech frames per phoneme class (mean and deviation over the trials). The results are given using the "pooled" condition (averaged on all the speakers). A large variation is observed between the phonemic classes: to withdraw nasal vowels (NV), nasal consonants (NC) or oral vowels (OV) leads to a loss of information compared to the "**Random**" case while the withdrawal of plosive (P), liquid (L) and fricative (F) does not seem to have an influence on the system accuracy.

This outcome corroborates results of (Amino et al., 2006; Antal et Toderean, 2006; Amino et al., 2012; Eatock et Mason, 1994), where nasals and vowels are found to be particularly speaker specific information. More precisely, nasal vowels appear in this study to be more informative than oral vowels. The result we obtained for the fricatives -the class exhibits low speaker discriminating properties- is clearly in conflict with (Gallardo et al., 2014b)'s finding. An explanation could be that (Gallardo et al., 2014b) uses a wide-band while in this investigation a narrow band (300-3400 Hz) is applied.

Table 9.1: *C_{llr} and C_{llr}^{min} for "Specific" and "Random" conditions (baseline results are: C_{llr}=0.12631 and C_{llr}^{min}=0.11779. Mean and SD of the duration per class are provided.*

	C_{llr}		C_{llr}^{min}		Duration (s)	
Category	**Withdrawn**		**Withdrawn**		**Mean**	**SD**
	Specific	**Random**	**Specific**	**Random**		
NV	0.14689	0.12941	0.13498	0.11975	3.14	1.56
NC	0.13713	0.12815	0.12728	0.11897	2.05	1.03
OV	0.15396	0.14689	0.14601	0.12819	13.00	5.50
L	0.12966	0.13032	0.12173	0.12029	4.03	1.96
P	0.13278	0.13431	0.12244	0.12228	7.72	3.40
F	0.12703	0.13238	0.12007	0.12135	5.84	2.68

Figure 9.2 is a stacked bar chart which shows the contribution of each phonemic class to the C_{llr}^{R}, depending on the speaker. The general tendency shown in Table 9.1 appears clearly here: C_{llr}^{R} results for nasal vowels and nasal consonants are negative and indicate that their withdrawal brings generally a degradation of FVC performance. But a large variability depending on the speaker is also present for all the phonemic classes. For example, speaker 2 shows a C_{llr}^{R} of $\sim -175\%$ when oral vowels are withdrawn while speaker 28 shows a C_{llr}^{R} of about 40% in a similar situation. Another time, the global tendencies are shadowing potential speaker-specific effects.

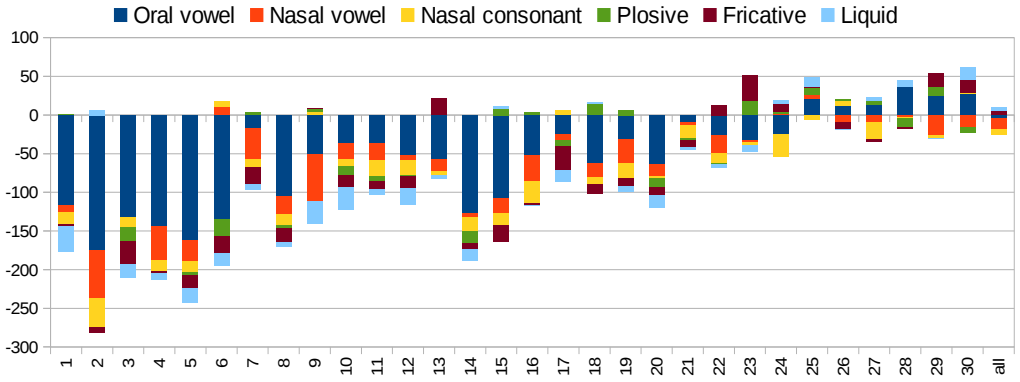

Figure 9.2: *Stacked bar chart of C_{llr}^{R} (computed on both target- and non-target trials) per speaker and for "all".*

These results reinforce our previously claim (in chapter 8), that -in the view of an ASpR-based voice comparison system- all the speakers do not behave the same way in response of similar condition changes. Indeed, the impact of a phonemic category on FVC depends on the speaker himself.

9.2.2 Phonemic content impact on FVC for non-target comparisons

Figure 9.3 presents a stacked bar chart which displays the impact of the phonemic classes per speaker, in terms of relative C_{llr}^{NON} (C_{llr}^{R} computed on C_{llr}^{NON}). The six phonemic classes appear to embed speaker discrimination power since ,in almost all the cases, their absence leads to a raise of C_{llr} value (i.e, a raise of the information loss) compared to the "Random" case. To withdraw oral vowels causes the largest accuracy loss, ranking in top this phonemic class in terms of speaker discrimination power with a large margin with the next class. After the oral vowels, nasals, vowels first and consonants second, appear to convey the most discrimination power. Liquid, fricative and plosive obtain similar results, at the lowest end of the speaker discrimination scale. The results are quite consistent between the 30 target speakers, with limited variations.

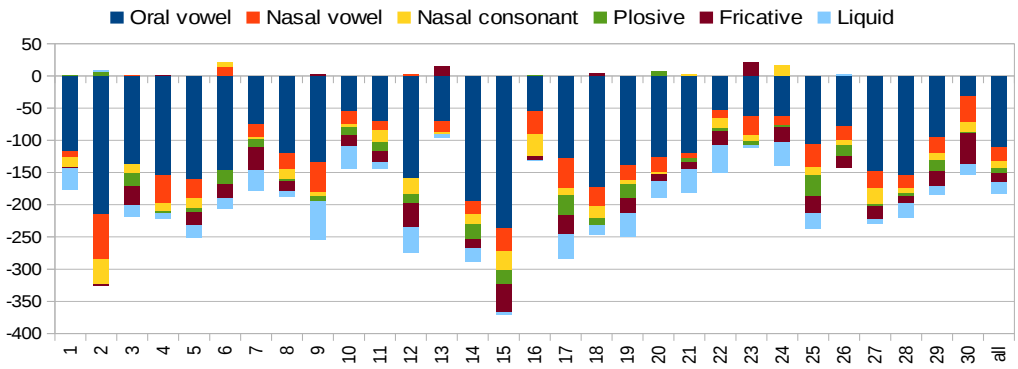

Figure 9.3: *Stacked bar chart of C_{llr}^{R} computed on C_{llr}^{NON} (non-target trials) per speaker and for "all".*

9.2.3 Phonemic content impact on FVC for target comparisons

Figure 9.4 presents the impact of the phonemic classes per speaker, in terms of relative C_{llr}^{TAR} (C_{llr}^{R} computed on C_{llr}^{TAR}), using a similar form than Figure 9.3. The first outcome differs significantly from the previous case: The withdrawal of oral vowels from the recordings leads to a decrease of C_{llr} value (i.e a decrease of the information loss) compared with the "random" case. For target trials, oral vowels are tied with C_{llr} degradation for about 70% of the speakers. Fricative, liquid and plosive classes have a similar behaviour than oral vowels. Instead, the nasals (and particularly the nasal vowels) still play a positive role for voice comparison: withdrawing these phonemes increases the C_{llr} in most cases.

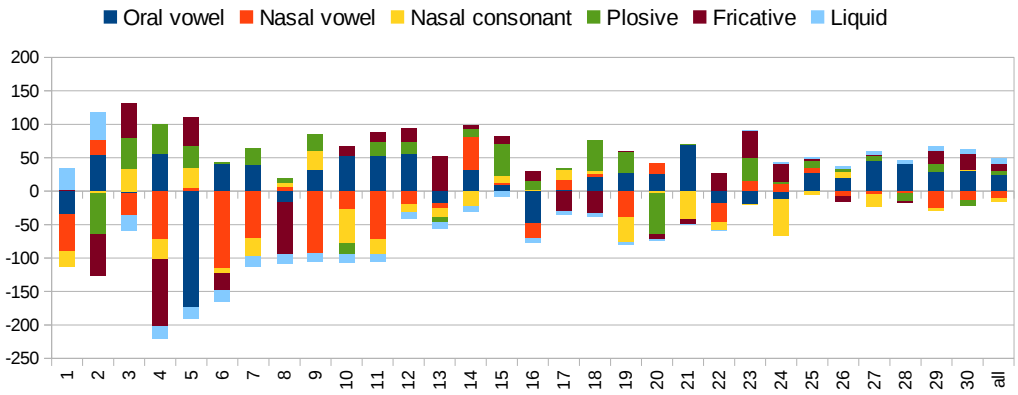

Figure 9.4: *Stacked bar chart of* C_{llr}^{R} *computed on* C_{llr}^{TAR} *(target trials) per speaker and for "all".*

Comments

Taken together, the results using relative C_{llr}^{TAR} and relative C_{llr}^{NON} bring to us some remarks:

- The nasal phonemes effectiveness for speaker comparison could be explained by the important contribution of nasal and paranasal cavities. This morphological aspect is difficult to control unintentionally or voluntarily and therefore allows low within-speaker variability (Stevens, 1999; Schindler et Draxler, 2013).
- A same phonemic class, the oral vowels, brings the largest part in terms of speaker discrimination but presents in the same time a large intra-speaker variability which conveys a significant part of the LR performance loss.

Remarks

In Figure 8.5, we showed that C_{llr}^{TAR} (may reflect intra-speaker variability) brings in general about two third of C_{llr} loss (0.66 vs 0.33). This proportion is significantly higher (up to 0.94 vs 0.06) for the speakers who present the largest contribution to the C_{llr} loss. It is interesting to link this finding with two facts: almost all studied phoneme classes are helping for speaker discrimination for all the speakers (Figure 9.3); all the speakers accept some phoneme classes which are degrading the target part of C_{llr} when some other classes are performing well (Figure 9.4). For the latter remark, it is interesting to notice that the same phoneme class could have a very different behavior depending on the speaker.

9.2.4 Statistical test evaluation

The difference between speakers in term of C_{llr} is large (Speaker 1 and speaker 29 have respectively a C_{llr} of about 10^{-5}bits and 0.8bits). A same relative C_{llr} value, C_{llr}^{R}, corre-

sponds to very different information losses (bits) according to two different speakers. For example, a difference of 10% of relative C_{llr}^{R} corresponds to 10^{-5}bits of performance loss for Speaker 1 while for Speaker 29, it corresponds to more than 0.1 bits. In order to investigate the significance of the differences observed when a specific phonemic category is withdrawn, we use here $\Delta C_{llr} = C_{llr}^{random} - C_{llr}^{specific}$ as an "*absolute*" measure of information loss/win between "random" and "specific" cases. Figure 9.5 presents ΔC_{llr} corresponding to the six phonemic categories for both target and non-target trials.

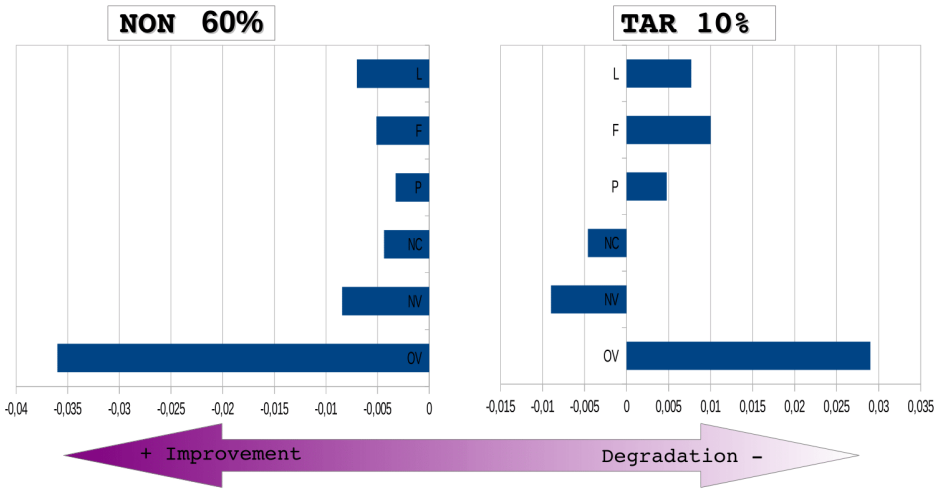

Figure 9.5: *ΔC_{llr} between the phonemic categories for target and non-target trials. "+" indicates the positive role of a specific phonemic category while "-" indicates a negative role of a specific phonemic category.*

An ANOVA with the phonemic category as fixed factor and the $\Delta C_{llr} = C_{llr}^{random} - C_{llr}^{specific}$ as the dependent variable is performed. Results are reported in Table 9.2.

Table 9.2: *Effect size of phonemic categories on both target (TAR) and non-target (NON) comparisons explained in terms of, Eta-square η^2. (*) represents the significance level. "bold", "italic" represent respectively high and medium effect.*

	Category	**p-value**
TAR	*10.2*	**
NON	**59.9**	***

Table 9.2 shows that differences observed on ΔC_{llr} between the phonemic categories are significant for both non-target trials and target trials and are more significant for non-target trials. Moreover, the phonemic category explains about 60% of the variance of ΔC_{llr} on non-target trials and 10.2% for target trials, thus indicating a large effect on non-target trials and a medium effect on target ones.

9.3 Phonemic content impact on speaker pairs

In this subsection, we look at the impact of phonological content for speaker pairs. Two subsets of speaker pairs are selected, according to speaker discrimination power, in order to better visualize our results. Figure 9.6 uses a form similar to Figure 9.3 in order to display the impact of the phonemic category for the 10 "best" speaker pairs in term of C_{llr}^{NON} value. Respectively Figure 9.7 displays the impact of the phonemic category for the 10 "worst" speaker pairs in term of C_{llr}^{NON} value.

Figure 9.6 shows that the six phonemic categories embed a speaker discrimination power of different extent. More precisely, oral vowels appear to convey the most important part of speaker-specific cues: the withdrawal of these phonemes causes the largest accuracy loss as shown for example for the pair "5-2" or "2-7". Another observation of note is that phonological information used to discriminate a pair of speakers depends on the speakers themselves. For example, for the pairs "1-7" or "1-22", fricatives appears to embed the largest speaker discrimination power.

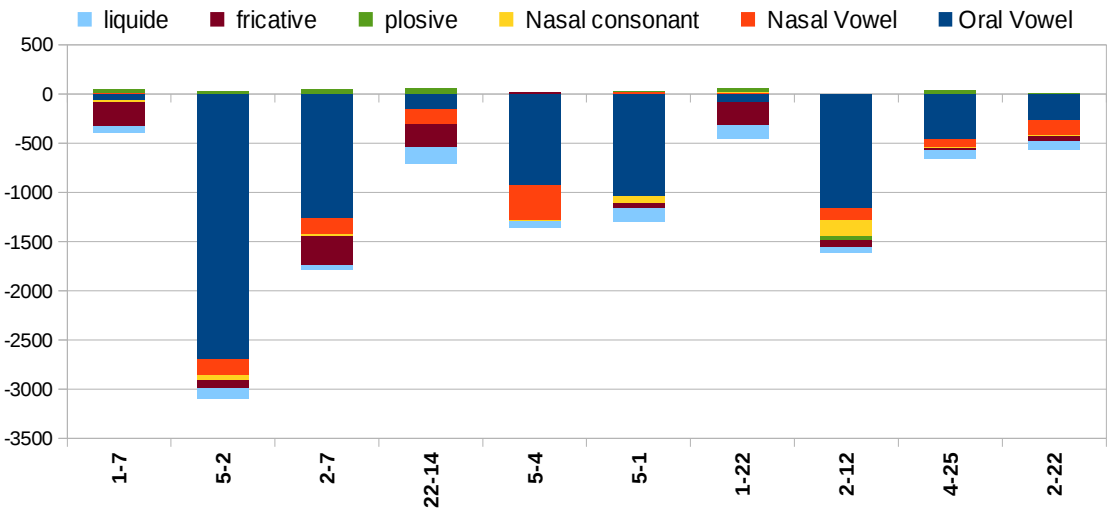

Figure 9.6: *Stacked bar chart of relative* C_{llr}^{NON} *computed on the 10 "best" speaker pairs,* spk^i-spk^j*, in term of discrimination power (Averaged* C_{llr}^{NON}=5.10^{-5}bits*).*

Figure 9.7 shows different outcomes: even if almost all phonemic categories still play a positive role in speaker discrimination, withdrawing nasals, vowels or consonant, from the recordings leads, surprisingly, to an improvement of the accuracy for most of speaker pairs. For example, the pairs "24-20" and "3-21", show a relative win of 40% and 25% respectively when nasals are withdrawn. This finding could be explained by the hypothesis of a nasal signature (Wright, 1986; Rossato et al., 1998) which corresponds to the transfer function of nasal cavities and sinuses. This nasal signature reflects mainly anatomical differences, as the speaker can only connect or not theses cavities to the vocal tract, without any controlled changes on them (Dang et Honda, 1996). Despite such inter-speaker anatomical differences, it may be possible that, for a pair of speakers, both acoustic spectra be similar. On a mathematical point of view, the

question was asked for a 2-D resonator in (Kac, 1966) and answered in (Gordon et al., 1992) where authors found two different shapes with the same acoustic spectra.

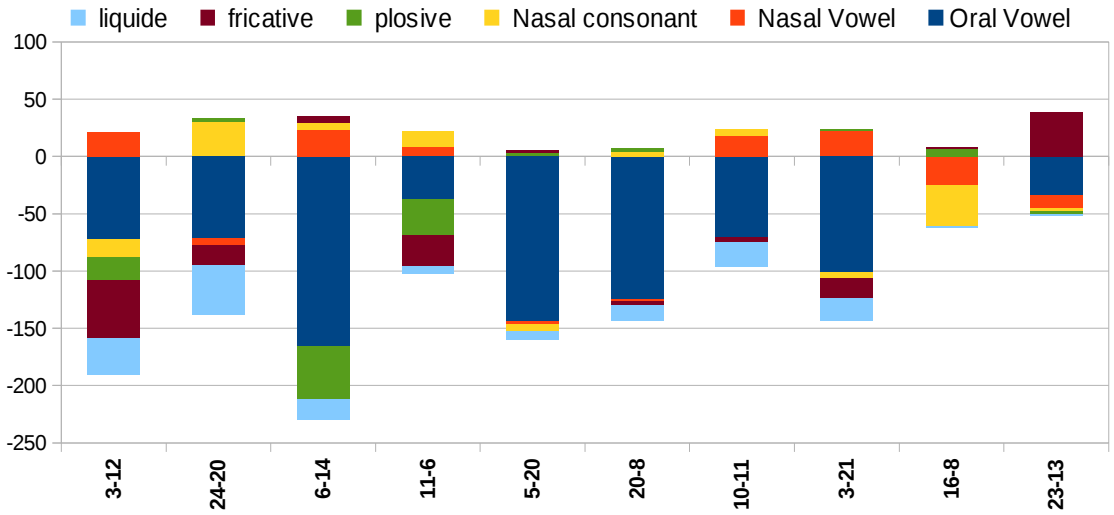

Figure 9.7: *Stacked bar chart of relative* C_{llr}^{NON} *computed on the 10 "worst" speaker pairs,* spk^i-spk^j*, in term of discrimination power (Averaged* C_{llr}^{NON}=0.52bits*).*

For the pair "23-13", fricatives appear to convey a significant part of the LR performance loss. This could be explained by the use of a narrow band which exclude fricative's speaker-specific-cues in high frequencies. Another time, the global tendencies are shadowing potential speaker-specific effects when only two speakers are involved.

9.4 Influence of Band-width in Forensic Voice Comparison

This subsection is mainly dedicated to study the phonological content impact of both voice recordings on FVC when a large bandwidth is used. This analysis follows the same logic as the one presented in the section 9.2.

9.4.1 Inter-speaker differences in genuine and impostor LRs

We repeat the experiments using a PLDA i-vector system but using a full band parametrization (0-8000 Hz). The global C_{llr} (computed using all the trial subsets put together) is equal to 0.10941bits and the corresponding global EER is 2.54%. Compared to the system where a narrow band (300-3400 Hz) is applied (EER is 2.88%), using the full band shows better performance, as expected.

In the following, all the new results are obtained using PLDA i-vector system and 0-8000 Hz bandwidth.

Figure 9.8 presents C_{llr} estimated individually for each *T* speaker. Results for Speaker 16 will not be discussed as some recordings are filtered 0-4 kHz. In this figure, C_{llr} is divided into C_{llr}^{TAR} and C_{llr}^{NON}. Compared with Figure 8.5, we obtain similar results:

- A deeper look at the relative weight of target and non-target trials in the global C_{llr} shows that target trials bring in general about two third of C_{llr} loss (0.67 vs 0.33). This proportion is significantly higher (up to 0.94 vs 0.06) for the speakers who present the largest C_{llr} loss contribution.
- Results also show that information loss related to non-target trials presents a quite small variation regarding speakers while there is a huge inter-speaker variation of the information loss related to target trials.

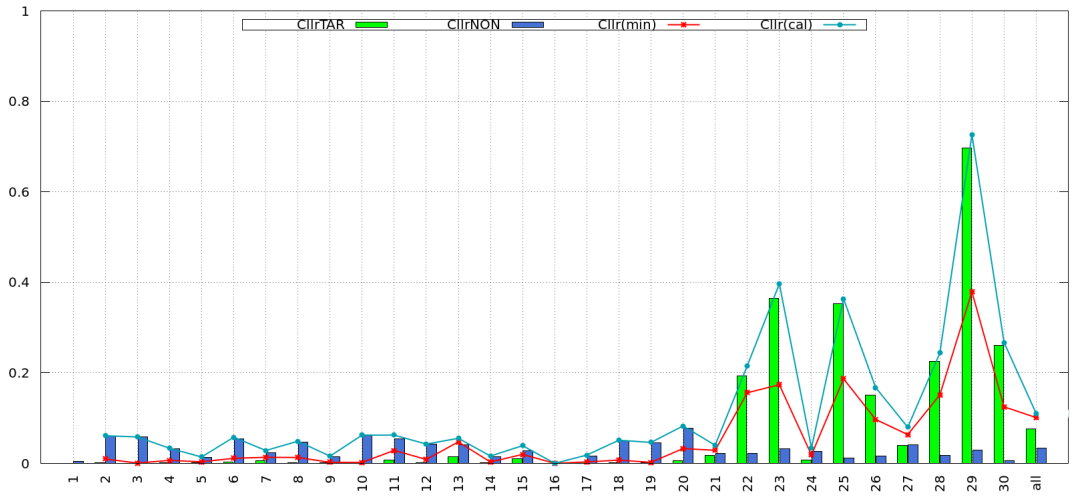

Figure 9.8: C_{llr}, C_{llr}^{min}, C_{llr}^{TAR}, C_{llr}^{NON} *per speaker and for "all" (data from all the speakers are pooled together). Experiments are performed taking advantage of the full bandwidth (0-8000 Hz).*

9.4.2 Phonemic impact on FVC

Table 9.3 shows the impact of the six phonemic categories on C_{llr} for "Specific" and "Random" conditions (C_{llr}^{min} results are also provided for comparison purposes). It gives also the amount of speech frames per phoneme class (mean and deviation over the trials). The results are given using the "pooled" condition (averaged on all the speakers). A large variation is observed between the phonemic classes: to withdraw nasals, vowels first then consonants, and fricatives leads to a significant loss of information compared to the "**Random**" case. To withdraw plosive leads to a small degradation compared to "**Random**" case while the absence of liquid or oral vowels does not seem to have an influence on the system accuracy.

