Université du Québec
Institut national de la recherche scientifique
Centre Énergie Matériaux Télécommunications

TOWARDS MULTIPLE VOCAL EFFORT SPEAKER VERIFICATION: EXPLORING SPEAKER-DEPENDENT INVARIANT INFORMATION BETWEEN NORMAL AND WHISPERED SPEECH

By

Milton Orlando Sarria Paja

A thesis submitted in fulfillment of the requirements for the degree of
Doctorate of Sciences, Ph.D
in Telecommunications

Evaluation Committee

Internal evaluator and committee president:	Prof. Long Le
External evaluator 1:	Prof. Patrick Cardinal École de Technologie Supérieure
External evaluator 2:	Prof. John Hansen University of Texas at Dallas
Research advisor:	Prof. Tiago H. Falk

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Prof. Tiago H. Falk for the continuous support and encouragement during my Ph.D, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I would also like to thank my committee members, professor Long Le, professor Patrick Cardinal and professor John Hansen for taking time off their busy schedules to serve on my Thesis Examination Committee and also for their valuable comments.

I thank Dr. Mohammed Senoussaoui for his insightful comments and ideas that helped me in one of the most difficult stages of my Ph.D studies. I thank my friends and lab partners, especially João F. Santos, for the encouragement, comments, ideas, discussions and loyal friendship offered during all these years.

My sincere thanks also goes to Dr.-Ing. Martin Heckmann, Dr. Andreas Richter and all the members of the *Sensory Processing & Learning* group from Honda Research Institute Europe, who provided me an opportunity to join their team as intern. That opportunity allowed me to learn many valuable lessons, as well as to be a better professional and researcher.

I would not have been here today if it were not for the love and care of my family. Words cannot express how grateful I am to my parents for all of the sacrifices that they have made on my behalf. They taught me to be a sincere learner and a good human being, to follow my dreams and to be persistent no matter the difficulties.

Abstract

Speech-based biometrics is one of the most effective ways for identity management and one of the preferred methods by users and companies given its flexibility, speed and reduced cost. In speaker recognition, the two most popular tasks are speaker identification (SI) and speaker verification (SV). Commonly, SV exhibits greater practical applications related to SI, especially in access control and identity management applications. Current state-of-the-art SV systems are known to be strongly dependent on the condition of the speech material provided as input and can be affected by unexpected variability presented during testing, such as environmental noise or changes in vocal effort. In this thesis, SV using whispered speech is explored, as whispered speech is known to be a natural speaking style that despite its reduced perceptibility, still contains relevant information regarding the intended message (i.e., intelligibility), as well as the speaker identity and gender. However, given the acoustic differences between whispered and normally-phonated speech, speech applications trained on the latter but tested with the former exhibit unacceptable performance levels. Within an automated speaker verification task, previous research has shown that *i*) conventional features (e.g., mel-frequency cepstral coefficients, MFCCs) do not convey sufficient speaker discrimination cues across the two vocal efforts, and *ii*) multi-style training, while improving the performance for whispered speech, tends to deteriorate the performance for normal speech. Moreover, by exploring the performance boundaries achievable with whispered speech for *speaker verification*, it was found that the lack of sufficient data to train whispered speech speaker models is a major limiting factor. As such, this thesis aims to address these three shortcomings and proposes *i*) innovative features that are less sensitive to changes from normal to whispered speech, thus helping to reduce the mismatch gap, *ii*) fusion strategies that improve accuracy for *both* normal and whispered speech, and *iii*) machine learning principles that overcome the limited resources/data problem.

To properly address the task at hand, we first perform a comparative analysis among some of the most common feature based approaches and training techniques reported in related fields that have shown benefits in the mismatch problem induced by vocal effort variations, including whispered speech. This allowed us to narrow down candidate solutions and strategies that could be useful for whispered speech speaker verification after some careful fine tuning. Next, by combining these insights, with those previously reported in the literature, as well as statistical analysis tools, innovative features are proposed that provide not only invariant information across the two vocal efforts, but also features that provide complementary information. Subsequently, in order to take advantage of the complementary information extracted from the different feature representations, we explore fusion schemes at different levels. When using a combination of normal and whispered speech data during parameter estimation, improved multi vocal effort speaker verification is achieved and relative improvements as high as 68% and 70% for normal and whispered speech, respectively, could be seen relative to a baseline system based on MFCC features. When including whispered speech during enrollment, further improvements are achieved for whispered speech without severely

affecting performance for normal speech. These results show that the proposed strategies allow to efficiently use the limited resources available, and achieve high performance levels for whispered speech inline with performance obtained for normal speech.

Index terms— Whispered speech, speaker verification, modulation spectrum, mutual information, system fusion, mismatch problem, neural networks.

Contents

List of Tables

List of Figures

List of acronyms

AAMF	Auditory-inspired Amplitude Modulation Features
AAMF(FS)	AAMF with Feature Selection
ASR	Automatic Speech Recognition
ASV	Automatic Speaker Recognition
BNF	Bottleneck Features
DCT	Discrete Cosine Transform
DET	Detection Error Tradeoff
DFT	Discrete Fourier Transform
DNN	Deep Neural Network
DTW	Dynamic Time Warping
EER	Equal Error Rate
EM	Expectation Maximization
FBBNF	Filterbank Bottleneck Features
GMM	Gaussian Mixture Model
KBA	Knowledge-Based Authentication
LDA	Linear Discriminant Analysis
LMFCC	Limited band MFCC
LRBNFi	Bottleneck Features in the i-th layer using limited and residual filterbank outputs
MANOVA	Multivariate Analysis Of Variance
MAP	Maximum a Posteriori
MHEC	Mean Hilbert Envelope Coefficients
MI	Mutual Information
PCA	Principal Component Analysis

PLDA	Probabilistic Linear Discriminant Analysis
RMFCC	Residual MFCC
SI	Speaker Identification
STDFT	Short-Time Discrete Fourier Transform
STG	Short-Time Gaussianization
SV	Speaker Verification
SVM	Support Vector Machine
TTS	Text-To-Speech synthesis
UBM	Universal Background Model
WIF	Weighted Instantaneous Frequencies
WSS	Whisper Sensitive Scale

Synopsis

0.1 Introduction

La parole humaine est un mode de communication naturel et flexible transmettant non seulement un message, mais aussi des traits tels que l'identité, l'âge, le sexe, l'origine, ainsi que les états émotionnels et ceux de la santé. Les systèmes de traitement de la parole sont devenus, sous des conditions contrôlées, utiles pour certains domaines. Par exemple, la reconnaissance automatique de la parole a ouvert la porte à l'utilisation de la parole comme une interface homme-machine fiable, permettant aux humains de contrôler les téléviseurs, les smartphones et les systèmes stéréo de voiture, sans oublier l'interaction avec des services automatisés de support à la clientèle. Les progrès des technologies de reconnaissance des locuteurs ont permis d'utiliser la voix humaine pour des fins d'authentification, à titre d'exemple, dans un système téléphonique automatisé des services bancaires. Ce domaine présente un grand potentiel. En effet, des rapports récents prévoient que le marché mondial de la technologie de la parole dépassera 31 milliards de dollars en 2017, principalement en raison de trois applications de la parole: la reconnaissance automatique de la parole (automatic speech recognition, ASR), la vérification automatique du locuteur (automatic speaker verification, ASV) et la synthèse vocale (text-to-speech synthesis, TTS) [1]. Une bonne partie de ce marché a été motivée par la prolifération des smartphones et des tablettes à travers le monde. À titre d'exemple, plusieurs applications qui ont vu le jour permettent aux utilisateurs d'utiliser leur voix pour interagir avec leurs appareils (par exemple l'application Siri d'Apple), de se connecter à des services sécurisés (par exemple, le service d'identification vocale de Bell Canada) ou même de déverrouiller leurs appareils mobiles, (exemple du service de vérification des locuteurs du centre de recherche Baidu-I^{2}R). Toutefois, ces systèmes commerciaux existants ont été développés pour des scénarios "typiques": des voix adultes claires avec des accents limités de langue étrangère et

peu de bruits ambiant. Ces hypothèses et conditions sont toutefois difficiles à satisfaire dans des environnements du monde réel et dans des applications à utilisation quotidienne.

Dans cette dissertation, le problème de reconnaissance du locuteur est d'un intérêt particulier. Le potentiel des applications de reconnaissance automatique du locuteur réside particulièrement dans les applications de contrôle d'accès et de gestion d'identité où l'authentification basée sur la connaissance (knowledge-based authentication, KBA) demeure le mode dominant d'authentification des utilisateurs [2]. KBA requiert des informations personnelles des utilisateurs afin d'accorder l'accès à un service; les détails représentatifs peuvent inclure une combinaison de mots de passe, de noms d'utilisateurs et de questions personnelles. Cette méthode présente cependant de nombreux inconvénients qui peuvent mettre à risque des informations critiques, ainsi que des problèmes de confidentialité lors de la collecte d'informations. À titre d'exemple, avec l'utilisation généralisée des réseaux sociaux, de nombreux utilisateurs ne sont pas conscients que l'information qu'ils partagent peut être utilisée à des fins de vérification pour autoriser des transactions (par exemple la date de naissance). En outre, les méthodes d'authentification existantes sont sujettes à de nombreuses attaques documentées. Des méthodes plus sûres et plus fiables ont fait l'objet de recherches récentes [3, 4, 5].

La méthode d'authentification basée sur la biométrie est une alternative à KBA. Elle combine les mathématiques et les techniques de traitement numérique du signal pour analyser les caractéristiques physiologiques. Cette méthode obtient une attention significative. Contrairement à KBA, la biométrie est basée sur l'identité de l'utilisateur, plutôt que sur ce que l'utilisateur connait. Ces technologies sont en pleine expansion dans le domaine de la gestion des identités. En effet, elles n'exigent pas de numéros personnels d'identification ni de mots de passe ou de questions de sécurité [6]. Malgré les nombreux avantages de l'utilisation de la biométrie à des fins d'identification, des défis et des problèmes non résolus persistent, entravant ainsi l'utilisation généralisée. Par exemple, la reconnaissance biométrique est compromise par des facteurs externes qui peuvent altérer les modèles qui sont analysés (par exemple, des coupures et les brûlures aux doigts dans les systèmes à empreinte digitale; le bruit ambiant dans les solutions basées sur la voix), ainsi que par des facteurs physiologiques humains naturels tels que le vieillissement et la maladie (par exemple, dans les systèmes basés sur les caractères faciaux et sur la voix) [6, 7].

Les préférences des utilisateurs jouent également un rôle important lors de la décision de la méthode à utiliser pour l'authentification. Selon des statistiques récentes, la biométrie basée sur la parole se classe très bien dans la préférence des clients, dépassant les méthodes à empreintes digitales et les solutions de numérisation de l'iris (iris scanning solutions) [8, 9]. En raison de l'utilisation répandue des smartphones dans le monde entier, la biométrie basée sur la parole gagne rapidement du terrain en terme de popularité, particulièrement dans les institutions financières [8]. Dans de telles applications, les clients peuvent accéder à leurs services sécurisés bancaires et d'assurance en parlant simplement dans leurs téléphones. Pour les institutions financières, cette facilité d'utilisation améliore la satisfaction du client tout en réduisant les coûts de prise en charge des clients et ceci grâce aux taux d'automatisation élevés. D'autre part, étant donné la flexibilité de la communication basée sur la parole et étant dans une situation particulière, les clients peuvent poser de grands défis à ces types d'applications en changeant par exemple leur effort vocal de chuchotement à des cris selon l'environnement. Ceci, ainsi que le bruit ambiant, ont posé de gros problèmes à la performance des applications vocales. Le bruit ambiant a des effets néfastes sur les systèmes biométriques basés sur la parole, en particulier ceux formés avec des caractéristiques classiques, tels que les coefficients MFCC (mel-frequency cepstral coefficients). À titre d'exemple, des précisions aussi faibles que 7% ont été rapportées dans des environnements très bruyants [10]. Au fil des années, plusieurs algorithmes d'amélioration de la parole ont été proposés pour des applications de reconnaissance de la parole dans des environnements robustes [11]. D'autre part, le fait de varier les efforts vocaux peut également entraîner des effets néfastes sur la performance du processus de vérification du locuteur. Par exemple, des précisions aussi faibles que 20% ont été rapportées pour l'identification de locuteur à voix basse [12] dans des *conditions propres*, alors que des exactitudes aussi basses que 8.71% ont été rapportées pour le cas d'une voix crieé [13]. En effet, il est très probable que des clients utilisant une application mobile de services bancaires sur leurs smartphones vont murmurer des informations sensibles ou bien parler plus fort, voire crier, si'ils perçoivent que leurs paroles ne sont pas entendues.

Dans la reconnaissance automatique des locuteurs, il existe deux tâches classiques qui peuvent être effectuées: l'identification du locuteur (speaker identification, SI) et la vérification du locuteur (speaker verification, SV) [14]. Généralement, SV a plus d'applications pratiques par rapport à SI, particulièrement dans des applications de contrôle d'accès et de gestion d'identité. Les systèmes basés sur le modèle GMM (Gaussian mixture model), avec la reconnaissance automatique des

locuteurs, utilisant l'adaptation du maximum a posteriori (MAP) et MFCC comme vecteurs caractéristiques, ont été pendant de nombreuses années l'approche dominante pour la reconnaissance du locuteur indépendante du texte[15]. Les avancées récentes ont introduit le concept de *supervector*, qui, combiné avec les machines à vecteurs de support (SVM), ont conduit à des résultats améliorés de reconnaissance [16]. Récemment, les approches inspirées des variables latentes (latent variable) pour l'extraction de caractéristiques telles que l'analyse factorielle conjointe (joint factor analysis, JFA), seules ou combinées avec la SVM, se sont révélées être une approche efficace pour la compensation des effets des canaux et la réduction du temps de notation [17, 18]. Selon le Framework JFA, les systèmes actuels SR sont basés sur l'extraction de vecteurs d'identité (*i-vectors*) avec une notation basée sur la distance cosinus ou sur l'analyse probabiliste discriminante linéaire (probabilistic linear discriminant analysis, PLDA) [19, 20].

En outre, étant donné les progrès récents des deux domaines d'apprentissage profond et des réseaux de neurones, de nouvelles approches basées sur les fonctionnalités de goulot d'étranglement (bottleneck features, BNF) [21] émergent. Les premières expériences montrent des performances améliorées par rapport aux caractéristiques conventionnelles tels que les MFCC [22, 23]. Toujours dans le domaine de l'apprentissage profond, des études récentes ont également commencé à explorer l'utilisation de réseaux de neurones afin de remplacer les GMMs dans le calcul des statistiques nécessaires lors de l'extraction i-vectorielle [22].

0.2 Description du problème

Nous tâcherons à attirer l'attention, le long de cette dissertation, sur l'un des nouveaux défis que les développeurs d'applications vocales automatisées rencontrent: *variation des efforts vocaux*. Dans [24], Traunmuller et Eriksson caractérisent l'effort vocal par: "*la quantité que les locuteurs ordinaires varient quand ils adaptent leur discours aux exigences d'une distance de communication accrue ou réduite.*" Même s'il s'agit d'une mesure subjective, nous identifions cinq niveaux d'effort vocal: 1) chuchotement, 2) voix douce, 3) voix normale, 4) voix forte et 5) des cris. Les deux extrêmes, c'est-à-dire la parole chuchotée et les cris, sont les deux efforts vocaux qui produisent des changements majeurs dans les caractéristiques acoustiques ainsi que dans les caractéristiques dynamiques générales du signal de la parole comparées à la voix normale [24]. Ces changements affectent de manière significative les performances des systèmes de la reconnaissance automatique de la parole

ainsi que des locuteurs, surtout si seulement la parole normale a été utilisée lors de l'entraînement (c'est-à-dire, les conditions d'inadéquation de entraînement/test) [12, 13, 25, 26, 27]. Cela pose un grand défi pour la plupart des tâches de reconnaissance à être développé à l'aide de la biométrie vocale; principalement les tâches qui tiennent compte uniquement des différences entre les locuteurs telles que les tâches de reconnaissance des locuteurs.

Parmi les cinq efforts vocaux, le discours chuchoté attire une grande attention pour les applications de sécurité ces derniers temps. Malgré sa perception et son intelligibilité réduites, la parole chuchotée est un mode naturel de production de la parole transmettant des informations pertinentes et utiles pour de nombreuses applications. Par exemple, tout comme la parole avec une voix normale, la parole chuchotée transmet non seulement un message, mais aussi des traits tels que l'identité, le sexe, l'émotion, et létat de santé, pour n'en nommer que quelques-uns [25, 28, 29, 30, 31, 32]. Comme mentionné auparavant, le discours chuchoté est couramment utilisé dans des situations publiques où des informations privées ou discrètes doivent être échangées, à titre d'exemple, lorsqu'on fournit un numéro de carte de crédit, un numéro de compte bancaire ou d'autres renseignements personnels. Malgré la quantité d'informations présentes dans la voix chuchotée, certaines caractéristiques rendent ce style de la parole difficile à traiter par les applications vocales. Citons l'exemple de la caractéristique la plus importante pour un discours chuchoté qui est l'absence de vibration du pli vocal. De plus, quand une personne chuchote, plusieurs changements se produisent dans la configuration de la voie vocale, altérant ainsi non seulement la source d'excitation, mais aussi le taux syllabique et les caractéristiques générales de dynamique temporelle du signal de la parole généré [25, 33]. En conséquence, on s'attend à ce que les méthodes classiques conçues pour la caractérisation vocale à voix normale échouent lorsqu'elles sont testées dans des scénarios atypiques, telle que la parole chuchotée [12, 25, 26, 27].

Malgré le nombre limité de recherche dans ce domaine, plusieurs approches visant à surmonter certains de ces inconvénients ont été signalées, particulièrement dans des conditions d'inadéquation entre l'entraînement et le test où les modèles de locuteurs étaient formés avec la parole normale et testés avec des discours murmurés [12, 26, 34, 35, 36, 37]. Dans ces travaux, on a signalé une faible précision dans des conditions incompatibles d'entraînement et de test. Aussi, certaines stratégies visant à remédier aux limitations intrinsèques des systèmes actuels de la parole pour traiter la parole chuchotée ont été proposées. Parmi les exemples représentatifs, nous pouvons citer: les caractéristiques robustes telles que les coefficients cepstraux linéaires modifiés (linear-

frequency cepstral coefficients, LFCC) et le mappage des caractéristiques [34], la déformation des caractéristiques sur MFCC (feature warping over MFCC) et la combinaison de scores au niveau des trames [35], les schémas d'adaptation de modèles tels que la régression linéaire maximale (MLLR) et la transformation des caractéristiques -de la parole normale à la parole chuchotée- à utiliser lors de l'adaptation du modèle (MAP-adaptation) [37]. Malgré les nombreux efforts déployés, les améliorations relatives vont de 8% à 46% et, pour tous les cas, les résultats de classification obtenus ne sont toujours pas utiles pour des applications réelles [26, 37].

En général, pour les applications vocales, telles que la reconnaissance de la parole, deux stratégies principales sont utilisées pour gérer le problème de désadaptation, à savoir, (*i*) reconnaisseur de modèle multiple, où des modèles vocaux dédiés sont obtenus pour différents efforts vocaux [27] et (*ii*) modèles à styles multiples (multi-style), où chaque modèle est obtenu à partir d'une combinaison de la parole normale et de petites quantités de parole de divers efforts vocaux [25, 27]. Néanmoins, les deux différentes méthodes ont montré avoir des avantages et des inconvénients. Par exemple, bien que les deux améliorent la performance de la parole chuchotée [26, 27], l'entraînement de plusieurs modèles nécessite des quantités significatives de données vocales murmurées afin d'obtenir des modèles de locuteurs, ce qui peut être difficile à obtenir dans la pratique. Pour les systèmes basés sur le multi-style, malgré qu'ils exigent moins de données provenant de discours chuchoté pour former les modèles, ils échangent des gains dans la parole chuchotée en des pertes dans la précision normale de la parole, souvent par la même quantité [27]. En ce qui concerne le problème spécifique que nous abordons dans cette thèse, dans le domaine de reconnaissance de locuteurs, les auteurs ont également signalé que la stratégie la moins coûteuse et la plus efficace est d'ajouter de petites quantités de discours chuchoté de locuteurs cibles lors de l'inscription [26, 35], c-à-d utiliser des modèles multi-styles.

L'approche multi-style convient plus à la tâche à accomplir, car les systèmes actuels SV exigent des quantités importantes de données pour l'estimation des paramètres. La grande variabilité entre les intervenants dans l'ensemble d'apprentissage est nécessaire afin de garantir un échantillon représentatif de l'ensemble des locuteurs qui sont prévus lors de la phase de test. Comme nous le décrivons dans le chapitre suivant, dans les bases de données existantes disponibles au public, le nombre de locuteurs avec des enregistrements de parole chuchotée est relativement faible comparé au nombre de locuteurs avec une parole normale. Comme tel, le défi lors de l'estimation des paramètres est d'apprendre autant de variabilité que possible à partir des deux styles de la parole

et ceci afin de modéliser correctement l'information discriminative entre les locuteurs, mais avec des ressources limitées disponibles à partir d'un effort vocal (c'est-à-dire chuchoté). Avec les fonctionnalités actuelles de goulot d'étranglement, ceci devient un problème encore plus difficile, car il est nécessaire de former un système automatique de reconnaissance vocale à grande échelle et d'utiliser sa sortie pour former un réseau de neurones profond [22]. Au meilleur de notre connaissance, aucun corpus à grande échelle avec discours chuchoté annoté n'est disponible pour entraîner un système ASR. Ainsi, si des fonctionnalités de goulot d'étranglement doivent être explorées pour la tâche à accomplir, des solutions innovantes sont nécessaires pour minimiser les effets de ressources limitées sur l'efficacité du vecteur caractéristique extrait. En résumé, alors que les modèles multi-style peuvent offrir une alternative viable à la parole vocale murmurée SV, des techniques adéquates doivent encore être mises en place afin de surmonter les problèmes qui découlent de cette stratégie (par exemple, ceux rapportés dans [25, 27] pour ASR).

Dans cet esprit, les problèmes que nous abordons dans cette thèse peuvent être résumés comme suit:

1. Dans le domaine de la reconnaissance des locuteurs, le discours chuchoté n'a été étudié que dans le cadre du problème *d'identification des locuteurs*, avec une exploration limitée des limites de performances réalisables avec la parole chuchotée pour la *vérification des locuteurs* dans différents scénarios d'entraînement/tests et en utilisant des techniques standard développées pour la parole normale.
2. Problème d'inadéquation d'entraînement/test: au cours de la phase de test, quand il n'y a pas de données vocales murmurées disponibles pour l'entraînement ou pour inscrire les locuteurs cibles, la parole chuchotée peut induire des effets néfastes sur la performance de reconnaissance du locuteur. En fait, même si des données vocales chuchotées sont disponibles pour l'estimation des paramètres, le problème d'incompatibilité peut toujours être présent, car les variations de l'effort vocal peuvent être considérées comme une variation "à l'intérieur du locuteur" et cette variation n'est pas bien représentée dans les échantillons d'inscription des locuteurs cibles.
3. Ressources limitées: des systèmes typiques de pointe sont formés sur de grands ensembles de données, avec des de milliers d'intervenants, de plusieurs heures d'enregistrement, d'une variété de canaux et de sessions d'enregistrement différentes. Cependant, un large éventail de variations de l'effort vocal n'est toujours pas inclus dans les tâches d'évaluation à grande

échelle. Les systèmes standards de vérification des locuteurs doivent être compensés par des techniques permettant d'utiliser efficacement la quantité limitée d'informations disponibles des efforts vocaux différents de la parole vocale normale, tout en maintenant des niveaux élevés de performance dans la parole normale.