Figure 9.9 is a stacked bar chart which shows the contribution of each phonemic class to the C_{llr}^{R}, depending on the speaker. The same general tendency than in table 9.3 appears clearly: The C_{llr}^{R} for nasal vowels and nasal consonants are negative and indi-

cates that their absence brings generally a degradation of FVC performance. The same behaviour is observed for fricative. But a large variability depending on the speaker is also present for all the phonemic classes. For example, speaker 5 shows a relative loss of 150% when oral vowels are withdrawn while speaker 30 shows a relative win of about 40% in a similar situation.

Table 9.3: *C_{llr} and C_{llr}^{min} for "Specific" and "Random" conditions using the full bandwidth (0-8000 Hz) (baseline results are: C_{llr}=0.10941 and C_{llr}^{min}=0.10169). Mean and SD of the duration per class are provided.*

	C_{llr}		C_{llr}^{min}		Duration (s)	
Category	**Withdrawn**		**Withdrawn**		**Mean**	**SD**
	Specific	**Random**	**Specific**	**Random**		
NV	0.11878	0.11119	0.10863	0.10276	3.14	1.56
NC	0.11508	0.11041	0.10609	0.10227	2.05	1.03
F	0.11738	0.11334	0.10713	0.10394	5.84	2.68
P	0.11552	0.11434	0.10617	0.10432	7.72	3.40
OV	0.11845	0.12216	0.11127	0.10824	13.00	5.50
L	0.11028	0.11141	0.10310	0.10285	4.03	1.96

Nasals and fricatives appear to convey a high speaker-specific information. This finding is coherent with (Amino et al., 2006) and (Gallardo et al., 2014b). The positive effect of nasals on FVC is also observed when a wide band is applied but for fricatives the situation is different: Fricatives only appear to be speaker specific information when a wide band is applied while when using a narrow band fricative bring a significant part of the LRs degradation.

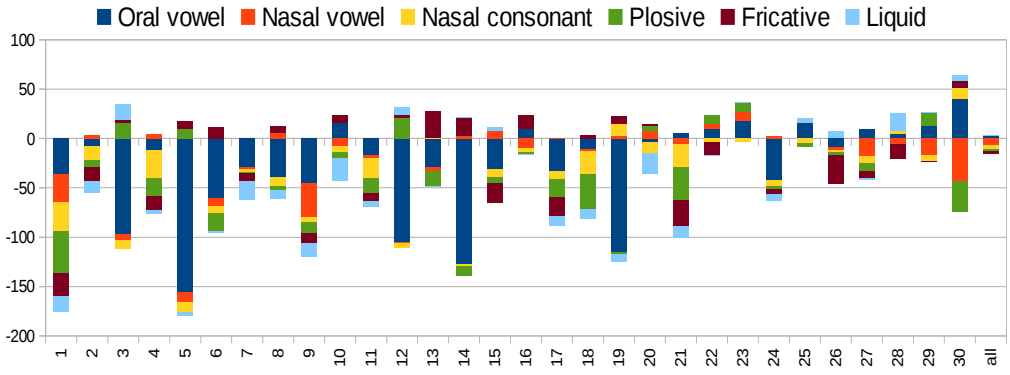

Figure 9.9: *Stacked bar chart of C_{llr}^{R} computed on C_{llr} per speaker and for "all". Experiments are performed taking advantage of the full bandwidth (0-8000 Hz).*

Results on oral vowels is in conflict with (Amino et al., 2006). To interpret this result,

a deeper look on target and non-target trials is mandatory.

9.4.3 Phonemic content impact on FVC for non-target comparisons in wide-band context

Figure 9.10 is a stacked bar chart which displays the impact of the phonemic classes per speaker, in terms of relative C_{llr}^{NON} (C_{llr}^{R} computed on C_{llr}^{NON}).

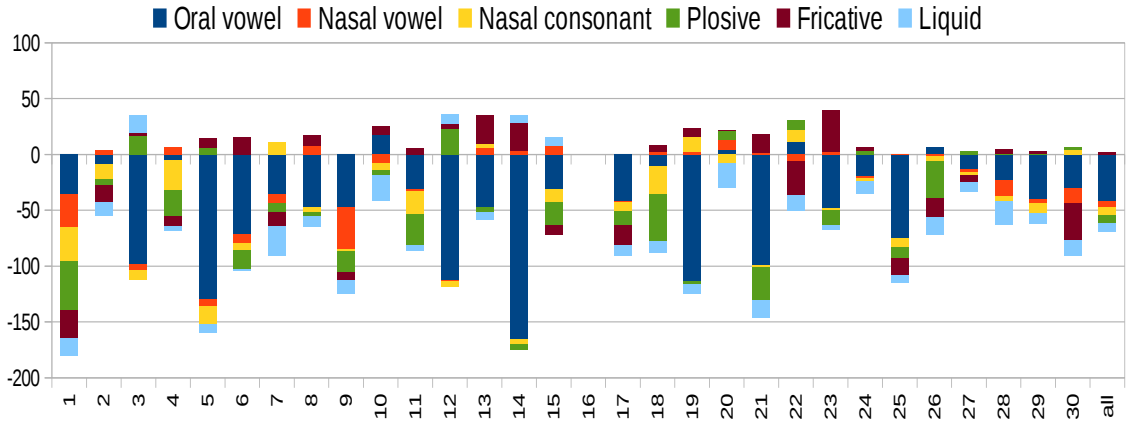

Figure 9.10: *Stacked bar chart of C_{llr}^{R} computed on C_{llr}^{NON} (non-target trials) per speaker and for "all". Experiments are performed taking advantage of the full bandwidth (0-8000 Hz).*

All the phonemic classes, except fricatives, appear to embed speaker discrimination power since their absence leads, in almost all the cases, to a degradation in C_{llr}. To withdraw the oral vowels causes the largest accuracy loss,. This phonemic class is ranked in the top in terms of speaker discrimination power with a large margin with the next classes. Nasals together, consonants first and vowels second, appear to convey the most discrimination power after the oral vowels. Liquids and plosives obtain similar results, at the end of the speaker discrimination power scale. The fricative class appears to be responsible of a small part of LR losses (for "all" and for some speakers such as speaker 23 where withdrawing fricatives leads to a relative win of 40%) since its absence leads to an improvement of the system accuracy. The results present a high variability across speakers (This time, oral vowels is not always the best class in term of speaker discrimination). For example, for speaker 1, nasals are the best class while for speaker 22, fricatives appear to convey the largest part of speaker specific information.

9.4.4 Phonemic impact on FVC for target comparisons in wideband context

Figure 9.11 presents the impact of the phonemic classes per speaker, in terms of relative C_{llr}^{TAR} (C_{llr}^{R} computed on C_{llr}^{TAR}). The first outcome differs significantly from the previous case: to withdraw the oral vowels from the recordings leads to an improvement of C_{llr}

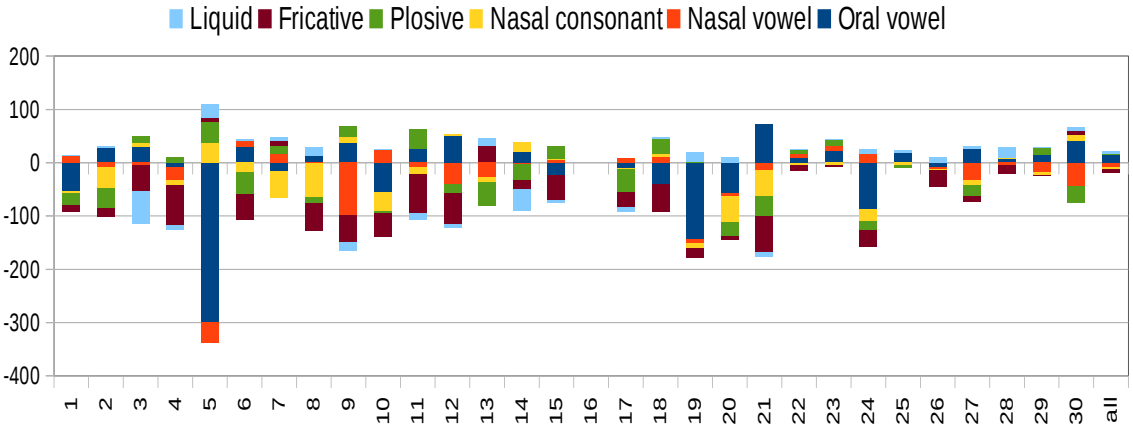

Figure 9.11: *Stacked bar chart of C_{llr}^{R} computed on C_{llr}^{TAR} (target trials) per speaker and for "all". Experiments are performed taking advantage of the full bandwidth (0-8000 Hz).*

for about 70% of the speakers. Liquid and plosive classes have a same behaviour than oral vowels. On the other hand, the nasals (and particularly the nasal vowels) and then fricatives still play a positive role: to withdraw these phonemes increases the C_{llr} compared to the "random" case.

Some remarks could be drawn from the results using relative C_{llr}^{TAR} and relative C_{llr}^{NON} taken together:

- The nasal phonemes show a positive effect for both target and non-target comparisons when a narrow-band or a wideband is applied.
- The fricative class brings a quite low information loss for non-target trials while it has proved a quite positive effect for target trials. This positive effect is not observed when a mobile phone channel (300-3400 Hz) is used (Ajili et al., 2016). This outcome agrees with (Schindler et Draxler, 2013) where authors asserted that when the frequencies above 4 kHz were removed, the fricative consonants were less useful. This finding could be explained by the acoustic properties of fricatives (Gordon et al., 2002).
- A same phonemic class, the oral vowels, brings the largest part in terms of speaker discrimination but presents in the same time a large intra-speaker variability which conveys a significant part of the *LR* performance loss. This effect is observed when a narrow or a wideband is applied.

Conclusion

This chapter is dedicated to investigate the impact of phonemic content on voice comparison process. It uses an ASpR system as measurement instrument and, more particularly, the C_{llr} variations. We analysed the influence of six phonemic classes: nasal

vowel, oral vowel, nasal consonant, fricative, plosive and liquid.

In a first step, we investigated the impact of each phonemic class on voice comparison performance using C_{llr} in order to evaluate the information loss. The results showed that oral vowels, nasal vowels and nasal consonants bring more speaker specific information than averaged phonemic content in terms of voice comparison performance. The fricatives do not seem to perform better than an averaged content, which is surprising compared to the literature, but this result is explained by the restricted bandwidth.

In a second step, we explored target and non target parts of C_{llr}. For non-target comparisons, we showed that all the phonemic content play a role in terms of speaker discrimination power. The oral vowels are the largest contributors, followed by nasals and liquids and this finding is consistent among most speakers. When we focused on target comparisons, oral vowels appeared to be tied with a high variability and thereby this phonemic class is responsible of a high information loss.

We saw previously that this phonemic class was bringing a large part of the speaker discrimination power but it also appears to be responsible of intra-speaker variability. In contrast, nasals showed a high capacity for speaker discrimination and at the same time appeared to be robust for intra-speaker variability.

In a third step, the analysis of the phonemic content impact on FVC is pushed further. At this level, we focused on speaker pairs. We showed that the phonological information used to discriminate a pair of speakers depends on the speakers themselves. A deep analysis was dedicated to the 10 "best" and "worst" pairs in terms of speaker discrimination power. We showed that: (i) For the "best" pairs, all the phonemic content still play a positive role in speaker discrimination. (ii) For the "worst" speaker pairs, it appears that sometimes nasals or fricatives convey a significant part of LR performance loss.

Finally, we explored the impact of the bandwidth on FVC. The analysis follows the same logic as the analysis performed when a narrow band (300-3400 Hz) is applied. Similar results are obtained compared to the first investigation but with several interesting variations. The main point here concerns fricatives. Indeed, this phonemic class has proven a positive effect for target trials with a wideband (0-8000 Hz) while this effect is not observed using a phone channel (300-3400 Hz). This outcome was expected as many studies asserted that when the frequencies above 4 kHz are removed, the fricative consonants are less useful.

In this chapter, we highlighted at several steps the importance of speaker factor. We observed large variations of C_{llr} and C_{llr}^{TAR} between our 30 speakers. We also observed large variations per speaker of the system's responses to different phonemic classes, in terms of relative C_{llr}^{TAR}.

As a consequence of these findings, the main takeaway of the present chapter is the fact that ASpR common evaluation protocols are mainly selecting the best features in terms of speaker discrimination (C_{llr}^{NON}) and are largely missing intra-speaker variability when the latter is a key factor for numerous application scenarios. This is particu-

larly true for FVC scenario and it appears mandatory to work more on intra-speaker variability as well as on speaker factor in order to estimate the reliability of a solution in this domain.

Chapter 10

Acoustic study of Oral Vowels

Contents

Introduction

In this chapter, we investigate the amount of speaker specific information carried by each vowel. This study is performed based on formant parameters since they are considered as good descriptors for vowels (Fant, 1970) and they are often used in acoustic-phonetic FVC approaches (Gold et French, 2011). ANOVA with speaker as fixed factor are adopted in this study in order to compare the variability in formant values that can be attributed to inter-speakers versus intra-speaker variability. It is important to note that the study carried in this chapter is performed using oral vowel formant values and thus it does not use ASpR system.

10.1 Motivation

Features or cues employed to discriminate speakers are influenced by the acoustic variability (McGehee, 1937). Therefore, understanding the variation of a specific phonemic class from an acoustic standpoint is an interesting investigation. In this chapter, an acoustic analysis for the *oral vowel* class is conducted. We choose oral vowels for many reasons:

- French language has a rich vowel inventory, as emphasized by (Crothers, 1978). Furthermore, oral vowel is the most present phonemic category in speech. This phonemic class represents in average more than 36% of Fabiole speech materials.
- Oral vowels have shown different behaviour for target and non-target trials as detailed in Section 9.2 of Chapter 9: even though these phonemes are the most contributors in speaker discrimination, they are tied with a high intra-speaker variability that leads to a large accuracy degradation.
- In FVC based on acoustic-phonetic approach, the first 4 formant frequencies are rather important parameters. They are typically measured on oral vowels and compared in actual forensic casework (Becker et al., 2008).

10.2 Acoustic features

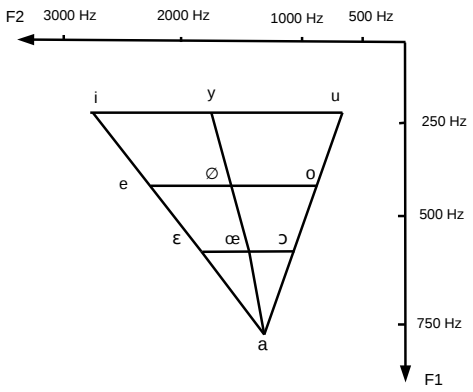

Figure 10.1: *Vowel chart of French oral vowel -F1 and F2-.*

The primary acoustic characteristic of vowels is the location of the formant frequencies, specifically the first three formants (F1-F3) which are considered as a critical determinant of vowel quality (Reetz et Jongman, 2011) (details in Chapter 10 "*Acoustic Characteristics of Speech Sounds*"). The first formant (F1) in vowels is inversely related to vowel height: the higher the formant frequency, the lower the vowel height while the second formant (F2) is somewhat related to the degree of backness and roundness: the

more front the vowel, the higher the second formant. F1 and F2 can also be somewhat affected by lip rounding (Vaissière, 2007). According to the first two formant values, F1 and F2, the 10 French oral vowels are distributed as shown in Figure 10.1.

For French language, F3 is related to the roundness of the anterior vowels. On the other side, F4 seems to be more speaker specific and may therefore provide information about the identity of the speaker rather than the vowel itself (Kahn et al., 2011).

Differences that could be observed between formant values of each vowel may reflect the different vocal tract shapes involved in the production of the vowel. Consequently, strength of intra- or inter-speaker variability could be studied based on these formant values.

In this study, all the formant measures have been estimated using Praat (Boersma et Weenink, 2008). These values are used without any post-processing (filter out erroneous or outlier formants).

10.3 Speaker and vowel effect on formant values

In this section, univariate and multivariate analyses of variance are performed in order to estimate the influence of the speaker and the vowel timbre on each formant.

10.3.1 Speaker and vowel factors

To study the effect of speaker and vowel on the acoustic parameters, experiments are carried out on two steps: First, a two way ANOVA is conducted where the first factor is the speaker, the second factor is the phoneme and the dependent variable is the formant in-interest. Formant values are estimated at the middle of the vowel. Second, a MANOVA is performed where the dependent variable is the 4 formants put together, thus providing a 4-dimensional vector. Results are summarized in Table 10.1.

Table 10.1: *Effect size of speaker and vowel factors for the first 4 formants taken separately and for the multivariate case, explained in term of eta square η^2. "bold", "italic" and "normal" represent respectively high, medium and small effect.*

factor	F1	F2	F3	F4	(F1-F4)
Speaker	3.24	2.90	3.32	*11.01*	*6.44*
Vowel	**25.40**	**63.95**	**42.49**	**27.26**	**30.05**

In this experiment, all observed differences on the formant values, between speakers or vowel timbres, are significant with p-value < 0.001. η^2 shows that an effect of both vowel and speaker factors is observed on formant values. For all the cases, the interaction between the formant value and the vowel is higher than the interaction between

the formant and the speaker. For example, for the multivariate case, the interaction between vowel and formant accounts for about 30% of the total formant variability while interaction between speaker and formant explains only 6.44%.

ANOVA using the vowel as a fixed factor indicates that, for most speakers, vowels have a large effect on the different formants taken separately. These effects are slightly stronger on F2 and F3 compared to F1 and F4. On the other side, speaker factor has shown the highest effect on F4 (indicating by a high eta-square value, η^2=11%), while speaker has a small effect on F1, F2 and F3 (η^2 for F1, F2 or F3 is smaller than 3%). This outcome confirms (Lavner et al., 2000) which indicates that the three first formants are mostly linked to the timbre of the vowel while the 4^{th} formant has more links with the speaker itself.

10.3.2 Speaker impact on each oral vowel

To quantify the effect of speaker on each vowel, two experiments are carried out. First, one-way ANOVA is performed with speaker as fixed factor and formant values as the dependent variable. Each oral vowel is taken separately. Second, one-way MANOVA is applied where the dependent variable is 4-dimensional vector corresponding to the four first formants (F1 to F4) put together. Results are shown in Table 10.2 and illustrated in Figure 10.2 for better visualization.

Table 10.2: *Eta square calculated for 10 oral vowels separately across 30 speakers for the 4 first formants, F1, F2, F3 and F4, and for the multivariate case (F1-4). "bold", "italic" and "normal" represent respectively high, medium and small effect.*

Vowel	**F1**	**F2**	**F3**	**F4**	**F1-4**
/i/	1.65	*6.16*	*8.09*	**14.08**	*7.72*
/y/	2.98	5.95	5.91	*11.68*	*7.77*
/u/	2.1	2.50	*6.51*	3.98	3.68
/e/	1.83	**17.46**	*9.72*	**20.56**	**14.02**
/ø/	5.79	*7.9*	4.13	**14.86**	*10.57*
/o/	*12.2*	*13.22*	*8.1*	8.10	*11.38*
/ɛ/	2.8	*11.30*	*10.75*	**18.48**	*12.27*
/œ/	*10.88*	*7.60*	*8.48*	**23.04**	*13.52*
/ɔ/	*12.51*	*10.56*	*7.34*	*13.18*	*11.48*
/a/	*12.85*	4.0	*13.21*	**19.82**	*13.75*

All the differences between speakers observed on formant values are significant with a p-value < 0.001. Concerning the speaker factor, an effect is observed with different extents. This effect depends on both the formant itself and the pronounced vowel. Indeed, for F1, η^2 varies from 1.83% (/*e*/) to 12.85% (/ɑ/). For F2, η^2 varies from 2.5% (/*u*/) to 17.46% (/*e*/). For F3, η^2 varies from 4.13% (/ø/) to 13.21% (/ɑ/) while for F4, η^2 varies from 3.98% (/*u*/) to 23.04% (/œ/).

The speaker effect is large on some specific vowels such as /e/ where the speaker

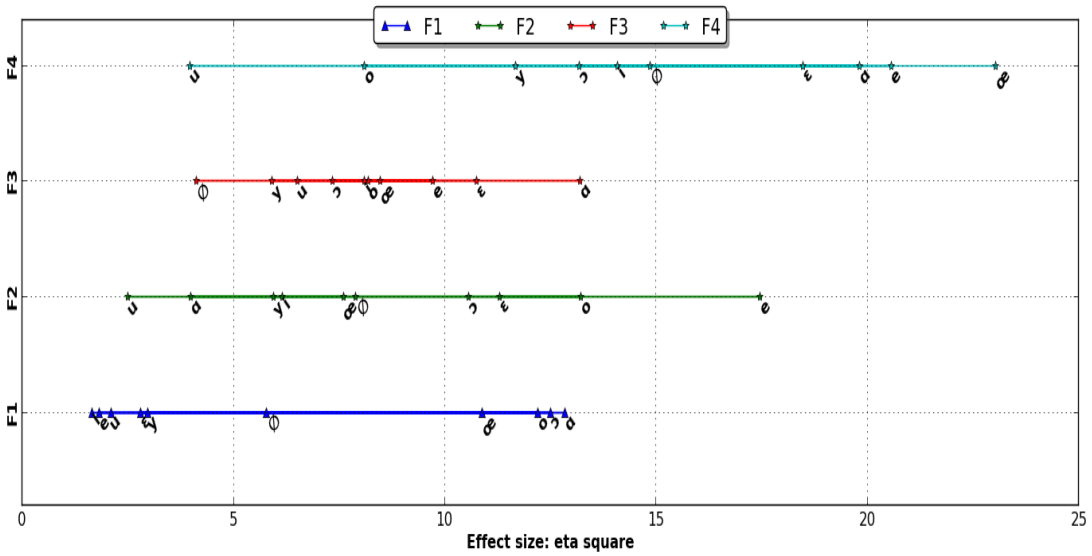

Figure 10.2: *Effect size of the 10 French oral vowels for the 4 first formant F1, F2, F3 and F4.*

explains about 14.02% of the variance of the formant in the multivariate case. On the univariate analyses, the speaker effect explains about 23.04% of the total variability of the 4^{th} formant on /œ/ vowel.

Table 10.2 shows also that the effect size for the 4^{th} formant is, most of the time, higher than the effect size of the 3 first formants (F1, F2, F3). Once more, this outcome confirms results in (Lavner et al., 2000).