4. Effets négatifs dans les modèles multi-style: dans les systèmes de reconnaissance de la parole, des effets négatifs ont été observés lors de la combinaison de données issues de différents efforts vocaux. Plus précisément, alors que la précision de la reconnaissance d'une parole chuchotée s'est améliorée, la précision de la parole normale a diminué [25, 27]. L'impact de la combinaison de données de deux efforts vocaux différents pour la tâche à accomplir n'est pas clair. Les stratégies pour surmonter ce compromis de perte de gain avec un discours murmuré-normal restent toujours nécessaires.

0.3 Contributions de thèse

Le but de cette thèse est d'aborder les quatre problèmes mentionnés auparavant avec un accent particulier sur deux efforts vocaux: normal et chuchoté. Bien qu'il y ait eu un certain travail dans la littérature traitant de la question de *l'identification de locuteur* avec une parole chuchotée, nous nous concentrons ici sur la *vérification du locuteur*, pour lequel il y a plus d'applications pratiques. En particulier, cette thèse aborde plusieurs aspects du pipeline de vérification des locuteurs, y compris l'extraction de caractéristiques, où des fonctionnalités novatrices sont proposées contenant des informations indépendantes du locuteur et invariantes à travers les efforts vocaux, ainsi que différents schémas de fusion pour obtenir une précision exacte de reconnaissance de l'effort multi-vocal. Plus précisément, les principales contributions de cette thèse peuvent être résumées comme suit:

1. Des limites de performances réalisables avec un discours chuchoté pour la vérification du locuteur ont été signalées pour la première fois. Nous avons constaté que la parole chuchotée peut contenir autant d'informations spécifiques au locuteur que la parole normale, mais les approches standard conçues pour la parole normale ont tendance à échouer pour la parole chuchotée. L'un des problèmes limitant l'utilisation généralisée de ce style de parole pour les applications d'authentification est le manque de données suffisantes à l'entraînement de modèles. À cet égard, nous avons développé des stratégies qui utilisent efficacement les

ressources limitées disponibles, comme le nombre limité de locuteurs avec des enregistrements vocaux chuchotés; à utiliser pendant l'estimation des paramètres. Les résultats expérimentaux montrent des niveaux de performances élevés conformes aux performances obtenues pour la parole normale. Ces résultats ont été rapportés dans les publications #2, #5, #6, et #7 listées dans la section 1.3.

2. La proposition de fonctionnalités innovantes et hautement informatives pour améliorer la vérification des locuteurs multi-vocaux à l'aide d'informations acoustiques et d'analyse statistique. Plus précisément, trois algorithmes d'extraction de caractéristiques ont été proposés: *i*) caractéristiques de la modulation d'amplitude inspirée par l'audit et *ii*) deux variantes des coefficients classiques MFCC utilisant des versions modifiées du pipeline de traitement du signal utilisé pour le calcul des MFCC. Le premier cas a été motivé par des preuves montrant que le fait de varier lentement l'enveloppe des signaux passe-bande permet de transmettre des informations importantes sur le locuteur, utiles pour les tâches de reconnaissance des locuteurs. De plus, l'information mutuelle (MI) a été utilisée comme mesure d'analyse pour identifier l'information invariante entre les caractéristiques de modulation d'amplitude de la voix normale et de la parole chuchotée. Cela permet de ne pas tenir compte des canaux dont les valeurs de MI sont faibles tout en préservant des informations importantes sur les locuteurs. Les caractéristiques extraites à l'aide de la dernière méthode ont montrés pour extraire des informations complémentaires qui réduisent l'impact négatif lors de tests avec un discours chuchoté. Les variantes MFCC se sont aussi avérés être utiles, et les informations relatives à ces caractéristiques ont également été utiles lors de l'entraînement des réseaux de neurones profonds pour l'extraction de caractéristiques de goulot d'étranglement. Dans les deux cas, les systèmes formés avec les caractéristiques nouvellement proposées ont surpassé les systèmes formés avec les MFCC classiques. Ces résultats ont été rapportés dans les publications #1, #3, et #4, listées dans la section 1.3.
3. La proposition d'un Framework pour explorer des schémas de fusion à différents niveaux, à savoir: *i*) trame, *ii*) i-vecteur et *iii*) niveau de score. Ce Framework nous permet d'explorer la complémentarité des différents ensembles de caractéristiques et d'étudier comment utiliser plus efficacement les informations encodées dans les différentes représentations de caractéristiques, réduisant ainsi le besoin de locuteurs supplémentaires avec des enregistrements chuchotés. En utilisant ce cadre de travail, nous avons appuié les résultats précédemment rapportés montrant que pour la parole normale, la fusion au niveau de la trame est plus

efficace [23]. Cependant, pour la parole chuchotée, nous avons trouvé qu'il est préférable de former des systèmes séparés et de les fusionner à des niveaux plus élevés, soit en concaténant les i-vecteurs, soit en utilisant la fusion au niveau du score. Ces résultats ont été rapportés dans les publications #3, #4, #7 et #8 listées dans la section 1.3.

4. En résumé, des performances améliorées ont été obtenues pour la vérification de l'effort vocal multiple avec une disponibilité de ressources limitées provenant d'un effort vocal (c'est-à-dire chuchoté) et en utilisant des modèles multi-styles et des schémas de fusion. Des stratégies de fusion telles que la concaténation de i-vecteurs et la fusion au niveau des scores des variantes MFCC, le goulot d'étranglement et les caractéristiques de la modulation d'amplitude inspirée par l'audit se sont avérées être efficaces pour résoudre des problèmes tels que l'état d'incompatibilité entraînement/test. Pour les discours chuchotés, par exemple, des gains aussi élevés que 61% ont été observés lors de l'utilisation d'enregistrements vocaux chuchotés pendant la phase d'entraînement, tandis qu'en combinant des enregistrements vocaux normaux et chuchotés à partir de locuteurs cibles durant l'inscription, on pouvait voir des gains aussi élevés que 71% par rapport à un système de référence basé sur les caractéristiques MFCC. D'autre part, pour la parole normale, ces stratégies aident non seulement à réduire les effets négatifs observés lorsque les enregistrements de parole avec deux types d'efforts vocaux différents ont été combinés durant l'entraînement et l'inscription, mais aussi à obtenir des gains aussi élevés que 79% par rapport au système de référence. Ces résultats ont été rapportés dans les publications #3, #4, et #8 listées dans la section 1.3.

0.4 Résumé

Cette thèse de doctorat a abordé un problème très important, mais pas suffisamment exploré, de la vérification des locuteurs dans les scénarios de test d'efforts vocaux multiples. En particulier, nous avons centré l'attention sur deux styles de la parole, c'est-à-dire, la parole normale et celle chuchotée. Nous avons constaté que la parole chuchotée peut contenir autant d'informations spécifiques au locuteur que la parole normale. Toutefois, les approches standard conçues pour la parole normale ont tendance à échouer pour la parole chuchotée. À cet égard, nous avons développé des stratégies pour intégrer ce style de parole dans les scénarios de tests des systèmes standards de vérification des locuteurs. Ces stratégies permettent d'utiliser efficacement les ressources limitées disponibles ainsi

que d'atteindre des niveaux de performance élevés pour la parole chuchotée conformément avec les performances obtenues pour la parole normale.

0.4.1 Chapitre 2: Contexte de travail

Le chapitre 2 fournit le contexte sur le discours chuchoté, en soulignant les principales différences avec la parole normale. Il présente les principaux points de vue des études de recherche perceptuelle et acoustique, ainsi qu'une revue des principales applications pratiques où la parole chuchotée a été utilisée. Ce chapitre décrit également le problème de la vérification des locuteurs et donne un aperçu général des principales techniques du traitement de la parole, de l'extraction des caractéristiques et de l'apprentissage automatique impliquées dans les blocs de construction d'un système de vérification automatique des locuteurs, ainsi qu'une description des bases de données vocales utilisées pour les expériences présentées ici.

0.4.2 Chapitre 3: Analyse comparative pour la vérification de l'effort vocale multiple

Dans le chapitre 3, nous avons exploré les avantages des différentes méthodes de prétraitement existantes, des stratégies de déformation de fréquences, des représentations de caractéristiques et des configurations des systèmes SV. Cela nous permet principalement d'étudier certaines stratégies déjà proposées dans la littérature. L'objectif général de ce chapitre est d'explorer l'enveloppe de performance réalisable avec la parole chuchotée, en particulier dans le cadre d'une tâche de vérification de locuteur (SV) à petite échelle, guidant ainsi les orientations de recherche des chapitres suivants pour des applications à plus grande échelle.

Ces expériences ont été réalisées dans un scénario idéal, en utilisant un nombre limité de locuteurs et un ensemble pour la vérification des locuteurs en utilisant des enregistrements de parole de locuteurs cibles aussi pour l'estimation des paramètres. Compte tenu de la quantité limitée de locuteurs et d'enregistrements vocaux, un système de classification fondé sur des modèles de mélange gaussiens (Gaussian mixture models) et une adaptation a posteriori maximale étaient plus appropriés dans ce scénario. Pour le système décrit, les coefficients MFCC, largement utilisés dans le domaine, ont été adoptés pour implémenter un système SV indépendant du texte.

Le corpus vocal **CHAINS** (Caractérisation des locuteurs individuels, Characterizing Individual Speakers) a été utilisé [12]. Le corpus contient les enregistrements de 36 locuteurs obtenus sur deux sessions différentes dans un interval de deux mois, il ya trois accents différents: 28 locuteurs d'Irlande (16 hommes), 5 locuteurs des états unis (2 hommes) et 3 locuteurs du Royaume-Uni (2 hommes). Des détails supplémentaires sur la base de données peuvent être trouvés dans [12]. Le stimulus de la parole a été généré selon six conditions de parole, à savoir solo (lecture à taux naturel), récapitulation sans contraintes de temps, lecture synchrone à deux personnes, imitation synchrone répétitive, lecture à vitesse accélérée et lecture chuchotée.

Selon les résultats expérimentaux; et en utilisant une déformation (warping) de fréquence différente de l'échelle de Mél; en limitant la bande de fréquences pour éliminer les différences spectrales entre les deux styles de parole et en utilisant des représentations de caractéristiques alternatives, des améliorations entre 13% et 38% relatif ont eté obtenus. Ces améliorations, toutefois, ont entraîné une pénalité sévère pour le scénario adapté, et ceci a été observé pour tous les ensembles de caractéristiques évalués. En ce qui concerne la combinaison de caractéristiques, d'entités ou la fusion au niveau de la trame, les résultats ont montré que cette stratégie n'aide pas à obtenir d'autres améliorations de la performance lors d'incompatibilité entraînement/test . Cependant, certaines combinaisons de caractéristiques peuvent aider à maintenir les performances conformes avec le système de base pour la condition correspondante, tout en réalisant des améliorations modeste dans l'état d'incompatibilité (environ 21% par rapport au système de base).

Des modèles à styles multiples ont également été explorés, c'est-à-dire l'utilisation de la parole normale et chuchotée durant l'ntraînement et l'adaptation du modèle comme ce qui a été fait dans les études précédentes sur l'identification du locuteur [26, 35]. Cela permet de modéliser correctement les informations; spécifiques au locuteur; présentes dans les traits vocales chuchotés. Selon nos résultats, en augmentant progressivement la durée de la parole chuchotée, la performance du système s'améliore aussi progressivement, confirmant ainsi les résultats précédents [26, 35]. En outre, on a également observé que l'augmentation de la quantité de parole chuchotée augmente légèrement le taux d'erreur égal (equal error rate, EER) pour la parole normale. Par exemple, en utilisant uniquement la parole normale pour l'entraînement, un EER de 2,13% a été signalé. En utilisant la même quantité de données pour les deux efforts vocaux, un EER de 3,05 % (c'est-à-dire, 43 % plus élevé) a été atteint. Selon ces résultats, pour une tâche de vérification de locuteur (SV)

réel, des performances améliorées peuvent être obtenues lors de tests avec des discours chuchotés, mais au prix d'une performance inférieure pour la parole normale.

Conclusions

Dans ce chapitre, les limites de performance d'un système standard de vérification des locuteurs GMM–UBM ont été obtenues en utilisant plusieurs stratégies, telles que la déformation de fréquence, l'analyse sous-bande, les représentations de fonctions alternatives, la combinaison de traits ainsi que des modèles multi styles. Notre évaluation expérimentale montre que les conditions de discordance *entraînement/test* peuvent affecter fortement les performances d'un système SV, indépendamment de la représentation de caractéristique utilisée. Comme dans les études précédentes, il a été démontré que pour qu'un système de SV puisse traiter à la fois la parole normale et chuchotée pour des applications pratiques, l'entraînement d'un modèle de locuteur doit impliquer des données contenant les deux types d'efforts. Une telle approche, cependant, a entraîné des performances de vérification plus faibles pour la parole normale. Dans l'ensemble, les représentations de caractéristiques évaluées ici ont été principalement proposées pour des applications vocales à voix normale, ce qui suggère que d'autres représentations de fonctionnalités, réglées pour la vérification de locuteur de voix chuchotée, sont toujours nécessaires.

0.4.3 Chapitre 4: Feature Mapping et schémas de fusion

Dans le chapitre 4 nous avons adopté un schéma d'évaluation plus réaliste en incluant des ensembles de données supplémentaires enregistrés dans des conditions différentes, à savoir les bases de données TIMIT [38] et wTIMIT [39]. De plus, en suivant les protocoles d'évaluation standard pour la vérification des locuteurs [14, 40], une distinction claire est établie entre les locuteurs en arrière-plan et les locuteurs ou clients cibles. Nous explorons tout d'abord ce qui est réalisable avec les caractéristiques standards MFCC et des caractéristiques dérivés du modèle AM-FM, en utilisant l'approche actuelle de vérification de locuteurs. En utilisant les stratégies de feature mapping (mappage de caractéristiques), nous avons verifié si les caractéristiques spécifiques des locuteurs peuvent être mappées à des domaines de caractéristiques spécifiques afin de compenser le manque de données vocales murmurées à partir de locuteurs cibles. Ensuite, la complémentarité des caractéristiques dérivées des modèles AM-FM par rapport au MFCC conventionnel est explorée en utilisant trois schémas de

fusion. Les résultats montrent que, dans le contexte de modèles multi-styles, les stratégies de fusion sont plus efficaces que les stratégies de mappage des caractéristiques. Plus de recherches devraient être faites dans cette direction.

Tout d'abord, nous avons exploré les effets de l'ajout de discours chuchoté lors de l'estimation des paramètres. Pour ces expériences, les données d'entraînement contiennent des enregistrements provenant des deux styles de parole, mais pour le langage normal, le nombre de locuteurs est significativement plus grand que le nombre de locuteurs pour la parole chuchotée. Nous avons constaté que l'ajout de la parole chuchotée lors de l'estimation de la matrice de variabilité totale (T-matrix) peut ajouter des gains de performance d'environ 30% lors de tests avec discours chuchoté, mais aussi de petites pertes ont été observées lors de tests avec la parole normale. Il est important de noter que l'extracteur i-vecteur et le système SV en général, peuvent apprendre une certaine variabilité à partir des enregistrements de parole qui ont été inclus lors de l'estimation des paramètres. Ceci n'est toutefois pas suffisant, en effet, un écart de performance d'environ 17% subsiste entrela parole normale et murmurée.

Ensuite, deux techniques de mappage des caractéristiques ont été évaluées dans nos expériences. La première est une technique classique GMM [41], proposée à l'origine pour la synthèse texte-parole. La deuxième technique est basée sur des réseaux de neurones, qui s'est avérée utile dans la littérature de conversion vocale [42]. Selon nos résultats, les deux mappages de caractéristiques ajoutent quelques gains lors de tests avec discours chuchoté, avec des améliorations relatives jusqu'à 37%. L'approche de mappage GMM semble être optimale pour compenser lorsque le discours chuchoté est présent pendant le test. Néanmoins, dans ce cas, l'écart de performance reste encore considérable, environ 14% (EER) entre pour la parole normale et chuchotée.

Enfin, trois schémas de fusion ont été étudiés dans ce chapitre, deux au niveau de l'entrée et un au niveau de la sortie, à savoir: *i) fusion de niveau de trame*, *ii) concaténation de i-vecteurs* et *iii) fusion au niveau du score*. En comparant les trois schémas de fusion, nous avons observé des différences significatives. Par exemple, la fusion au niveau de la trame semble être la manière la moins efficace de combiner les informations des deux ensembles de caractéristiques. La fusion aux niveaux supérieurs, tels que i-vecteur ou niveau-score, se révèle être de meilleures options. Il est nécessaire de tenir compte du fait que cette fusion au niveau du score nécessite l'entraînement d'un système de fusion, alors que pour la concaténation du i-vecteur, il n'est pas nécessaire de

entraîner de systèmes supplémentaires. Cependant, en terme de performance, la fusion au niveau du score présente des résultats légèrement supérieures que la concaténation du i-vecteur. En fait, les meilleurs EER globaux sont atteints avec ce schéma. En comparaison avec le système de référence, des améliorations relatives de 44% et de 42% on été obtenus pour la parole normale et murmurée respectivement.

Conclusions

Dans ce chapitre, nous avons abordé la question de la vérification des locuteurs basée sur la parole chuchotée dans un scénario plus réaliste. Trois bases de données ont été regroupées afin d'augmenter le nombre de locuteurs et d'ajouter plus de souplesse à l'évaluation expérimentale. L'ajout de données chuchotées durant l'entraînement, afin d'ajouter des informations sur la variabilité de la parole chuchotée, combinée avec des techniques de mappage des caractéristiques, pour compenser le manque de données vocales murmurées à partir de locuteurs cibles, nr suffit pas à améliorer la performance de vérification du locuteur pour la parole chuchotée. Nous avons également exploré la complémentarité des informations extraites des ensembles de caractéristiques WIF et MFCC par l'intermédiaire de trois schémas de fusion, à savoir: *i*) niveau de trame, *ii*) concaténation i-vecteur et *iii*) niveau de score. On obtient des gains aussi élevés que 42% et 44% pour un discours murmuré et normal, par rapport à un système de base basé sur i-vecteurs/PLDA + MFCC sans discours chuchoté dans l'ensemble de entraînement.

Dans l'ensemble, nous avons observé que les fonctionnalités existantes (par exemple, MFCC) ne transmettent pas suffisamment d'informations fiables sur l'identité du locuteur à travers différents efforts vocaux. Étant donné le manque de locuteurs suffisants pour entraîner des modèles indépendants et dédiés à la parole chuchotée, des techniques telles que le mappage des caractéristiques demeurent insuffisantes pour améliorer les performances et les schémas de fusion semblent être plus efficaces. Néanmoins, le problème d'inadéquation est toujours présent et l'écart de performance reste considérable entre la parole normale et chuchotée.

0.4.4 Chapitre 5: Exploration de l'information invariante dépendante du locuteur à travers les efforts vocaux

Dans le chapitre 5, nous avons continué à explorer les schémas de fusion afin de combiner les avantages des différents systèmes et représentations des caractéristiques. Ici, pour compléter les stratégies décrites dans les chapitres 3 et 4, nous explorons le calcul de caractéristiques innovantes qui extraient des informations invariantes intégrées dans les deux styles de parole. *i*) Nous présentons des preuves sur la façon dont nous pouvons calculer des variantes des caractéristiques standard MFCC pour extraire des informations complémentaires et réduire l'impact négatif lors de tests avec discours chuchoté, et ceci en utilisant des connaissances acoustiques. *ii*) Calcul des caractéristiques de modulation d'amplitude inspirée par l'audition (Auditory Inspired Amplitude Modulation Spectrum based features, AAMF). Ceci est motivé par des résultats antérieurs qui ont montré dans le passé comment les caractéristiques basées sur le spectre de modulation peuvent séparer de façon précise la parole des composants basés sur l'environnement (par exemple le bruit et la réverbération)[43], ce qui ajoute de la robustesse aux systèmes de reconnaissance des locuteurs. De même, dans le chapitre 4, l'enveloppe (lentement variable) des signaux de bande passante combinée à l'information de phase a montré porter des informations importantes dépendantes du locuteur ainsi que des informations complémentaires utiles pour les schémas de fusion.

Pour le cas spécifique des caractéristiques basées sur le spectre de modulation, nous utilisons l'information mutuelle (MI) comme mesure d'analyse afin d'identifier l'information invariante entre les paires de caractéristiques de la voix normale et la parole chuchotée. MI permet d'analyser les dépendances statistiques linéaires et non linéaires entre les deux ensembles de caractéristiques. Cette mesure s'est avérée être un moyen efficace de mesurer la pertinence et la redondance entre les fonctionnalités pour la sélection des fonctionnalités ou même la caractérisation [44, 45, 46]. Ceci, combiné à un système de fusion, contribue non seulement à réduire les taux d'erreur lorsqu'il n'y a pas d'enregistrements vocaux chuchotés par les locuteurs cibles, mais aussi à réduire l'impact négatif observé de l'ajout de la parole chuchotée durant l'estimation des paramètres.

Nous avons tout d'abord évalué les caractéristiques individuelles afin de caractériser leur performance dans une tâche de vérification des locuteurs. Ces expériences ont été réalisées en utilisant des modèles multi-style comme ce qui a été fait lors du chapitre 4. Ensuite, en évaluant les deux variantes MFCC, nous avons constaté que dans le premier cas, le calcul des MFCC, sur le signal

résiduel présentait des performances médiocres pour les deux styles de parole, de sorte que les systèmes basés uniquement sur cet ensemble de caractéristiques ne devraient pas fonctionner au même niveau que les MFCCs, mais ils peuvent fournir des informations complémentaires. Ensuite, en calculant les MFCCs sur une bande de fréquences limitée, on a observé que cet ensemble de caractéristiques fonctionne aussi bien que les MFCCs pour la parole normale, mais améliore le taux d'erreur de vérification du locuteur à voix chuchoté de 20% à 16,67% EER. Enfin, les caractéristiques basées sur le spectre de modulation d'amplitude inspirées par l'audition (AAMF) ont présenté des performances supérieures lors des tests avec la parole normale. Toutefois, l'écart de performance entre la parole normale et la voix chuchotée demeure supérieur de 16% EER.