The study of the speaker effect on each vowel shows that some vowels, such as /e/, are largely influenced by the speaker. It differs with the analysis reported in subsection 10.3.1 which does not take into account characteristics of each vowel. In 10.3.1, the speaker effect size is medium for F4 and very small for F1, F2 and F3. This difference shows that averaging the speaker effect on several factors could shadow his effect on each individual factor. Therefore, speaker effect should be studied on each factor separately.

10.3.3 Speaker impact on formant dynamics

In the two previous subsections, 10.3.1 and 10.3.2, we have focused on "static" formant frequencies[1] in order to estimate the influence of the speaker in each vowel. In this subsection, we investigate the effect of speaker factor on the dynamics[2] of the first four formants. Following the procedures described in McDougall (McDougall, 2007), time-normalised formant measurements were taken at +10% steps across the duration of

[1]Formant frequencies are estimated at the midpoint of the vowel.
[2]It is so called also "formant trajectory" or "formant contour".

each vowel as described in Figure 10.3.

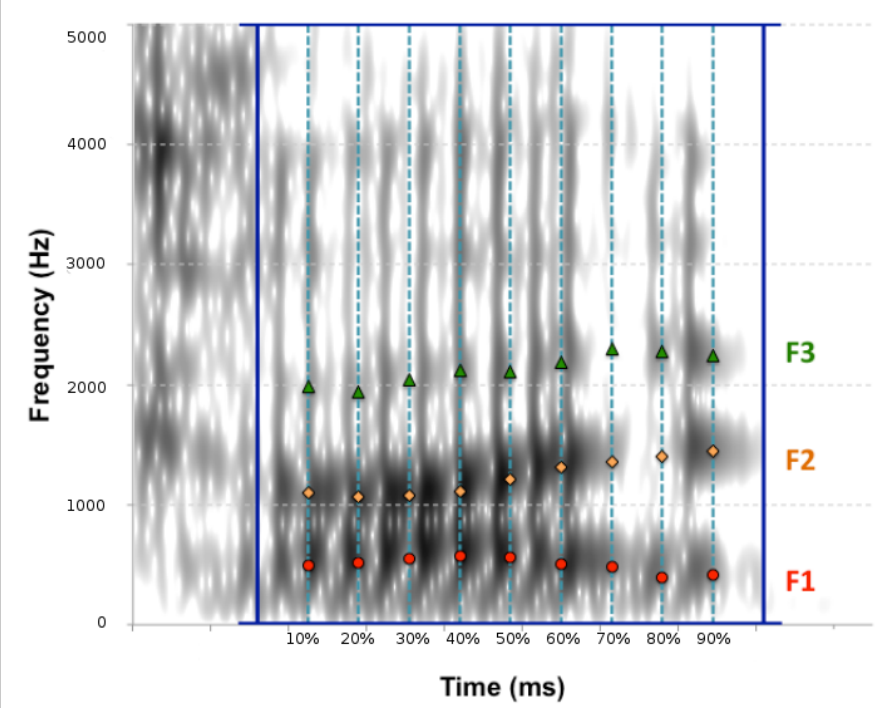

Figure 10.3: *Example of points for time-normalised dynamic formant analysis with measurements taken at +10% steps (Adapted from (Hughes, 2014)) on a syllable. Each formant is then represented by a 9-dimensional vector.*

A one-way MANOVA is performed with speaker as fixed factor and formant dynamic frequencies as the dependent variable. Table 10.3 summarizes the speaker effect size on the different vowels explained in terms of eta-square η^2.

Table 10.3: *Eta square calculated for* 10 *oral vowels separately across* 30 *speakers for the first 4 formants,F1 , F2, F3 and F4. "bold", "italic" and "normal" represent respectively high, medium and small speaker effect.*

Vowel	F1	F2	F3	F4
/i/	1.61	1.29	1.52	2.67
/y/	1.33	1.50	1.81	3.31
/u/	1.19	0.90	1.66	1.42
/e/	0.70	2.92	1.92	3.39
/ø/	1.55	1.55	0.9	2.69
/o/	2.80	2.91	1.74	1.76
/ɛ/	1.02	2.05	2.06	3.20
/œ/	3.10	2.18	2.21	4.28
/ɔ/	2.19	2.0	1.43	2.26
/a/	2.08	0.70	2.11	3.19

A significant speaker effect is observed on all vowels and on all formant values (p-value < 0.001). However, this effect is small for all cases:

- for F1, η^2 varies between 0.7 (/e/) and 3.10 (/œ/).
- for F2, η^2 varies between 0.7 (/a/) and 2.92 (/e/).
- for F3, η^2 varies between 0.9 (/ø/) and 2.21 (/œ/).
- for F4, η^2 varies between 1.42 (/u/) and 4.28 (/œ/).

This outcome could be explained by the fact that vowel formant dynamics are highly influenced by the "*phonetic context*"[3] (or also co-articulation aspects). This hypothesis is corroborated by (Hillenbrand et al., 2001) where authors studied whether a close relationship between "*vowel identity*" and "*spectral change patterns*" (i.e formant dynamics) is maintained when the consonant environment varies and showed significant consonantal context effect on the vowel. In order to better investigate the speaker factor effect on formant dynamics, the vowel phonetic context should be controlled by fixing the initial and final consonants of the *CVC* syllable.

10.4 Intra-speaker variability in the acoustic space

In order to show the weight of intra-speaker variability, we estimate the first 4 formant values (F1, F2, F3, F4) at the middle of each vowel. These measures are estimated speaker by speaker on the 100 speech recordings corresponding to each T speaker.

10.4.1 Intra-speaker variability for vowel /a/

Table 10.4 presents the mean frequency and the standard deviation of the first 4 formants, for each speaker, estimated on the vowel /a/.

The mean frequencies vary substantially across the 30 speakers. The mean formant frequency varies:

- from 503 Hz to 729 Hz for F1.
- from 1390 Hz to 1588 Hz for F2.
- from 2413 Hz to 2703 Hz for F3.
- from 3338 Hz to 3882 Hz for F4.

On the other hand, standard deviations of the formant frequencies vary from person to person:

- On F1, speaker 2 shows the largest intra-speaker variation (159 Hz), while speaker 3 shows the smallest value (74 Hz).

[3] The vowel context /c_iVc_f/ where c_i is the initial consonant, c_f the final consonant and V is the vowel.

- On F2, speaker 2 shows the largest intra-speaker variation (269 Hz) while speaker 9 shows the smallest one (177 Hz).

- On F3, speaker 12 shows the highest intra-speaker variation (300 Hz), while speaker 24 shows the smallest variation (110 Hz).

- On F4, speaker 28 shows the largest intra-speaker variation (412 Hz), while speaker 22 shows the smallest variation (120 Hz).

Table 10.4: *Mean frequencies and standard deviation of the first four formants for 30 speakers estimated on /aa/ phoneme. 'bold" and "italic" represent respectively the "max" and "min" values.*

Speakers	F1		F2		F3		F4	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	512	80	1463	185	2545	150	3350	240
2	558	**159**	1532	**269**	2581	211	3583	278
3	542	*74*	1486	179	2655	294	3775	267
4	564	154	1559	250	2542	297	3493	220
5	620	105	1481	246	2633	137	3733	208
6	617	134	1499	201	2526	173	3698	279
7	665	105	1503	247	2622	153	3708	174
8	585	88	1472	204	2542	170	3658	200
9	*503*	75	1487	*177*	2660	130	3499	184
10	607	101	**1588**	228	2641	170	3593	167
11	642	157	1510	216	2551	228	3625	203
12	602	77	1449	196	2465	**300**	3736	247
13	652	149	1519	232	2605	210	3722	283
14	573	99	1515	191	2425	250	3758	216
15	603	101	1570	198	2516	125	3686	228
16	551	86	*1390*	216	2591	153	3460	208
17	597	113	1460	217	2482	165	3475	182
18	613	111	1482	184	*2413*	211	3597	263
19	637	109	1567	227	2651	196	3540	330
20	596	92	1488	222	**2703**	136	3614	228
21	602	100	1522	185	2593	175	**3882**	381
22	720	140	1534	204	2667	144	3509	*120*
23	591	130	1509	244	2558	167	3613	182
24	607	88	1565	198	2654	*110*	3486	219
25	568	93	1575	223	2682	171	3935	282
26	574	106	1544	250	2589	204	3564	220
27	585	87	1460	204	2518	175	3508	226
28	569	102	1542	211	2560	128	*3338*	**412**
29	628	141	1585	221	2658	164	3609	231
30	**729**	128	1429	184	2687	189	3654	180
All(*)	600	109	1509	214	2584	183	3613	235

Table 10.4 shows that F2 and F4 are accepting a high variation compared to F1 and F3.

10.4.2 Intra-speaker variability estimation using acoustic distance

In this section, in order to estimate the strength of intra-speaker variation, we use an acoustic distance computed across all the recording pairs of each speaker. We expect that the larger the distance is, the higher the intra-speaker variability is.

For each speech recording, formant parameters are extracted over the 10 oral vowels. Each vowel will be represented by a 4-dimensional vector corresponding to the formant mean values as shown in Equation 10.1.

$$V_\phi = \begin{bmatrix} \overline{F1}_\phi & \overline{F2}_\phi & \overline{F3}_\phi & \overline{F4}_\phi \end{bmatrix} \tag{10.1}$$

Where $\phi \in \{$ /i/, /y/, /u/, /e/, /ø/, /o/, /ɛ/, /œ/, /ɔ/, /a/ $\}$

The speech recording, S, is therefore modelled by a 10×4 dimensional matrix, M_S corresponding to the 10 oral vowel's V_ϕ vectors stacked vertically as shown in Equation 10.2.

$$M_S = \begin{bmatrix} V_{/i/} \\ V_{/y/} \\ V_{/u/} \\ V_{/e/} \\ V_{/ø/} \\ V_{/o/} \\ V_{/ɛ/} \\ V_{/œ/} \\ V_{/ɔ/} \\ V_{/a/} \end{bmatrix} \tag{10.2}$$

The acoustic distance used in this study is the Euclidean distance (also called "*Frobenius norm*" when applied on matrices). Given two speech recording S_1 and S_2, the Euclidean distance is calculated as shown in Equation 10.3.

$$D(S_1, S_2) = ||M_{S_1} - M_{S_2}||_2 \tag{10.3}$$

Figure 10.4 displays the mean and the standard deviation of the distribution of Euclidean distances for each speaker. In this case, distance uses only F1 and F2 (i.e the two first columns of the matrix are used).

Figure 10.4 shows large acoustic differences between the speakers. For example, speaker 3 presents a low intra-speaker variability equal to 306Hz while speaker 2 shows a high intra-speaker variability of 835Hz. It is interesting to remind from Figure 9.4 that

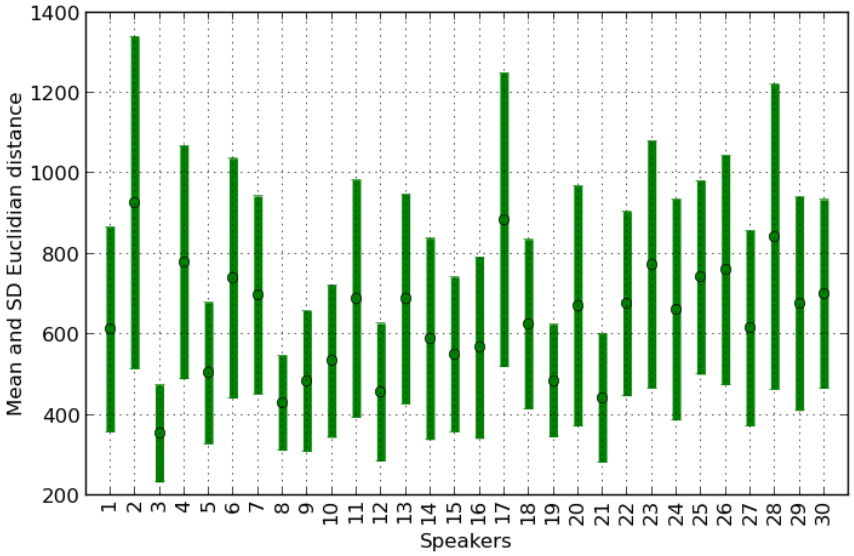

Figure 10.4: *Mean and standard deviation of euclidean distance calculated on all pairs of comparison for each speaker of set T (Hz) using only the two first formants, F1 and F2.*

withdrawing vowels leads to a decrease of C_{llr} for Speaker 2 while for the Speaker 3, the same option leads to an increase of C_{llr}.

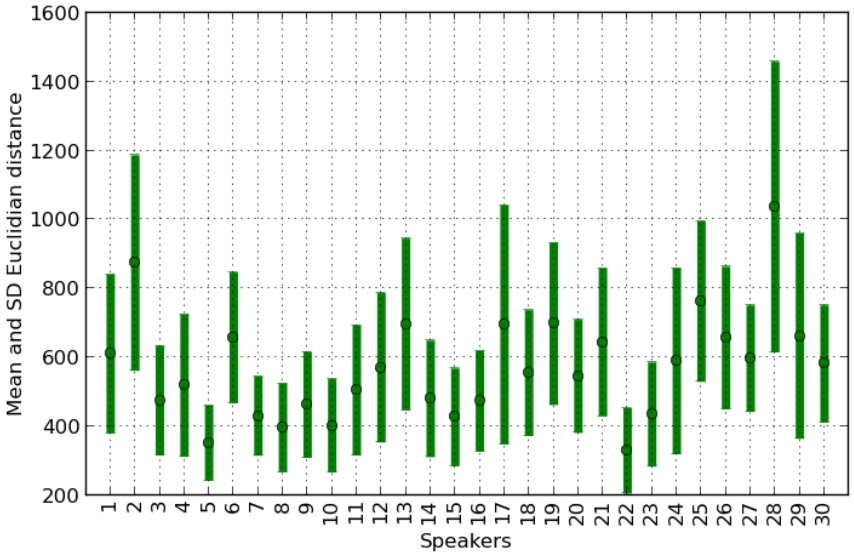

Figure 10.5: *Mean and standard deviation of euclidean distance for each speaker of set T (Hz) using only F4.*

Figure 10.5 presents the mean and standard deviation of Euclidean distance in a form corresponding to Figure 10.4 but in this case, only the 4^{th} formant, F4, is used to compute the distance.

Figure 10.5 shows a large variability of the Euclidean distance between speakers. For example, speaker 22 presents a low acoustic intra-speaker variability equal to 326Hz while speaker 28 shows a high acoustic intra-speaker variability of 1036Hz. At this level, it is interesting to remind that oral vowels play a positive (respectively negative) role in speaker discrimination for speaker 22 (respectively speaker 28), as shown in Figure 9.4. This finding may reveal a potential relation between the size of acoustic variability and the effect of oral vowels on FVC.

10.4.3 Euclidean distance as a predictor of oral vowels behaviour

In FVC, oral vowels have showed two different behaviour:

- A positive behaviour: The oral vowel class appears to convey a significant part of speaker-specific information.
- A negative behaviour: The oral vowel class appears to be responsible of a "big" part of the LR information loss.

The sign of C_{llr}^{R} is enough to detect the behaviour of the the oral vowels category. In order to quantify this effect in terms of quantity of information, we propose ΔC_{llr}[4]= C_{llr}^{random}-$C_{llr}^{specific}$.

In order to investigate whether the strength of formant intra-speaker variability has an impact on oral vowel behaviour, we study the correlation between: (i) ΔC_{llr} and (ii) mean Euclidean distance estimated across the 30 Fabiole speakers. Table 10.5 presents three correlation coefficients corresponding to R^2, ρ Spearman and Kendall's τ. In this case, Euclidean distance is calculated using only F1 and F2 as they may reflect different speaking style (F1 and F2 are linked directly to the aperture and backness of vowels).

Table 10.5: *Correlation between Euclidean distance calculated using the first two formants (F1, F2) and ΔC_{llr} -information win/loss- (explained in terms of bits). (*) indicates the significance level, n.s represents non significance.*

	Correlation coefficient	**p-value**
R^2	0.232	n.s
ρ **Spearman**	0.087	n.s
Kendall's τ	0.059	n.s

Table 10.5 shows that F1 and F2 intra-speaker variability does not have a noticeable

[4] ΔC_{llr} differs from C_{llr}^{R} only on the normalisation factor. Here, a reminder of C_{llr}^{R} formula: $C_{llr}^{R}=\frac{C_{llr}^{random}-C_{llr}^{specific}}{C_{llr}^{random}}\times 100\%$

impact on oral vowels behaviour. It is confirmed by a non significant low correlation, evaluated by a R^2 equal to 0.23 (p-value=0.22).

It seems that to focus only on F1 and F2 intra-speaker variability could not predict the oral vowels behaviour. This could be explained by the fact that these formants are mainly linked to the vowel itself and less to the speaker identity.

Table 10.6 shows that F4 intra-speaker variability has a noticeable impact on the oral vowels behaviour. This result is confirmed by a relatively large positive correlation coefficients, R^2=0.567 respectively Spearman's ρ and Kendall's τ indicating a high correlation between F4 intra-speaker variability and ΔC_{llr}= C_{llr}^{random}-$C_{llr}^{specific}$. This correlation is statistically significant (p-value < 0.01 for all cases).

Table 10.6: *Correlation between acoustic intra-speaker variability calculated using the fourth formant (F4) and information win/loss (explained in terms of bits). (*) indicates the significance level.*

	Correlation coefficient	**p-value**
R^2	0.567	**
ρ **Spearman**	0.473	**
Kendall's τ	0.374	**

This result suggests that there is a link between F4 intra-speaker variability and the oral vowel behaviour.

Conclusion

In this chapter, we carried out a deep study on oral vowels based on formant parameters. We chose formants as they are considered as good descriptors for vowels and they are often used in acoustic-phonetic FVC approaches. Our study aimed to :

- quantify the amount of speaker specific information for each vowel.
- study the influence of formant intra-speaker variability on the vowel behaviour.

The main takeaways of this investigation could be summarized in three points.

- Speaker and vowel factor effects are observed on formant values. The interaction between the formant value and the vowel is higher than for the formant and the speaker. We showed also that the first 3 formants are more linked to the vowel than to the speaker itself. On the other hand, the 4^{th} formant is mainly linked to the speaker.
- We showed that the amount of speaker specific information in each vowel presents a high variability. Indeed, the speaker effect depends on both the formant itself and the pronounced vowel.

- The study of the influence of the formant intra-speaker variability revealed a potential relation between the size of acoustic variability and the oral vowels behaviour in FVC.

The results presented in this chapter could put into question the use of formant measures in acoustic-phonetic FVC.

Chapter 11

Homogeneity measure for Forensic voice comparison

Contents

Introduction

Issues of validity and reliability of evidence evaluation are of great concern in forensic science and more particularly in FVC (Bonastre et al., 2003; National Research Council, 2009; Campbell et al., 2009; Daubert, 1993; Saks et Koehler, 2005; Morrison, 2009a; Rose, 2006; Morrison et al., 2011; Morrison, 2011b). Several factors could influence the quality of the LR including, without limitation, the case (i) when there is a lack of discriminative information in both voice recordings or (ii) when there is a mismatch between elements used by the system (UBM, total variability matrix, PLDA,...) and the pair of voice records S_A-S_B (Greenberg et al., 2011, 2013; Kahn et al., 2010). In this chapter, we propose a methodology in order to cope with the first case (i). Our objective is to define a "*confidence measure*[1]" (CM) able to estimate the amount of "speaker discriminant cues[2]" and the homogeneity of this information between the pair of voice recordings S_A-S_B. So, this measure is estimated only from the two in-interest voice records. This measure is expected to be linked to the LR accuracy.

[1] Or also reliability measure.

[2] In this chapter, speaker discriminant information refers to the speech and does not take into account the speaker discrimination power of each phoneme.

11.1 An information theory data based homogeneity measure

In this paragraph, we define an information theory (IT) based Homogeneity Measure denoted HM. Its objective is to calculate the amount of acoustic information that appertains to the same class between the two voice records. The set of acoustic frames gathered from the two files S_A and S_B is decomposed into acoustic classes thanks to a Gaussian mixture Model (GMM) clustering. Then the homogeneity is first estimated in term of bits as the amount of information embedded by the respective "number of acoustic frames" of S_A and S_B linked to a given acoustic class. Each acoustic class is represented by the corresponding Gaussian component of the GMM model. The occupation vector could be seen as the number of acoustic frames of a given recording belonging to each class m. It is noted: $[\gamma_{g_m}(s)]_{m=1}^{M}$.

Given a Gaussian g_m and two posterior probability vectors of the two voice records S_A and S_B, $[\gamma_{g_m}(A)]_{m=1}^{M}$ and $[\gamma_{g_m}(B)]_{m=1}^{M}$, we define:

- $\chi_A \cup \chi_B = \{x_{1A}, ..., x_{NA}\} \cup \{x_{1B}, ..., x_{NB}\}$ the full data set of S_A and S_B with cardinality N=N_A +N_B
- $\gamma(m)$ and $\omega(m)$ are respectively the occupation and the prior of Gaussian m where $\omega(m) = \frac{\gamma(m)}{\sum_{k=1}^{M}\gamma(k)} = \frac{\gamma(m)}{N}$
- $\gamma_A(m)$ (respectively $\gamma_B(m)$) is the partial occupations of the m^{th} component due to the voice records S_A (respectively S_B).
- p_m is the probability of the Bernoulli distribution of the m^{th} bit (due to the m^{th} component), $B(p_m)$. $p_m = \frac{\gamma_A(m)}{\gamma(m)}$, $\overline{p}_m = 1 - p_m = \frac{\gamma_B(m)}{\gamma(m)}$.
- $H(p_m)$ the entropy of the m^{th} Gaussian (the unit is bits) given by: $H(p_m) = -p_m log_2(p_m) - \overline{p}_m log_2(\overline{p}_m)$.

The class entropy, $H(p_m)$, has some interesting properties in the context of an homogeneity measure:

* $H(p_m)$ belongs to $[0, 1]$.
* $H(p_m) = 0$ if $p_m = 0$ or $p_m = 1$. It means that when the repartition of the example of a given class m is completely unbalanced between S_A and S_B, $H(p_m)$ is zero (i.e. $H(p_m)$ goes to zero when p_m is close to 0 or 1).
* $H(p_m) = 1$ when $p_m = 0.5$. $H(p_m)$ is maximal when the examples belonging to a given class are perfectly balanced between between S_A and S_B (i.e. $H(p_m)$ goes to the maximum value 1 when the repartition goes to the balanced one).

Two measures based on $H(p_m)$ are proposed in the following of this paragraph. The first measure ignores the size of the frame sets (i.e. the duration of the recordings) when the second one takes this aspect into account.

Normalized HM

The HM is calculated as shown in Equation 11.1. It measures the Bit Entropy Expectation (BEE) with respect to the multinomial distribution defined by GMM's priors $\{\omega(m)\}_{i=1}^{M}$.

$$\begin{aligned} HM_{BEE} &= \sum_{m=1}^{M} \frac{\gamma(m)}{N} H(p_m) \\ &= \sum_{m=1}^{M} \omega_m H(p_m) \end{aligned} \tag{11.1}$$

By definition HM_{BEE} contains the percentage of the data-homogeneity between S_A and S_B. It does not take into account the quantity of the homogeneous information between the two speech extracts. For example, assuming a trial S_c-S_D that verifies $D_c = 2 \times D_A$ and $D_D = 2 \times D_B$, where D_A,D_B,D_C and D_D are respectively the length of signal S_A, S_B, S_C and S_D. If their vectors of occupation are related by: $[\gamma_{g_m}(C)]_{m=1}^{M} = 2 \times [\gamma_{g_m}(A)]_{m=1}^{M}$ and $[\gamma_{g_m}(D)]_{m=1}^{M} = 2 \times [\gamma_{g_m}(B)]_{m=1}^{M}$. So, $H_{A,B}(m) = H_{C,D}(m)$ because they have the same Bernoulli distribution.

$$\begin{aligned} p_m(C,D) &= \frac{\gamma_C(m)}{\gamma_C(m) + \gamma_D(m)} \\ &= \frac{2 \times \gamma_A(m)}{2 \times \gamma_A(m) + 2 \times \gamma_B(m)} \\ &= p_m(A,B) \end{aligned} \tag{11.2}$$

As $\omega(m)$ is the same for the two examples, $HM_{BEE}(A,B) = HM_{BEE}(C,D)$.

Non-normalized HM

We propose here a variant of HM, *Non-normalized Homogeneity Measure* (NHM). NHM calculates the quantity of homogeneous information between the two voice records as shown in Equation 11.3. The amount of information is defined in term of (equivalent) number of acoustic frames. NHM measures the BEE with respect of the quantity of information present in each acoustic class $\{\gamma(m)\}_{i=1}^{M}$.

$$\begin{aligned} NHM_{BEE} &= \sum_{m=1}^{M} (\gamma_A(m) + \gamma_B(m)) H(p_m) \\ &= \sum_{m=1}^{M} \gamma(m) H(p_m) \end{aligned} \tag{11.3}$$

11.2 Homogeneity impact on LR accuracy and reliability

In this section, we use the NIST SRE 2008 det1 protocol in our experiments in order to evaluate the proposed homogeneity measure impact on the LR accuracy. First, we apply it on all the trials of the evaluation set and sort the set accordingly. We are expecting that lowest values of homogeneity are correlated with the lowest LRs accuracy (computed using C_{llr}), as well as the opposite behaviour for high values.

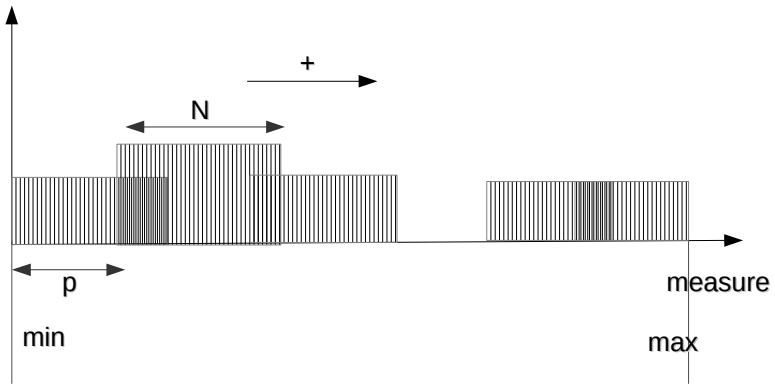

Figure 11.1: *Algorithm to evaluate the homogeneity measure impact on the LRs. (i) Homogeneity measure sorted from the lowest (min) to the highest value (max). (ii) use of a sliding window of size N and step p. For each window, C_{llr} is averaged on all the trials.*

In order to compute the C_{llr} corresponding to a given NHM value, we apply on the trials sorted by homogeneity values a sliding window of size N, moved using a step of p values as shown in Figure 11.1. On each window, we compute the averaged C_{llr} to be compared with the NHM value, computed here as the median value on the window (FA and FR could also be provided for comparison purposes). If a C_{llr} could be computed for a given trial, it makes sense to average the values on a reasonably large set of trials.

HM_{BEE} impact on the LRs accuracy

Table 11.1 presents the impact of the HM_{BEE} on the LR accuracy. p and N are fixed to 1000 and 1500 respectively. We remind that NIST SRE 2008 det1 protocol is used in this study.

Table 11.1: HM_{BEE} *versus* C_{llr}^{min}. *False alarm rate (FA), false rejection rate (FR) as well as the number of target trials (TAR) and non target trials (NON) are also provided.* C_{llr}^{min} *baseline system is equal to* 0.2241bits

HM_{BEE}	**#TAR**	**#NON**	**FR%**	**FA%**	C_{llr}^{min}
0.56	50	1450	64	0	0.417
0.61	88	1412	38.63	0.35	0.223
0.64	143	1357	24.47	0.51	0.21
0.67	195	1305	24.10	1.22	0.2924
0.69	248	1252	22.98	1.83	0.2939
0.71	301	1199	14.61	3.25	0.2579
0.72	306	1194	12.09	4.52	0.2365
0.74	374	1126	8.02	6.39	0.2350
0.75	455	1045	5.93	9.56	0.2677
0.76	549	951	4.73	13.24	0.2709
0.78	744	756	3.09	19.04	0.2551
0.80	971	529	2.78	30.05	0.2701
0.82	1250	250	1.52	42.4	0.2495
0.83	1340	65	0.52	55.38	0.2784

Table 11.1 summarizes information corresponding to each application of the window including HM_{BEE}, number of target and non-target trials, FR% and FA% and finally the C_{llr}^{min} value. Figure 11.2 shows the behavioural curve of HM_{BEE}.

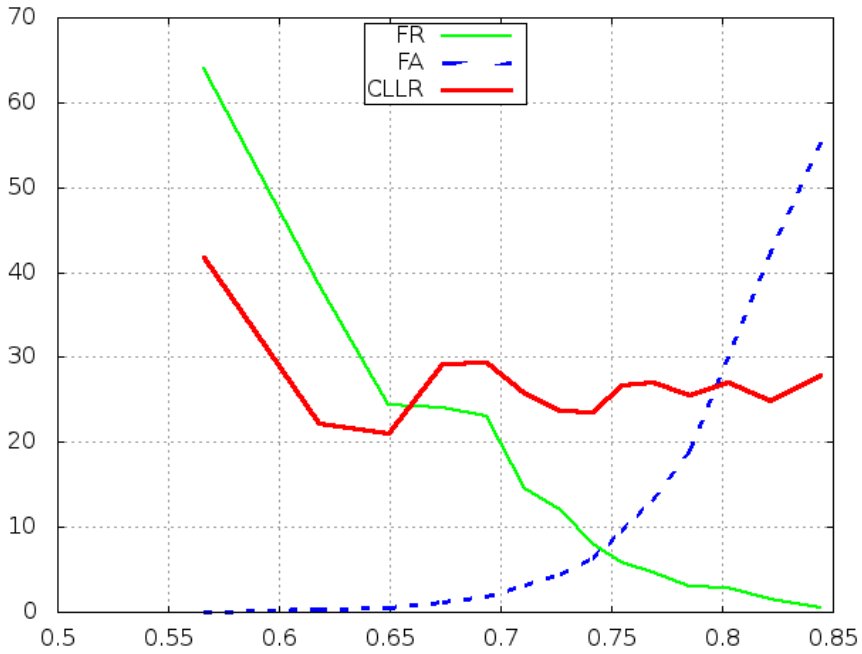

Figure 11.2: HM_{BEE} *behaviour, estimated with GMM-AB.*

From Table 11.1, HM_{BEE} value does not have a noticeable correlation with C_{llr}^{min}. Moreover, in most cases, the C_{llr}^{min} for each chunk is higher than the C_{llr}^{min} of the baseline system. We observe also that non target trials are concentrated in HMs lower values whereas target trials are concentrated in HMs higher values. This fact seems to indicate that HM_{BEE} is correlated with the system outputs. This is confirmed by a R^2 equal to 0.73.

Figure 11.2 shows that HM_{BEE} does not have a monotonic shape. All C_{llr}^{min} values except the first one (there is very few target trials in the first chunk) are between 0.2 and 0.3. The discrimination loss is almost the same no matter the HM values.

NHM impact on the LRs accuracy

Table 11.2 shows the experimental results obtained using NHM_{BEE}. This version uses a GMM learnt only on the pair of speech signals, GMM A-B.

Table 11.2: NHM_{BEE} *(GMM A-B) versus* C_{llr}^{min}*. False alarm rate (FA), false rejection rate (FR) as well as the number of target trials (TAR) and non target trials (NON) are also provided.*

NHM_{BEE}	**TAR**	**NON**	**FR%**	**FA%**	$\mathbf{C_{llr}^{min}}$
4689	129	1371	31	1.45	0.309
5099	196	1304	28.57	1.91	0.3452
5357	226	1274	22.12	2.19	0.3122
5574	258	1242	14.72	3.30	0.2759
5751	295	1205	11.86	5.06	0.2654
5916	355	1145	11.54	5.67	0.277
6071	408	1092	9.55	6.68	0.2823
6224	489	1011	6.95	7.61	0.2304
6373	514	986	5.83	7.80	0.2127
6529	624	876	4.32	8.9	0.1930
6699	671	829	3.57	11.09	0.2074
6891	798	702	2.93	15.09	0.1925
7136	954	546	1.88	16.84	0.1811
7579	1130	275	0.7	17.09	**0.1227**

Figure 11.3 presents NHM_{BEE} behavioural curve. The shape of the curve is interesting with C_{llr}^{min} varying from 0.309 to 0.122. It seems that NHM_{BEE} brings new information compared to the system outputs. The result is confirmed with a R^2 of 0.55, to be compared with a R^2 equal to 0.73 in the case of HM_{BEE}.

In Table 11.3 and Figure 11.4, we present the results obtained using NHM_{BEE} like in the previous case but with a different representation of the acoustic classes. Here, we use an UBM[3] in order to cluster the acoustic frames of the pair of speech recordings.

[3]UBM used in the NIST SRE I-vector system.

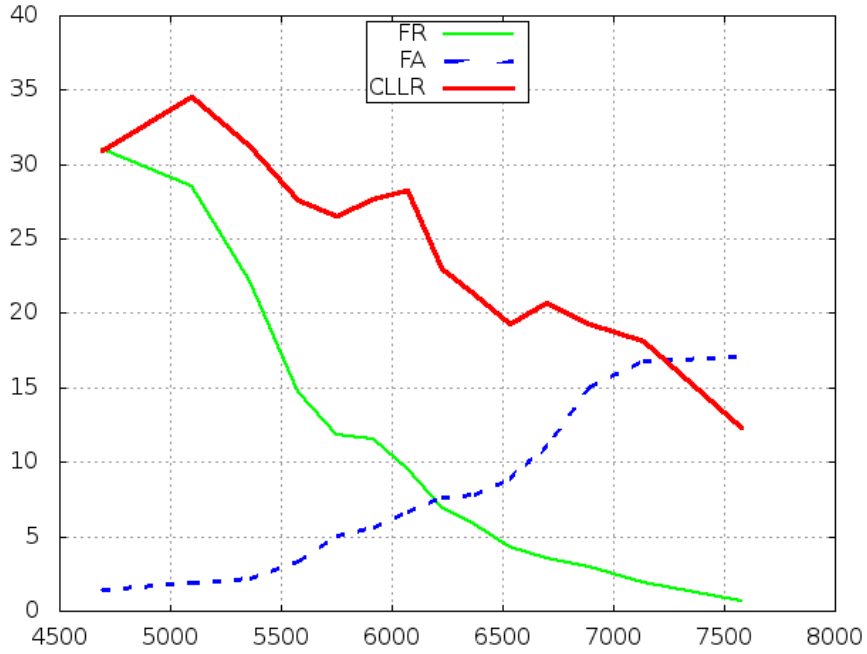

Figure 11.3: NHM_{BEE} *behaviour, estimated with GMM-AB.*

Table 11.3: NHM_{BEE} *(UBM) versus* C_{llr}^{min}*. False alarm rate (FA), false rejection rate (FR) as well as the number of target trials (TAR) and non target trials (NON) are also provided.*

NHM_{BEE}	**TAR**	**NON**	**FR%**	**FA%**	$\mathbf{C_{llr}^{min}}$
6341	333	1167	14.71	4.79	0.3
6685	338	1162	12.72	5.42	0.303
6900	370	1130	12.97	4.70	0.306
7058	371	1129	9.43	7.08	0.284
7204	382	1118	8.90	5.81	0.238
7338	415	1085	7.95	4.05	0.22
7462	461	1039	7.59	5.48	0.237
7576	511	989	6.45	5.25	0.199
7700	516	984	5.81	5.38	0.1756
7826	574	926	5.92	8.09	0.20
7967	600	900	3.5	8.55	0.188
8135	662	838	4.07	9.66	0.195
8358	723	777	2.35	8.49	0.145
8762	826	579	0.96	5.87	**0.089**

With a C_{llr}^{min} varying between 0.3 and 0.089 and its high correlation with NHM_{BEE}, evaluated to R^2 equal to 0.9 ($p < 0.001$), this variant seems to outperform the previous one. Two explanations to this result could be proposed. Firstly, it is reasonable to think that as it is learnt on a very large data set, the estimation quality of the UBM is

higher than the GMM A-B (learnt only on the two speech recordings). Second, as the UBM models the whole acoustic space, it embeds also some knowledge about speaker specificities.

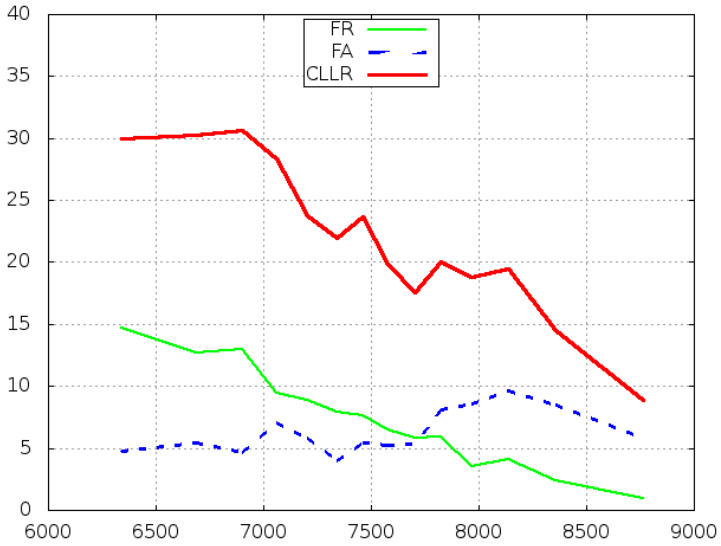

Figure 11.4: $\mathrm{NHM}_{\mathrm{BEE}}$ *behaviour, estimated with UBM.*

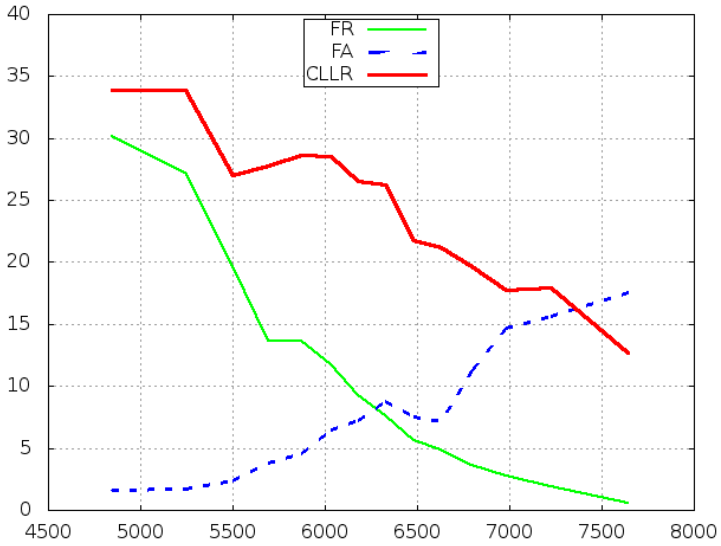

Figure 11.5: *NHM behaviour using GMM A-B initialized with UBM.*

Two more experiments are performed. In Figure 11.5, we report results when the GMM A-B is initialized with the UBM (case A), and in Figure 11.6, we use the UBM mean-adapted (using MAP) by the two speech recordings S_A and S_B (case B). In both

cases, NHM is highly correlated with the C_{llr}^{min} (A: $R^2 = 0.927$, $p < 0.001$; B: $R^2 = 0.94$, $p < 0.001$).

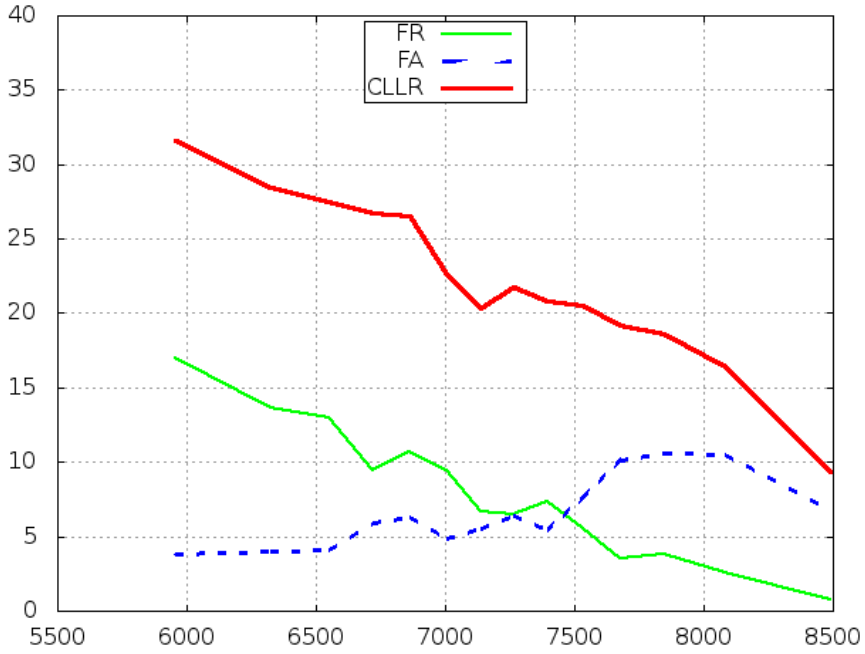

Figure 11.6: *NHM behaviour estimated using the UBM adapted by* S_A *and* S_B.

Remarks

Throughout this section, we used several variants of homogeneity measure in order to predict the LR accuracy. We showed that focusing on BEE ignoring the involved quantity of examples (the case of HM) does not allow to build an homogeneity measure with the desired characteristics. On the other hand, all the NHM variants appeared to be highly linked to the LR accuracy with different extents. Making a compromise between the capacity of prediction (indicated by the magnitude of correlation coefficients) and the time needed to process, we select NHM based on the UBM.

11.3 Homogeneity impact on target and non-target trials

In order to study the impact of the homogeneity measure on target and non-target trials, we come back to Fabiole database. Figure 11.7 is presenting the size of information loss relative to target and non-target trials using Fabiole main protocol.

Figure 11.8 presents, for all the trials, the homogeneity measure, NHM, in function of C_{llr} (as well as C_{llr}^{min} and C_{llr}^{cal}). In order to compute the C_{llr} corresponding to a given

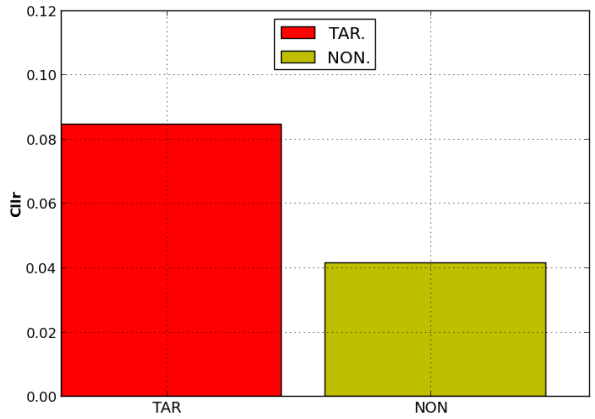

Figure 11.7: *Information loss relative to target and non-target trials.*

NHM value, we follow the same procedure described in 11.2.

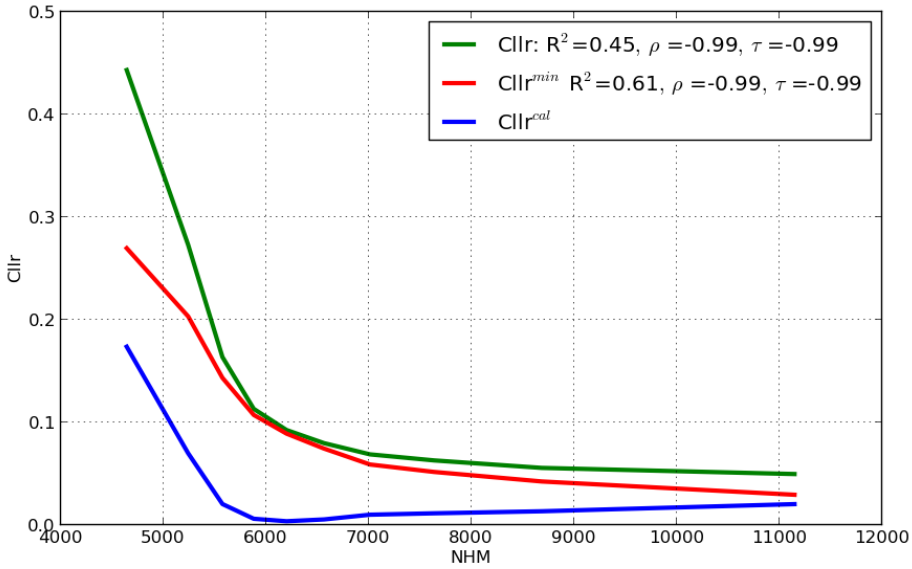

Figure 11.8: *NHM behavioral curve for the pooled condition (all the comparison tests taken together)*

The shape of the curve brings interesting comments. The C_{llr} is decreasing in function of *NHM* with a quite consistent evolution from (*NHM*=4650, C_{llr}=0.44) to (*NHM*=11140, C_{llr}=0.04bits). A large significant correlation between the homogeneity values *NHM* and the C_{llr} is observed, confirmed by a large correlation coefficients, Spearman ρ=-0.99, Kendall's τ=-0.99 and a p-value < 0.001 (R^2 is also provided for comparison purpose).