Avant d'explorer les schémas de fusion, et compte tenu du nombre de combinaisons possibles, une analyse statistique a été effectuée afin d'explorer les contributions de chaque ensemble de caractéristiques. Dans ce cas, nous avons effectué une analyse en utilisant comme caractéristiques les scores de sortie des systèmes formés sur les ensembles de caractéristiques proposés. Pour l'analyse, nous utilisons la statistique de Lawley-Hotelling [47], une mesure communément utilisée dans l'analyse MANOVA (analyse de variance multivariée) quand nous voulons comparer les vecteurs moyens de k groupes d'échantillons pour vérifier s'il y a des différences significatives. À partir de l'analyse statistique, nous avons trouvé les combinaisons de caractéristiques qui peuvent offrir une meilleure performance pour les deux styles de paroles. Ces prédictions ont été vérifiées lors de la comparaison des résultats obtenus avec le système de base. On a observé que des améliorations relatives variant de 19% à 39% pouvaient être atteintes pour la parole normale, tandis que pour la parole chuchotées, les améliorations se situaient entre 26% et 57%, en utilisant dans les deux cas les ensembles de caractéristiques proposés. D'autre part, en comparant les schémas de fusion, on a observé que la fusion au niveau du score obtenait la meilleure performance pour la parole chuchotée, tandis que la meilleure solution pour la parole normale était la fusion au niveau de la trame. D'autre part, la concaténation de i-vecteur était le schéma montrant un compromis entre la performance et la charge de calcul, car un entraînement supplémentaire n'est pas nécessaire, comme c'était le cas avec la fusion au niveau du score.

Conclusions

Dans ce chapitre, nous avons décrit trois ensembles de caractéristiques innovantes fournissant des informations invariantes à travers les efforts vocaux et des informations complémentaires aux fonctionnalités existantes d'une tâche SV. Les caractéristiques proposées ont été construites à partir des informations recueillies dans les chapitres précédents, ainsi que de celles rapportées dans la littérature. Deux variantes du MFCC ont été proposées, l'une focalisée principalement sur le résidu LP, soulignant ainsi les similitudes des segments de parole non vocalisés entre les deux efforts vocaux. La deuxième variante est basée sur la sous-bande de 1,2-8 kHz qui est moins affectée par les chuchotements. Il a été montré que les deux variantes MFCC fournissent des informations complémentaires au MFCC classique et fournissent des gains aussi élevés que 39% et 41% respectivement pour la parole normale et chuchotée. Un troisième ensemble de caractéristiques a été construit à partir de la preuve du chapitre 4 montrant que les enveloppes de sous-bandes lentement variées véhiculent des informations utiles pour l'effort vocal croisé SV. En utilisant le critère de l'information mutuelle, un masque binaire a été développé pour sélectionner des canaux acoustiques/ de modulation qui sont invariants aux changements de l'effort vocal. Lorsque les trois ensembles de caractéristiques ont été combinés, des améliorations de 66% et de 63% par rapport à un systeme de base basée sur MFCC ont été obtenues pour la parole normale et chuchotée, respectivement. Alors que l'écart entre l'EER de la parole normale et chuchotée était considérablement réduit, les niveaux atteints pour la parole chuchotée peuvent encore être considérés comme élevés à environ 10% EER. Le chapitre suivant va explorer l'utilisation de ces nouvelles caractéristiques proposées comme contribution aux approches actuelles des réseaux de neurones.

0.4.5 Chapitre 6: Approches d'apprentissage profond pour la vérification des locuteurs multi-vocaux

Dans le chapitre 6, nous avons continué à explorer l'information invariante relative au locuteur à travers les efforts vocaux en utilisant des approches d'apprentissage profond. Les systèmes actuels de pointe reposent sur l'extraction de i-vecteurs [19]. Les techniques les plus récentes remplacent le MFCC classique, comme caractéristique acoustique, par des approches basées sur l'apprentissage profond afin d'extraire les caractéristiques dites de goulot d'étranglement (BNF). Cependant, la robustesse de ces approches, n'a pas été testée sous différents styles de parole tel que la parole

chuchotée. Dans ce chapitre, nous visons à combler cette lacune. Tout d'abord, en explorant une configuration de réseaux de neuronnes de goulot d'étranglement standard, où les entrées sont les sorties classiques du filtre log mel-scale. Et deuxièmement, lorsque l'entrée au réseau de neuronnes est un ensemble de vecteurs concaténés associés aux ensembles proposés dans le chapitre 5. Ces vecteurs permettent de réduire l'impact de la condition d'inadéquation entre entraînement/test lors de tests avec discours chuchoté et en absence d'information de ce style de parole dans l'ensemble d' entraînement.

Les réseaux de neurones avec goulot d'étranglement sont des réseaux de neurones profonds (DNN, deep neural networks) avec une topologie particulière, où l'une des couches cachées a une dimension significativement plus faible que les couches voisines; cette couche est connue sous le nom de couche de goulot d'étranglement. Un vecteur de caractéristiques de goulot d'étranglement (bottleneck feature, BNF) est obtenu en envoyant un vecteur de caractéristiques d'entrée primaire à travers le DNN et en lisant le vecteur de valeurs au niveau de la couche du goulot d 'étranglement [21]. Dans nos expériences, les cibles pour le DNN ont été obtenues à l'aide d'un système automatique de reconnaissance de la parole (ASR) entraîné avec kaldi [48]. Dans notre cas le nombre de cibles est de 4121. Les données d' entraînement correspondent à 460 heures extraites de l'ensemble de données LibriSpeech [49]. Pour les expériences présentées ici, deux approches ont été testées, d'abord les caractéristiques d'entrée DNN sont des contextes temporels concaténés de 15 trames, chaque trame étant représentée par 27 sorties de filtres à échelle de Mel (Mel-scale filterbank), en utilisant le même réglage que celui utilisé dans le calcul des caractéristiques MFCC. La troisième couche cachée est la couche de goulot d'étranglement et nous allons référer à cette fonctionnalité par FBBNF (filterbank bottleneck features). Pour la deuxième approche, nous avons concaténé des caractéristiques de treize trames consécutives: *i*) Treize MFCC, ces caractéristiques sont destinées à la tâche originale pour former le DNN, classification des unités sub-phonétiques. *ii*) 27 sorties de filtre à échelle de Mel, les filtres triangulaires sont espacés entre 1,2 kHz et 8 kHz. *iii*) 27 sorties de filtre à échelle de Mel, extraites du résidu LP. Nous avons également varier l'emplacement de la couche de goulot d'étranglement, de la deuxième couche à la quatrième couche. Nous allons référer à cet ensemble de caractéristiques par LRBNFi, où i représente la couche où le goulot d'étranglement est situé, et LR représente les sorties de filtre limité et résiduel. Pour tous les cas, la couche goulot d'étranglement a 80 neurones.

Tout d'abord, nous avons évalué les ensembles de caractéristiques de goulots d'étranglement individuels et les avons comparé au système de référence MFCC/PLDA. Nous avons constaté que dans notre environnement expérimental, le système standard MFCC surpasse le système basé sur FBBNF. Ceci est dû à la faible variabilité phonétique présente dans les énoncés de courte durée utilisés pour notre tâche d'évaluation. Toutefois, les caractéristiques FBBNF sont plus robustes contre les changements dans l'effort vocal, les différences relatives sont plus de 45% comparées au MFCC, les tests sont effectués avec une parole chuchotée. Ensuite, en évaluant le système avec l'entrée proposée au DNN et en faisant varier la couche de goulot d'étranglement, c'est-à-dire les ensembles de caractéristiques LRBNFi, on a observé que les deux ensembles de caractéristiques utilisant la couche de goulot d'étranglement plus proche de l'entrée, ont un meilleur rendement que celui qui se rapproche de la sortie, ce qui renforce les observations dans [50]; cette constatation s'applique aux deux styles de parole. Lors de la comparaison des ensembles de caractéristiques extraits dans la troisième couche, c'est-à-dire les ensembles de caractéristiques FBBNF et LRBNF3, il était clair que le DNN formé avec le schéma d'entrée proposé, c'est-à-dire les informations d'entrée concaténées provenant des MFCC, montre un compromis entre les deux styles de parole.

Ensuite, en utilisant les caractéristiques du goulot d'étranglement dans les schémas de fusion et les tests avec la parole normale, on a observé qu'indépendamment du schéma de fusion, les meilleurs résultats globaux ont été obtenus en utilisant des systèmes combinés formés avec les ensembles de caractéristiques proposés au Chapitre 5 ainsi que les caractéristiques de goulot d'étranglement extraites du schéma d'entrée proposé. En outre, il a été observé que les différences de performances entre les différents paramètres étaient minimes. Cela suggère que les ensembles de caractéristiques proposés contiennent des informations fortement discriminatives relatives au locuteur, capables de compenser les limitations dont souffrent les caractéristiques de goulot d'étranglement standard de notre cadre expérimental. Pour la parole normale, la concaténation i-vector est la meilleure approche de fusion car elle n'obtient pas seulement les taux d'erreur les plus faibles (EER = 0,63%), mais ne nécessite aucune donnée supplémentaire pour former le schéma de fusion. D'une autre part, pour le discours chuchoté, les caractéristiques basées sur le spectre de modulation sans sélection de caractéristique, c'est-à-dire l'ensemble de caractéristiques antérieur à l'analyse de l'information mutuelle, combinées aux caractéristiques de goulot d'étranglement au niveau du score ont atteint les taux d'erreur les plus faibles.

Enfin, nous avons évalué les effets de l'ajout de discours chuchoté lors de l' entraînement en augmentant progressivement le nombre d'énoncés vocaux chuchotés par les locuteurs cibles et en comparant la concaténation i-vecteur et la fusion au niveau du score. Nous avons observé que les deux schémas de fusion utilisant les ensembles de caractéristiques proposés, seulement en ajoutant une énonciation la performance pour la parole chuchotée était déjà conforme avec la performance du système de base (système basé sur MFCC/PLDA) lors de l'utilisation de huit énoncés. L'autre aspect à mettre en évidence est la dégradation de la performance pour la parole normale avec l'ajout de prononcés de discours chuchoté au cours de l'inscription. Il s'agit d'un problème qui affecte à la fois les schémas de fusion de référence et celui proposé, mais est plus visible dans le cas du système de référence. Dans l'ensemble, les schémas de fusion proposés maintiennent le taux d'erreur inférieur à 2% pour la parole normale, ce qui est en fait meilleur que la performance obtenue par le système de base MFCC/PLDA sans discours chuchoté. Pour ces résultats, la concaténation de i-vecteur semble être le schéma qui montre un compromis entre la performance et la charge de calcul, etant donné qu'aucun entraînement supplémentaire est nécessaire.

Conclusions

Dans ce chapitre, nous avons abordé le problème de la recherche d'informations invariantes dépendantes du locuteur à travers les efforts vocaux en utilisant des réseaux de neurones profonds. En plus de cela, nous avons continué à explorer les avantages de deux schémas de fusion (niveau de score et niveau i-vecteur) pour surmonter les défis existants, à savoir: *i*) les énoncés de courte durée (4,5 secondes en moyenne), *ii*) absence de donnée vocale murmurée disponible durant l'entraînement à partir de locuteurs cibles, et *iii*) les effets négatifs observés lors de l'ajout d'enregistrements vocaux chuchotés durant l' entraînement.

Dans les travaux précédents, il a été démontré que les performances des systèmes de vérification des locuteurs dépendent fortement du type de parole fourni comme entrée [51]. En caractérisant les systèmes de référence (Tableau 6.1) pour la parole normale, il est devenu évident que les MFCCs et les caractéristiques standards BNF ont atteint des valeurs EER plus élevés que ce qui est généralement rapporté dans la littérature [49]. Ceci est probablement dû à la courte durée de la parole qui limite la variabilité phonétique présente dans l'ensemble d'entraînement [52, 53]. Néanmoins, les caractéristiques proposées du LRBNF*i* semblent réduire cet effet négatif, en particulier lorsque la

couche de goulot est plus proche de l'entrée. Pour les discours chuchotés, d'autre part, les caractéristiques BNF surpassent, significativement, tous les ensembles de caractéristiques individuelles précédemment étudiés. On suppose que cela était dû aux capacités supérieures du DNN pour modéliser les informations invariantes lors de la comparaison de la parole normale et chuchotée. Dans l'ensemble, si un seul système de référence devait être utilisé, celui basé sur le BNF en utilisant comme entrée la combinaison de MFCC, les sorties résiduelles et à bande limitée de log-filtre ainsi qu'une couche de goulot d'étranglement plus proche de l'entrée (LRBNF2) donneraient les meilleurs résultats globaux de l'effort vocal multiple.

Les stratégies de fusion ont clairement démontré de nombreux avantages. Elles ont non seulement réduit les taux d'erreur lorsqu'il n'y avait pas d'enregistrements vocaux chuchotés par les locuteurs cibles, mais ont également contribué à réduire l'impact négatif de l'ajout de mots chuchotés lors de l'estimation des paramètres. En général, l'ensemble de fonctionnalités AAMF proposé s'est avéré être le plus informatif pour la parole normale et chuchotée, à utiliser dans un schéma de fusion avec des fonctionnalités de goulot d'étranglement quand il n'y a pas de données vocales chuchotées provenant des locuteurs cibles. Cependant, cette configuration ne montre pas les meilleurs résultats lorsque des enregistrements vocaux chuchotés provenant de locuteurs cibles ont été inclus, et des fonctionnalités alternatives telles que LRBNF3, LMFCC, RMFCC et AAMF(FS) se sont avérées être un meilleur choix dans une tâche de vérification de locuteurs à voix multiple. Lors de la comparaison des schémas de fusion, le schéma de concaténation i-vecteur s'avère être la meilleure stratégie de fusion à utiliser. Cela n'est pas seulement justifié par les résultats obtenus, mais aussi par le fait qu'il n'est pas nécessaire d'entraîner un système de fusion supplémentaire, comme c'est le cas pour la fusion au niveau du score. Dans l'ensemble, avec le schéma de fusion proposé, il est démontré que seulement 4,5 secondes (approximativement) de données d'entraînement chuchotées sont nécessaires pour obtenir les mêmes performances que le système de référence, qui nécessitait à son tour 22,5 secondes (approximativement) de données d'inscription chuchotées. En conséquence, les systémes proposées sont efficace pour gérer les variations de l'effort vocal et les tâches de vérification des locuteurs à faible ressources.

Chapter 1

Introduction

Human speech is a natural and flexible mode of communication that not only conveys a message, but also traits such as identity, age, gender, social and region of origin, emotional, and health states, to name a few. Under controlled conditions, speech processing systems have become useful across a number of domains, as depicted by Figure 1.1. For example, automatic speech recognition has opened doors for speech to be used as a reliable human-machine interface, letting humans control things such as televisions, smartphones, and car stereo systems, not to mention interact with automated customer support services. Advances in speaker recognition technologies, in turn, have allowed humans to use their voice to e.g., authenticate themselves into their bank's automated phone system. Such is the potential in this field, that recent reports suggest that the global speech technology market is expected to surpass the $31 billion mark by the end of 2017, mostly due to three speech applications: automatic speech recognition (ASR), automatic speaker verification (ASV), and text-to-speech synthesis (TTS) [1]. A good portion of this market has been driven by the proliferation of smartphones and tablets across the globe. As examples, a number of applications have emerged that allow people to use their voices to interact with their devices (e.g., Apple's Siri), login to secure services (e.g., Bell Canada's Voice Identification Service), or even unlock their mobile devices (e.g., Baidu-I^{2}R Research Centre's Speaker Verification Service). Notwithstanding, these existing commercial systems have been developed for "typical" scenarios, such as clear adult voices with limited foreign language accents and small amounts of ambient noise. These assumptions and conditions, however, are difficult to satisfy in real-world environments and in everyday applications.

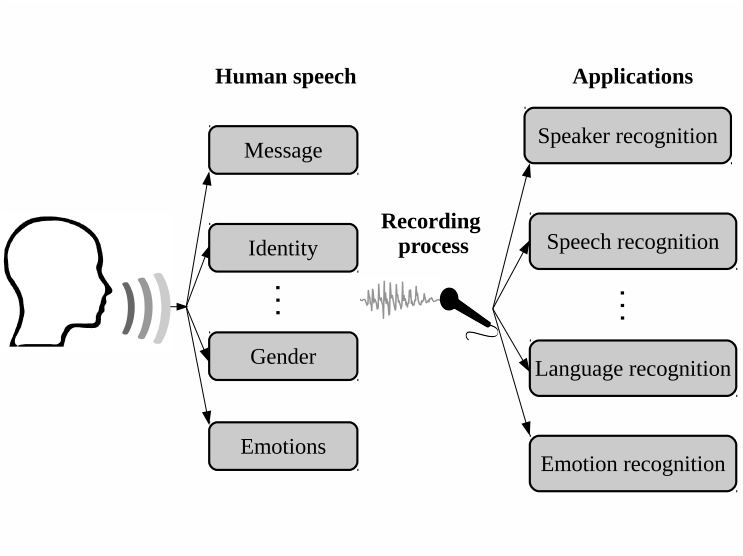


Figure 1.1 – Diagram of representative information sources available with human speech and their potential applications.

Particularly, in this dissertation the speaker recognition problem is of special interest. The potential of automatic speaker recognition applications lies especially in access control and identity management applications where knowledge-based authentication (KBA) is still the dominant way of authenticating users [2]. KBA requires users personal information in order to grant access to a service; representative details can include: passwords, user names, personal questions, or a combination of them. This method, however, has many drawbacks that can put at risk critical information, as well as concerns regarding privacy while collecting information. As an example, with the widespread use of social networks, many users are not aware of the information they share, which in many situations can be the same information used for verification purposes when authorizing a transaction (e.g. birth date). Moreover, existing authentication methods are prone to many documented attacks, thus more secure and reliable methods have been the focus of recent research [3, 4, 5].

As an alternative to KBA, biometrics based authentication, which combines mathematics and digital signal processing techniques to analyze physiological characteristics, is one such domain that has gained significant attention. In contrast to KBA, biometrics is based on who the user is, instead of what the user knows. Such technologies are burgeoning for identity management as they eliminate the need for personal identification numbers, passwords, and security questions [6]. Despite the several advantages of using biometrics for identification purposes, challenges and unresolved problems still remain, thus hampering widespread usage. For example, biometrics recognition is

compromised by external factors that may alter the patterns that are being analyzed (e.g., cuts and burns to the finger in fingerprint-based systems; ambient noise in speech-based solutions), as well as by natural human physiological factors, such as aging and disease (e.g., in facial and speech-based systems) [6, 7].

User preference also plays an important role when deciding the method to be used for authentication. In this regard and according to recent statistics, speech-based biometrics have ranked highly in costumer preference, outranking fingerprint and iris scanning solutions [8, 9]. Due to widespread usage of smartphones worldwide, speech-based biometrics are quickly gaining popularity, particularly in financial institutions [8]. Within such applications, customers can gain access to their secure banking and insurance services by simply speaking into their phones. For financial institutions, this ease-of-use enhances customer satisfaction, whilst reducing customer care costs through increased automation rates. Costumers, on the other hand, given the flexibility of speech based communication and being in a particular situation, can pose big challenges to these types of applications by changing e.g. their vocal effort from whispering to shouting according to the environment. This, together with ambient noise, have posed serious threats to speech enabled applications performance in general. Ambient noise has detrimental effects on speech based biometrics systems, particularly those trained with conventional features, such as mel-frequency cepstral coefficients (MFCC). As an example, speaker identification accuracy as low as 7% have been reported in very noisy environments [10]. As such, over the years several speech enhancement algorithms have been proposed for environment-robust speaker recognition applications [11]. Varying vocal efforts, on the other hand, also can induce severe detrimental effects on speaker verification performance. For example, whispered-speech speaker identification accuracy as low as 20% have been reported [12] in *clean conditions*, whereas accuracies as low as 8.71% have been reported for shouted speech [13]. In fact, it is highly likely that customers utilizing a mobile banking application on their smartphones will whisper sensitive information or to speak louder, even shout, if the user perceives that his/her spoken words are not being heard.

In automatic speaker recognition there are two classical tasks that can be performed: speaker identification (SI) and speaker verification (SV). Identification is the task of deciding, given a speech sample, who among a set of speakers said it. This is an N–Class problem (given N speakers), and the performance measure is usually the classification rate or accuracy. Verification, in turn, is the task of deciding, given a speech sample, whether the specified speaker really said it or not. The

SV problem is a two class problem of deciding if it is the same speaker or an impostor requesting verification. Commonly, SV exhibits greater practical applications related to SI, especially in access control and identity management applications [14]. With automatic speaker recognition, Gaussian mixture model (GMM) based systems using maximum a posteriori (MAP) adaptation and MFCC as feature vectors were for many years the dominant approach for text-independent speaker recognition [15, 54]. Recent advances have introduced the concept of *supervector*, which combined with support vector machines (SVM), have lead to improved recognition results [16]. More recently, latent variable inspired approaches for feature extraction such as joint factor analysis (JFA), alone or combined with SVM, showed to be an effective approach to compensate channel effects and reduce scoring time [17, 18]. Following the JFA framework, current state-of-the-art SR systems are based on identity vectors (*i-vectors*) extraction with cosine distance or probabilistic linear discriminant analysis (PLDA) based scoring [19, 20]. Moreover, given the recent advances in deep learning and deep neural networks, new approaches based on the so-called bottleneck features (BNF) [21] are emerging and first experiments are showing improved performance over conventional MFCC features (e.g., [22, 23]). Still within the deep learning realm, recent studies have also started exploring the use of deep neural networks to replace the GMMs in the computation of necessary statistics during i-vector extraction [22].

1.1 Problem description

In this dissertation we want to focus attention on one of the emerging challenges for developers of automated speech-enabled applications: *varying vocal efforts*. Traunmuller and Eriksson in [24] characterized vocal effort as: "*the quantity that ordinary speakers vary when they adapt their speech to the demands of an increased or decreased communication distance*". Even though it is a subjective measure, we can identify five vocal effort levels: 1) whisper, 2) soft voice, 3) normal voice, 4) loud voice and 5) shouting. The two extremes, i.e. whispered and shouted speech, are the two vocal efforts that produce major changes in the acoustic and the general dynamic characteristics of the speech signal when comparing with normal voice [24]. These changes have proven to affect significantly the performance of automatic speech recognition and speaker recognition systems, especially if only normal speech was used during training (i.e., training/testing mismatch conditions) [12, 13, 25, 26, 27, 55]. This poses a big challenge for most recognition tasks to be developed using voice biometrics, mostly in those that account only for differences between speakers such as speaker

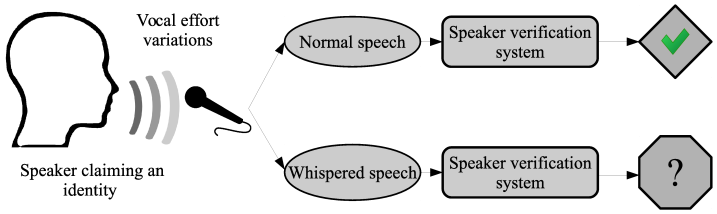


Figure 1.2 – Diagram of typical train/test mismatch issue encountered with whispered speech.

recognition tasks. For the sake of clarity, *normal voice* is also referred to as *neutral speech* in different publications [26, 37], and it refers to the case when a speaker uses a natural and comfortable level of speech in typical communication environments (e.g. a conversation, a phone call, etc.), and within this thesis we will refer constantly to it as *normal speech* (short for *normally-phonated speech*).