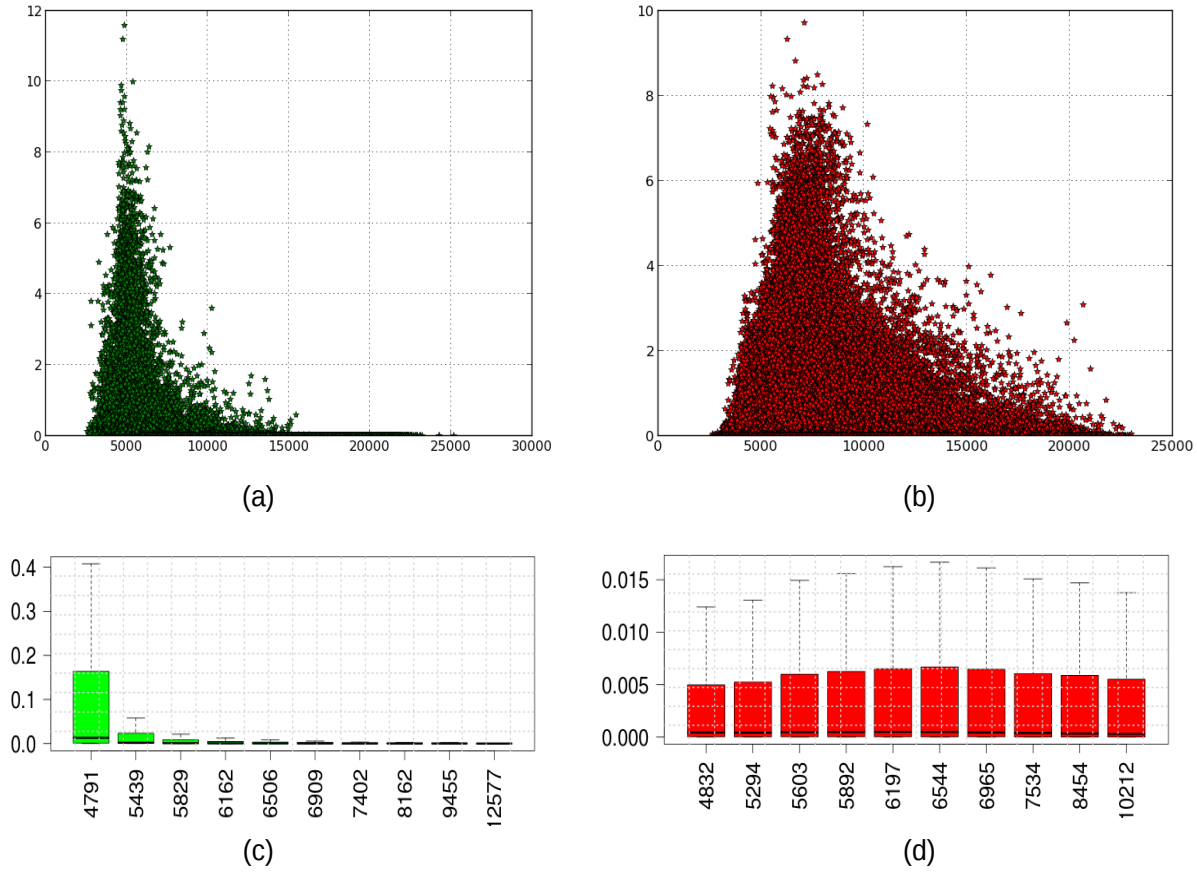

Figure 11.9: C_{llr} *against homogeneity criterion. (a)/(c) are respectively a scatter plot and box-plot (10 bins without outliers) of Cllr against NHM for target trials. (b)/(d) present the same information for non-target trials.*

Figure 11.8, shows also that the calibration error, measured by C_{llr}^{cal}, decreases in function of *NHM* and reaches its minimum value when *NHM* is about 6000. For larger NHM, C_{llr}^{cal} shows a tiny degradation remaining quite low (C_{llr}^{cal} does not exceed 0.02 bits). It is important to note that the largest calibration error, C_{llr}^{cal}=0.16 bits, is observed at the lowest homogeneity value.

Another time, *NHM* appears to be able to predict correctly the performance class of a FVC system in terms of C_{llr} using only the two speech recordings of a voice comparison trial.

In Figure 11.8, the impact of homogeneity on genuine and impostor LRs is not visible. To investigate this point, we present in Figure 11.9 a scatter plot corresponding to (NHM,C_{llr}) as well as box-plot for better visualization. The visualization is proposed separately for target (a and c) and non-target (b and d) trials. Several important comments could be extracted from Figure 11.9 :

- For target trials, it is interesting to notice that both C_{llr} average values and stan-

dard deviation are decreasing when the homogeneity is increasing. The general shape of the curve is clearly exponential. Indeed, C_{llr} average value and standard deviation ($\overline{C_{llr}}$, SD) vary quite consistent between (0.406, 1.06) and (0.004, 0.04). Starting on "bin 4", C_{llr} is becoming infinitesimal as *NHM* increases. This observation indicates that the LR accuracy and reliability are directly linked to the acoustic homogeneity between the two files of a voice comparison trial.

- For non-target trials, the situation is less easy to interpret. First, the right part of the plot shows a behaviour comparable to the previous case: after a specific value, the C_{llr} is decreasing when *NHM* is increasing. But the left part of the curve does not show at all this direct relation between NHM and C_{llr}. Interestingly, these homogeneity values correspond to the largest calibration losses shown in Figure 11.8. For non-target trials, if homogeneity is still a required parameter, it seems that it is not a sufficient criterion to predict the reliability of a voice comparison trial. We could hypothesize that the presence of the same acoustic criterion in both audio files is needed in order to separate two speakers, but is not sufficient: cues able to discriminate these two speakers are also needed. This hypothesis is corroborated by our analysis in Subsection 9.3 where we showed that the phonological information used to discriminate a pair of speakers depends on the speakers themselves.

Conclusion and Discussion

In this chapter, we proposed a family of information theory (IT) based data Homogeneity Measures (HM) in order to predict the LR accuracy. All our measures belong to the Bit Entropy Expectation (BEE). They use a GMM view of the couple of speech recordings corresponding to a voice comparison trial.

The first measure, denoted HM_{BEE}, is directly issued from the degree of homogeneity between the two voice records, expressed in term of bit entropy. This measure showed a significantly high correlation with the speaker recognition outputs. It seems that BEE which ignores the involved quantity of data does not allow to build an homogeneity measure with the desired characteristics.

In order to solve this issue, we proposed a non-normalized version of HM_{BEE}, denoted NHM_{BEE}. Here, the quantity of samples is taken into account so the measure depends on the duration of the two in-interest speech segments. 4 variants of this measure are proposed. The only difference between these variants is the way to cluster the pair of speech recordings.

- The first variant uses the same GMM modelling of the pair of recordings than HM_{BEE}. It showed interesting properties with a nice relation between the homogeneity values and the C_{llr}^{min}, varying quite consistently from (HM=4689,C_{llr}^{min}=0.309) to (HM=7579, C_{llr}^{min}=0.1227).
- A second variant of NHM_{BEE} uses directly the UBM model in order to cluster

the pair of speech recordings (without training or adaptation of the UBM). This version has a similar behaviour than the previous one but outperformed it with a behavioural curve moving from (HM=6341, C_{llr}^{min}=0.3) to (HM=8762, C_{llr}^{min}=0.089).

- The third variant uses GMM A-B initialized with the UBM model. This variant appeared to be highly correlated with C_{llr}^{min}. This was confirmed by a significant high correlation coefficient, R^2=0.92.
- A forth variant uses the mean adapted UBM model. This variant was highly correlated with C_{llr}^{min}. The R^2 is equal to 0.94 and seems to outperform all the other variants.

NHM_{BEE} measures showed a low correlation with the scores issued by the speaker recognition system. The behavioural curve of NHM_{BEE} and this low correlation encourage us strongly to conclude that NHM_{BEE} is a good candidate in order to measure the data homogeneity between a pair of speech recordings, in the view of voice comparison reliability.

The impact of homogeneity was also examined separately on genuine and impostor trials taking advantage of Fabiole database dedicated to these kind of study. For target trials, we found that a good NHM is directly linked to a low level of C_{llr} loss. It says that an acceptable amount of homogeneous acoustic information is enough to authorize the system to evaluate if the two speech recordings come from the same source. For the 70% highest NHM, the corresponding C_{llr} are $\approx$0. For non-target trials, the same behaviour than for target trials is observed for the 70% highest NHM. For the 30% lowest NHM, there is no clear link between NHM and C_{llr}. We could hypothesized that for non-target trials having common acoustic material is not enough and the presence of adequate cues for the in-interest pair of speakers is mandatory. Further work is needed, on a larger number of speakers, in order to precise the scope of this behaviour.

Part IV

Conclusions and Perspectives

Chapter 12

Conclusions and perspectives

Contents

12.1 Conclusions

Forensic Voice Comparison (FVC) is undergoing a "*coming paradigm shift*" (Morrison, 2009a; Gonzalez-Rodriguez et Ramos, 2007) towards emulating other forensic disciplines[1] in quantifying the value of the evidence based on transparent, reliable and testable means of comparing suspect and criminal's speech samples. In this paradigm, the Likelihood ratio (LR) is denoted as the logically and theoretically sounded framework to model and represent forensic expertise. The LR has being increasingly used by the experts and quite often required by "best practice guides" issued by the expert's associations. In FVC, *Automatic Speaker Recognition* (ASpR) is considered as one of the most appropriate solutions when LR framework is involved.

In this thesis, we have addressed the reliability of FVC using automatic speaker recognition system. To this end, three steps were realized.

[1] We refer to forensic sciences using "DNA gold standard".

Understanding the system behaviour

In chapter 8, we decided to not limit ourselves by the traditional evaluation protocols which look only at the global performance of an ASpR system (i.e the ASpR system is always summarized in only one number, EER or other performance measure). A more exploratory phase was performed in which we explored how an ASpR system behaves in different scenarios. Indeed, the interest was the "speaker factor", the "speaker behaviour" as viewed by an ASR system. This investigation took advantage of FABIOLE, a database designed specifically for this kind of investigations.

In a first step, we showed the significant difference of performance across speakers, thus confirming the presence of the "speaker factor". Indeed, when the global C_{llr} (computed using all the speaker subsets put together) is equal to 0.12631 bits, we observed that more than half of the speakers obtain low C_{llr} values (lower than 0.05 bits) and about 10% of the speakers present significantly higher costs (higher than $0.4bits$) compared to the average cost.

In a second step, we pushed further our analysis by looking separately at the information loss relative to target and non-target LRs. These two losses were quantified based on C_{llr}^{TAR} and C_{llr}^{NON} respectively, two measures we derived from C_{llr}. We showed that LR target trials bring in general about two third of C_{llr} loss (0.67% vs 0.33%). This proportion varies significantly when we look at the speaker level (up to 0.94% vs 0.06% for the speaker who presents the largest C_{llr} loss). On the other hand, the information loss for non-target LRs appeared to present a small variation across speakers compared to the target loss. This finding (i.e the huge inter-speaker variation of the information loss related to target trials) suggests strongly the presence of a high intra-speaker variability effect in FVC. This factor should be taken into account in reliability evaluation of the LR.

Therefore, it is of utmost necessity to quantify this effect and deal with it when forensic voice comparison is performed. In other words, experts should be able to tell whether the differences observed between two voice recordings are the origin of within-speaker differences or differences speakers variation.

Finally, we explored deeply the differences in performance by focusing on speaker pairs. We showed that the vast majority of pairs (more than 90%) presents a very low C_{llr}^{NON} ($<$0.01) while few pairs present a quite large C_{llr}^{NON}. This finding raises many questions about the information used to discriminate a pair of speakers.

The main conclusion of this chapter is that averaging the system behaviour over a high number of factors hides many important details. For a better understanding of an ASpR system, we should look at the details (*The devil lies in the details*).

Phonetic impact on FVC and the search of speaker biometric cues

In chapter 9, we investigated the impact of phonemic content on voice comparison process. We used an ASpR system as measurement instrument and, more particularly, the

C_{llr} variations. We analysed the influence of six phonemic classes: nasal vowel, oral vowel, nasal consonant, fricative, plosive and liquid. The experiments were performed using FABIOLE database, a corpora dedicated to voice comparison reliability experiments with a large number of speech extracts per speaker.

In a first step, we investigated the impact of each phonemic class on voice comparison performance using C_{llr} in order to evaluate the information loss. The results showed that oral vowels, nasal vowels and nasal consonants bring more speaker specific information than averaged phonemic content in terms of voice comparison performance. The fricatives do not seem to perform better than an averaged content, which is surprising compared to the literature, but this result is explained by the restricted bandwidth. In order to confirm our intuition, we explored the impact of the bandwidth on FVC. The analysis follows the same logic as the previous one but here we took advantage of the full band (0-8000 Hz). Similar results are obtained compared to the first investigation but with several interesting variations. As expected, the fricatives have shown a positive effect especially for target trials. This outcome suggests that when frequencies above 4 kHz were removed, the fricative consonants became less useful.

In a second step, we explored target and non target parts of C_{llr}. For non-target comparisons, we showed that all the phonemic content play an important role in terms of speaker discrimination power. The oral vowels are the largest contributors, followed by nasals and liquids and this finding is consistent among most speakers. When we focused on target comparisons, oral vowels appeared to be tied with a high variability and thereby this phonemic class is responsible of a high information loss. We saw previously that this phonemic class was bringing a large part of the speaker discrimination power but it appears also very sensitive to intra-speaker variability. In contrast, nasals showed a high capacity for speaker discrimination and at the same time appeared to be robust for intra-speaker variability.

In a third step, we explored the phonetic impact by focusing on speaker pairs. We showed that the phonetic information used to discriminate a pair of speakers depends on the speakers themselves. A deep analysis was dedicated for the 10 "best" and "worst" pairs in terms of speaker discrimination power. We showed that: (i) For the "best" discriminated pairs, all the phonemic content still play a positive role in speaker discrimination. (ii) For the "worst" speaker pairs, it appears that sometimes nasals or fricatives convey a significant part of LR performance loss.

In this study, we highlighted at several steps the importance of speaker factor which is a denomination that reflects mainly intra-speaker variability and differences between the speakers according to this variability. It also includes differences linked to the speaker of ASpR systems responses to a same stimulus. We observed large variations per speaker of the system's responses to different phonemic classes, in terms of relative C_{llr}^{TAR}.

As a consequence of these findings, the main takeaway is the fact that ASpR usual evaluation protocols are mainly selecting the best features in terms of speaker discrimination (C_{llr}^{NON}) and are largely missing intra-speaker variability when the latter is a key factor for numerous application scenarios. This is particularly true for FVC sce-

nario and it appears mandatory to work more on intra-speaker variability as well as on speaker factor in order to estimate the reliability of a solution in this domain.

In chapter 10, we carried a deep study on oral vowels based on formants parameters. This investigation was mainly motivated by the different behaviour of oral vowels shown for target and non-target comparison. Our study aimed to quantify the amount of speaker specific information for each vowel and to study the influence of formant intra-speaker variability on the vowel behaviour. Three main results could be drawn from this investigation. First, speaker and vowel factors effect are observed on formant values. The interaction between the formant value and the vowel is higher than for the formant and the speaker. We showed also that the first 3 formants are more linked to the timbre of the vowel than to the speaker itself. Only the 4^{th} formant are linked to the speaker. Second, we showed that the amount of speaker specific information in each vowel presents a high variability. Indeed, the speaker effect depends on both the formant itself and the pronounced vowel. Third, the study of the influence of the formant intra-speaker variability revealed a potential relation between the size of acoustic variability and the oral vowels behaviour in FVC. The results presented in this chapter could put in question the use of formant measures in acoustic-phonetic FVC.

Homogeneity measure

In chapter 11, we proposed a family of information theory (IT) based data Homogeneity Measures (HM) in order to predict the LR accuracy. All our measures belong to the Bit Entropy Expectation (BEE). They use a GMM view of the pair of speech recordings corresponding to a voice comparison trial.

The first measure, denoted HM_{BEE}, is directly issued from the degree of homogeneity between the two voice records, expressed in term of bit entropy. This measure showed a significantly high correlation with the speaker recognition outputs. It seems that BEE which ignores the involved quantity of data does not allow to build an homogeneity measure with the desired characteristics.

In order to solve this issue, we proposed a non-normalized version of HM_{BEE}, denoted NHM_{BEE}. Here, the quantity of samples is taken into account so the measure depends on the duration of the two in-interest speech segments. 4 variants of this measure are proposed. The only difference between these variants is the way to cluster the pair of speech recordings.

- The first variant uses the same GMM modelling of the pair of recordings than HM_{BEE}.
- A second variant of NHM_{BEE} uses directly a UBM in order to cluster the pair of speech recordings.
- The third variant uses GMM A-B initialized with the UBM model.
- A forth variant uses the mean adapted UBM model.

NHM_{BEE} measures showed a low correlation with the scores issued by the speaker recognition system. The behavioural curve of NHM_{BEE} and this low correlation encourage us strongly to conclude that NHM_{BEE} is a good candidate in order to measure the data homogeneity between a pair of speech recordings, in the view of voice comparison reliability.

The impact of homogeneity was also examined separately on genuine and impostor trials. For target trials, we found that a good NHM is directly linked to a low level of C_{llr} loss. It says that an acceptable amount of homogeneous acoustic information is enough to authorize the system to evaluate if the two speech recordings come from the same source. For non-target trials, the same behaviour than for target trials is observed for the 70% highest NHM. For the 30% lowest NHM, there is no clear link between NHM and C_{llr}. We could hypothesized that for non-target trials having common acoustic material is not enough and the presence of adequate cues for the in-interest pair of speakers is mandatory.

Main takeaways

The aims of this Thesis are mainly to analyse the limits of LR-based FVC and to propose some ways to reinforce its reliability.
We hope that the findings of this Thesis will have a number of implications for LR-based FVC, and potentially other areas of forensic science by extension. The reported results may allow academic and practitioners to better understand and acknowledge the effects of intra-speaker variability[a] on FVC and incite them to pay attention to the different sources of trial variation encountered throughout LR-based analyses. The outcomes reported on the phonetic content impact on the resulting *LR* may help analysts determine which phonetic information to use, based on the magnitude of the speaker-specific-information of each phonemic category. The results reported of the impact of the homogeneity measure on the LR accuracy may make the numerical LR more practically viable for the analysis of FVC evidence in casework.

[a] The numerator of the LR is a similarity term which reflects the probability that the suspect is the origin of the piece of evidence. Generally, this value calls for an assessment of the intra-variability of the system. Ideally, it would approach a value close to 1.

12.2 Perspectives

The research described in this Thesis presents a variety of opportunities for further study:

12.2.1 Phonetic content

Speaker specific information for each phoneme

In this Thesis, we analysed the influence of six phonemic classes on FVC in order to "quantify" the speaker specific information in each class. Or the individual speaker specific information investigated for oral vowels showed a high variability between each vowel. In future work, it is important to consider individual phonemes instead of phoneme classes.

Reliable LR estimation based on phonetic information

The analysis of the phonetic content impact on FVC will contribute towards the integration of this information in the LR estimation process. Indeed, the the speaker-specific-information size of each phonemic category will be used in order to estimate a more reliable LR. This could be of particular interest as ASpR systems are commonly viewed as black boxes and are treated with suspicion by the courts and therefore, using the phonetic information of a FVC trial could provide more transparency to the FVC.

12.2.2 Prediction of speaker profiles

In this study, we showed that -in the view of an ASR-based voice comparison system- speakers do not behave the same way in response of similar condition changes. This observation suggests strongly that speakers could be classified into "speaker profiles". In the following study, it is of utmost interest to confirm or reject this hypothesis. If this hypothesis will be confirmed, it will be also useful to propose an automatic prediction of the speaker profile. A first step in this avenue will be to cross the speaker profiles with as many as possible speaker characteristics. But this work will require a significantly larger database in terms of speakers than FABIOLE (thousands of speakers with hundreds of examples per speakers). Such a database will also allow to experimentally confirm the preliminary results presented in this paper.

It is important to note that characterizing a "speaker profile" according to speaker's characteristics is not as simple as it seems to be. The main point is to study the inter-speaker differences in terms of performance variations when some factors are changing, and to classify correspondingly the speakers into the profiles. This process involves many variation factors as mentioned before such as speaking style and rate, emotion, speaker age or other factors such as speaker accent or dialect, sex, prosody and so forth.

The factors mentioned before should be studied deeply in order to define properly the speaker profile. It is surprising that much less attention were paid to study the effect of intrinsic speaker variability compared to extrinsic factors as noise, and channel or microphone effects.

12.2.3 Homogeneity measure

HM estimation process

This work will firstly extended by working on the GMM representation of the pair of recordings. In addition to this point, the behaviour of our measures depending on the session variability factors should be explored more deeply. Then, as expressed in the introduction of Chapter 11, data homogeneity is a mandatory first step for a voice comparison feasibility measure and we expect to explore this new avenue.

NHM speaker specific

Our homogeneity measure NHM will take greatly benefit to be revised or extended in order to focus on speaker specific information and take into account the speaker discrimination power of each phonetic category. The homogeneity measure should also take into account the trials conditions such as speech quality... This will be the core of our next work, by adding phonetic and voice quality labelling to the current acoustic classes.

NHM calibration at the score level

In a specific field such as forensic voice comparison, calibration is very important in order to make the scores produced by an ASpR system more reliable. Our homogeneity measure appeared to be lined to the LR accuracy. Therefore, it seems reasonable to calibrate the ASpR system scores by incorporating the homogeneity measure in the calibration model. To this end, an efficient straightforward method, inspired by the concept of quality measures presented in (Garcia-Romero et al., 2004), could be applied. The concept of including quality measures into score calibration for improving system performance can be found in numerous works in literature (Hasan et al., 2013; Mandasari et al., 2013). In this approach, the quality measures q_1, q_2,..., q_3 are modelled using a function Q(q_1, q_2,...). This function is integrated in the conventional score calibration as an additional term. Therefore, the general score transformation model is:

$$s^{ca} = w_0 + w_1 s^{raw} + Q(q_1, q_2, q_3...) \tag{12.1}$$

The quality measures could be duration, signal to noise ratio, ... or the homogeneity measure. Therefore, the transformation is modelled as follows:

$$s^{ca} = w_0 + w_1 s^{raw} + Q(HM) \tag{12.2}$$

12.2.4 Databases and protocols

Even though the research presented in this thesis goes deeper in the study of intra-speaker variability than usual, it remains limited by the used database.

FABIOLE offers an opportunity to open the door for research on intra-speaker variability, but has only 100 $\times$30s of speech per speaker, with few contextual variability and with only 30 male speakers. Based on the methodology proposed in this thesis, we aim to conduct in a near future similar analysis on a significantly larger database, or on significantly larger databases, showing several different contexts. An appropriate size for such a database is about one order of magnitude larger than FABIOLE (about 1000 recordings per speaker and hundreds of speakers), which seems realistic in terms of costs as this thesis showed that automatic processes could be largely used.

List of Figures

List of Tables

Bibliography

(AFCP, 2002) AFCP, 2002. Motion adoptée à l'unanimité par le bureau du gcp (groupe de la communication parlée) de la sfa, reconduite intégralement par le gfcp de la sfa en 1997 et par l'afcp en 2002. *(http://www.afcp-parole.org/doc/MOTION_1990.pdf)*.

(Ahuja et Orlin, 2001) R. K. Ahuja & J. B. Orlin, 2001. A fast scaling algorithm for minimizing separable convex functions subject to chain constraints. *Operations Research 49*(5), 784–789.

(Aitken et Leese, 1995) C. G. Aitken & M. Leese, 1995. *Statistics and the evaluation of evidence for forensic scientists*. J. Wiley.

(Aitken et Taroni, 2004) C. G. Aitken & F. Taroni, 2004. *Statistics and the evaluation of evidence for forensic scientists*, Volume 10. Wiley Online Library.

(Ajili et al., 2018) M. Ajili, J. Bonastre, & S. Rossato, 2018. Voice comparison and rhythm: Behavioral differences between target and non-target comparisons. Dans les actes de *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018.*, 1061–1065.

(Ajili et al., 2016) M. Ajili, J.-F. Bonastre, W. Ben Kheder, S. Rossato, & J. Kahn, 2016. Phonetic content impact on forensic voice comparison. Dans les actes de *Spoken Language Technology Workshop (SLT)*, 210–217. IEEE.

(Ajili et al., 2017a) M. Ajili, J.-F. Bonastre, W. B. Kheder, S. Rossato, & J. Kahn, 2017a. Homogeneity measure impact on target and non-target trials in forensic voice comparison. *Proc. Interspeech 2017*, 2844–2848.

(Ajili et al., 2017b) M. Ajili, J.-F. Bonastre, W. B. Kheder, S. Rossato, & J. Kahn, 2017b. Phonological content impact on wrongful convictions in forensic voice comparison context. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

(Ajili et al., 2016a) M. Ajili, J.-F. Bonastre, S. Rossato, & J. Kahn, 2016a. Inter-speaker variability in forensic voice comparison:a preliminary evaluation. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

(Ajili et al., 2016b) M. Ajili, J.-F. Bonastre, S. Rossato, J. Kahn, & G. Bernard, 2016b. Fabiole, a speech database for forensic speaker comparison. *International Conference on Language Resources, Evaluation and Corpora*.

(Ajili et al., 2015a) M. Ajili, J.-F. Bonastre, S. Rossato, J. Kahn, & I. Lapidot, 2015a. Homogeneity measure for forensic voice comparison: A step forward reliability. Dans les actes de *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, 135–142. Springer.

(Ajili et al., 2015b) M. Ajili, J.-F. Bonastre, S. Rossato, J. Kahn, & I. Lapidot, 2015b. An information theory based data-homogeneity measure for voice comparison. Dans les actes de *Interspeech 2015*.

(Ajili et al., 2018a) M. Ajili, J.-F. Bonastre, B. K. Waad, S. Rossato, & J. Kahn, 2018a. Comparaison des voix dans le cadre judiciaire: influence du contenu phonétique.

(Ajili et al., 2018b) M. Ajili, S. Rossato, D. Zhang, & J. Bonastre, 2018b. Impact of rhythm on forensic voice comparison reliability. Dans les actes de *Odyssey, The Speaker and Language Recognition Workshop (pp. 1-8)*.

(Alexander, 2005) A. Alexander, 2005. *Forensic automatic speaker recognition using Bayesian interpretation and statistical compensation for mismatched conditions*. Thèse de Doctorat, Institut de traitement des signaux SECTION DE GÉNIE ÉLECTRIQUE ET ÉLECTRONIQUE POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES PAR Bachelor of Technology in Computer Science and Engineering, Indian Institute of Technology, Madras.

(Alexander et al., 2004) A. Alexander, F. Botti, D. Dessimoz, & A. Drygajlo, 2004. The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications. *Forensic science international 146*, S95–S99.

(Alexander et al., 2005) A. Alexander, D. Dessimoz, F. Botti, & A. Drygajlo, 2005. Aural and automatic forensic speaker recognition in mismatched conditions. *International Journal of Speech Language and the Law 12*(2), 214.

(Alzqhoul, 2015) E. Alzqhoul, 2015. *Impact of the CDMA Mobile Phone Network on Speech Used for Forensic Voice Comparison*. Thèse de Doctorat, ResearchSpace@ Auckland.

(Alzqhoul et al., 2012) E. Alzqhoul, B. Nair, & B. J. Guillemin, 2012. Speech handling mechanisms of mobile phone networks and their potential impact on forensic voice analysis. *Proc. SST-12*.

(Alzqhoul et al., 2016) E. A. Alzqhoul, B. B. Nair, & B. J. Guillemin, 2016. Impact of background noise in mobile phone networks on forensic voice comparison. *J Forensic Leg Investig Sci 2*(007).

(Amino et al., 2012) K. Amino, T. Osanai, T. Kamada, H. Makinae, & T. Arai, 2012. Effects of the phonological contents and transmission channels on forensic speaker recognition. Dans les actes de *Forensic Speaker Recognition*, 275–308. Springer.

(Amino et al., 2006) K. Amino, T. Sugawara, & T. Arai, 2006. Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties. *Acoustical science and technology 27*(4), 233–235.

(Ando et al., 2016) A. Ando, T. Asami, Y. Yamaguchi, & Y. Aono, 2016. Speaker recognition in duration-mismatched condition using bootstrapped i-vectors. Dans les actes de *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific*, 1–4. IEEE.

(Antal et Toderean, 2006) M. Antal & G. Toderean, 2006. Speaker recognition and broad phonetic groups. Dans les actes de *SPPRA*, 155–159.

(Atal, 1976) B. S. Atal, 1976. Automatic recognition of speakers from their voices. *Proceedings of the IEEE 64*(4), 460–475.

(Auckenthaler et al., 2000) R. Auckenthaler, M. Carey, & H. Lloyd-Thomas, 2000. Score normalization for text-independent speaker verification systems. *Digital Signal Processing 10*(1), 42–54.

(Auckenthaler et al., 2001) R. Auckenthaler, M. J. Carey, & J. S. Mason, 2001. Language dependency in text-independent speaker verification. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Volume 1, 441–444. IEEE.

(Austin, 1970) J. L. Austin, 1970. Quand dire, c'est faire= how to do things with words.

(Bao et al., 2007) H. Bao, M.-X. Xu, & T. F. Zheng, 2007. Emotion attribute projection for speaker recognition on emotional speech. Dans les actes de *INTERSPEECH*, 758–761.

(Becker et al., 2008) T. Becker, M. Jessen, & C. Grigoras, 2008. Forensic speaker verification using formant features and gaussian mixture models. Dans les actes de *Interspeech*, 1505–1508.

(Bedau et Radelet, 1987) H. A. Bedau & M. L. Radelet, 1987. Miscarriages of justice in potentially capital cases. *Stanford Law Review*, 21–179.

(Bell, 2008) S. Bell, 2008. *Encyclopedia of forensic science*. Infobase Publishing.

(Ben Kheder et al., 2016) W. Ben Kheder, D. Matrouf, M. Ajili, & J.-F. Bonastre, 2016. Probabilistic approach using joint long and short session i-vectors modeling to deal with short utterances for speaker recognition. Dans les actes de *INTERSPEECH*, 1830–1834.

(Besacier et al., 2000) L. Besacier, J.-F. Bonastre, & C. Fredouille, 2000. Localization and selection of speaker-specific information with statistical modeling. *Speech Communication 31*(2), 89–106.

(Biedermann et al., 2016) A. Biedermann, S. Bozza, & F. Taroni, 2016. The decisionalization of individualization. *Forensic science international 266*, 29–38.

(Bigot et al., 2013) B. Bigot, G. Senay, G. Linares, C. Fredouille, & R. Dufour, 2013. Combining acoustic name spotting and continuous context models to improve spoken person name recognition in speech. Dans les actes de *INTERSPEECH*, 2539–2543.

(Bilmes et al., 1998) J. A. Bilmes et al., 1998. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute* *4*(510), 126.

(Boë, 2000) L.-J. Boë, 2000. Forensic voice identification in france. *Speech Communication* *31*(2), 205–224.

(Boë et al., 1999) L.-J. Boë, F. Bimbot, J.-F. Bonastre, & P. Dupont, 1999. De l'évaluation des systèmes de vérification du locuteur à la mise en cause des expertises vocales en identification juridique. *Langues* *2*(4), 270–288.

(Boersma et Weenink, 2008) P. Boersma & D. Weenink, 2008. Praat: doing phonetics by computer (version 5.0. 35)[computer program]. retrieved september 23, 2008.

(Bolle et al., 2003) R. Bolle, J. Connell, S. Pankanti, N. Ratha, & A. Senior, 2003. Guide to biometrics.

(Bolt et al., 1970) R. H. Bolt, F. S. Cooper, E. E. David Jr, P. B. Denes, J. M. Pickett, & K. N. Stevens, 1970. Speaker identification by speech spectrograms: a scientists' view of its reliability for legal purposes. *The Journal of the Acoustical Society of America* *47*(2B), 597–612.

(Bonastre et al., 2003) J.-F. Bonastre, F. Bimbot, L.-J. Boë, J. P. Campbell, D. A. Reynolds, & I. Magrin-Chagnolleau, 2003. Person authentication by voice: a need for caution. Dans les actes de *INTERSPEECH*.

(Bonastre et al., 2015) J. F. Bonastre, J. Kahn, S. Rossato, & M. Ajili, 2015. Forensic speaker recognition: Mirages and reality. Dans les actes de *In Fuchs, S., Pape, D., Petrone, C., Perrier, P. (2015). Individual Differences in Speech Production and Perception.*

(Bonastre et al., 2008) J.-F. Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. W. Evans, B. G. Fauve, & J. S. Mason, 2008. Alize/spkdet: a state-of-the-art open source software for speaker recognition. Dans les actes de *Odyssey*, 20.

(Bonastre et al., 2005) J.-F. Bonastre, F. Wils, & S. Meignier, 2005. Alize, a free toolkit for speaker recognition. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 737–740.

(Bousquet et al., 2014) P.-M. Bousquet, J.-F. Bonastre, & D. Matrouf, 2014. Exploring some limits of gaussian plda modeling for i-vector distributions. Dans les actes de *Odyssey: The Speaker and Language Recognition Workshop*.

(Bousquet et al., 2012) P.-M. Bousquet, A. Larcher, D. Matrouf, J.-F. Bonastre, & O. Plchot, 2012. Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis. Dans les actes de *Odyssey*, 157–164.

(Bousquet et al., 2011) P.-M. Bousquet, D. Matrouf, J.-F. Bonastre, et al., 2011. Intersession compensation and scoring methods in the i-vectors space for speaker recognition. Dans les actes de *Interspeech*, 485–488.

(Boves, 1998) L. Boves, 1998. Commercial applications of speaker verification: overview and critical succes factors.

(Broeders, 2007) A. Broeders, 2007. Some observations on the use of probability scales in forensic identification. *International Journal of Speech Language and the Law* *6*(2), 228–241.

(Brummer, 2007) N. Brummer, 2007. Focal toolkit. *Available in http://www. dsp. sun. ac. za/nbrummer/focal*.

(Brümmer et al., 2007) N. Brümmer, L. Burget, J. H. Černockỳ, O. Glembek, F. Grezl, M. Karafiat, D. A. Van Leeuwen, P. Matě, P. Schwarz, & A. Strasheim, 2007. Fusion of heterogeneous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006. *Audio, Speech, and Language Processing* *15*(7), 2072–2084.

(Brümmer et De Villiers, 2010) N. Brümmer & E. De Villiers, 2010. The speaker partitioning problem. Dans les actes de *Odyssey*, 34.

(Brümmer et du Preez, 2006) N. Brümmer & J. du Preez, 2006. Application-independent evaluation of speaker detection. *Computer Speech & Language* *20*(2), 230–275.

(Brummer et van Leeuwen, 2006) N. Brummer & D. A. van Leeuwen, 2006. On calibration of language recognition scores. Dans les actes de *Speaker and Language Recognition Workshop, 2006. IEEE Odyssey 2006: The*, 1–8. IEEE.

(BT Nair et al., 2014) B. BT Nair, E. A. Alzqhoul, & B. J. Guillemin, 2014. Impact of mismatch conditions between mobile phone recordings on forensic voice comparison. *The Journal of the Acoustical Society of America* *136*(4), 2083–2083.

(Buckleton et al., 2006) J. Buckleton, C. Triggs, & C. Champod, 2006. An extended likelihood ratio framework for interpreting evidence. *Science & Justice* *46*(2), 69–78.

(Buckleton et al., 2016) J. S. Buckleton, J.-A. Bright, & D. Taylor, 2016. *Forensic DNA evidence interpretation*. CRC press.

(Burget et al., 2011) L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matějka, & N. Brümmer, 2011. Discriminatively trained probabilistic linear discriminant analysis for speaker verification. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4832–4835. IEEE.

(Campbell, 1997) J. P. Campbell, 1997. Speaker recognition: A tutorial. *Proceedings of the IEEE* *85*(9), 1437–1462.

(Campbell, 2014) J. P. Campbell, 2014. Speaker recognition for forensic applications. *Odyssey Keynote: FSR 2*.

(Campbell et al., 2009) J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J.-F. Bonastre, & D. Matrouf, 2009. Forensic speaker recognition. Institute of Electrical and Electronics Engineers.

(Campbell et al., 2006) W. M. Campbell, D. E. Sturim, & D. A. Reynolds, 2006. Support vector machines using gmm supervectors for speaker verification. *IEEE signal processing letters 13*(5), 308–311.

(Castro, 2007) D. R. Castro, 2007. *Forensic evaluation of the evidence using automatic speaker recognition systems*. Thèse de Doctorat, Universidad autónoma de Madrid.

(Champod et al., 2001) C. Champod, I. W. Evett, & B. Kuchler, 2001. Earmarks as evidence: a critical review. *Journal of Forensic science 46*(6), 1275–1284.

(Champod et al., 2016) C. Champod, C. J. Lennard, P. Margot, & M. Stoilovic, 2016. *Fingerprints and other ridge skin impressions*. CRC press.

(Champod et Meuwly, 1998) C. Champod & D. Meuwly, 1998. The inference of identity in forensic speaker recognition. Dans les actes de *RLA2C Workshop: Speaker Recognition and its Commercial and Forensic Applications*, 125–135.

(Champod et Meuwly, 2000) C. Champod & D. Meuwly, 2000. The inference of identity in forensic speaker recognition. *Speech Communication 31*(2), 193–203.

(Chaudhari et al., 2003) U. V. Chaudhari, J. Navrátil, & S. H. Maes, 2003. Multi-grained modeling with pattern specific maximum likelihood transformations for text-independent speaker recognition. *IEEE Transactions on Speech and Audio Processing 11*(1), 61–69.

(Clifford, 1980) B. R. Clifford, 1980. Voice identification by human listeners: On earwitness reliability. *Law and Human Behavior 4*(4), 373.

(Cohen, 1977) J. Cohen, 1977. Statistical power analysis for the behavioral sciences (revised ed.).

(Cohen, 1988) J. Cohen, 1988. Statistical power analysis for the behavior science. *Lawrance Eribaum Association*.

(Cole, 2009) S. A. Cole, 2009. Forensics without uniqueness, conclusions without individualization: the new epistemology of forensic identification. *Law, Prob. & Risk 8*, 233.

(Cole, 2014) S. A. Cole, 2014. Individualization is dead, long live individualization! reforms of reporting practices for fingerprint analysis in the united states. *Law, Probability and Risk*, mgt014.

(COUNCIL et al., 1992) N. R. COUNCIL et al., 1992. Committee on dna technology in forensic science, dna technology in forensic science.

(Council et al., 1996) N. R. Council et al., 1996. *The evaluation of forensic DNA evidence*. National Academies Press.

(Crane et Ostrem, 1983) H. D. Crane & J. S. Ostrem, 1983. Automatic signature verification using a three-axis force-sensitive pen. *IEEE Transactions on Systems, Man, and Cybernetics* (3), 329–337.

(Crothers, 1978) J. Crothers, 1978. Typology and universals of vowel systems in phonology. *Universals of human language 2*, 95–152.

(Dang et Honda, 1996) J. Dang & K. Honda, 1996. Acoustic characteristics of the human paranasal sinuses derived from transmission characteristic measurement and morphological observation. *The Journal of the Acoustical Society of America 100*(5), 3374–3383.

(Das et al., 2013) B. Das, S. Mandal, P. Mitra, & A. Basu, 2013. Effect of aging on speech features and phoneme recognition: a study on bengali voicing vowels. *International journal of speech technology 16*(1), 19–31.

(Daubert, 1993) Daubert, 1993. Daubert v. merrell dow pharmaceuticals, inc.

(Davis et Mermelstein, 1980) S. Davis & P. Mermelstein, 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing 28*(4), 357–366.

(Dehak et al., 2007) N. Dehak, P. Dumouchel, & P. Kenny, 2007. Modeling prosodic features with joint factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing 15*(7), 2095–2103.

(Dehak et al., 2011) N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, & P. Ouellet, 2011. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing 19*(4), 788–798.

(Dessimoz et Champod, 2008) D. Dessimoz & C. Champod, 2008. Linkages between biometrics and forensic science. Dans les actes de *Handbook of biometrics*, 425–459. Springer.

(Doddington, 2012) G. Doddington, 2012. The role of score calibration in speaker recognition. Dans les actes de *Thirteenth Annual Conference of the International Speech Communication Association*.

(Doddington et al., 1998) G. Doddington, W. Liggett, A. Martin, M. Przybocki, & D. Reynolds, 1998. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. Rapport technique, DTIC Document.

(Doddington, 1985) G. R. Doddington, 1985. Speaker recognition—identifying people by their voices. *Proceedings of the IEEE 73*(11), 1651–1664.

(Doyle, 2013) A. C. Doyle, 2013. *Sherlock Holmes: The complete novels and stories*, Volume 1. Bantam classics.

(Drozd et al., 2016) L. Drozd, M. Saini, & N. Olesen, 2016. *Parenting plan evaluations: Applied research for the family court*. Oxford University Press.

(Drygajlo, 2007) A. Drygajlo, 2007. Forensic automatic speaker recognition. *IEEE Signal Processing Magazine 24*(2), 132–135.

(Drygajlo, 2012) A. Drygajlo, 2012. Automatic speaker recognition for forensic case assessment and interpretation. Dans les actes de *Forensic Speaker Recognition*, 21–39. Springer.

(Drygajlo, 2015) A. Drygajlo, 2015. Forensic evidence of voice. *Encyclopedia of Biometrics*, 1594–1602.

(Drygajlo et Haraksim, 2017) A. Drygajlo & R. Haraksim, 2017. Biometric evidence in forensic automatic speaker recognition. Dans les actes de *Handbook of Biometrics for Forensic Science*, 221–239. Springer.

(Drygajlo et al., 2015) A. Drygajlo, M. Jessen, S. Gfroerer, I. Wagner, J. Vermeulen, & T. Niemi, 2015. Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition. *European Network of Forensic Science Institutes*.

(Drygajlo et al., 2016) A. Drygajlo, M. Jessen, S. Gfroerer, I. Wagner, J. Vermeulen, & T. Niemi, 2016. *Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition*. Verlag für Polizeiwissenschaft.

(Drygajlo et al., 2003) A. Drygajlo, D. Meuwly, & A. Alexander, 2003. Statistical methods and bayesian interpretation of evidence in forensic automatic speaker recognition. Dans les actes de *INTERSPEECH*.

(Eatock et Mason, 1994) J. P. Eatock & J. S. Mason, 1994. A quantitative assessment of the relative speaker discriminating properties of phonemes. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Volume 1, I–133. IEEE.

(Eckert et Rickford, 2001) P. Eckert & J. R. Rickford, 2001. *Style and sociolinguistic variation*. Cambridge University Press.

(Enzinger et Morrison, 2015) E. Enzinger & G. S. Morrison, 2015. Mismatched distances from speakers to telephone in a forensic-voice-comparison case. *Speech Communication 70*, 28–41.

(Evett et al., 2000) I. Evett, G. Jackson, J. Lambert, & S. McCrossan, 2000. The impact of the principles of evidence interpretation on the structure and content of statements. *Science & Justice 40*(4), 233–239.

(Evett, 1983) I. W. Evett, 1983. What is the probability that this blood came from that person? a meaningful question? *Journal of the Forensic Science Society 23*(1), 35–39.

(Evett et Weir, 1998) I. W. Evett & B. S. Weir, 1998. *Interpreting DNA evidence: statistical genetics for forensic science*. Sunderland, Massachusetts.

(Fant, 1970) G. Fant, 1970. {Acoustic Theory of Speech Production}.

(Forst, 2004) B. Forst, 2004. *Errors of justice: Nature, sources and remedies*. Cambridge University Press.

(French et al., 2010) J. French, F. Nolan, P. Foulkes, P. Harrison, & K. McDougall, 2010. The uk position statement on forensic speaker comparison: a rejoinder to rose and morrison. *International journal of speech, language and the law 17*(1), 143–152.

(French et Harrison, 2007) P. French & P. Harrison, 2007. Position statement concerning use of impressionistic likelihood terms in forensic speaker comparison cases. *International Journal of Speech Language and the Law 14*(1), 137.

(Fritz et al., 2012) C. O. Fritz, P. E. Morris, & J. J. Richler, 2012. Effect size estimates: current use, calculations, and interpretation. *Journal of Experimental Psychology: General 141*(1), 2.

(Furui, 1981) S. Furui, 1981. Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing 29*(2), 254–272.

(Galibert et Kahn, 2013) O. Galibert & J. Kahn, 2013. The first official repere evaluation. Dans les actes de *SLAM@ INTERSPEECH*, 43–48.

(Gallardo et al., 2014a) L. F. Gallardo, M. Wagner, & S. Möller, 2014a. Advantages of wideband over narrowband channels for speaker verification employing mfccs and lfccs. Dans les actes de *INTERSPEECH*, 1115–1119.

(Gallardo et al., 2014b) L. F. Gallardo, M. Wagner, & S. Möller, 2014b. I-vector speaker verification based on phonetic information under transmission channel effects. Dans les actes de *INTERSPEECH*, 696–700.

(Galliano et al., 2006) S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, & K. Choukri, 2006. Corpus description of the ester evaluation campaign for the rich transcription of french broadcast news. Dans les actes de *Proceedings of LREC*, Volume 6, 315–320.

(Galliano et al., 2009) S. Galliano, G. Gravier, & L. Chaubard, 2009. The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. Dans les actes de *Interspeech*, Volume 9, 2583–2586.