Among the five vocal efforts, whispered speech has gained great attention for security applications lately. Despite its somewhat reduced perceptibility and intelligibility, whispered speech is a natural mode of speech production that still conveys relevant and useful information for many applications. For example, just as normal-voiced speech, whispered speech not only conveys a message, but also traits such as identity, gender, emotional, and health states, to name a few [25, 28, 29, 30, 31, 32]. As previously mentioned, whispered speech is commonly used in public situations where private or discrete information needs to be exchanged, for example, when providing a credit card number, bank account number, or other personal information. Despite the amount of information present in whispered speech, there are certain characteristics that make this speaking style challenging when presented as a possible input to speech enabled applications. As an example, the most salient characteristic of whispered speech is the lack of vocal fold vibration. Furthermore, when a person whispers, several changes occur in the vocal tract configuration, thus altering not only the excitation source, but also the syllabic rate and the general temporal dynamics characteristics of the generated speech signal [25, 33]. Hence, it is expected that classical methods designed for normal-voiced speech characterization will fail when tested in atypical scenarios including whispered speech [12, 25, 26, 27], as illustrated by Figure 1.2.

Despite the limited research done in this field, different approaches attempting to overcome some of these disadvantages have been reported, particularly within training/test mismatch conditions where speaker models were trained with normal speech and tested with whispered speech [12, 26, 34, 35, 36, 37]. In these works, low accuracy in mismatched training/testing conditions have been reported, and some strategies to address the intrinsic limitations of current speech enabled systems

Approach	Number of speakers and gender	Baseline Accuracy	Achieved accuracy - (Relative improvement)
Modified LFCC (MLFCC) [34]	10 (Male)	48%	58% (20.8%)
MLFCC and Feature Mapping [34]	10 (Male)	48%	68% (41.6%)
AM-FM based features [12]	36 (20 Male and 16 Female)	19%	30% (57%)
Feature Mapping using CMLLR [37]	28 (Female)	79%	82% (4.1%)
Feature Mapping using ConvTran [37]	28 (Female)	79%	88% (12%)
Multi-style approaches			
Addition of 4.5 secs. of Whsp. [26]	28 (Female)	79%	91% (14%)
Addition of 15 secs. of Whsp. [35]	22 (Male and Female)	20%	90% (350%)

Table 1.1 – Comparison among different approaches reported in the literature for speaker identification using whispered speech in mismatched train/test conditions. In the Table: CMLLR - Constrained Maximum Likelihood Linear Regression, ConvTran - Convolutional Transformation, Whsp - Whispered speech,

to process whispered speech have been proposed. Representative examples include: robust features such as modified linear cepstral coefficients (LFCC) and feature mapping [34], feature warping over MFCC and score combination at the frame level [35], model adaptation schemes such as maximum linear regression (MLLR), and feature transformation from normal to whispered speech to be used during map adaptation [37]. Despite the many different efforts, relative improvements range from 8% to 46% and for all cases, achieved classification results are still not useful for practical applications [26, 37].

Table 1.1 reports details from different approaches that have been reported in the literature for speaker identification using whispered speech in mismatched train/test conditions (top) as well as two examples showing the benefits of whispered speech addition during parameter estimation (multi-style models - bottom). In the Table we have also included the number of speakers, the gender, and, when available, the specific number of male and female speakers. As can be seen, the reported accuracy for female speakers is higher than the accuracy for male speakers, which signals that gender dependencies occur with whispered speech, as has been also reported for normally-phonated speech [56, 57]. Another key aspect to notice is that regardless of the lower accuracy for male speakers, relative improvements are higher when using only male speakers, and this also happens for gender independent experiments, i.e., mixed male and female speakers. These results from previous research also illustrate how challenging the task at hand is, and given the variability observed within a speaker identification task, it is hard to predict what effects will be seen within a speaker verification task.

Typically, in speech enabled applications such as speech recognition, two main strategies are used to handle the mismatch problem, namely, (*i*) multiple model recognizer, where dedicated speaker models are obtained for different vocal efforts [27] and (*ii*) multi-style models, where each model is obtained from a combination of normal speech and small amounts of speech of varying vocal efforts [25, 27]. Notwithstanding, the two different methods were shown to have their advantages and disadvantages. For example, while both improve the performance of whispered speech [26, 27], multiple model training requires significant amounts of whispered speech data to obtain the speaker models, which can be hard to obtain in practice. Multi-style based systems, in turn, despite requiring lower amounts of whispered speech to train the models, trade gains in whispered speech to losses in normal speech accuracy, often by the same amount [27]. More recent approaches in ASR have explored model adaptation by using artificially generated whispered speech samples from transcribed normal speech recordings (pseudo-whisper data) [58, 59]. It has been shown that by using this approach it is possible to outperform an ASR system that has been directly adapted to available transcribed whispered samples. This approach, however, for speaker recognition seems to not have similar benefits. As an example, in Table 1.1 it is reported that by using a Convolutional Transformation (ConvTran) to generate pseudo-whisper data, relative improvements of 12% were obtained [37], while by using 4.5 seconds of whispered speech data per target speaker, the relative gains were 14% [26], contrary to the ASR case. Therefore, it seems that while phonetic information can be mapped from whispered speech domain to normal speech domain for speech recognition purposes, mapping identity related information is a more challenging task. Hence, regarding the specific problem we are addressing in this thesis, in the speaker recognition field, the less expensive and most effective strategy is to add small amounts of whispered speech from target speakers during enrollment [26, 35], i.e., to use multi-style models.

The multi-style approach suits better the task at hand, as current state-of-the-art SV systems require significant amounts of data for parameter estimation. More importantly, high variability between speakers in the training set is required in order to guarantee a representative sample of the universe of speakers that are expected during the testing stage. As we will describe in the following chapter, in existing publicly available databases the number of speakers with whispered speech recordings is relatively small when compared to the number of speakers with normal speech. As such, the challenge during parameter estimation is to learn as much variability as possible from both speaking styles in order to properly model discriminative information between speakers, but

with limited available resources from one vocal effort (i.e., whispered). With current state-of-the-art bottleneck features, this becomes an even more challenging problem, as it is necessary to train a large scale automatic speech recognition system and use its output to train a deep neural network [22]. To the best of our knowledge, no large scale corpus with annotated whispered speech is available to train an ASR system. As such, if bottleneck features are to be explored for the task at hand, innovative solutions are needed to minimize the effects of limited resources on the efficiency of the extracted feature vector. In summary, while multi-style models can offer a viable alternative to whispered speech based SV, adequate techniques still need to be put in place in order to overcome issues that arise from such strategy (e.g., such as those reported in [25, 27] for ASR).

With this in mind, the problems we address in this dissertation can be summarized as follows:

1. In the speaker recognition field, many advances have been reported in the literature to tackle the train/test mismatch issue. Moreover, within the more closely related field of *speaker identification* a handful of techniques have also been proposed. As we show in Chapter 3, however, the gains achieved with these tools and techniques developed for normally-phonated speech and for whispered speech SID do not necessarily translate to the whispered speech *speaker verification* task. As such, new solutions are still needed within a SV scenario.
2. Train/test mismatch problem: during the testing stage, when there is no whispered speech data available for training or to enroll target speakers, whispered speech can induce severe detrimental effects on speaker recognition performance. In fact, even if whispered speech data is available for parameter estimation, the mismatch problem can still be present as changes in the vocal effort can be viewed as "within-speaker" variation, and such variation is not well represented in the enrollment samples from target speakers.
3. Limited resources: Typical state-of-the-art systems are trained on large datasets considering thousands of speakers, several hours of recordings, a variety of channels and different recording sessions. However, a wide spectrum of vocal effort variations is still not included in large scale evaluation tasks. Standard speaker verification systems need to be compensated with techniques allowing to efficiently use the limited amount of information available from vocal efforts different to normal voiced speech, while maintaining high performance levels in normal speech.
4. Negative effects in multi-style models: In speech recognition systems negative effects were observed when combining data from different vocal efforts. More specifically, while whispered

speech recognition accuracy improved, the accuracy for normally-phonated speech decreased [25, 27]. It is not clear what the impact will be when combining data from two different vocal efforts for the task at hand. Strategies to overcome this gain-loss tradeoff with whispered-normal speech, respectively, are still drastically needed.

1.2 Thesis contributions

The aim of this thesis is to address the four abovementioned problems with particular emphasis on two vocal efforts: normal and whispered. While there has been some significant advances in the literature addressing the issue of whispered *speaker identification*, which have given important insights on how to address the mismatch problem in speaker recognition, here we focus on *speaker verification*. These two tasks are closely related, but as previously mentioned, SI is a N-class problem while SV is a two class problem. Hence, granting that previously proposed strategies for SI are also expected to contribute for SV, given the fundamental difference in the definition of the tasks we consider relevant to explore SV in more detail and propose strategies to allow a standard SV to handle inputs from the two vocal efforts and give reliable decisions. In particular, this thesis addresses several aspects of the speaker verification pipeline, including feature extraction, where innovative features are proposed containing speaker-dependent information less sensitive to normal/whisper mismatch, as well as different fusion schemes to achieve accurate multi-vocal effort recognition accuracy. More specifically, the key thesis contributions can be summarized as:

1. Performance boundaries achievable with whispered speech for speaker verification have been explored and a comprehensive review of previously reported strategies in related areas have been put in place to analyze their contribution to the speaker verification task with whispered speech. We found that whispered speech can contain as much speaker specific information as normal speech, but standard approaches designed for normal speech tend to fail for whispered speech. One of the problems limiting the widespread usage of this speaking style for authentication applications is the lack of sufficient data to train the models. In this regard, we have developed strategies that efficiently use the limited resources available, such as limited number of background speakers with whispered speech recordings to be used during parameter estimation. Experimental results show high performance levels inline with perfor-

mance obtained for normal speech. These findings have been reported in publications #2, #5, #6, and #7 listed in Section 1.3.

2. The proposal of innovative and highly informative features for improved multi-vocal speaker verification using acoustical insights and statistical analysis. More specifically, three feature extraction algorithms were proposed: *i*) Auditory-Inspired Amplitude Modulation features, and *ii*) Two variants of the classical mel-frequency cepstral coefficients (MFCC) using modified versions of the signal processing pipeline used for MFCC computation. The former case was motivated by evidence showing that the slowly varying envelope of bandpass signals convey important speaker-dependent information useful for speaker recognition tasks. We propose an approach using short time Fourier transform instead of a time domain filter bank analysis as has been done before for modulation based features. In addition to this, mutual information (MI) was used as an analysis measure to identify invariant information between amplitude modulation features of normal-voiced and whispered speech. This allows channels with low MI values to be disregarded while preserving important speaker dependent information. Features extracted using the later, in turn, were shown to extract complementary information and reduce the negative impact during testing with whispered speech. The MFCC variants were shown to be useful on their own, and information related to these features also showed to be useful during training of deep neural networks for bottleneck feature extraction. In both cases, systems trained with the the newly-proposed features outperformed systems trained with the classical MFCCs. These findings have been reported in publications #1, #3, and #4, listed in Section 1.3.
3. The proposal of a framework to explore fusion schemes at different levels, namely: *i*) Frame, *ii*) i-vector, and *iii*) Scoring level. This framework allows us to explore the complementarity of the different feature sets and to study how to use more efficiently the information encoded in the different feature representations, thus reducing the need of additional speakers with whispered recordings. By using this framework, we corroborated previously reported results showing that for normal speech, fusion at frame level is more effective [23], however for whispered speech we found that it is better to train separate systems and fuse them at higher levels, either by concatenating i-vectors or using score level fusion. These findings have been reported in publications #3, #4, #7, and #8 listed in Section 1.3.
4. In summary, improved performance was achieved for multiple vocal effort speaker verification with limited available resources from one vocal effort (i.e., whispered) using multi-style

models and fusion schemes. Fusion strategies such as i-vector concatenation and fusion at the score level of MFCC variants, bottleneck and Auditory-Inspired Amplitude Modulation features showed to be effective to address problems such as the train/test mismatch condition. For whispered speech, for example, gains as high as 62% (absolute error rate reduction from 20.83% to 7.73 %) were observed when using whispered speech recordings only during the training stage, whilst by combining normal and whispered speech recordings from target speakers during enrollment gains as high as 71% (absolute error rate reduction from 8.25% to 2.35%) could be seen relative to a baseline system based on MFCC features. For normal speech, on the other hand, these strategies not only help to reduce the negative effects observed when speech recordings from two different vocal efforts were combined during training and enrollment, but also gains as high as 79% (from EER=3.13% to 0.63%) were observed relative to the baseline system. These findings were reported in publications #3, #4, and #8 listed in Section 1.3.

1.3 List of publications

1. Whispered Speech Detection in Noise Using Auditory-Inspired Modulation Spectrum Features, Sarria-Paja, M. and Falk, T.H., *IEEE Signal Processing Letters*, 2013, Vol. 20, No. 8, pp 783-786. [60].
2. Strategies to Enhance Whispered Speech Speaker Verification: A Comparative Analysis, Sarria-Paja, M. and Falk, T.H., *Journal of the Canadian Acoustical Association*, 2015, Vol. 43, No. 4, pp. 31-45. [61].
3. Fusion of Auditory Inspired Amplitude Modulation Spectrum and Cepstral Features for Whispered and Normal Speech Speaker Verification, Sarria-Paja, M. and Falk, T.H., *Computer Speech & Language*, Accepted with minor revisions. [62].
4. Bottleneck and Amplitude Modulation Features for Improved Whispered Speech Speaker Verification in Train/Test Mismatch Conditions, Sarria-Paja, M. and Falk, T.H., *Under preparation – to be submitted to Speech communications.* [63].
5. Whispered speaker verification and gender detection using Weighted Instantaneous Frequencies, Sarria-Paja, M., Falk, T.H. and O'Shaughnessy, D., *in proceedings of ICASSP* 2013. [64]

6. The effects of whispered speech on state-of-the-art voice based biometrics systems, M. Sarria-Paja and M. Senoussaoui and T. H. Falk, *in proceedings of IEEE CCECE* 2015. [65].
7. Feature mapping, score-, and feature-level fusion for improved normal and whispered speech speaker verification, Sarria-Paja, M., Senoussaoui, M., O'Shaughnessy, D. and Falk, T.H., *in proceedings of ICASSP* 2016. [66].
8. Variants of Mel-frequency Cepstral Coefficients for Improved Whispered Speech Speaker Verification in Mismatched Conditions, Sarria-Paja, M. and Falk, T.H., submitted to EUSIPCO 2017. [67].

1.4 Organization of this dissertation

Chapter 2 provides the background on whispered speech, emphasizing the main differences with normal speech, and presents the principal insights from perceptual and acoustic research studies, together with a review of the main practical applications where whispered speech has been used. This chapter also describes the speaker verification problem, and provides a general background on the main techniques from speech processing, feature extraction and machine learning involved in the building blocks of an automatic speaker verification system, as well as a description of the speech databases used for the experiments herein. Chapter 3 explores the performance envelope achievable with whispered speech, particularly within the scope of a small scale speaker verification (SV) task. To this end, we explore the benefits of different existing preprocessing methods, frequency warping strategies, feature representations, and SV system configurations. In Chapter 4 we adopt a more realistic evaluation scheme by including additional datasets recorded in different conditions, which also increases the number of speakers. By following standard evaluation protocols for speaker verification, we make a clear distinction between background speakers and target speakers or clients. In this chapter we explore the advantages of feature mapping alongside other mismatch compensation strategies, such as fusion schemes at different levels. Chapter 5 describes three innovative feature extraction approaches aiming at extracting invariant information embedded within both speaking styles. We present evidence on how to compute perceptually-relevant variants of the standard MFCC features to extract complementary information and reduce the negative impact during testing with whispered speech. We also describe how the auditory-inspired modulation spectrum can be used to extract discriminative features. Chapter 6 explores the effects of whispered

speech in a standard speaker verification system when using the current state-of-the-art bottleneck features. In this chapter, we propose an approach to compute more informative bottleneck features useful for tasks involving short length utterances and variations in vocal effort. In addition to this, we present an approach to implement dedicated models for each vocal effort to guarantee the best approach per speaking style is used during testing stage. Finally, we explore robustness of the proposed feature sets when the testing recordings have been contaminated with environmental noise. Lastly, Chapter 7 provides a general discussion, provides the conclusions of this thesis, as well as lists potential areas of future research.

Chapter 2

Background

2.1 Whispered speech

In normal speech, air from the lungs causes the vocal folds to vibrate, exciting the resonances of the vocal tract in a particular configuration. This configuration modulates the excitation source allowing the speaker to produce a great variety of voiced sounds [68]. In whispered speech, the glottis is opened and turbulent flow created by exhaled air passing through this glottal constriction provides a source of sound [25, 33]. From the very definition of whispered speech, there is a fundamental difference with normal phonated speech: the complete lack of vibration of the vocal folds being the main physical difference. However, when a person whispers, different changes occur in the vocal tract configuration. Besides the excitation source, the syllabic rate, and the general dynamic characteristics of the speech signal also differ from those of normal speech [69]. These differences have been analyzed from a perceptual and acoustical point of view, resulting in notable cues that give us insights about what is achievable by automated systems using whispered speech.

Perceptual studies have addressed important topics such as pitch perception and the correlation between perceived pitch and formant location, as well as the measurement of the formant shifts towards higher frequencies [70, 71]. For automatic speech recognition systems, subjective intelligibility tests would be necessary to analyze the possible losses and predict the performance of an automated system designed to recognize whispered speech. For instance, consonant discrimination and identifiability of vowels has been studied and whispered speech was shown to be highly intelli-

gible but still poorer than normally-phonated speech [29, 30]. These studies have also shown that whispered speech, despite its reduced perceptibility, conveys relevant speaker identity and gender information [28, 30, 72], which makes it feasible to also develop speaker recognition systems.

Acoustic studies have corroborated and complemented perceptual findings. For instance, whispered speech has a lower and flatter power spectral density [25]. In [33], it was found that the duration of consonants in whispered speech is prolonged by about 10% relative to normally-voiced speech. In addition to the duration increase, the intensity of the whispered consonants is lower by about 12 dB. Perceptual findings regarding the formant shifts in whispered mode were also corroborated in [73]. In addition to this, in [74] a statistical analysis was carried out to compare several acoustic and visual features between the two speaking styles. In total, 64 acoustic features referred as *low level descriptors* (LLDs) grouped in three categories, i.e., spectral LLDs, prosody LLDs and voice quality LLDs were compared and 56 showed to be statistically different. But not everything regarding normal-voiced and whispered speech is different; for example while it has been documented that characteristics of vowels and voiced consonants are significantly different, unvoiced consonants are relatively similar [33]. The above-mentioned insights have been used by the research community to tackle different challenges, such as reconstruction of normal speech from whispers [75, 76, 77], speech recognition [25, 27], and speaker identification [12, 26, 35, 37] with whispered speech.

To illustrate some of the significant differences between normal and whispered speech, their waveforms and spectrograms are depicted by Figure 2.1 (normal: left; whisper: right), for the utterance "*Here I was in Miami and Illinois*", these speech recordings correspond to a male speaker and were extracted from the CHAINS speech corpus (see Section 2.3). From Figure 2.1 (right), it can be observed that whispered speech is mostly turbulent noise modulated by the vocal tract with no clear structure. With normal speech (left), on the other hand, the glottal excitation is clear. Moreover, the time waveform for whispered speech is significantly lower in amplitude; in this particular case about 15 dB lower. Figure 2.2(a) in turn, illustrates the average power spectrum for the same utterance, using 25 ms windows and a 12th order linear predictive model to estimate the spectral envelope. From Figure 2.2(a), it is evident that the differences lie mostly in the low frequencies. For normal speech, most of the energy is concentrated below 1 kHz, whereas for whispered speech it is concentrated below 500 Hz, with frequency shifts in the spectral peaks and valleys. Between 1 kHz and 4 kHz the two spectral envelopes follow a similar trend, where spectral

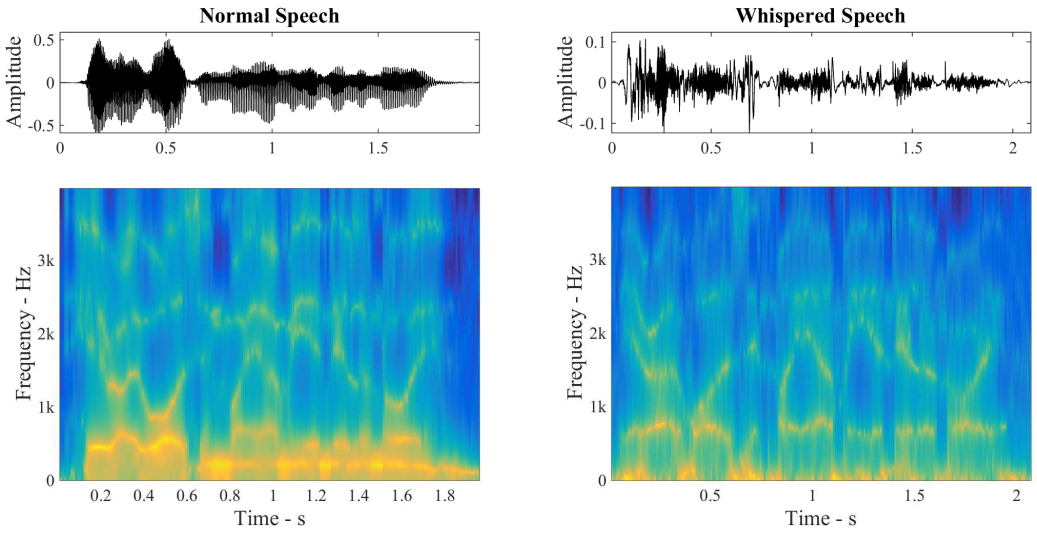
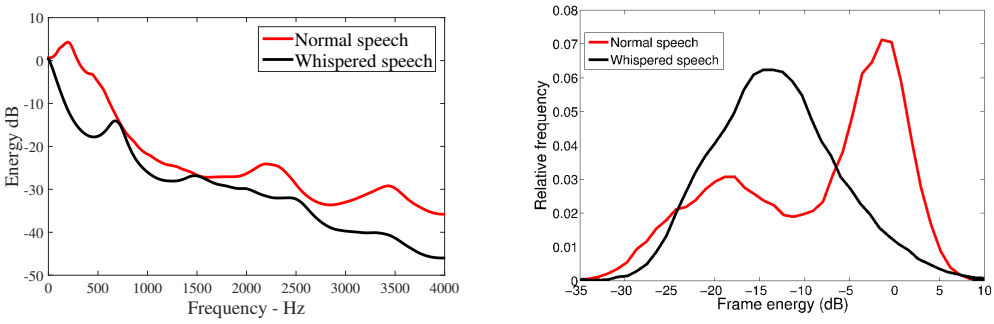


Figure 2.1 – Comparison of waveform and spectrogram of the speech signal "*Here I was in Miami and Illinois*" from the same speaker in: normal (left) and whispered (right) speech mode. Speech recordings were extracted from the CHAINS speech corpus (see Section 2.3)

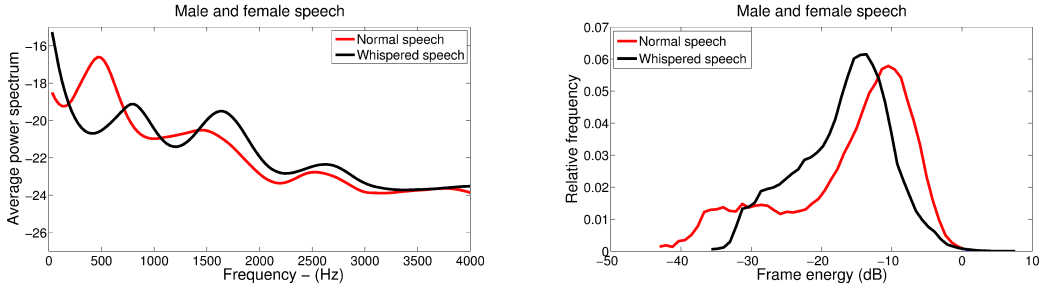
peaks and valleys are located in approximately the same frequency values, however the differences in magnitude are not constant. Regarding frame energy distribution, the histogram in Figure 2.2(b) was computed using male and female speech and utterances of about 55 s from 36 speakers and shows that the concentration of high-energy frames is higher for normal speech, with 60% of the frames having energy between -10 dB and 10 dB. For whispered speech, on the other hand, 70% of the frames have energy between -35 dB and -10 dB. Combined, these findings show that significant differences exist between whispered and normal-voiced speech in terms of temporal, spectral and energy dynamics. As such, it is expected that any speech-based technology trained on normal speech will perform poorly when tested on whispered speech unless clever strategies are put in place.

Differences related to frame energy distribution and average power spectrum can be reduced by using two pre-processing steps. First, each speech recording is normalized in amplitude, then pre-emphasized using a first order finite impulse response filter with constant a; this reduces the dynamic range of the speech spectrum and helps to model formants of differing intensity equally well. A typical value for the constant is $a = 0.97$. To illustrate the effects of pre-emphasizing and normalizing the speech recording, Figure 2.3(a) and 2.3(b) depict the average spectrum and frame energy distribution, respectively, of amplitude-normalized and pre-emphasized recordings using male and female speech. As can be seen, the gap between the two speaking styles seen in Figure 2.2 has been greatly diminished, although most of the differences remain below 1.2 kHz.



(a) (b)

Figure 2.2 – Plots of average power spectrum and frame energy distribution. (a) average power spectrum comparison of the utterance "*Here I was in Miami and Illinois*" spoken by same speaker and (b) frame energy distribution for normal and whispered speech using combined male and female data across 36 speakers.



(a) (b)

Figure 2.3 – Plots of (a) average power spectrum and (b) frame energy distribution after preprocessing for normal and whispered speech (averaged over 36 speakers).

Despite many potential applications where whispered speech can be used to increase privacy and improve identity management, most recent studies have argued that speaker dependent whispered training data is generally not available in real-world scenarios [26, 34, 37]. Notwithstanding, the interest in this speaking style is rising and now different researchers are collecting data and making it available for research, as is the case of the CHAINS [12] and the whispered TIMIT (wTIMIT) [39] databases. Availability of both, normal and whispered speech, allows techniques and strategies in small scale experiments to be explored, thus providing insights about the performance and what is achievable with whispered speech for future large scale applications.

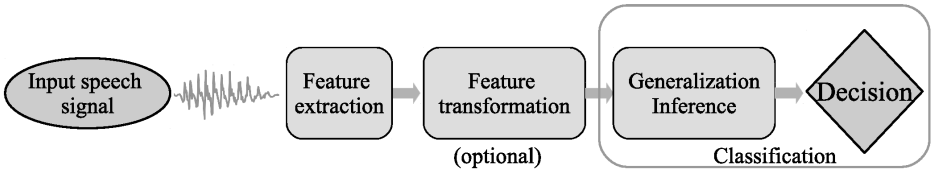


Figure 2.4 – Building blocks for a general purpose pattern recognition system that can be applied to speaker verification.

2.2 Automatic speaker recognition

Humans perform fairly well at identifying people based on their voice. This fact has lead to the idea that within the speech signal there are some biometric cues much like the fingerprint. After almost 30 years of research in this area, today it is possible to implement digital speech processing systems that can perform, with high reliability, tasks related to automatic speaker recognition [14, 40], thus replacing the human listener with a machine. Applications with voice biometrics are burgeoning as a secure method of authentication, which eliminates the common use of personal identification numbers, passwords, and security questions.

Figure 2.4 depicts a diagram with the basic building blocks for a general purpose pattern recognition system that can be applied to speaker verification, namely feature extraction, feature transformation (optional) and classification. More details of each block are given in the subsections below.

2.2.1 Feature extraction

Speech is produced from a time varying vocal tract system, which makes speech signals dynamic or time-varying in nature. Even though the speaker has control over many aspects of speech production, e.g., loudness, voicing, fundamental frequency and vocal tract configuration, much speech variation is not under speaker control and is random, e.g., vocal fold vibration is not truly periodic. These random variations also make speech sound more natural, thus do not affect intelligibility [68]. Most of the developments and tools from applied mathematics to study systems and signals, however, assume time invariant systems and time invariant excitations, i.e. stationary signals. Since speech is inherently an non-stationary signal, in order to be able to use these analysis tools, short

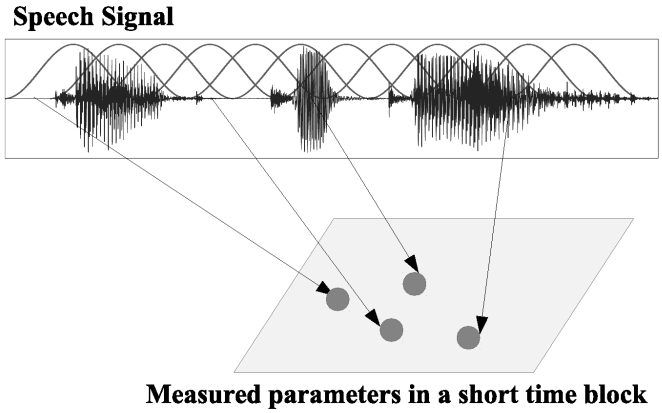


Figure 2.5 – Speech analysis over short time duration blocks to estimate parameters of interest such as formant location or energy.

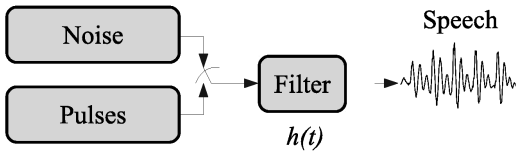


Figure 2.6 – Block diagram of the source filter model of speech production.

time duration blocks need to be used, as depicted by Figure 2.5. Such "short time" processing methods, as they are known, can be performed either in time or in frequency domain [68, 78].

Typical methods of feature extraction for speech enabled applications are based on the conventional model of speech production, the source-filter model. In this model it is possible to split the speech signal in two components, *i*) an excitation signal (known as the residual) that can be visualised as the combination of two different signal generators, one for voiced-speech and another for voiceless (noise-like) speech, and *ii*) a transfer function which models the vocal tract configuration and shapes the spectral envelope of the resulting speech [68], as depicted by Figure 2.6, where $h(t)$ represents the impulse response of the transfer function modelling the vocal tract configuration. The human auditory system appears to pay much more attention to spectral aspects of speech (e.g., amplitude distribution in frequency) than to phase or timing aspects. Thus, spectral or frequency analysis methods have been the preferred approaches to estimate most parameters from speech [68]. Next, we describe in more detail some of the typical feature extraction methods that will be explored within this research.

Mel Frequency Cepstral Coefficients

According to psychophysical studies, human perception of the frequency content of sounds can be characterized by what is known as *critical bands*. A critical band defines a frequency range in psychoacoustic experiments for which perception abruptly changes as a narrowband sound stimulus is modified to have frequency components beyond the band, and the frequency distribution of these critical bands follows a subjectively defined nonlinear scale [68]. Although many analytical expressions have been proposed to describe this nonlinear scale, over the years the *mel* scale, originally proposed in [79], has been the most widely used for speech characterization. The mapping from acoustical frequency to perceptual frequency resolution is approximately linear in frequency up to 1 kHz and logarithmic at higher frequencies [68], and is commonly defined as:

$$f_{mel} = 2595 \log_{10}(1 + \frac{f}{700}), \tag{2.1}$$

where f is the acoustic frequency and f_{mel} is the resulting mel scale frequency warping.

The most popular analysis method for automatic speech recognition combines cepstral analysis theory [80] with aspects related to the human auditory system [68]. The so-called mel-frequency cepstral coefficients (MFCC) are the classical frame based feature extraction method widely used in speech applications. Originally proposed for speech recognition, MFCC also became a standard for many speech enabled applications including speaker recognition. One of the reasons for widespread usage of MFCCs is that they provide an alternative and efficient representation for speech spectra which incorporates some aspects of audition. Still, some authors argue that MFCCs are suboptimal, and except for the first two coefficients, it is difficult to relate MFCC to any clear aspects of speech production or perception [68, 81].

For MFCC computation, each speech recording is pre-emphasized and windowed in overlapped frames of length τ using a Hamming window to smooth the discontinuities at the edges of the segmented speech frame. Let $x(n)$ represent a frame of speech that is pre-emphasized and Hamming-windowed. First, $x(n)$ is converted to the frequency domain by an N point discrete Fourier transform (DFT) and the resulting energy spectrum can be written as $|X(k)|^2$, with $1 \leq k \leq N$. Next, P triangular bandpass filters spaced according to the mel scale are imposed on the spectrum. These filters do not filter time domain signals, they instead apply a weighted sum across the frequency

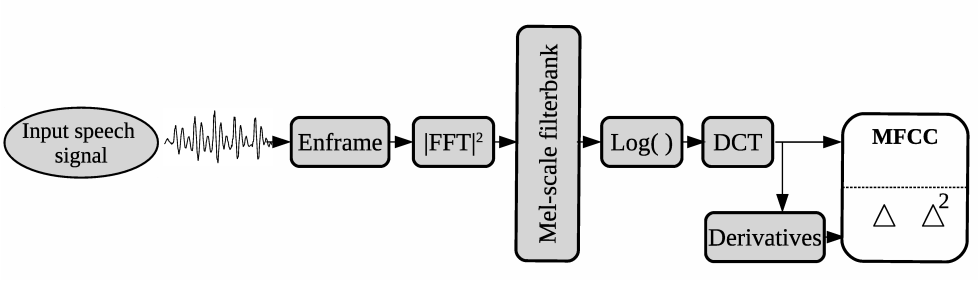


Figure 2.7 – General scheme of MFCC, Δ and Δ^2 computation.

indexes k, which allows to group the energy of frequency bands into a single value, resulting in P energy values $E(l)$ with $1 \leq l \leq P$. Finally, a discrete cosine transform (DCT) is applied to the log filter bank energies and the final MFCC coefficients can be written as:

$$MFCC_m = \sqrt{\frac{2}{P}} \sum_{l=0}^{P-1} \log\left[E(l+1)\right] \cos\left[m\left(\frac{2l-1}{2}\right)\frac{\pi}{P}\right], \tag{2.2}$$

where $0 \leq m \leq R-1$, and R is the desired number of cepstral coefficients. The elements of these coefficients are highly correlated. The DCT has the effect of decorrelating these elements which allows the use of diagonal covariance matrices in subsequent statistical modelling steps.

The temporal changes in adjacent frames play a significant role in human perception. To capture this dynamic information in the speech, first- and second-order difference features (Δ and ΔΔ MFCC) can be appended to the static MFCC feature vector. Dynamic or transitional features are computed by means of an anti-symmetric Finite Impulse Response (FIR) filter with an odd number of coefficients (e.g., five or nine) to avoid phase distortion of the temporal sequence. Figure 2.7 depicts a diagram with the basic building blocks for MFCC computation. Further details can be found in [82].

Alternatively, different frequency warping strategies have been proposed and can be used in lieu of the classical mel scale. These frequency warpings allow greater resolution to be placed at certain frequency ranges. Commonly used scales include: linear, exponential [34, 36] and the whisper sensitive scale (WSS) [83]. Table 2.1 shows the mappings between the original (f) and warped ($\hat{f}$) frequencies previously mentioned. Previous studies using the exponential and linear scales showed that relative improvements of around 20% could be achieved; however, for further improvements some knowledge about the speaking style was needed for testing [34, 36]. Furthermore,

Scale	**Frequency warping**
Linear	$\hat{f} = f$
Exp.	$\hat{f} = 10610 \times (10^{f/50000} - 1)$
WSS	$\hat{f} = \begin{cases} \frac{2475 f^4}{1220^4 + f^4}, & 0 < f < 2000 \\ 4100 - \frac{2000}{1+e^{(f-300)/310}}, & 2000 \leq f < 4000 \end{cases}$

Table 2.1 – List of frequency warping strategies used in the experiments. Cepstral coefficients derived are LFCC (linear), EFCC (exponential - Exp. in the table) and WSSCC (WSS).

the improvements were shown only for the whispered speech speaker identification task, thus there is no evidence about the effects of this front-end in the speaker verification task.

Bottleneck features - BNF

Before describing in detail this approach for feature extraction, some general background in neural network notation is presented. Given an input vector o, a neural network performs a sequence of N non-linear operations that can be expressed as follows [42]:

$$G(o) = \tilde{g}\left(W^{(N)} \dots g\left(W^{(2)} g\left(W^{(1)} o\right)\right)\right), \tag{2.3}$$

where $W^{(i)}$ denotes the weight matrix of i-th layer, $g(\cdot)$ is a non-linear operation denoting the activation function for the hidden layers, typically a sigmoid or hyperbolic tangent function, $\tilde{g}(\cdot)$ is the output activation function, which usually is a linear or identity function, but this depends on the specific task the neural network is being used for. The parameters are estimated to optimize a cost function which is also related to the task at hand. Typically, the mean square error is used for regression problems and the cross-entropy function for classification tasks [21, 84].

Most recent feature extraction techniques have replaced the classical MFCC as acoustic features by approaches based on deep learning to extract the so-called bottleneck features. Bottleneck Neural-Networks are deep neural networks (DNN) with a particular topology, where one of the hidden layers has significantly lower dimensionality than the surrounding layers; such layer is known as the bottleneck layer. A bottleneck feature (BNF) vector is obtained by forwarding a primary input feature vector through the DNN and reading off the vector of values at the bottleneck layer [21]. For speaker recognition purposes, the DNN for feature extraction needs to be trained first for

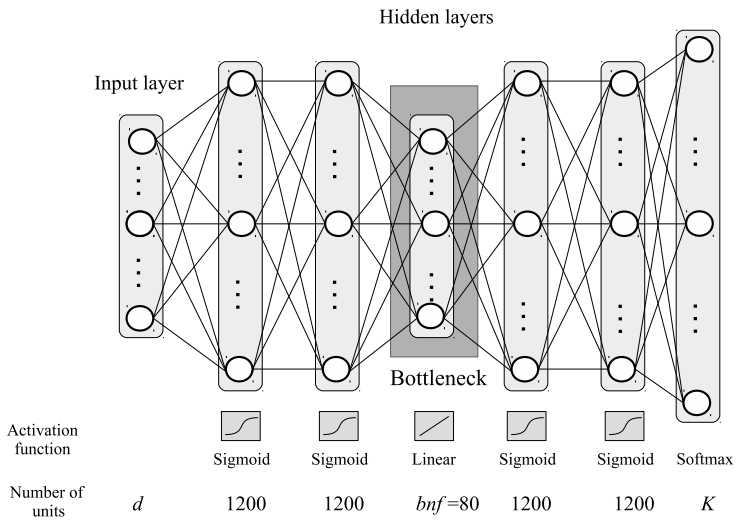


Figure 2.8 – Bottleneck Neural Network architecture used in this work.

a specific frame-by-frame classification task. According to previous reports, excellent results were observed with features extracted using DNN trained for a phone-like classification, specifically where the targets are sub-phonetic units known as "senones", by minimizing the cross-entropy function [22].

The typical configuration of a DNN used for BNF extraction is depicted by Figure 2.8, where d is the dimensionality of the input feature vector and K is the number of target labels defined by the output transcription file given by an automatic speech recognition (ASR) system. In this case the number of units in the bottleneck layer or dimensionality of the feature vector was fixed to $bnf = 80$, all displayed values are typical in the normal speech speaker verification literature [21, 22, 23].

Alternate feature representations

Different types of low-level features have been proposed in the speaker recognition area with the motivation to improve the performance of MFCC baseline systems under noisy/reverberant conditions, or to provide complementary information to MFCCs [85]. Some of these features are extracted from slowly varying subband envelopes, as an example, features derived from the AM-FM [86] signal representation have proven to be more robust in noisy conditions and perform at the same level as cepstral coefficients in clean conditions [12, 87]. The main difference is that cepstral coefficients are

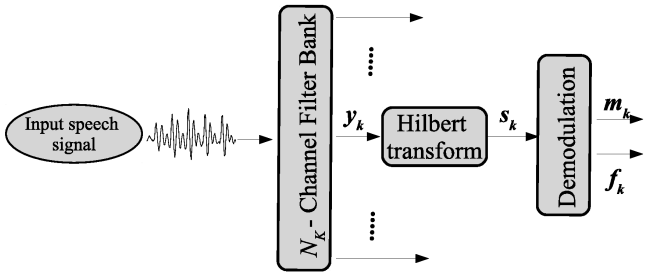


Figure 2.9 – AM-FM signal representation. Block diagram to decompose the speech signal in bandpass channels and compute the low frequency modulator and the instantaneous frequency per channel.

based on power spectrum estimation (i.e., frequency domain) whilst features derived from the AM-FM signal representation are computed in the time domain. More specifically, the AM-FM model decomposes the speech signal into bandpass channels and characterizes each channel in terms of its envelope and phase (instantaneous frequency) [12, 88]. The speech signal $s(n)$ is filtered through a bank of N_K filters, resulting in the bandpass signal $y_k(n) = s(n) * h_k(n)$, where $h_k(n)$ corresponds to the impulse response of the k-th filter. After filtering, each analytic sub-band signal $s_k(n)$ is uniquely related to a real–valued bandpass signal $y_k(n)$ by the relation:

$$s_k(n) = y_k(n) + j \cdot \hat{y}_k(n), \tag{2.4}$$

where $\hat{y}_k(n)$ stands for Hilbert transform of $y_k(n)$. There are two approaches to decompose each analytic signal in terms of its envelope and phase: *i*) the Hilbert envelope approach (non–coherent demodulation) and *ii*) coherent demodulation [88]. The main difference between these two approaches is in the allocation of phase between the envelope and carrier. Whereas the Hilbert envelope places all of the sub-band phase in the carrier, coherent demodulation makes the important distinction between carrier and modulator phase. For the sake of notation, let $m_k(n)$ denote the low–frequency modulator and $f_k(n)$ the instantaneous frequency for each bandpass signal. Figure 2.9 depicts the general process to decompose the speech signal into bandpass channels and their respective modulator and instantaneous frequencies when using the Hilbert envelope approach.

Here, two features are explored based on the AM-FM signal decomposition. The first is the so called Weighted Instantaneous Frequencies (WIF). These features are computed by combining the values of $m_k(n)$ and $f_k(n)$ using a short-time approach [12]:

$$F_k = \frac{\sum_{i=n_0}^{n_0+\tau} f_k(i) \cdot m_k^2(i)}{\sum_{i=n_0}^{n_0+\tau} m_k^2(i)}, \quad k = 1, \ldots, N_K, \tag{2.5}$$

where τ, as before, represents the length of the time frame. F_k is calculated over the full length of each $m_k(n)$ with increments of $\tau/2$.

The second feature set is the mean Hilbert envelope coefficients (MHEC) proposed in [87] and shown to perform better than traditional MFCC features under noisy conditions for normal speech for speaker verification. In this case, the envelope $m_k(n)$ is blocked into frames and the mean Hilbert envelope for a specific frame in channel k is calculated as:

$$E_k = \frac{\log\left(\frac{1}{\tau}\sum_{i=n_0}^{n_0+\tau} w(i-n_0+1) \cdot m_k(i)\right)}{\bar{E_k}}, \quad k = 1, \ldots, N_K, \tag{2.6}$$

where $w(n)$ is a Hamming window of length τ, and the term $\bar{E_k}$ represents the long-term average in each channel which normalizes the values of E_k. Finally, for a specific frame and using all 23 E_k values, a DCT is applied to produce the MHEC features [87].

These two feature sets, WIF and MHEC, are just some examples of modulation based features. Other authors have proposed related features, including the Medium Duration sub-band Speech Amplitudes (MMeDuSA) [89], nonlinear Teager energy operator (TEO) derived features [90], and Gammatone Filterbank (GFBs) Energies [91], to name a few. After pilot experiments, WIF and MHEC were the feature sets that showed best performance for the task at hand. Hence, experiments in Chapter 3 are presented using these two feature sets.

2.2.2 Feature transformation

This stage relies on a suitable change (simplification or enrichment) of a representation, e.g. by a reduction of the number of features, relations or primitives describing objects, or some non-linear transformation of the features, to enhance the class or cluster descriptions [84]. Typically, for speech processing applications, such transformation are designed to mitigate the effects of linear channel mismatch, and to add robustness to the overall system. Feature normalization techniques such as Short-Time Mean and Variance Normalization (STMVN), Short-Time Mean and Scale Normalization (STMSN) and Short-Time Gaussianization (STG) techniques are a standard pre-processing procedure in the state-of-the-art speaker verification systems [92]. But it is not only limited to normalization algorithms, if we assume the output of the triangular filterbanks in the MFCC pipeline to be the features, then the DCT can be seen as a feature transformation process, which reduces the dimensionality and decorrelates the variables, thus resulting in a more compact and informative feature vector. Techniques such as principal component analysis (PCA) or linear discriminant analysis (LDA) are commonly used for more general classification tasks [84]. Current state-of-the-art SV systems use latent variable inspired approaches to map typical variable length frame based representation to a fixed dimensional feature vector. We will describe such methods later in Section 2.2.3, as different modeling techniques needed for transformation should first be introduced.

Moreover, in some specific tasks it is necessary to perform more elaborate transformations, for example to map from a feature space to a different one. This is particularly useful when there is a mismatch between the training data and what the model encounters in real life. The model is trained on a specific source distribution, but during testing, it receives data from a different, target distribution. Two typical techniques in speech applications to address this problem have been used: the first one relies on Gaussian mixture models based regression and the second on neural networks [13, 41, 42]. The use of these techniques in the context of whispered speech speaker verification will be described in detail in Chapter 4.

2.2.3 Classification - generalization and inference

Once we have a set of features or parameters to describe the speech recordings, another important stage, when implementing a speaker recognition system, is the generalization/inference stage. In

this stage, a classifier/identifier is trained. The training process involves the parameter tuning of models to describe training samples, i.e., features extracted from speech recordings. The learning process requires assumptions on the general form of model or the classifier, and use the training samples to estimate the unknown parameters of the model. Then an algorithm is applied in order to reduce the error on a set of training data or in general terms, optimize a cost function related to the task at hand [93]. In this regard, for speaker modelling, the well known Gaussian mixture model has succeeded to remain in the scope of speaker recognition research for many years. It is considered an extension of an earlier approach known as vector quantization that allows the modelling of probability density functions by the distribution of prototype vectors, i.e., a feature vector is assigned to the nearest prototype vector (cluster). In Gaussian mixture models, the clusters are overlapping, and each cluster is described by a single Gaussian density function [40]. The use of Gaussian mixture models for speaker recognition is motivated by their capability to model arbitrary densities, and the individual components of a model are interpreted as broad acoustic classes [40, 94].