(Ganapathy et al., 2011) S. Ganapathy, J. Pelecanos, & M. K. Omar, 2011. Feature normalization for speaker verification in room reverberation. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4836–4839. IEEE.

(Garcia-Romero et Espy-Wilson, 2011) D. Garcia-Romero & C. Y. Espy-Wilson, 2011. Analysis of i-vector length normalization in speaker recognition systems. Dans les actes de *Interspeech*, 249–252.

(Garcia-Romero et al., 2004) D. Garcia-Romero, J. Fiérrez-Aguilar, J. Gonzalez-Rodriguez, & J. Ortega-Garcia, 2004. On the use of quality measures for text-independent speaker recognition. Dans les actes de *ODYSSEY04-the speaker and language recognition workshop*.

(Gauvain et Lee, 1994) J.-L. Gauvain & C.-H. Lee, 1994. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE transactions on speech and audio processing* *2*(2), 291–298.

(Gay, 1968) T. Gay, 1968. Effect of speaking rate on diphthong formant movements. *The Journal of the Acoustical Society of America* *44*(6), 1570–1573.

(Gay, 1978) T. Gay, 1978. Effect of speaking rate on vowel formant movements. *The journal of the Acoustical society of America* *63*(1), 223–230.

(GFCP, 1999) GFCP, 1999. Pétition pour l'arrêt des expertises vocales tant qu'elles n'auront pas été validées scientifiquement. *(http://www.afcp-parole.org/doc/petition.pdf)*.

(Gialamas, 2000) D. Gialamas, 2000. Criminalistics. *Encyclopedia of Forensic Sciences, Three-Volume set*, 471–477.

(Giannelli et al., 1993) P. C. Giannelli, E. J. Imwinkelried, A. Roth, & J. C. Moriarty, 1993. *Scientific evidence*. Michie Company.

(Gibbon et al., 1997) D. Gibbon, R. Moore, & R. Winski, 1997. *Handbook of standards and resources for spoken language systems*. Walter de Gruyter.

(Glembek et al., 2009) O. Glembek, L. Burget, N. Dehak, N. Brummer, & P. Kenny, 2009. Comparison of scoring methods used in speaker recognition with joint factor analysis. Dans les actes de *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 4057–4060. IEEE.

(Gold et French, 2011) E. Gold & P. French, 2011. An international investigation of forensic speaker comparison practices. Dans les actes de *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, China*, 1254–1257.

(Gonzalez-Rodriguez et al., 2006) J. Gonzalez-Rodriguez, A. Drygajlo, D. Ramos-Castro, M. Garcia-Gomar, & J. Ortega-Garcia, 2006. Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech & Language* *20*(2), 331–355.

(Gonzalez-Rodriguez et Ramos, 2007) J. Gonzalez-Rodriguez & D. Ramos, 2007. Forensic automatic speaker classification in the "coming paradigm shift". Dans les actes de *Speaker Classification I*, 205–217. Springer.

(Gonzalez-Rodriguez et al., 2007) J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, & J. Ortega-Garcia, 2007. Emulating dna: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *Audio, Speech, and Language Processing* *15*(7), 2104–2115.

(Gordon et al., 1992) C. Gordon, D. L. Webb, & S. Wolpert, 1992. One cannot hear the shape of a drum. *Bulletin of the American Mathematical Society* *27*(1), 134–138.

(Gordon et al., 2002) M. Gordon, P. Barthmaier, & K. Sands, 2002. A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association 32*(02), 141–174.

(Gravier et al., 2012) G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel, O. Galibert, et al., 2012. The etape corpus for the evaluation of speech-based tv content processing in the french language. *International Conference on Language Resources, Evaluation and Corpora*.

(Greenberg et al., 2011) C. S. Greenberg, A. F. Martin, B. N. Barr, & G. R. Doddington, 2011. Report on performance results in the nist 2010 speaker recognition evaluation. Dans les actes de *Twelfth Annual Conference of the International Speech Communication Association*.

(Greenberg et al., 2013) C. S. Greenberg, V. M. Stanford, A. F. Martin, M. Yadagiri, G. R. Doddington, J. J. Godfrey, & J. Hernandez-Cordero, 2013. The 2012 nist speaker recognition evaluation. Dans les actes de *INTERSPEECH*, 1971–1975.

(Grimaldi et Cummins, 2009) M. Grimaldi & F. Cummins, 2009. Speech style and speaker recognition: a case study. Dans les actes de *INTERSPEECH*, 920–923.

(Gupta et Chatterjee, 2012) S. Gupta & S. Chatterjee, 2012. Text dependent voice based biometric authentication system using spectrum analysis and image acquisition. *Advances in Computer Science, Engineering & Applications*, 61–70.

(Hansen et Hasan, 2015) J. H. Hansen & T. Hasan, 2015. Speaker recognition by machines and humans: A tutorial review. *IEEE Signal processing magazine 32*(6), 74–99.

(Hasan et al., 2013) T. Hasan, R. Saeidi, J. H. Hansen, & D. A. van Leeuwen, 2013. Duration mismatch compensation for i-vector based speaker recognition systems. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7663–7667. IEEE.

(Hatch et al., 2006) A. O. Hatch, S. S. Kajarekar, & A. Stolcke, 2006. Within-class covariance normalization for svm-based speaker recognition. Dans les actes de *Interspeech*.

(Hébert et Heck, 2003) M. Hébert & L. P. Heck, 2003. Phonetic class-based speaker verification. Dans les actes de *INTERSPEECH*.

(Hermansky et Morgan, 1994) H. Hermansky & N. Morgan, 1994. Rasta processing of speech. *IEEE transactions on speech and audio processing 2*(4), 578–589.

(Hill, 1996) R. Hill, 1996. Retina identification. *Biometrics: Personal Identification in Networked Society*, 123–141.

(Hillenbrand et al., 2001) J. M. Hillenbrand, M. J. Clark, & T. M. Nearey, 2001. Effects of consonant environment on vowel formant patterns. *The Journal of the Acoustical Society of America 109*(2), 748–763.

(Hofker, 1977) U. Hofker, 1977. Auros-automatic recognition of speakers by computers: phoneme ordering for speaker recognition. Dans les actes de *Proc. 9th International Congress on'Acoustics, Madrid*, 506–507.

(Hong et al., 2015) Q. Hong, L. Li, M. Li, L. Huang, L. Wan, & J. Zhang, 2015. Modified-prior plda and score calibration for duration mismatch compensation in speaker recognition system. Dans les actes de *INTERSPEECH*, 1037–1041.

(Huber et Headrick, 1999) R. A. Huber & A. M. Headrick, 1999. *Handwriting identification: facts and fundamentals*. CRC press Boca Raton.

(Hughes, 2014) V. Hughes, 2014. *The definition of the relevant population and the collection of data for likelihood ratio-based forensic voice comparison*. Thèse de Doctorat.

(Hyon et al., 2012) S. Hyon, H. Wang, J. Wei, & J. Dang, 2012. An investigation of dependencies between frequency components and speaker characteristics based on phoneme mean f-ratio contribution. Dans les actes de *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, 1–4. IEEE.

(Inman et Rudin, 2000) K. Inman & N. Rudin, 2000. *Principles and practice of criminalistics: the profession of forensic science*. CRC Press.

(Ishihara et al., 2008) S. Ishihara, Y. Kinoshita, et al., 2008. How many do we need? exploration of the population size effect on the performance of forensic speaker classification.

(Jackson et al., 2015) G. Jackson, C. Aitken, & P. Roberts, 2015. *Case assessment and interpretation of expert evidence: guidance for judges, lawyers, forensic scientists and expert witnesses*.

(Jackson et al., 2006) G. Jackson, S. Jones, G. Booth, C. Champod, & I. Evett, 2006. Scientific and technical articles-the nature of forensic science opinion–a possible framework to guide thinking and practice in investigations and in court proceedings. *Science and Justice 46*(1), 33–44.

(Jain et al., 2007) A. Jain, P. Flynn, & A. A. Ross, 2007. *Handbook of biometrics*. Springer Science & Business Media.

(Jain et al., 2005) A. Jain, D. Maltoni, D. Maio, & J. Wayman, 2005. Biometric systems technology, design and performance evaluation. *Springer-Verlag London limited, 2005*.

(Jain et al., 2011) A. Jain, A. A. Ross, & K. Nandakumar, 2011. *Introduction to biometrics*. Springer Science & Business Media.

(Jain et al., 2000) A. K. Jain, R. P. W. Duin, & J. Mao, 2000. Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence 22*(1), 4–37.

(Jain et al., 2016) A. K. Jain, K. Nandakumar, & A. Ross, 2016. 50 years of biometric research: Accomplishments, challenges, and opportunities. *Pattern Recognition Letters 79*, 80–105.

(Jain et al., 2004) A. K. Jain, A. Ross, & S. Prabhakar, 2004. An introduction to biometric recognition. *IEEE Transactions on circuits and systems for video technology 14*(1), 4–20.

(James et al., 2002) S. H. James, J. J. Nordby, & S. Bell, 2002. *Forensic science: an introduction to scientific and investigative techniques*. CRC press.

(Jessen et al., 2007) M. Jessen, O. Koster, & S. Gfroerer, 2007. Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech Language and the Law 12*(2), 174–213.

(Jongman et al., 2000) A. Jongman, R. Wayland, & S. Wong, 2000. Acoustic characteristics of english fricatives. *The Journal of the Acoustical Society of America 108*(3), 1252–1263.

(Juliette, 2011) K. Juliette, 2011. *Parole de locuteur: performance et confiance en identification biométrique vocale*. Thèse de Doctorat, Université d'Avignon et des Pays de Vaucluse.

(Junqua et al., 1991) J.-C. Junqua, B. Reaves, & B. Mak, 1991. A study of endpoint detection algorithms in adverse conditions: incidence on a dtw and hmm recognizer. Dans les actes de *Second European Conference on Speech Communication and Technology*.

(Kac, 1966) M. Kac, 1966. Can one hear the shape of a drum? *The american mathematical monthly 73*(4), 1–23.

(Kahn et al., 2011) J. Kahn, N. Audibert, J.-F. Bonastre, & S. Rossato, 2011. Inter and intraspeaker variability in french: an analysis of oral vowels and its implication for automatic speaker verification. Dans les actes de *International Congress of Phonetic Sciences (ICPhS)*, 1002–1005.

(Kahn et al., 2010) J. Kahn, N. Audibert, S. Rossato, & J.-F. Bonastre, 2010. Intra-speaker variability effects on speaker verification performance. Dans les actes de *Odyssey*, 21.

(Kahn et al., 2012) J. Kahn, O. Galibert, L. Quintard, M. Carré, A. Giraudel, & P. Joly, 2012. A presentation of the repere challenge. Dans les actes de *10th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 1–6. IEEE.

(Kajarekar et al., 2006) S. S. Kajarekar, H. Bratt, E. Shriberg, & R. De Leon, 2006. A study of intentional voice modifications for evading automatic speaker recognition. Dans les actes de *2006 IEEE Odyssey-The Speaker and Language Recognition Workshop*, 1–6. IEEE.

(Karlsson et al., 2000) I. Karlsson, T. Banziger, J. Dankovicová, T. Johnstone, J. Lindberg, H. Melin, F. Nolan, & K. Scherer, 2000. Speaker verification with elicited speaking styles in the verivox project. *Speech Communication 31*(2), 121–129.

(Karlsson et al., 1998) I. Karlsson, T. Bänziger, J. Dankovicova, T. Johnstone, J. Lindberg, H. Melin, F. Nolan, & K. R. Scherer, 1998. Within-speaker variability due to speaking manners. Dans les actes de *ICSLP*.

(Kashyap, 1976) R. Kashyap, 1976. Speaker recognition from an unknown utterance and speaker-speech interaction. *IEEE Transactions on Acoustics, Speech, and Signal Processing 24*(6), 481–488.

(Kay, 2005) R. Kay, 2005. Biometric authentication. *Computerworld 39*(14), 26.

(Kelly et al., 2012) F. Kelly, A. Drygajlo, & N. Harte, 2012. Speaker verification with long-term ageing data. Dans les actes de *5th IAPR International Conference on Biometrics (ICB)*, 478–483. IEEE.

(Kelly et Hansen, 2015) F. Kelly & J. H. Hansen, 2015. The effect of short-term vocal aging on automatic speaker recognition performance.

(Kelly et Harte, 2011) F. Kelly & N. Harte, 2011. Effects of long-term ageing on speaker verification. Dans les actes de *European Workshop on Biometrics and Identity Management*, 113–124. Springer.

(Kelly et al., 2014) F. Kelly, R. Saeidi, N. Harte, & D. A. van Leeuwen, 2014. Effect of long-term ageing on i-vector speaker verification. Dans les actes de *INTERSPEECH*, 86–90.

(Kennedy, 2003) D. Kennedy, 2003. Forensic science: oxymoron? *Science 302*(5651), 1625–1625.

(Kenny, 2010) P. Kenny, 2010. Bayesian speaker verification with heavy-tailed priors. Dans les actes de *Odyssey*, 14.

(Kenny et al., 2007) P. Kenny, G. Boulianne, P. Ouellet, & P. Dumouchel, 2007. Speaker and session variability in gmm-based speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing 15*(4), 1448–1460.

(Kenny et Dumouchel, 2004) P. Kenny & P. Dumouchel, 2004. Disentangling speaker and channel effects in speaker verification. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Volume 1, I–37. IEEE.

(Kenny et al., 2003) P. Kenny, M. Mihoubi, & P. Dumouchel, 2003. New map estimators for speaker recognition. Dans les actes de *INTERSPEECH*.

(Kersta, 1962) L. G. Kersta, 1962. Voiceprint identification. *The Journal of the Acoustical Society of America 34*(5), 725–725.

(Kheder et al., 2016a) W. B. Kheder, M. Ajili, P.-M. Bousquet, D. Matrouf, & J.-F. Bonastre, 2016a. Lia system for the sitw speaker recognition challenge. Dans les actes de *INTERSPEECH*, 848–852.

(Kheder et al., 2016b) W. B. Kheder, D. Matrouf, M. Ajili, & J.-F. Bonastre, 2016b. Iterative bayesian and mmse-based noise compensation techniques for speaker recognition in the i-vector space. Dans les actes de *Proc. Speaker Lang. Recognit. Workshop*, 60–67.

(Kheder et al., 2016c) W. B. Kheder, D. Matrouf, M. Ajili, & J.-F. Bonastre, 2016c. Local binary patterns as features for speaker recognition. Dans les actes de *Odyssey 2016, Proceedings of The Speaker and Language Recognition Workshop Odyssey 2016*, 346–351.

(Kheder et al., 2016d) W. B. Kheder, D. Matrouf, M. Ajili, & J.-F. Bonastre, 2016d. Probabilistic approach using joint clean and noisy i-vectors modeling for speaker recognition. Dans les actes de *INTERSPEECH*, 3638–3642.

(Kheder et al., 2018) W. B. Kheder, D. Matrouf, M. Ajili, & J.-F. Bonastre, 2018. A unified joint model to deal with nuisance variabilities in the i-vector space. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 26*(3), 633–645.

(Kheder et al., 2015) W. B. Kheder, D. Matrouf, J.-F. Bonastre, M. Ajili, & P.-M. Bousquet, 2015. Additive noise compensation in the i-vector space for speaker recognition. Dans les actes de *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 4190–4194. IEEE.

(Kheder et al., 2014) W. B. Kheder, D. Matrouf, P.-M. Bousquet, J.-F. Bonastre, & M. Ajili, 2014. Robust speaker recognition using map estimation of additive noise in i-vectors space. Dans les actes de *International Conference on Statistical Language and Speech Processing*, 97–107. Springer.

(Kheder et al., 2017) W. B. Kheder, D. Matrouf, P.-M. Bousquet, J.-F. Bonastre, & M. Ajili, 2017. Fast i-vector denoising using map estimation and a noise distributions database for robust speaker recognition. *Computer Speech & Language 45*, 104–122.

(Kinnunen et Li, 2010) T. Kinnunen & H. Li, 2010. An overview of text-independent speaker recognition: From features to supervectors. *Speech communication 52*(1), 12–40.

(Kinoshita et al., 2014) Y. Kinoshita, S. Ishihara, et al., 2014. Background population: how does it affect lr-based forensic voice comparison? *The International Journal of Speech, Language and the Law 21*(2), 191–224.

(Kirk, 1953) P. L. Kirk, 1953. Crime investigation; physical evidence and the police laboratory.

(Kirk, 1963) P. L. Kirk, 1963. The ontogeny of criminalistics. *The Journal of Criminal Law, Criminology, and Police Science 54*(2), 235–238.

(Kockmann et al., 2011) M. Kockmann, L. Ferrer, L. Burget, & J. Černockỳ, 2011. ivector fusion of prosodic and cepstral features for speaker verification. Dans les actes de *Twelfth Annual Conference of the International Speech Communication Association*.

(Künzel, 1994) H. J. Künzel, 1994. Current approaches to forensic speaker recognition. Dans les actes de *Automatic Speaker Recognition, Identification and Verification*.

(Künzel, 2001) H. J. Künzel, 2001. Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics 8*(1), 80–99.

(Kwan, 1977) Q. Y. Kwan, 1977. *Inference of identity of source*. Thèse de Doctorat, University of California, Berkeley.

(Labov, 1972) W. Labov, 1972. *Sociolinguistic patterns*. Numéro 4. University of Pennsylvania Press.

(Lakens, 2013) D. Lakens, 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Frontiers in psychology 4*, 863.

(Larcher et al., 2013) A. Larcher, J.-F. Bonastre, B. G. Fauve, K.-A. Lee, C. Lévy, H. Li, J. S. Mason, & J.-Y. Parfait, 2013. Alize 3.0-open source toolkit for state-of-the-art speaker recognition. Dans les actes de *INTERSPEECH*, 2768–2772.

(LaRiviere, 1975) C. LaRiviere, 1975. Contributions of fundamental frequency and formant frequencies to speaker identification. *Phonetica 31*(3-4), 185–197.

(Lavner et al., 2000) Y. Lavner, I. Gath, & J. Rosenhouse, 2000. The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Communication 30*(1), 9–26.

(Lee et al., 2017) K.-A. Lee, V. Hautamäki, T. Kinnunen, A. Larcher, C. Zhang, A. Nautsch, T. Stafylakis, G. Liu, M. Rouvier, W. Rao, et al., 2017. The i4u mega fusion and collaboration for nist speaker recognition evaluation 2016.

(Lee et al., 2016) K. A. Lee, H. Sun, S. Aleksandr, W. Guangsen, et al., 2016. The i4u submission to the 2016 nist speaker recognition evaluation. Dans les actes de *Proc. NIST SRE 2016 Workshop*.

(Levine et Hullett, 2002) T. R. Levine & C. R. Hullett, 2002. Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research 28*(4), 612–625.

(Li et al., 1966) K. Li, J. Dammann, & W. Chapman, 1966. Experimental studies in speaker verification, using an adaptive system. *The Journal of the Acoustical Society of America 40*(5), 966–978.

(Li et al., 2002) Q. Li, J. Zheng, A. Tsai, & Q. Zhou, 2002. Robust endpoint detection and energy normalization for real-time speech and speaker recognition. *IEEE Transactions on Speech and Audio Processing 10*(3), 146–157.

(Li et Jain, 2015) S. Z. Li & A. Jain, 2015. *Encyclopedia of biometrics*. Springer Publishing Company, Incorporated.

(Li et al., 2017) X. Li, L. Wang, & J. Zhu, 2017. Snr-multiconditon approaches of robust speaker model compensation based on plda in practical environment. *International Conference Artificial Intelligence, ICAI*, 146–150.

(Linares et al., 2007) G. Linares, P. Nocéra, D. Massonie, & D. Matrouf, 2007. The lia speech recognition system: from 10xrt to 1xrt. Dans les actes de *International Conference on Text, Speech and Dialogue*, 302–308. Springer.

(Lindblom et Sundberg, 1971) B. E. Lindblom & J. E. Sundberg, 1971. Acoustical consequences of lip, tongue, jaw, and larynx movement. *The Journal of the Acoustical Society of America* *50*(4B), 1166–1179.

(Lindley, 1977) D. Lindley, 1977. Probability and the law. *The Statistician*, 203–220.

(Linville et Rens, 2001) S. E. Linville & J. Rens, 2001. Vocal tract resonance analysis of aging voice using long-term average spectra. *Journal of Voice* *15*(3), 323–330.

(Loevinger, 1995) L. Loevinger, 1995. Science as evidence. *Jurimetrics*, 153–190.

(Luck, 1969) J. E. Luck, 1969. Automatic speaker verification using cepstral measurements. *The Journal of the Acoustical Society of America* *46*(4B), 1026–1032.

(Lucy, 2013) D. Lucy, 2013. *Introduction to statistics for forensic scientists*. John Wiley & Sons.

(Ma et Meng, 2004) B. Ma & H. Meng, 2004. English-chinese bilingual text-independent speaker verification. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Volume 5, V–293. IEEE.

(Ma et al., 2007) B. Ma, H. M. Meng, & M.-W. Mak, 2007. Effects of device mismatch, language mismatch and environmental mismatch on speaker verification. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Volume 4, IV–301. IEEE.

(Maaten et Hinton, 2008) L. v. d. Maaten & G. Hinton, 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* *9*(Nov), 2579–2605.

(Magrin-Chagnolleau et al., 1995) I. Magrin-Chagnolleau, J.-F. Bonastre, & F. Bimbot, 1995. Effect of utterance duration and phonetic content on speaker identification usind second order statistical methods. Dans les actes de *Proceedings of EUROSPEECH*.

(Mak et al., 2016) M.-W. Mak, X. Pang, & J.-T. Chien, 2016. Mixture of plda for noise robust i-vector speaker verification. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* *24*(1), 130–142.

(Maltoni et al., 2009) D. Maltoni, D. Maio, A. Jain, & S. Prabhakar, 2009. *Handbook of fingerprint recognition*. Springer Science & Business Media.

(Mandasari et al., 2012) M. I. Mandasari, M. McLaren, & D. A. van Leeuwen, 2012. The effect of noise on modern automatic speaker recognition systems. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4249–4252. IEEE.

(Mandasari et al., 2013) M. I. Mandasari, R. Saeidi, M. McLaren, & D. A. van Leeuwen, 2013. Quality measure functions for calibration of speaker recognition systems in various duration conditions. *Audio, Speech, and Language Processing* *21*(11), 2425–2438.

(Markova et al., 2016) D. Markova, L. Richer, M. Pangelinan, D. H. Schwartz, G. Leonard, M. Perron, G. B. Pike, S. Veillette, M. M. Chakravarty, Z. Pausova, et al., 2016. Age-and sex-related variations in vocal-tract morphology and voice acoustics during adolescence. *Hormones and behavior 81*, 84–96.

(Martin et al., 1997) A. Martin, G. Doddington, T. Kamm, M. Ordowski, & M. Przybocki, 1997. The det curve in assessment of detection task performance. Rapport technique, DTIC Document.

(Martin et Greenberg, 2009) A. F. Martin & C. S. Greenberg, 2009. Nist 2008 speaker recognition evaluation: Performance across telephone and room microphone channels. Dans les actes de *Tenth Annual Conference of the International Speech Communication Association*.

(Martin et Przybocki, 2001) A. F. Martin & M. A. Przybocki, 2001. The nist speaker recognition evaluations: 1996-2001. Dans les actes de *Proc. of SPIE Vol*, Volume 7324, 732411–1.

(Marzinzik et Kollmeier, 2002) M. Marzinzik & B. Kollmeier, 2002. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. *IEEE Transactions on Speech and Audio Processing 10*(2), 109–118.

(Marzotti et Nardini, 2006) M. Marzotti & C. Nardini, 2006. Biometric authentication using voice.

(Matějka et al., 2016) P. Matějka, O. Glembek, O. Novotnỳ, O. Plchot, F. Grézl, L. Burget, & J. H. Cernockỳ, 2016. Analysis of dnn approaches to speaker identification. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5100–5104. IEEE.

(Matrouf et al., 2015) D. Matrouf, W. B. Kheder, P.-M. Bousquet, M. Ajili, & J.-F. Bonastre, 2015. Dealing with additive noise in speaker recognition systems based on i-vector approach. Dans les actes de *Signal Processing Conference (EUSIPCO), 2015 23rd European*, 2092–2096. IEEE.

(Matrouf et al., 2007) D. Matrouf, N. Scheffer, B. G. Fauve, & J.-F. Bonastre, 2007. A straightforward and efficient implementation of the factor analysis model for speaker verification. Dans les actes de *INTERSPEECH*, 1242–1245.

(Matveev, 2013) Y. Matveev, 2013. The problem of voice template aging in speaker recognition systems. Dans les actes de *Speech and Computer*, 345–353. Springer.

(McDougall, 2007) K. McDougall, 2007. Speaker-specific formant dynamics: an experiment on australian english/ai. *International Journal of Speech Language and the Law 11*(1), 103–130.

(McGehee, 1937) F. McGehee, 1937. The reliability of the identification of the human voice. *The Journal of General Psychology 17*(2), 249–271.

(Meester et Sjerps, 2003) R. Meester & M. Sjerps, 2003. The evidential value in the dna database search controversy and the two-stain problem. *Biometrics 59*(3), 727–732.

(Meuwly, 2001) D. Meuwly, 2001. *Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique*. Université de Lausanne-Faculté de droit-Institut de police scientifique et de criminologie.

(Meuwly, 2006) D. Meuwly, 2006. Forensic individualisation from biometric data. *Science & Justice 46*(4), 205–213.

(Meuwly et Drygajlo, 2001) D. Meuwly & A. Drygajlo, 2001. Forensic speaker recognition based on a bayesian framework and gaussian mixture modelling (gmm). Dans les actes de *2001: A Speaker Odyssey-The Speaker Recognition Workshop*.

(Meuwly et al., 1998) D. Meuwly, M. El-Maliki, & A. Drygajlo, 1998. Forensic speaker recognition using gaussian mixture models and a bayesian framework. Dans les actes de *8th COST 250 workshop: Speaker Identification by Man and by Machine: Directions for Forensic Applications*, 52–55.

(Misra et Hansen, 2014) A. Misra & J. H. Hansen, 2014. Spoken language mismatch in speaker verification: An investigation with nist-sre and crss bi-ling corpora. Dans les actes de *Spoken Language Technology Workshop (SLT), 2014 IEEE*, 372–377. IEEE.

(Morrison et al., 2015) G. Morrison, C. Zhang, E. Enzinger, F. Ochoa, D. Bleach, M. Johnson, B. Folkes, S. De Souza, N. Cummins, & D. Chow, 2015. Forensic database of voice recordings of 500+ australian english speakers. *URL:) http://databases. forensic-voice-comparison. net*.

(Morrison, 2009a) G. S. Morrison, 2009a. Forensic voice comparison and the paradigm shift. *Science & Justice 49*(4), 298–308.

(Morrison, 2009b) G. S. Morrison, 2009b. Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs a. *The Journal of the Acoustical Society of America 125*(4), 2387–2397.

(Morrison, 2011a) G. S. Morrison, 2011a. A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: Multivariate kernel density (mvkd) versus gaussian mixture model–universal background model (gmm–ubm). *Speech Communication 53*(2), 242–256.

(Morrison, 2011b) G. S. Morrison, 2011b. Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice 51*(3), 91–98.

(Morrison et al., 2016) G. S. Morrison, F. H. Sahito, G. Jardine, D. Djokic, S. Clavet, S. Berghs, & C. G. Dorny, 2016. Interpol survey of the use of speaker identification by law enforcement agencies. *Forensic science international 263*, 92–100.

(Morrison et al., 2011) G. S. Morrison, C. Zhang, & P. Rose, 2011. An empirical estimate of the precision of likelihood ratios from a forensic-voice-comparison system. *Forensic science international 208*(1), 59–65.

(Moye, 1979) L. Moye, 1979. Study of the effects on speech analysis of the types of degradation occurring in telephony. *Harlow: Standard Telecommunication Laboratories Monograph 1*.

(Nair et al., 2016) B. B. Nair, E. A. Alzqhoul, & B. J. Guillemin, 2016. Impact of the gsm and cdma mobile phone networks on the strength of speech evidence in forensic voice comparison. *j ournal of Forensic Research*.

(National Research Council, 2009) W. D. National Research Council, Committee on Identifying the Needs of the Forensic Science Community. National Academies Press, 2009. Strengthening forensic science in the united states: A path forward.

(Nautsch et al., 2016) A. Nautsch, R. Saeidi, C. Rathgeb, & C. Busch, 2016. Robustness of quality-based score calibration of speaker recognition systems with respect to low-snr and short-duration conditions. Odyssey.

(Nemer et al., 2001) E. Nemer, R. Goubran, & S. Mahmoud, 2001. Robust voice activity detection using higher-order statistics in the lpc residual domain. *IEEE Transactions on Speech and Audio Processing 9*(3), 217–231.

(NIST, 2010) S. NIST, 2010. Evaluation plan, 2010.

(Nolan, 1980) F. Nolan, 1980. *The phonetic bases of speaker recognition*. Thèse de Doctorat, University of Cambridge.

(Nolan, 1991) F. Nolan, 1991. Forensic phonetics. *Journal of Linguistics 27*(02), 483–493.

(Nolan, 2001) F. Nolan, 2001. Speaker identification evidence: its forms, limitations, and roles. Dans les actes de *Proceedings of the Conference "Law and Language: Prospect and Retrospect", Levi (Finnish Lapland)*, 1–19.

(Nolan et Grigoras, 2005) F. Nolan & C. Grigoras, 2005. A case for formant analysis in forensic speaker identification. *International Journal of Speech Language and the Law 12*(2), 143.

(Ortega-García et González-Rodríguez, 1996) J. Ortega-García & J. González-Rodríguez, 1996. Overview of speech enhancement techniques for automatic speaker recognition. Dans les actes de *Fourth International Conference on Spoken Language, 1996. ICSLP 96. Proceedings*, Volume 2, 929–932. IEEE.

(Park et al., 2016) S. J. Park, C. Sigouin, J. Kreiman, P. A. Keating, J. Guo, G. Yeung, F.-Y. Kuo, & A. Alwan, 2016. Speaker identity and voice quality: Modeling human responses and automatic speaker recognition. Dans les actes de *INTERSPEECH*, 1044–1048.

(Pelecanos et Sridharan, 2001) J. Pelecanos & S. Sridharan, 2001. Feature warping for robust speaker verification.

(Pereira et Watson, 1998) C. Pereira & C. I. Watson, 1998. Some acoustic characteristics of emotion. Dans les actes de *ICSLP*.

(Pigeon et al., 2000) S. Pigeon, P. Druyts, & P. Verlinde, 2000. Applying logistic regression to the fusion of the nist'99 1-speaker submissions. *Digital Signal Processing 10*(1-3), 237–248.

(Prince et Elder, 2007) S. J. Prince & J. H. Elder, 2007. Probabilistic linear discriminant analysis for inferences about identity. Dans les actes de *IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007*, 1–8. IEEE.

(Pruzansky, 1963) S. Pruzansky, 1963. Pattern-matching procedure for automatic talker recognition. *The Journal of the Acoustical Society of America 35*(3), 354–358.

(Ramos, 2007) D. Ramos, 2007. Forensic evaluation of the evidence using automatic speaker recognition systems.

(Ramos et al., 2008) D. Ramos, J. Gonzalez-Rodriguez, J. Gonzalez-Dominguez, & J. J. Lucena-Molina, 2008. Addressing database mismatch in forensic speaker recognition with ahumada iii: a public real-casework database in spanish. Dans les actes de *Interspeech*. International Speech Communication Association.

(Ramos et al., 2017) D. Ramos, R. P. Krish, J. Fierrez, & D. Meuwly, 2017. From biometric scores to forensic likelihood ratios. Dans les actes de *Handbook of Biometrics for Forensic Science*, 305–327. Springer.

(Reetz et Jongman, 2011) H. Reetz & A. Jongman, 2011. *Phonetics: Transcription, production, acoustics, and perception*, Volume 34. John Wiley & Sons.

(Reubold et al., 2010) U. Reubold, J. Harrington, & F. Kleber, 2010. Vocal aging effects on f0 and the first formant: a longitudinal analysis in adult speakers. *Speech Communication 52*(7), 638–651.

(Reynolds et al., 2003a) D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, et al., 2003a. The supersid project: Exploiting high-level information for high-accuracy speaker recognition. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Volume 4, IV–784. IEEE.

(Reynolds et al., 2003b) D. Reynolds, J. Campbell, B. Campbell, B. Dunn, T. Gleason, D. Jones, T. Quatieri, C. Quillen, D. Sturim, & P. Torres-Carrasquillo, 2003b. Beyond cepstra: exploiting high-level information in speaker recognition. Dans les actes de *Proceedings of the Workshop on Multimodal User Authentication*, 223–229.

(Reynolds, 1995) D. A. Reynolds, 1995. Speaker identification and verification using gaussian mixture speaker models. *Speech communication 17*(1), 91–108.

(Reynolds, 2003) D. A. Reynolds, 2003. Channel robust speaker verification via feature mapping. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Volume 2, II–53. IEEE.

(Reynolds et al., 2000) D. A. Reynolds, T. F. Quatieri, & R. B. Dunn, 2000. Speaker verification using adapted gaussian mixture models. *Digital signal processing 10*(1-3), 19–41.

(Rhodes, 1956) H. T. F. Rhodes, 1956. *Alphonse Bertillon, father of scientific detection*. Abelard-Schuman.

(Ribas et al., 2015) D. Ribas, E. Vincent, & J. R. Calvo, 2015. Full multicondition training for robust i-vector based speaker recognition. Dans les actes de *Interspeech 2015*.

(Robertson et al., 2016) B. Robertson, G. A. Vignaux, & C. E. Berger, 2016. *Interpreting evidence: evaluating forensic science in the courtroom*. John Wiley & Sons.

(Rose, 2003) P. Rose, 2003. *Forensic speaker identification*. CRC Press.

(Rose, 2006) P. Rose, 2006. Technical forensic speaker recognition: Evaluation, types and testing of evidence. *Computer Speech & Language 20*(2), 159–191.

(Rose et al., 2003) P. Rose et al., 2003. The technical comparison of forensic voice samples. Dans les actes de *Expert Evidence*. The Law Book Company.

(Rose et al., 2009) P. Rose, G. Morrison, et al., 2009. A response to the uk position statement on forensic speaker comparison. *The international journal of speech, language and the law 16*(1), 139.

(Rosenberg, 1976) A. E. Rosenberg, 1976. Automatic speaker verification: A review. *Proceedings of the IEEE 64*(4), 475–487.

(Rossato et al., 1998) S. Rossato, G. Feng, & R. Laboissière, 1998. Recovering gestures from speech signals: a preliminary study for nasal vowels. Dans les actes de *ICSLP*.

(Rossato et al., 2018) S. Rossato, D. Zhang, M. Ajili, & J.-F. Bonastre, 2018. Suivre le rythme de tes paroles. Dans les actes de *XXXIIe Journées d'Études sur la Parole*.

(Rother et al., 2002) P. Rother, B. Wohlgemuth, W. Wolff, & I. Rebentrost, 2002. Morphometrically observable aging changes in the human tongue. *Annals of Anatomy-Anatomischer Anzeiger 184*(2), 159–164.

(Rouvier et al., 2016) M. Rouvier, P.-M. Bousquet, M. Ajili, W. B. Kheder, D. Matrouf, & J.-F. Bonastre, 2016. Lia system description for nist sre 2016. *arXiv preprint arXiv:1612.05168*.

(Rouvier et al., 2015) M. Rouvier, P.-M. Bousquet, & B. Favre, 2015. Speaker diarization through speaker embeddings. Dans les actes de *Signal Processing Conference (EUSIPCO), 2015 23rd European*, 2082–2086. IEEE.

(Rozi et al., 2016) A. Rozi, L. Li, D. Wang, & T. F. Zheng, 2016. Feature transformation for speaker verification under speaking rate mismatch condition. Dans les actes de *Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 1–4. IEEE.

(Sadjadi et Hansen, 2010) S. O. Sadjadi & J. H. Hansen, 2010. Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions. Dans les actes de *Eleventh Annual Conference of the International Speech Communication Association*.

(Saedi et al., 2013) R. Saedi, K. A. Lee, T. Kinnunen, T. Hasan, B. Fauve, P.-M. Bousquet, E. Khoury, P. L. S. Martinez, J. M. K. Kua, C. You, et al., 2013. I4u submission to nist sre 2012: A large-scale collaborative effort for noise-robust speaker verification. Dans les actes de *Interspeech*, Numéro EPFL-CONF-192763.

(Saks et Koehler, 2005) M. J. Saks & J. J. Koehler, 2005. The coming paradigm shift in forensic identification science. *Science 309*(5736), 892–895.

(Saks et Koehler, 2008) M. J. Saks & J. J. Koehler, 2008. The individualization fallacy in forensic science evidence.

(Sallavaci, 2014) O. Sallavaci, 2014. *The Impact of Scientific Evidence on the Criminal Trial: The Case of DNA Evidence*. Routledge.

(Sambur, 1975) M. Sambur, 1975. Selection of acoustic features for speaker identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing 23*(2), 176–182.

(Sarkar et al., 2012) A. K. Sarkar, D. Matrouf, P.-M. Bousquet, & J.-F. Bonastre, 2012. Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification. Dans les actes de *Interspeech*, 2662–2665.

(Scheffer et Lei, 2014) N. Scheffer & Y. Lei, 2014. Content matching for short duration speaker recognition. Dans les actes de *INTERSPEECH*, 1317–1321.

(Scherer, 1986) K. R. Scherer, 1986. Vocal affect expression: a review and a model for future research. *Psychological bulletin 99*(2), 143.

(Scherer et al., 2000) K. R. Scherer, T. Johnstone, G. Klasmeyer, & T. Bänziger, 2000. Can automatic speaker verification be improved by training the algorithms on emotional speech? Dans les actes de *INTERSPEECH*, 807–810.

(Schindler et Draxler, 2013) C. Schindler & C. Draxler, 2013. The influence of bandwidth limitation on the speaker discriminating potential of nasals and fricatives. *International Association for Forensic Phonetics and Acoustics (IAFPA)*.

(Shriberg, 2007) E. Shriberg, 2007. Higher-level features in speaker recognition. Dans les actes de *Speaker Classification I*, 241–259. Springer.

(Shriberg et al., 2005) E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, & A. Stolcke, 2005. Modeling prosodic feature sequences for speaker recognition. *Speech Communication 46*(3), 455–472.

(Shriberg et al., 2008) E. Shriberg, M. Graciarena, H. Bratt, A. Kathol, S. S. Kajarekar, H. Jameel, C. Richey, & F. Goodman, 2008. Effects of vocal effort and speaking style on text-independent speaker verification. Dans les actes de *INTERSPEECH*, 609–612.

(Siddiq et al., 2012) S. Siddiq, T. Kinnunen, M. Vainio, & S. Werner, 2012. Intonational speaker verification: a study on parameters and performance under noisy conditions. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4777–4780. IEEE.

(Solomonoff et al., 2005) A. Solomonoff, W. M. Campbell, & I. Boardman, 2005. Advances in channel compensation for svm speaker recognition. Dans les actes de *International Conference Acoustics, Speech, and Signal Processing (ICASSP).*, Volume 1, I–629. IEEE.

(Soong et al., 1985) F. Soong, A. Rosenberg, L. Rabiner, & B. Juang, 1985. A vector quantization approach to speaker recognition. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Volume 10, 387–390.

(Stevens, 1999) K. Stevens, 1999. Acoustic phonetics. 1998.

(Stevens, 1971) K. N. Stevens, 1971. Sources of inter-and intra-speaker variability in the acoustic properties of speech sounds. *Proceedings of the Seventh International Cons. Phonetic Sciences*, 206–232.

(Stoney, 1991) D. A. Stoney, 1991. What made us ever think we could individualize using statistics? *Journal of the Forensic Science Society 31*(2), 197–199.

(Swart et Brümmer, 2017) A. Swart & N. Brümmer, 2017. A generative model for score normalization in speaker recognition. *Proc. Interspeech 2017*, 1477–1481.

(Tanyer et Ozer, 2000) S. G. Tanyer & H. Ozer, 2000. Voice activity detection in nonstationary noise. *IEEE Transactions on speech and audio processing 8*(4), 478–482.

(Taroni et al., 1998) F. Taroni, C. Champod, & P. Margot, 1998. Forerunners of bayesianism in early forensic science. *Jurimetrics 38*(2), 183–200.

(Thompson et al., 2013) W. C. Thompson, S. O. Kaasa, & T. Peterson, 2013. Do jurors give appropriate weight to forensic identification evidence? *Journal of Empirical Legal Studies 10*(2), 359–397.

(Tistarelli et al., 2014) M. Tistarelli, E. Grosso, & D. Meuwly, 2014. Biometrics in forensic science: challenges, lessons and new technologies. Dans les actes de *International Workshop on Biometric Authentication*, 153–164. Springer.

(Tosi et Tosi, 1979) O. Tosi & O. Tosi, 1979. *Voice identification: theory and legal applications*. University Park Press Baltimore.

(Trauring, 1963) M. Trauring, 1963. Automatic comparison of finger-ridge patterns. *Nature 197*, 938–940.

(Triggs et Buckleton, 2004) C. M. Triggs & J. S. Buckleton, 2004. Comment on: Why the effect of prior odds should accompany the likelihood ratio when reporting dna evidence. *Law, Probability and Risk 3*(1), 73–82.

(Tucker, 1992) R. Tucker, 1992. Voice activity detection using a periodicity measure. *IEE Proceedings I (Communications, Speech and Vision) 139*(4), 377–380.

(Tversky et Kahneman, 1975) A. Tversky & D. Kahneman, 1975. Judgment under uncertainty: Heuristics and biases. Dans les actes de *Utility, probability, and human decision making*, 141–162. Springer.

(Ulatowska, 1985) H. K. Ulatowska, 1985. *The aging brain: Communication in the elderly*. College-Hill.

(Vacher et al., 2015) M. Vacher, B. Lecouteux, J. S. Romero, M. Ajili, F. Portet, & S. Rossato, 2015. Speech and speaker recognition for home automation: Preliminary results. Dans les actes de *Speech Technology and Human-Computer Dialogue (SpeD), 2015 International Conference on*, 1–10. IEEE.

(Vaissière, 2007) J. Vaissière, 2007. Area functions and articulatory modeling as a tool for investigating the articulatory, acoustic and perceptual properties of sounds across languages. *Experimental approaches to phonology*, 54–71.

(van der Vloed et al., 2014) D. van der Vloed, J. Bouten, & D. A. van Leeuwen, 2014. Nfi-frits: A forensic speaker recognition database and some first experiments.

(Variani et al., 2014) E. Variani, X. Lei, E. McDermott, I. L. Moreno, & J. Gonzalez-Dominguez, 2014. Deep neural networks for small footprint text-dependent speaker verification. Dans les actes de *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4052–4056. IEEE.

(Wang et Xu, 2011) L. Wang & M. Xu, 2011. Sdbm-based speaker recognition for speaking style variations.

(Ward Jr, 1963) J. H. Ward Jr, 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association 58*(301), 236–244.

(Wertheim, 2001) K. Wertheim, 2001. The weekly detail, dec. 17, 2001 http://www.clpex.com/legacy/thedetail/1-99/thedetail17.htm.

(Wolf, 1972) J. J. Wolf, 1972. Efficient acoustic parameters for speaker recognition. *The Journal of the Acoustical Society of America 51*(6B), 2044–2056.

(Wright, 1986) J. T. Wright, 1986. The behavior of nasalized vowels in the perceptual vowel space. *Experimental phonology*, 45–67.

(Wu et al., 2005) T. Wu, Y. Yang, & Z. Wu, 2005. Improving speaker recognition by training on emotion-added models. Dans les actes de *International Conference on Affective Computing and Intelligent Interaction*, 382–389. Springer.

(Wu et al., 2006) W. Wu, T. F. Zheng, M.-X. Xu, & H. Bao, 2006. Study on speaker verification on emotional speech. Dans les actes de *INTERSPEECH*.

(Xue et Hao, 2003) S. A. Xue & G. J. Hao, 2003. Changes in the human vocal tract due to aging and the acoustic correlates of speech productiona pilot study. *Journal of Speech, Language, and Hearing Research 46*(3), 689–701.

(Zadrozny et Elkan, 2002) B. Zadrozny & C. Elkan, 2002. Transforming classifier scores into accurate multiclass probability estimates. Dans les actes de *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 694–699. ACM.

(Zetterholm, 1998) E. Zetterholm, 1998. Prosody and voice quality in the expression of emotions. Dans les actes de *ICSLP*.

(Zhang et Hansen, 2007) C. Zhang & J. H. Hansen, 2007. Analysis and classification of speech mode: whispered through shouted. Dans les actes de *INTERSPEECH*, Volume 7, 2289–2292.

(Zhang et al., 2013) C. Zhang, G. S. Morrison, E. Enzinger, & F. Ochoa, 2013. Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison–female voices. *Speech Communication* *55*(6), 796–813.

(Zhang et al., 2011) C. Zhang, G. S. Morrison, & T. Thiruvaran, 2011. Forensic voice comparison using chinese/iau. Dans les actes de *Proceedings of the 17th International Congress of Phonetic Sciences*, Volume 17, 21.

(Zunkel, 1996) R. L. Zunkel, 1996. Hand geometry based verification. Dans les actes de *Biometrics*, 87–101. Springer.