Adapted Gaussian Mixture Models

Herein, we describe the most popular approach based on Gaussian mixture models (GMM), which for many years was the dominant approach for text-independent speaker verification: Adapted Gaussian Mixture Models using Maximum a Posteriori (MAP) adaptation. First of all, we will describe the generalities of a GMM model and how, using training samples, the parameters are tuned.

A GMM is composed of a finite mixture of multivariate Gaussian components and the set of parameters denoted by λ. It is characterized by a weighted linear combination of C unimodal Gaussian densities by the function:

$$p\left(o\,|\lambda\right) = \sum_{i=1}^{C} \alpha_i \mathcal{N}\left(o, \mu_i, \mathbf{\Sigma}_i\right), \tag{2.7}$$

where o is a D-dimensional observation or feature vector, α_i is the mixing weight (prior probability) of the i-th Gaussian component, and $\mathcal{N}(\cdot)$ is the D-variate Gaussian density function with mean vector μ_i and covariance matrix $\mathbf{\Sigma}_i$.

Let $\mathcal{O} = \{o_1, \cdots, o_K\}$ be a training sample with K observations. Training a GMM consists of estimating the parameters $\lambda = \{\alpha_i, \mu_i, \mathbf{\Sigma}_i\}_{i=1}^{C}$ to fit the training sample $\mathcal{O}$ while optimizing a cost function. The typical approach is to optimize the average log-likelihood (LL) of $\mathcal{O}$ with respect to the model λ and is defined as [84]:

$$LL\left(\mathcal{O}, \lambda\right) = \log p\left(\mathcal{O}|\lambda\right) = \frac{1}{K}\sum_{k=1}^{K}\log\sum_{i=1}^{C}\alpha_i\mathcal{N}\left(o_k, \mu_i, \mathbf{\Sigma}_i\right). \tag{2.8}$$

The higher the value of LL, the higher the indication that the training sample observations originate from the model λ. Although gradient-based techniques are feasible, the popular expectation-maximization (EM) algorithm is used for maximizing the likelihood with respect to a given data. The interested reader is referred to [84] for more complete details.

For speaker recognition applications, first a speaker-independent *world model* or *universal background model* (UBM) is trained using several speech recordings gathered from several speakers. Regarding the training data for the UBM, selected speech recordings should reflect the expected alternative speech to be encountered during recognition. This applies to both the type and the quality of speech, as well as the composition of speakers. Next, the speaker models are derived by updating the parameters in the UBM using a form of Bayesian adaptation [15]. In this way, the model parameters are not estimated from scratch, with prior knowledge from the training data being used instead. It is possible to adapt all the parameters, or only some of them from the background model. For instance, adapting the means only has been found to work well in practice [15, 40].

Consider an enrollment sample $\mathcal{O}_j$, with K observations from a new speaker. The first step to obtain a GMM model for this speaker is to compute the sufficient statistics for the weight, mean, and variance parameters using the UBM and the sample $\mathcal{O}_j$. Then, these sufficient statistics are used to update the old UBM parameters. For instance, the adapted mean vector for the i-th mixture component is given by:

$$\hat{\mu}_i = \kappa_i\mathbb{E}_i\left(o_k\right) + (1 - \kappa_i)\mu_i, \tag{2.9}$$

where

$$\mathbb{E}_i\left(o_k\right) = \frac{1}{N_i}\sum_{k=1}^{K}p(i|o_k)o_k, \tag{2.10}$$

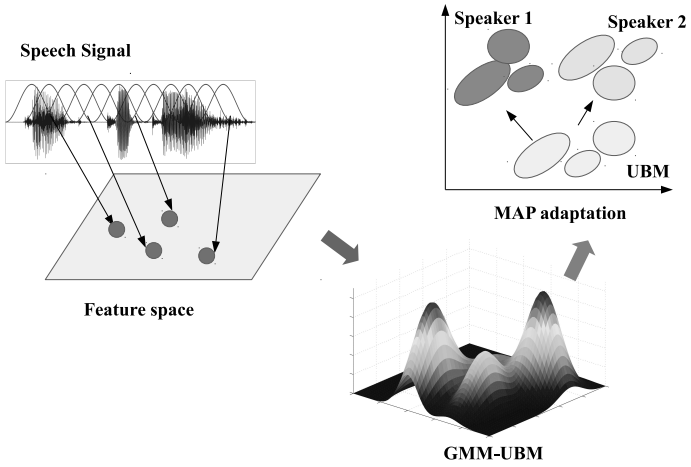


Figure 2.10 – General scheme of MAP adaptation using target speakers enrollment data.

$$N_i = \sum_{k=1}^{K} p(i|o_k), \tag{2.11}$$

$$p(i|o_k) = \frac{\alpha_i \mathcal{N}(o_k, \mu_i, \mathbf{\Sigma}_i)}{\sum_{j=1}^{C} \alpha_j \mathcal{N}(o_k, \mu_j, \mathbf{\Sigma}_j)}, \tag{2.12}$$

$$\kappa_i = \frac{N_i}{N_i + \rho}. \tag{2.13}$$

The relevance parameter ρ controls the effect of the sample $\mathcal{O}_j$ on the resulting model with respect to the UBM. And the new resulting model is denoted as λ_{target}. Figure 2.10 illustrates how using MAP adaptation, the Gaussian components of the universal background model are adapted using target speakers enrollment data. Equation 2.12 represents the probabilistic alignment of the enrollment sample $\mathcal{O}_j$ into the UBM mixture components (posteriors). Equation 2.11 represents the *zero*-order statistics, and it accumulates the probability of observations o_k being generated by the i-mixture component. Finally, Equation 2.10 represents the *first*-order statistics, are the weighted sum of the means per a component [15].

During testing, in a verification scenario, we consider a testing sample $\mathcal{O}_t$ and a hypothesized speaker with model λ_{hyp}, the task of the speaker verification system is to determine if $\mathcal{O}_t$ matches the speaker model. There are two possible hypotheses: 1) $\mathcal{O}_t$ is from the hypothesized speaker

and 2) $\mathcal{O}_t$ is not from the hypothesized speaker. The decision can be made by computing the log-likelihood (score) between the two hypotheses, which is given by $s = LL(\mathcal{O}_t, \lambda_{hyp}) - LL(\mathcal{O}_t, \lambda_{UBM})$ [15]. If s is greater than a decision threshold then hypothesis 1) is accepted otherwise the hypothesis 2) is accepted.

i-vectors/PLDA approach

Current state-of-the-art speaker recognition systems are based on identity vectors (*i-vectors*) extraction [19] and matching between a test utterance and a target speaker is done using either a fast scoring method based on cosine distance between i-vectors or probabilistic linear discriminant analysis (PLDA) [20] based scoring. Next, we describe the most common approach, the i-vectors/PLDA approach. *i-vectors* extraction can be considered as a feature transformation stage, as depicted by Figure 2.4, prior to the generalization/inference block.

The i-vectors extraction technique was proposed to map a variable length frame based representation of an input speech recording to a small-dimensional feature vector while retaining most relevant speaker information. First, a C-Component GMM is trained as an universal background model (UBM) using the Expectation – Maximization (EM) algorithm and the data available from all speakers from the train set or background data, as described in the previous section. Speaker and session-dependent supervectors of concatenated GMM means are modeled as:

$$M = m + T\phi, \tag{2.14}$$

where m is the speaker- and channel-independent supervector, $T \in \mathbb{R}^{CF \times D}$ is a rectangular matrix of low rank covering the important variability (total variability matrix) in the supervector space. C, F and D represent, respectively, the number of Gaussians in the UBM, the dimension of the acoustic feature vector and the dimension of the total variability space. Finally $\phi \in \mathbb{R}^{D \times 1}$ is a random vector with density $\mathcal{N}(0, I)$ and referred to as the identity vector or *i-vector* [19]. A typical i-vector extractor can be expressed as a function of the zero- and first-order statistics (Equations 2.11 and 2.10) generated using the GMM-UBM model, and it estimates the Maximum a Posteriori (MAP) point estimate of the variable ϕ. This procedure is complemented with some post-processing techniques such as linear discriminant analysis (LDA), whitening, and length normalization [95].

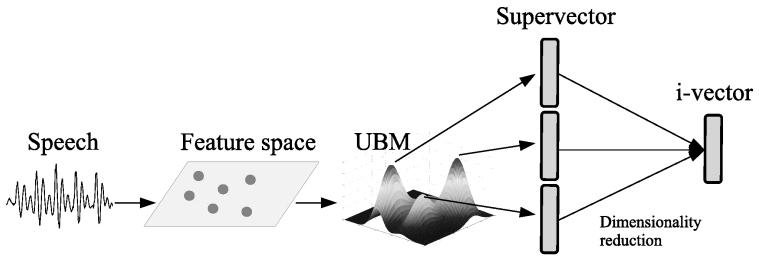


Figure 2.11 – i-vector extraction from a speech recording.

These techniques can be used to remove nuisance effects in the total variability space. For the experiments herein, an i-vector is computed per enrollment utterance and then they were averaged to obtain a single i-vector per target speaker. The interested reader is referred to [19, 95] for more complete details. Figure 2.11 depicts a diagram with the basic steps for i-vector computation.

The cosine distance is a fast and efficient method of scoring which eliminates the need of enrolling and model parameter estimation and is commonly used with i-vectors. Given a target or hypothesized speaker (ϕ_{hyp}) and the test (ϕ_{test}) feature vectors, the cosine distance is given by:

$$s = \frac{\langle \phi_{hyp}, \phi_{test} \rangle}{(\|\phi_{hyp}\| \, \|\phi_{test}\|)}, \tag{2.15}$$

where $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$ correspond to dot product and magnitude respectively. Finally, a decision is made based on thresholding [19].

The PLDA model [20, 96], on the other hand, splits the total data variability into within-individual and between-individual variabilities, both residing in small-dimensional subspaces. Originally introduced for face recognition, PLDA has become a standard in speaker recognition. PLDA was formulated in [96] as:

$$\phi_{ij} = \mu + V y_i + U x_{ij} + \varepsilon_{ij}, \tag{2.16}$$

where ϕ_{ij} is the i-th feature vector associated to the j-th speaker, the matrices $V \in \mathbb{R}^{D \times P}$ and $U \in \mathbb{R}^{D \times M}$ span the between- and within- individual spaces, μ is a global mean, $y_i \sim \mathcal{N}(0, I)$ and $x_{ij} \sim \mathcal{N}(0, I)$ are hidden variables in the spaces spaned by V and U, respectively, and the residual $\varepsilon_{ij} \sim \mathcal{N}(0, \Sigma)$ is defined to be Gaussian with zero mean and diagonal covariance Σ. In a verification scenario, there are two possible hypotheses: 1) ϕ_{test} and ϕ_{enrol} share the

same class, and 2) ϕ_{test} and ϕ_{enrol} are from different classes. Lastly, the corresponding score can be obtained by computing the log-likelihood between the two hypotheses, which is given by $s = \ln(P(\phi_{test}, \phi_{enrol})) - \ln(P(\phi_{test})P(\phi_{enrol}))$; details can be found in [20, 96]. For the experiments herein, the dimensionality of matrices V and U were set to $P = dim_{LDA}$ and $M = 0$, where dim_{LDA} represents the dimensionality of the LDA model, which is tuned accordingly per feature set.

2.2.4 Variants of the general structure of the pattern recognition system

Features described in Section 2.2.1 place emphasis on different aspects of the signal (e.g., temporal, spectral, phase), thus likely contain complementary information. This hypothesis has motivated the exploration of fusion at different levels to combine the strengths of feature representations extracting complementary information [14, 97]. Referring to the general scheme illustrated by Figure 2.4, it is possible to perform fusion at two different levels: *i*) Fusion at the input, i.e., at the feature level, where different feature representations can be concatenated in order to obtain and enriched representation, and *ii*) Fusion at the output, i.e., at the score level, where the outputs of systems trained on different feature representation are combined in a new feature vector and feed to a new model for decision making. Notwithstanding, the overall general scheme still follows the same structure as depicted by Figure 2.4.

Recent literature on SI and ASR has recommended the use of speaking-style dependent models [25, 27, 35], as depicted by Figure 2.12. The method builds on the previously described general pattern recognition system and takes into account the different subclasses that can be modelled in order to build a complete automated system. This scheme is useful for gender- or speaking style dependent systems and is commonly known as multiple model recognizer. This approach has shown to improve the performance of whispered speech recognition [25, 27], and has been used in speaker verification tasks for gender dependent models [15]. Multiple model training, however, requires significant amounts of representative data per sub-class to properly estimate models, which can be hard to obtain in practice for some applications, as is the case for whispered speech.

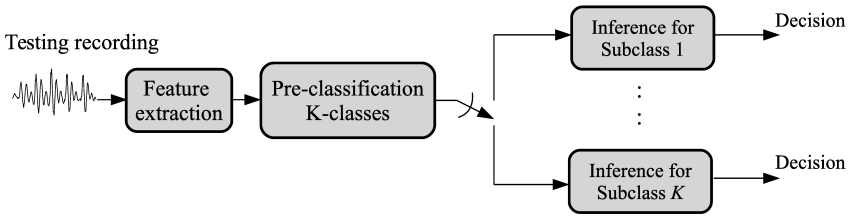


Figure 2.12 – Multimodel framework for automatic classification using a K-class model selector

2.3 Speech databases

Publicly available speech corpora containing normal and whispered speech are not common. In fact, several have been reported in the literature but have not been made available to the public, such as the UT-Vocal Effort I and II datasets [98]. In this section we describe the publicly available speech databases used the experiments herein for speaker verification purposes, namely:

TIMIT: This database is largely known in speech processing related fields. TIMIT contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. It comprises 6300 speech recordings (approximately five hours), recorded using 16 bits precision at 16 kHz. Even though the TIMIT corpus of read speech has been designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems, given the relatively large number of speakers, it is reasonable to be used also for speaker recognition applications [38].

wTIMIT: The whispered TIMIT (wTIMIT) corpus is designed for the study and construction of large vocabulary speech recognizers. This corpus contains recordings of 48 speakers, where each speaker utters 450 phonetically balanced sentences of the TIMIT prompt set in both normal-voiced and whispered speech, recorded using 16 bits precision at 16 kHz. The corpus has two accents (Singaporean-English, and North American English), with roughly 20 to 28 speakers from each accent group. This speech corpus is gender balanced [39].

CHAINS: the **Characterizing Individual Speakers** speech corpus contains the recordings of 36 speakers obtained in two different sessions with a time separation of about two months, there are three different accents: 28 speakers from Ireland (16 male), 5 speakers from the USA (2 male) and 3 speakers from the United Kingdom (2 male). Additional details about the database can be found in [99]. Speech stimuli was generated under six speaking conditions, namely solo (natural rate

Database	Num. of speakers		recordings/speaker	
	Female	Male	Norm.	Whsp.
TIMIT	192	438	10	–
wTIMIT	24	24	450	450
CHAINS	16	20	37	37

Table 2.2 – Details about the three databases used in our experiments.

reading), retelling without time constraints, two-person synchronous reading, repetitive synchronous imitation, accelerated-rate reading, and whispered. All recordings are available as 16 bit PCM encoded files with a sampling rate of 44.1 kHz [99].

As can be seen, the TIMIT database contains a large number of speech recordings from different speakers only in normal speech mode, while the CHAINS and wTIMIT databases contain normal and whispered speech. Table 2.2 presents details about the number of speakers and recordings per speaker available in the datasets.

LibriSpeech: This is a Large-scale corpus of read English speech, and contains approximately 1000 hours of speech derived from read audiobooks from the LibriVox project [49]. The speech is recorded using a sampling rate of 16 kHz using 16 bits precision. The data has been carefully segmented and aligned which makes it a suitable database for training ASR systems. In total, training data corresponds to 400 hours of continuous speech. The aim of including this dataset is solely to train the BNF extractor system which comprises two stages, first training an ASR system which generates the target labels or "senones", next, training the bottleneck neural network using as input acoustic features extracted from the training speech recordings and as targets the senones generated by the ASR system.

2.4 Summary

This chapter has presented a general overview of whispered speech. We have presented the main insights from perceptual and acoustic studies which also helps to illustrate some of the main differences with normally-phonated speech. In addition to this, we have presented a brief description of the building blocks needed to implement a standard speaker verification system. In the next chapter, we will explore what is achievable for whispered speech speaker verification using tech-

niques described above, within an ideal experimental setup. Ultimately, by performing experiments in such limited experimental conditions will help us to understand the performance envelope of existing solutions and guide where significant efforts need to be placed.

Chapter 3

Comparative Analysis for Normal and Whispered Speech Speaker Verification

3.1 Preamble

Results presented in this chapter were published in publications #2 and #5 listed in Section 1.3 [61, 64]. The overarching goal of this chapter is to explore the performance envelope achievable with whispered speech, particularly within the scope of a small scale speaker verification (SV) task, thus guiding the research directions of subsequent chapters for larger-scale applications. To this end, we explore the benefits of different existing preprocessing methods, frequency warping strategies, feature representations, and SV system configurations.

3.2 Introduction

In the past, whispered speech has only been explored within the SI problem [12, 26, 34, 35, 36, 37], where the use of the accuracy metric does not give a clear picture of the actual impact of mismatch conditions between training and testing [14]. In addition, it is not clear whether the strategies proposed for SI systems can also be useful for SV systems. Currently, state-of-the-art SV

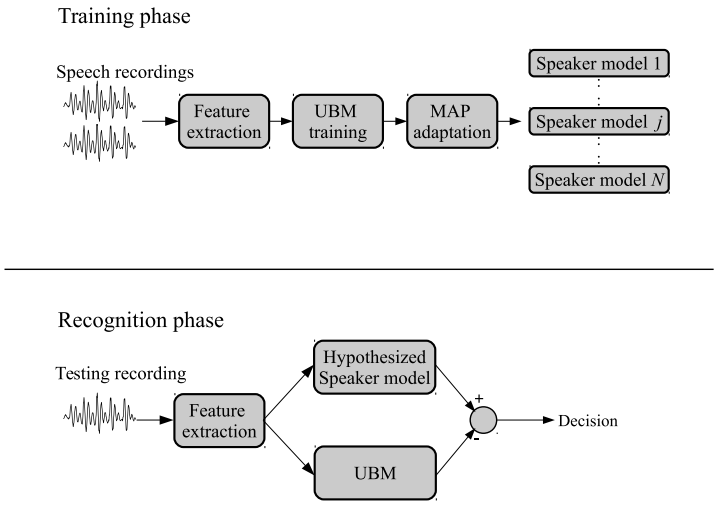


Figure 3.1 – Block diagram of a general SV system. Top and bottom diagrams represent the training and testing stages, respectively, for a GMM-UBM SV based system

systems based on normal speech use highly elaborate techniques, such as i-vectors [19]. However, to properly train such systems, large amounts of training data are required [40, 100]. Reliable training of an i-vector extractor requires datasets with large number of speakers as well as samples per speaker, which is not the case for the experimental setup we want to explore in this chapter. Furthermore, these methods are heavily dependent of the data, i.e., the nature of the testing data should be the same with the one the i-vector extractor was trained on [51]. According to our experiments, a classification system based on Gaussian mixture models (GMM) and maximum a posteriori (MAP) adaptation, as depicted by Figure 3.1, was more suitable for the small amount of speakers and recordings we have dedicated for these preliminary experiments and dealing with mismatched scenarios. For the described system, the widely-used mel-frequency cepstral coefficients (MFCC) are used to implement a text–independent SV system [15, 40]. First an C-Component GMM is trained as an universal background model (UBM) using the Expectation – Maximization (EM) algorithm and the training data available from all speakers. Then, a GMM for each speaker is obtained using MAP adaptation, as depicted by top half diagram in Figure 3.1. During the recognition phase (bottom half of Figure 3.1), the hypothesized speaker model is scored against the UBM and a decision is made based on thresholding as described in Section 2.2.3. More details can be found in [15].

3.3 Baseline performance characterization in matched and mismatched conditions

For the experiments described in this chapter only the CHAINS speech corpus was used. A complete description can be found in Section 2.3 and in [12]. In particular, two speaking styles were used - solo and whispered - where the same text was read in both conditions. We used the speech stimuli generated from reading the paragraph of the *Cinderella story* (average duration: 55 seconds, minimum duration: 48 seconds) for training, and kept the stimuli generated from reading the *Rainbow Text* (average duration: 30 seconds; minimum duration: 23 seconds) segmented in short sentences of 3 seconds, plus 32 individual sentences (nine selected from the CSLU Speaker Identification corpus and 23 from the TIMIT corpus) for testing. Data was originally recorded at 44.1 kHz sample rate but downsampled to 8 kHz. The sampling rate is motivated by results reported in [12], where using the same dataset in a speaker identification task they found that MFCC computed on the acoustic band from 0 to 4 kHz where more robust to the mismatched train/test condition.

Prior to feature extraction, and motivated by Figure 2.3, in our experiments we normalized the speech data to -26 dBov (dB overload) using the ITU-T P.56 speech voltmeter [101], and pre-emphasized using a first order FIR filter with constant $a = 0.97$. Then 19 MFCC were computed on a per-window basis excluding the 0–th order cepstral coefficient, using a 32 ms window with 50% overlap and 24 triangular bandpass filters. Delta coefficients were also included to convey temporal dynamics information. Delta coefficients were computed by means of an anti-symmetric Finite Impulse Response (FIR) filter of length nine to avoid phase distortion of the temporal sequence. For all experiments herein, the training data was fixed to 35 seconds per speaker, and the number of Gaussian components per model was fixed to $C = 32$, showing a tradeoff between performance and computational burden. Parameters such as window length and overlap, and number of cepstral coefficients are motivated by previous research works in speaker identification [26, 36, 34] using similar configurations.

Table 3.1 reports the Equal Error Rate (EER %) obtained with the baseline system under different *train/test* conditions. In the table, 'c' stands for cepstral coefficients and 'Δ' for delta coefficients. As can be seen, for normal speech in the *normal/normal* (train/test) matched condition inclusion of delta coefficients did not provide any advantage over using only MFCCs. In fact, in the *normal/whisper* and *whisper/whisper* scenarios, inclusion of delta parameters had a negative

Training	Testing	EER(%)	
		c	$c+\Delta$
Normal	Normal	**2.13**	2.33
Normal	Whisper	**35.75**	38.62
Whisper	Normal	29.81	28.18
Whisper	Whisper	2.90	3.12

Table 3.1 – EER(%) comparison for different *training/testing* conditions after power normalization and pre-emphasis. Results in bold represent the baseline systems with which the tested improvements will be gauged against.

impact on system performance, as previously reported by [26] for a speaker identification task. Only in the mismatch *whisper/normal* condition, was an improvement in EER with the inclusion of Δ parameters observed. Differences, however, between the two experimental conditions were modest and we can not considerate this as a significant advantage or disadvantage and draw definitive conclusions on the basis of these results. Typically, for large scale speaker verification evaluations, first- and second-order time derivative estimates are included as it is considered that they still convey useful speaker-specific information [22, 15, 40].

Overall, it can be seen that significant performance degradation occurs in the mismatch conditions. When testing with whispered speech, the obtained EER for the mismatch condition was more than 10 times greater than in the matched condition. Moreover, a gap of approximately 6 – 9% can be seen in mismatched cases, depending on what speaking style is used for training. As can be seen, lower EER is achieved when training with whispered speech and testing with normal. This was expected, as in our dataset, approximately 70/30% of the normal-speech training data was comprised of voiced/unvoiced speech segments. When training with normal speech, it is likely the GMMs became biased towards voiced characteristics which are not present in whispered speech. On the other hand, when training with whispered speech, the GMMs could more accurately represent unvoiced normal-speech segments, as only small differences have been observed between unvoiced consonants in whispered and normal speech modes [33]. To better illustrate this point, Figure 3.2 shows the plots of the scores distribution for target speakers and impostors under the two training conditions. Continuous lines represent the speaking style used for training (i.e., normal speech in subplot (a) and whispered speech in subplot (b)).

Similar experiments were performed by Xing Fan [26] for SID. Since the results reported in [26] relied on only a subset of 28 females as target speakers, direct comparisons cannot be made, but the obtained trends can be compared. For example, in both cases, higher performance levels were

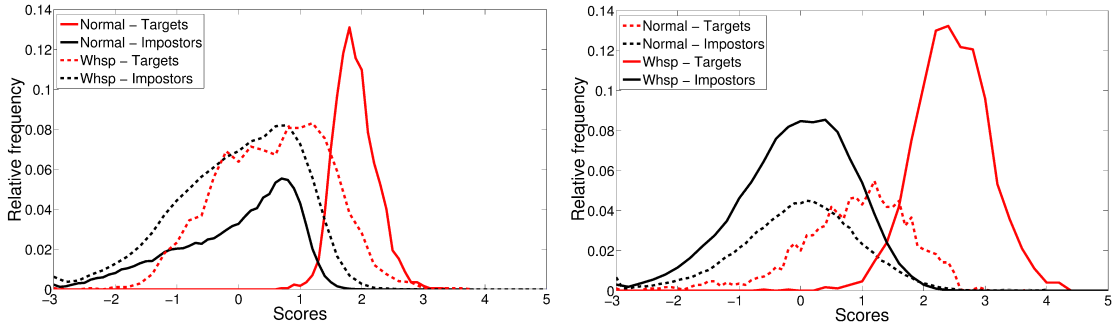


Figure 3.2 – Plots of score distributions for target and impostor speakers using normal and whispered speech files. The scores were computed using two different systems, the system in (a) was trained only with normal speech and the system in (b) was trained only with whispered speech. Continuous lines are representative of the speaking style used for training.

achieved in the train/test matched conditions, with *normal/normal* outperforming *whisper/whisper*. In the mismatch conditions, however, [26] reported speaker identification accuracies of almost 80% in the *normal/whisper* condition, but of around 10% with *whisper/normal*. In our case, we obtained opposite trends and showed that for a speaker verification task involving both male and female speakers, the *whisper/normal* condition resulted in slightly lower error rates than training with normal speech and testing with whispers. Such differences motivate further analysis to truly gauge the benefits that previously-proposed methods for whispered SID may have on whisper SV.

Figure 3.2(a) shows that by using normal speech for training the scores of normal speech are less scattered than those for whispered speech, which, in turn, show a high degree of overlap. Figure 3.2(b), on the other hand, shows the scores obtained when training only with whispered speech. As can be seen, scores from whispered speech testing recordings are still more scattered than those for normal speech, but the overlap has been reduced. Overall, as expected the matched *normal/normal* scenario resulted in the lowest EER. Together these findings suggest that alternate strategies are needed to improve the performance of SV systems based on whispered speech, particularly in mismatched cases. This is the focus of the sections to follow.

A final and important aspect is the difference between reported accuracy for a speaker identification system using the same dataset and a similar experimental setup, and the error rates reported herein for speaker verification. While accuracy using MFCC computed in the 0-4kHz frequency band for normal speech was lower than 80% (error rate higher than 20%) according to results reported in [12], in our experiments the error rate is less than 3%. In mismatched condition, on

the other hand, the reported accuracy when testing with whispered speech was around 20% (80% error) in [12] and in our case the error rate is around 38%. As can be seen, speaker verification and speaker identification are related areas within the scope of speaker recognition, but they are not directly comparable and we do not expect that techniques that have shown to work in speaker identification to present similar results or be equally effective in speaker verification. Hence the need for the comparative analysis presented in this chapter.

3.4 Comparative analysis using different system configurations

Using the same settings as before, 19 cepstral coefficients were computed using the frequency warping strategies described in Table 2.1, along with the mel scale and the delta coefficients. More specifically, cepstral coefficients derived are MFCC (mel), EFCC (exponential), WSSCC (WSS), and LFCC (linear). This experiment allows us to determine which frequency warping strategy can better reduce the negative impact of train/test mismatch. Additionally, to mitigate the effects of linear channel mismatch, a widely accepted method is called *feature warping*, which maps the distribution of the cepstral features to a normal distribution ($\mathcal{N}(0,1)$) by using a 3-second sliding window, also known as short-time Gaussianization (STG) [102]. For the sake of comparison, the different feature sets are evaluated in the two possible scenarios: with and without STG.

Results are shown in Table 3.2 where two *training/testing* conditions are evaluated, namely *normal/normal* and *normal/whisper* (represented in the table as N/N and N/W, respectively). Whilst the negative impact of mismatch is still evident, all frequency warping strategies have improved the MFCC performance. As an example, by using the whisper sensitive scale and appending delta coefficients it is possible to reduce the EER by approximately 13% relative to the baseline in mismatch condition without using feature warping. Furthermore, STG can result in additional improvements in the mismatch condition, leading to improvements up to 31% relative to the baseline. Notwithstanding, one disadvantage of frequency and feature warping is the drop in performance obtained in the matched N/N condition. For example, with MFCCs the EER doubles after STG. The other frequency warping strategies, on the other hand, resulted in more stable results after STG. As before, no significant advantages were observed by appending the delta coefficients.

Cepstral Coefficients	without STG				with STG			
	c		$c+\Delta$		c		$c+\Delta$	
	N/N	N/W	N/N	N/W	N/N	N/W	N/N	N/W
MFCC	**2.13**	35.75	2.33	38.62	5.08	32.23	4.78	35.23
LFCC	4.88	31.04	4.60	30.20	4.17	**24.33**	5.20	25.82
EFCC	5.09	31.36	5.21	30.10	4.18	24.57	5.26	25.64
WSSCC	6.01	31.02	6.21	29.08	6.17	25.70	7.50	27.26

Table 3.2 – EER(%) comparison for matched and mismatched *training/testing* condition, using different frequency warping strategies and comparing the effects of using STG as feature warping. N/N and N/W correspond to training with normal speech and testing with normal or whispered speech, respectively. All feature representations where computed from the full 0 to 4 kHz band. EER values in bold highlight the best performance achieved in matched and mismatched conditions.

3.4.1 Frequency sub-band analysis

Results presented in Tables 3.1 and 3.2 suggest that whispered speech conveys information highly related to each speaker, but significant differences are still present between the two speaking styles. Motivated by the results in Figure 2.3(a), we also explore the use of only a sub-band of the speech signal in which their difference is minimized. According to Figure 2.3(a), this sub-band ranges from approximately 1.2 kHz to 4 kHz. As such, the frequency-warpings are calculated between 1.2 and 4 kHz. This frequency band comprises mostly information from the second and third formants (F2 and F3). EER performance results are shown in Table 3.3. As observed, further gains are obtained in the mismatch condition, but at the cost of reduced performance in the matched scenario. Notwithstanding, these findings corroborate previously-reported cues showing a significant amount of speaker-specific information in the second and third formants [103, 104]. An additional advantage of focusing within this sub-band is that for whispered speech, shifts in F2 of 2 - 24% and in F3 of 1 - 10% have been observed relative to normal-voiced speech [73]. This is a rather low variation when compared with the shift for F1 that can be 50% or higher [73]. The most relevant improvement in mismatch condition is achieved using MFCC; when comparing with the results in Table 3.2, a relative reduction in the error rate of approximately 38% is achieved using STG and without appending delta coefficients. It is important to emphasize that in the matched condition the error rate is three times higher than that reported in Table 3.2. Together, these results show the high relevance of speaker identity information contained below 1.2 kHz, particularly for normal speech.

Cepstral Coefficients	without STG				with STG			
	c		$c+\Delta$		c		$c+\Delta$	
	N/N	N/W	N/N	N/W	N/N	N/W	N/N	N/W
MFCC	8.64	26.50	9.02	26.82	**7.14**	**21.81**	9.20	24.51
LFCC	9.58	27.54	9.53	25.96	7.44	21.81	9.62	22.89
EFCC	9.39	27.18	9.45	26.24	7.74	22.47	9.38	23.43
WSSCC	8.36	27.75	8.85	26.93	8.89	24.87	11.62	25.58

Table 3.3 – EER(%) comparison for matched and mismatched *training/testing* condition using the sub-band from 1.2 kHz to 4 kHz to compute the different feature sets with different frequency warping strategies and comparing the effects of using STG as feature warping. N/N and N/W correspond to training with normal speech and testing with normal or whispered speech, respectively. EER values in bold highlight the best performance achieved in matched and mismatched conditions.

3.4.2 Alternate feature representations

Features described in Section 2.2.1 were used for the following experiments. First, as there are different approaches for filter design that have been used in speech applications, for the experiments herein, two approaches were tested: a gammatone filterbank [105], and the Gabor filterbank [12], each with 23 channels. Filter center frequencies range from 50 Hz to 3528 Hz and their bandwidths are characterized by the mel frequency scale. Originally, it was proposed to use a 80-channel Gabor filterbank [12], however, according to our experiments it is not necessary to have such a high resolution and for extracting speaker dependent information a $N_K = 23$-channel filterbank suffices.

Second, since there are two approaches to decompose each analytic signal in terms of its envelope and phase, i.e., *i*) the Hilbert envelope approach (non–coherent demodulation) and *ii*) coherent demodulation, in a pilot experiment we explored the performance of different features in the matched testing condition (i.e., train/test on whispered speech). As a result, features that were computed using the the Hilbert envelope approach achieved performance inline with those obtained with the classical MFCC features. As such, for feature extraction purposes, only the Hilbert envelope approach is used for the following experiments. Detailed results appeared in publication # 5 listed in Section 1.3 [64].

Table 3.4 reports the EER obtained with the different filterbank characterizations, considering both the full band and the limited sub-band (1.2–4 kHz) components. In the matched condition, MHEC and WIF perform better than cepstral coefficients without STG and at the same level using STG. However, in mismatched condition both WIF [12] and MHEC [87] achieve error rates similar to the ones achieved with cepstral coefficients. These results suggest that the information present in

	Filter Bank	EER–Full band		EER–limited band	
		N/N	N/W	N/N	N/W
WIF	Gammatone	**1.63**	33.73	5.87	24.63
	Gammatone + STG	4.48	29.48	7.86	23.19
	Gabor	2.18	35.65	6.53	24.27
	Gabor + STG	4.17	30.92	7.99	**22.77**
MHEC	Gammatone	2.06	42.24	9.80	26.72
	Gammatone + STG	5.51	41.34	10.71	28.78
	Gabor	**1.57**	36.73	9.13	**26.24**
	Gabor + STG	4.23	34.09	11.62	26.78

Table 3.4 – EER(%) comparison for matched and mismatched *training/testing* conditions, using features derived from the AM-FM signal representation. Limited band corresponds to 1.2–4 kHz. Norm/Norm and Norm/Whsp correspond to training with normal speech and testing with normal or whispered speech, respectively. For each feature representation (WIF and MHEC) EER values in bold highlight the best performance per train/test condition.

the slowly varying envelope of the bandpass signals is highly discriminative, but extremely sensitive to changes in the vocal effort. By limiting the analysis frequency band to 1.2–4 kHz, a significant reduction of approximately 36% could be achieved relative to the baseline system in mismatched condition (see Table 3.1). This, however came at a severe penalty for the matched scenario, as was similarly observed with the cepstral coefficients (see Table 3.3).

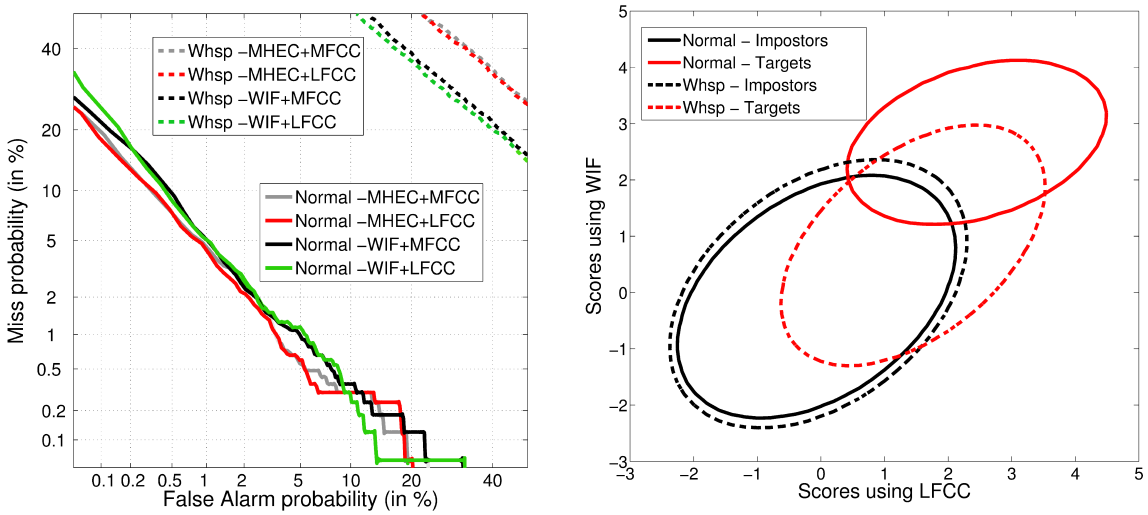
3.4.3 Feature combination

As described in Section 2.2, different feature representations can extract complementary information, and one way to combine the strengths of these features is by combining them at the frame level. For this experiment, and based on the results presented in Table 3.3, the mel and linear scales were selected to compute the cepstral coefficients in the 1.2–4kHz sub-band with STG. Moreover, motivated by results in Table 3.4, the WIF features using the Gammatone filter bank and the MHEC features using the Gabor filter bank were selected as they showed to be more effective in the matched condition without STG.

Results for feature combination are shown in Figure 3.3(a) and Table 3.5. Figure 3.3(a) depicts the Detection Error Tradeoff (DET) curves comparing different feature combinations and testing with normal and whispered speech. In the table, the features labeled in the columns are combined with the features labelled in the rows to produce a new feature space and the EER corresponding to each testing condition is presented in the respective intersection. According to these results, feature combination does not help to obtain further reductions of the EER in mismatch condition

Cepstral Coefficients	WIF		MHEC	
	N/N	N/W	N/N	N/W
MFCC	2.17	29.35	2.29	36.96
LFCC	2.29	**28.16**	**2.05**	36.60

Table 3.5 – EER(%) comparison with different feature combination, where the best features from Tables 3.3 and 3.4 were selected. EER values in bold represent the best performance per train/test condition.



(a) (b)

Figure 3.3 – Plots of (a) DET curves for feature combination and (b) contours of an estimated Gaussian distribution for the scores of testing utterances. These Plots were obtained by using only normal speech for training and normal and whispered speech for testing.

(N/W). Notwithstanding, combining WIF and LFCC and comparing the results with the baseline system, this combination can help to maintain the performance inline with the baseline system for the match condition, whilst achieving relative reduction of the EER in the mismatch condition by approximately 21%. To extend the analysis, the scores of target speakers and impostors were calculated separately using WIF and LFCC. These scores were used to estimate the parameters of a 2 dimensional full covariance Normal distribution. The contours of the distributions are depicted by Figure 3.3 (b) with continuous lines for normal speech and dashed lines for whispered speech. As can be seen, the overlap between target speakers and impostors for normal speech is minimum, however for whispered speech the scores are more scattered and higher overlap exists. As such, any decision boundary minimizing the error rate for normal speech will not necessarily be optimal for whispered speech. Such findings suggest the need for speaking-style dependent models, as will be described in Section 3.4.5.

3.4.4 Training with combined *normal/whisper* data

Results presented so far have shown that reliable performance can be achieved in matched conditions, but significant drop in performance occurs in mismatched conditions. As an alternate solution, here we explore the use of both normal and whispered speech during training and model adaptation as has been done in previous studies for SI [26, 35]. This allows speaker-specific information represented in whispered speech features to be properly modeled. Since whispered speech training data can be sparse, it is not clear how much whispered speech material is necessary to achieve acceptable performance levels for practical applications. In order to be able to perform a comparison with the baseline system, we investigate the effects of adding small amounts of whispered speech to the training set, using a MFCC–GMM system (without delta coefficients). Experiments were conducted using a fixed duration length of normal speech (35 seconds per speaker) and different duration lengths of whispered speech for training.

Results of these experiments are illustrated in Figure 3.4 and Table 3.6. As can be seen, there is significant improvement by adding as little as 5 seconds of whispered speech per speaker relative to the mismatch performance reported in Table 3.1. By gradually increasing the duration length of whispered speech, the performance of the system also gradually improves, thus corroborating previous speaker identification findings [26, 35]. Nevertheless, using the same amount of data (35 s) for both vocal efforts shows that improved performance is still achieved with normal speech with respect to whispered speech (11% lower EER). In addition, it is necessary to pay attention to the slight losses induced by the addition of whispered speech, which slightly increases the EER for normal speech. For example, using only normal speech for training, an EER of 2.13 % was reported in Table 3.1. Here, in the case of using the same amount of data for both vocal efforts, an EER of 3.05 % (i.e., 43% higher) was found. According to these results, for a practical SV verification task, improved performance can be achieved for whispered test speech, but at the cost of lower performance for normal test speech. These results agree with previously reported research work for speech recognition, where multi-style models can offer a feasible alternative to tackle the mismatch problem [25, 27].

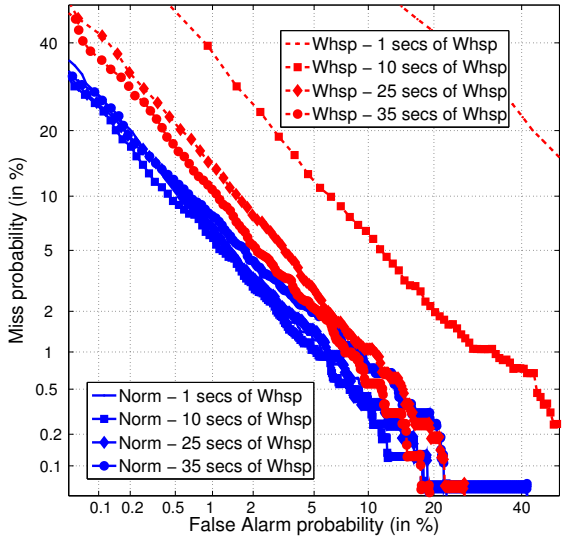


Figure 3.4 – DET curves exploring the effects of adding different amounts of whispered speech to the 35 s of normal speech during the training phase.

Amount of whispered training data (s)	EER(%)	
	Normal	Whispered
1	2.54	30.97
5	2.53	13.25
10	2.49	7.91
15	2.60	5.47
20	2.62	4.24
25	2.66	3.94
30	2.63	3.52
35	3.05	3.45

Table 3.6 – Effects of adding different amounts of whispered speech to the normal speech training set.

3.4.5 Speaking–style dependent SV systems

Up to now speaking-style *independent* SV systems have been described to handle both vocal efforts. In this section, two classes are investigated: normal and whispered modes. In order to develop a speaking-style dependent SV system, a classification stage is needed in order to detect specific speaking styles. With speaking style dependent systems, the concept of "mismatch" shifts from one of *train/test* mismatch to one of errors in speaking style classification. In order to analyze the benefits of having dedicated speaker models for each speaking style, this first set of experiments will assume an "oracle" system in which perfect *normal/whisper* classification is achieved. Within this scenario, we are particularly interested in the performance obtained with the whispered test speech

Cepstral coefficients	EER(%)	
	c	$c+\Delta$
MFCC	2.90	3.12
LFCC	2.90	3.08
EFCC	3.12	4.15
WSSCC	4.22	6.02

Table 3.7 – EER(%) comparison in W/W condition using speaking style dependent models. Results are for whispered test files and using different warping strategies to compute cepstral coefficients.

Filter Bank	AM-FM features	
	WIF	MHEC
Gammatone	**2.55**	3.10
Gabor	2.62	**2.60**

Table 3.8 – EER(%) comparison in W/W condition using speaking style dependent models. Results are for whispered test files and using AM-FM based features. Highlighted results are the best EER values per feature representation.

files. Tables 3.7 and 3.8 show the EER comparison for different frequency warpings and AM-FM feature representations, respectively. As can be seen from Table 3.7, inclusion of delta coefficients degrades performance of the system. Overall, the Linear-Frequency Cepstral Coefficients (LFCC) and MFCC showed to be the two sets of feature vectors that can achieve the lowest error rates, outperforming the WSS scale, which was developed specifically for whispered speech [83]. From Table 3.8, in turn, it can be seen that the AM-FM based features provide a modest improvement over the cepstral-based features. When using the gammatone filterbank, WIF features outperformed the MHEC ones. The opposite behaviour was observed with the Gabor filter bank. In both cases (cepstral and AM-FM based features), the EER results obtained with whispered test speech files are slightly higher than those obtained with the normal-voiced files in Table 3.2, where an EER of 2.13% was reported with MFCCs.

Subsequently, feature combination was explored. Motivated by the results presented in Tables 3.7 and 3.8, the mel and linear scales were chosen to compute the MFCC and LFCC features, respectively. The gammatone filterbank was used to compute the WIF features and the Gabor filterbank to compute the MHEC features. Since the inclusion of delta coefficients did not present any advantage for the considered feature sets, they were not included in this feature combination analysis. Results are shown in the Table 3.9. According to these results, significant improvements can be achieved by combining features, thus corroborating their complementarity. A relative reduc-

Cepstral Coefficients	AM-FM features	
	WIF	MHEC
MFCC	1.79	2.03
LFCC	1.91	1.85

Table 3.9 – EER(%) comparison in W/W condition with different feature combination, where the best features from Tables 3.7 and 3.8 were selected.

tion of the EER of approximately 33% can be seen when comparing the best results from Tables 3.7 and 3.8, and outperforming those for normal speech reported in Table 3.1.

These experiments show that whispered speech carry important speaker dependent information, and by using the adequate feature representations it is possible to achieve high performance in speaker verification tasks. As an example, by comparing results presented in Table 3.1 in the *Whispered/Whisper* condition with best results presented in Table 3.9, a relative EER reduction of 38% can be achieved by combining MFCC and WIF. It is important to emphasize, however, that dedicated whispered speaker models for large-scale applications will be more challenging to be developed, thus limiting the potential applications of speaking-style aware solutions.

3.5 Discussion

There is evidence based on subjective studies suggesting that invariant speaker identity across different vocal efforts exists [30], i.e., a listener can recognize a speaker without training, using only the experience with normally voiced speech of the same speaker. Despite different strategies, such as frequency warping, preprocessing, and alternate feature representations, our results suggest that the invariant information between normal and whispered speech is not sufficient to achieve reliable performance in an SV task for *both* speaking styles. A compromise must be kept in order to guarantee system performance in normal and whispered speech. Notwithstanding, for most of the cases evaluated herein, improvements in the mismatched condition were accompanied with reduced performance in the matched scenario. Moreover, the strategies that performed better for normal speech did not exhibit the same benefits for whispered speech. This makes it difficult to find a speaker feature representation that stores speaker identity information invariant across both vocal efforts. More research is needed to find vocal effort invariant features.

Frequency warping strategies, in the matched condition for whispered speech showed interesting results. Simple approaches such as mel and linear scales showed to outperform the WSS scale, which was designed specifically for whispered speech. This WSS scale divides the frequencies into several critical bands from 0 Hz to 4 kHz giving more emphasis to the frequencies where the resonance peaks of F1 and F3 are located. We found that the only advantage given by this strategy is an error rate reduction in the mismatched condition. While the mel scale places emphasis on lower frequencies around F1 and F2, WSS can better handle the mismatch condition due to the lower variation of the third formant between normal and whispered speech relative to F1 and F2 [73].

According to our results, two techniques have shown promising results. First, feature combination, or fusion at the input level, helps to maintain the performance for normal speech inline with the baseline system, whilst achieving gains in the relative EER reduction for the mismatched condition. Second, multi-style modeling is the most effective way to actually bring down error rates for whispered speech to comparable levels with normal speech. This however, is not enough for practical applications, as according to our results negative effects were observed for normal speech speaker verification when combining data from both speaking styles. Furthermore, a gap in performance is still expected when comparing the two speaking styles, thus signaling the need of additional strategies to compensate the expected losses and close the gap in performance. Experiments with multi-style models, however, showed promising results.

3.6 Conclusions

In this chapter, the speaker verification (SV) task based on whispered speech recordings was addressed. More specifically, the performance bounds of a standard GMM–UBM SV system were obtained using several strategies, such as frequency warping, sub-band analysis, alternate feature representations, feature combination, as well as class-dependent modeling (i.e., speaking-style). Our experimental evaluation shows that mismatch *train/test* conditions can highly affect the performance of a SV system, independent of the feature representation used. As in previous studies in adjacent areas, it was shown that in order for a SV system to handle both normal and whispered speech for practical applications, speaker model training had to involve data of both vocal efforts. Such approach, however, resulted in poorer verification performance for normal speech. Overall, feature representations evaluated here have been mainly proposed for normal-voiced speech appli-

cations, thus suggesting that alternate feature representations, tuned for whispered speech speaker verification, are still needed.

Experiments in the subsequent chapter will focus on multi-style models and fusion schemes in a more realistic scenario. This will be done by including additional speakers from different datasets and using more recent proposed approaches for speaker verification. More specifically, by following results presented in this chapter, we will focus on specific feature representations such as the classical MFCC and WIF as they showed a good tradeoff between performance in matched and mismatched conditions for both speaking styles.

Chapter 4

Feature Mapping and Fusion Schemes

4.1 Preamble

Results presented in this chapter were published in publications #6 and #7 listed in Section 1.3 [65, 66]. In this chapter, we first explore what is achievable with standard mel-frequency cepstral coefficients and features derived from the AM-FM model. By using feature mapping strategies, we explore if speaker specific characteristics affected when the speaker changes the speaking style can be mapped to specific feature domains in order to compensate for the lack of whispered speech data from target speakers. Next, complementarity of features derived from AM-FM models over conventional MFCC is explored using three fusion schemes. Results show that in the context of multi-style models, fusion strategies are more effective than feature mapping strategies and more research should be done in this direction.

4.2 Introduction

Here, we build on the insights presented in Chapter 3 and compare the performance of different speaker verification systems trained and tested under different scenarios and using different feature representations. Furthermore, experiments in the previous chapter were carried out in an ideal scenario, using a limited number of speakers, and a closed set scheme for speaker verification by using speech recordings from target speakers also for parameter estimation. In this chapter, we adopt

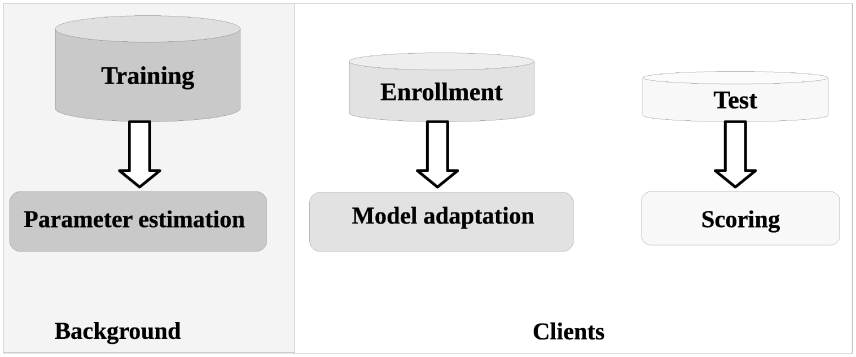


Figure 4.1 – Different data recordings involved during training, enrollment and testing of a speaker verification system.

a more realistic evaluation scheme by including additional datasets recorded in different conditions, which also increases the number of speakers. Besides, following standard evaluation protocols for speaker verification, we make a clear distinction between background speakers and target speakers or clients. Figure 4.1, shows the protocol typically followed during training, enrollment and testing stages. As can be seen, three different sets of speech recordings are needed. First, large amounts of speech data are needed to train e.g., the so-called GMM universal background model (UBM) and estimate other parameters needed for i-vector extraction (e.g., the T matrix estimation). During enrollment, a separate set is needed from each target speaker to allow for e.g., maximum a posteriori adaptation in GMM-based systems or for i-vector extraction to match with testing samples. Lastly, a third unseen data set is needed for system accuracy calculation. In the case of multi-style training, whispered speech data can be available in one or multiple datasets [14, 40].

Results in Chapter 3 have focused on exploration of different features for whispered speaker verification. As shown in Figure 2.4, one optional block includes feature transformation or feature mapping. Such strategy was not explored in Chapter 3, thus is investigated here. A recent study showed that such an approach can be helpful in speaker identification scenarios when the input presented is shouted speech [13]. For feature mapping, neural networks and Gaussian mixture models have been widely used in the voice conversion and voiced speech reconstruction literature (from whispered to normal-voiced speech) [42, 106, 107]. It is not clear, however, if such mappings can alter speaker identity information relevant for automated speaker recognition when using whispered speech. This chapter explores the advantages of feature mapping alongside other mismatch compensation strategies, namely fusion at the i-vector level.

	Num. speakers/Database			Total record.	
	TIMIT	wTIMIT	CHAINS	Norm.	Whsp.
UBM estimation	462	0	0	3696	0
T-matrix estimation	462	14	0	9996	6300
LDA and PLDA training	462	14	0	9996	6300
Enrollment	100	24	36	1280	480
Testing	100	24	36	320	120
Fusion system	68	10	0	780	230

Table 4.1 – Number of speakers and total number of recordings per database for training, enrollment and testing, and train the fusion system at score level.

Moreover, results presented previously have focused on a small scale baseline system. Here, a larger scale is performed, thus a more relevant baseline is needed with which performances can be compared to. Prior to investigating the benefits of feature mapping and new fusion schemes, accurate characterization of the baseline is needed, as detailed next.

4.3 Baseline SV system characterization

In this section we describe how the datasets described in Section 2.3 are used for SV system training and a baseline system is presented and characterized.

4.3.1 Task design

Speakers from the three databases were divided into three disjoint sets, one for training (to be used as background data), a second for enrollment and testing, and a third one to train the score-level fusion system. Recordings from 462 speakers from the TIMIT database and 14 speakers from wTIMIT, 476 in total, are included in the training set. Recordings from 100 speakers from TIMIT, 24 speakers from wTIMIT and 36 speakers from CHAINS are included in the enrollment and testing set. Average duration of each speech recording is 4.5 seconds, thus are rather short utterances with limited phonetic variability compared to typical NIST datasets for normal SV (which are around 120 seconds). To characterize the baseline system we included only normal speech recordings for training and enrollment, for testing we used two recordings per speaker, and if there are whispered speech recordings available then two additional sentences were included per speaker. Details can be found in Table 4.1.

Since fusion at the score-level requires training of the fusion system, we selected an independent set of speakers, namely 68 from the TIMIT database and 10 from the wTIMIT database, to create a new evaluation list. For enrollment, a configuration similar to the one used for the original evaluation list was used, including eight additional recordings of whispered speech for the 10 speakers of wTIMIT. For the new evaluation list, in order to have approximately the same amount of target and impostor scores from each speaking style, two recordings of normal speech and 15 recordings of whispered speech per speaker were used. For i-vector fusion, on the other hand, training of an additional system was not required, thus represents an advantage of such fusion scheme.

4.3.2 Settings for feature extraction and parameter estimation

For all databases in this study and prior to feature extraction, each speech recording was down-sampled to 16 kHz and the signal values were normalized to the range $[-1, 1]$. Feature vectors were computed on a per-window basis using a 25 ms window with 40% overlap. In particular for MFCC features, 27 triangular bandpass filters spaced according to the mel scale were used in the computation of 13 MFCC features including the 0–th order cepstral coefficient (log-energy) motivated by [40]. A 13 dimensional MFCC feature vector was shown to be an optimal setting for i-vector extraction, opposed to the 19 dimensional feature vector used for the GMM-UBM + MAP adaptation approach. Dynamic or transitional features (Δ and $\Delta\Delta$ MFCC) were computed by means of an anti-symmetric Finite Impulse Response (FIR) filter of length nine to avoid phase distortion of the temporal sequence. After dropping frames where no vocal activity was detected, cepstral mean and variance normalization was applied per recording to remove linear channel effects. The other feature set considered is the WIF, in this case and according to results presented in Chapter 3, a gammatone filterbank [43] with 27 channels was used. Filter center frequencies (fc_k) range from 100 Hz to 7000 Hz and their bandwidths are characterized by the mel scale. Pre-emphasis filter and feature normalization are not used for this feature set as they were shown to perform better without these pre-processing stages. Parameters such as sampling rate, window length, window overlap, number of filters and number of cepstral coefficients were selected motivated by [37, 40, 43].

For the UBM, different number of Gaussians were tested, i.e., $C = \{128, 256\}$, and results presented in the baseline characterization are those that turned out to be the best ones. The same

SV system	EER	
Baseline system (LFCC) [97]	2.68	
	MFCC	WIF
GMM-UBM + MAP adaptation	2.38	1.33
i-vector/cosine kernel	2.91	2.94
i-vector/PLDA	1.79	1.37

Table 4.2 – EER comparison with the baseline system using only the TIMIT database. For these results $C = 256$, and $D = 400$.

GMM-UBM was later used for different purposes such as for adapting to speakers specific models by using MAP adaptation or to compute the Baum-Welch statistics during T matrix estimation. For the T matrix, different dimensions were evaluated, i.e., $D = \{200, 300, 400\}$.

4.3.3 Baseline results

To characterize the performance of a valid baseline system to compare performances to, we follow the steps suggested in [97]. In [97], the authors provided the lists for background, enrollment and test sets using the TIMIT dataset. By using the same lists we report equal error rate (EER) results in Table 4.2 using two scoring strategies, i.e., cosine kernel and PLDA based scoring, and as feature vectors we used MFCC and WIF.

According to our results, and referring only to cepstral coefficients for this particular task, MFCC seems to be a better choice than LFCC when comparing the two SV approaches, i.e., the system reported in [97] which uses LFCC and the system we implemented using MFCC. It is important to notice that WIF outperformed MFCCs with the GMM-UBM + MAP adaptation, while for the PLDA based schemes, the two feature vectors have similar performance. Also, in the context of i-vectors, the PLDA based system is preferable to the cosine distance based system. Hence, in the experiments to follow we evaluated only the PLDA based system.

Next we performed similar experiments by using the three databases and the configuration described in Section 4.3.1. Results are presented in Table 4.3. In the table we also included the error rates when the system is evaluated with whispered speech.

As can be seen, for a standard speaker verification system based on the classical GMM+MFCC paradigm, addition of new speakers during enrollment whose recordings have been obtained in

	MFCC		WIF	
SV system	Norm	Whsp	Norm	Whsp
GMM + MAP	4.38	25.83	2.50	29.17
i-vector/PLDA	2.81	27.31	2.19	25.28

Table 4.3 – EER comparison between MFCC and WIF for the GMM+MAP adaptation based system with train/test mismatch where $C = 256$, and the i-vectors/PLDA based system with $C = 256$ and $T = 400$. Recordings from three databases were combined in these experiments, CHAINS, wTIMIT and TIMIT.

different conditions can highly affect the performance of the system. Additionally, the performance is highly affected if we evaluate the system with whispered speech. When we used WIF features, on the other hand, the error rates are affected as well for normal speech, but this feature set seems to handle in a better way the nuisance effects of the addition of new speakers. In the case of the i-vector/PLDA based system, both feature vectors resulted in high performance when testing with normal speech, in both cases better than the GMM+MAP adaptation based system. When testing with whispered speech, again, the negative effects of the mismatch train/test condition are evident. In the following sections strategies such as multi-style models, feature mapping and fusion schemes will be explored.

Lastly, we performed an additional experiment in the context of the i-vectors/PLDA based system by using only static MFCC. Results of these experiments were EER of 3.44% for normal speech and 32.5% for whispered speech, which justify the inclusion of dynamic or transitional features for MFCC in the experiments described above.

4.4 Multi-style model training

Previous results have shown the need to find strategies to compensate for the negative effects when whispered speech is considered into the possible testing scenarios. If we assume that whispered speech recordings are not available from target speakers, but there is data available from speakers in the training set (background speakers), two strategies are possible: *i*) include whispered recordings in the training set such that those recordings can be used during parameter estimation. From the baseline experiments it is clear that the total variability matrix can map to a highly discriminative space as long as there is sufficient statistics to learn from. This is the case for normal-voiced speech, but not for whispered speech. The lack of sufficient whispered speech recordings for parameter

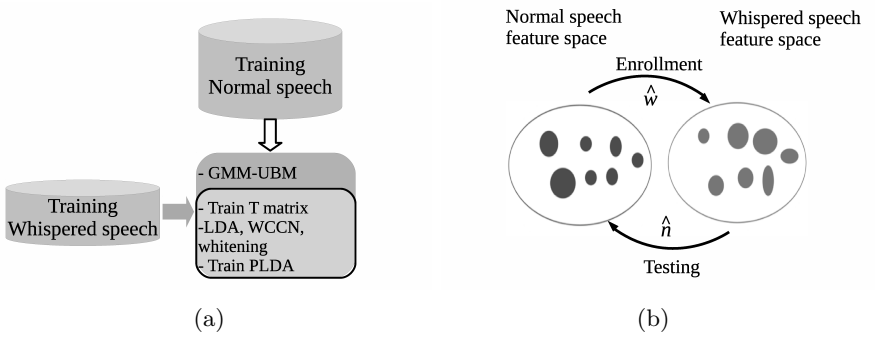


Figure 4.2 – Use of background data to train *i*) multi style models and *ii*) feature mapping.

estimation is one of the problems that has been discussed before. Even if a large number of recordings were collected, it would not suffice as long as the number of speakers is small. This is the case for the experiments herein, however, this scenario allows us also to evaluate how efficiently a system uses the data available during training or parameter estimation. *ii*) Feature mapping, i.e., by using training pairs of whispered and normal speech we can learn a mapping function and then apply this transformation either from normal to whispered speech to create artificial whispered enrollment observations or from whispered to normal speech to be used during testing, and compensate for the differences between training and testing data. Figure 4.2 illustrates how to use the speech recordings from background speakers in these two possible solutions.

For the following experiments we will use only MFCC feature vectors; this allows us to better illustrate different configurations of a classical SV system and point out the need for better or complementary feature representations such as WIF. First, the objective is to explore the effects of adding whispered speech during parameter estimation. As mentioned previously, training data contains recordings from both speaking styles, however for normal speech the number of speakers is significantly larger than the number of speakers for whispered speech. In a pilot experiment we found that for the parameter estimation of the GMM-UBM it sufficed to use only normal speech recordings to estimate the parameters of the Gaussian components because the role of the GMM model in this context is to cluster the acoustic features into broad acoustic classes. When adding whispered speech recordings during T-matrix estimation, as illustrated in Figure 4.2 (a), significant differences were observed.

For MFCC feature vectors, results presented in Table 4.4 show that the addition of whispered speech during T matrix estimation can add gains in performance of about 30% when testing with whispered speech, but also small increments are observed when testing with normal speech. It

Feature Set	UBM (C)	Normal			Whispered		
		T matrix dimension					
		200	300	400	200	300	400
MFCC	Only normal speech in T-matrix						
	128	3.23	3.44	3.38	30.00	29.17	28.56
	256	3.05	2.92	2.81	29.43	28.52	27.31
	Norm. and Whsp. speech in T-matrix						
	128	3.78	3.69	3.36	20.99	21.11	20.83
	256	3.18	3.44	3.13	20.00	21.91	20.83
WIF	Only normal speech in T-matrix						
	128	1.54	1.88	2.63	26.34	24.04	26.67
	256	1.38	1.56	2.19	25.83	24.17	25.28
	Norm. and Whsp. speech in T-matrix						
	128	2.71	3.13	3.44	18.81	19.13	21.53
	256	1.88	2.81	3.41	18.72	19.17	18.42

Table 4.4 – EER comparison using MFCC and WIF feature vectors, between using only normal speech recordings for parameter estimation and using normal and whispered speech for parameter estimation.

is important to notice that the i-vector extractor, and the SV system in general, can learn some variability from the speech recordings that were included during parameter estimation, but it is not enough, and there is a gap in performance of about 17% between EER for normal and whispered speech. This gap in performance is expected as the mismatch problem is still present. While we have added some general structure of the whispered speech feature space which is being learned by the recognition system, during enrollment there are still within-speaker variations that are not well represented in the samples used from the target speakers. Thus, the mismatch problem is present as session variability within the target speaker that the recognition system is not able to handle.

For the sake of completeness, we perform a similar experiment as the one presented for MFCC using WIF feature vectors. Results are also presented in Table 4.4. As can be seen, WIF perform better than MFCC in the matched and mismatched conditions. These results are in line with the preliminary experiments we performed in Chapter 3 and points towards the idea that the information present in the slowly varying envelope of the bandpass signals is highly discriminative, but we cannot disregard the phase of these signals and WIF are a feature set that combines the information from the envelope and the phase resulting in a feature vector highly discriminative. The remainder of this chapter will explore the potential of feature mapping and alternate fusion schemes as strategies to reduce this gap between normal and whispered speech, without the need for whispered speech recordings from target speakers.

4.4.1 Feature mapping

Two feature mapping techniques were evaluated in our experiments. The first approach is the classical Gaussian mixture model (GMM) regression [41], originally proposed for text-to-speech synthesis. Such method models both the source and the target feature vectors using a joint density GMM of time aligned target and source features. Model parameters are estimated using the standard expectation-maximization (EM) algorithm. With the estimated parameters a mapping function is formulated to compute the minimum mean square error estimate of the target feature vectors. Let $\mathcal{X}$ be the source feature space, and $\mathcal{Y}$ the target feature space, then the feature mapping operation is denoted by $\hat{y} = f_\Theta(x) = \mathbb{R}^d \rightarrow \mathbb{R}^d$, where d is the dimensionality of the feature vectors, Θ denotes the model parameters, and $\hat{y}$ is the estimated target feature vector from the input x. The GMM is trained using the stacked feature vectors $z_t = \left[x_t^{\mathrm{T}}, y_t^{\mathrm{T}}\right]^{\mathrm{T}}$ of dimensionality $2d$. The joint probability density function is given by:

$$p\left(z_t|\Theta\right) = \sum_{i=1}^{C} \alpha_i \mathcal{N}(z_t, \mu_i^{(z)}, \Sigma_i^{(z)}), \tag{4.1}$$

where $\mu_i^{(z)} = \begin{pmatrix} \mu_i^{(x)} \\ \mu_i^{(y)} \end{pmatrix}$, and $\Sigma_i^{(z)} = \begin{pmatrix} \Sigma_i^{(xx)} & \Sigma_i^{(xy)} \\ \Sigma_i^{(yx)} & \Sigma_i^{(yy)} \end{pmatrix}$ are the mean vector and covariance matrix, respectively. The four sub-matrices in $\Sigma_i^{(z)}$ are full covariance matrices. Once the parameters of the joint probability density function have been estimated via the EM algorithm, the mapping function can be written as [41]:

$$\hat{y} = f_\Theta(x) = \sum_{i=1}^{C} p(i|x) \left(\mu_i^{(y)} + \Sigma_i^{(xx)} \left(\Sigma_i^{(xx)}\right)^{-1} (x - \mu_i^x)\right), \tag{4.2}$$

where is $p(i|x)$ the posterior probability that the i-th Gaussian component generated x, and can be calculated as described by Equation 2.12. Additional details can be found in [41]. This mapping can be used to transform whispered to normal speech features or vice-versa, by properly defining source and target feature vectors.

The second technique is based on neural networks, which have been shown to be useful in the voice conversion literature [42]. Here, we explore the use of emerging deep neural networks (DNN), which have achieved state-of-the-art results across several research domains. We explore

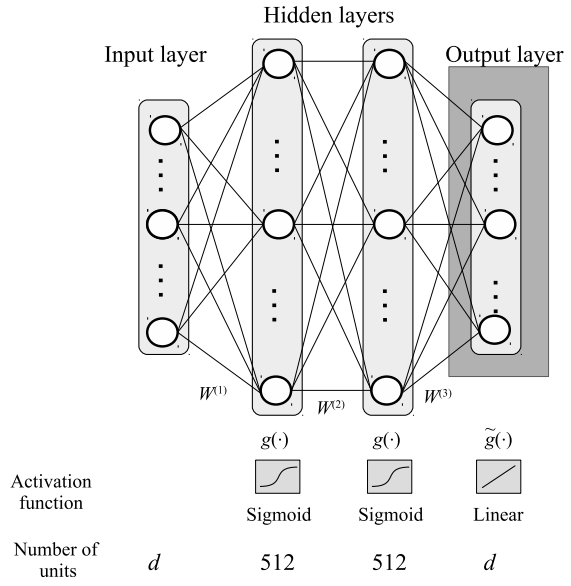


Figure 4.3 – Deep neural network architecture for feature mapping.

their flexibility in learning the direct mappings between whispered and normal speech features. By using the same notation as before, the feature mapping $\hat{y} = f_\Theta(x) = \mathbb{R}^d \rightarrow \mathbb{R}^d$ in this case is a sequence of non-linear operations that can be expressed using the same notation as used in Equation 2.3 [42]:

$$\hat{y} = G(x) = \tilde{g}\left(W^{(3)} g\left(W^{(2)} g\left(W^{(1)} x\right)\right)\right). \tag{4.3}$$

For the experiments herein, two stacked pre-trained autoencoders [108] with 512 hidden units each were used in our experiments, with $g(\cdot)$ as a sigmoid function and $\tilde{g}(\cdot)$ the identity function. Figure 4.3 illustrates the DNN architecture used in our experiments. This technique is also used to transform whispered to normal speech features or vice-versa, depending on the specific setting.

Alignment of normal and whispered speech features

Before learning any mapping it is necessary to guarantee the correspondence in time of the sequences to be used during training. To align the time sequences we used a similar approach as the one in [13], where alignment is achieved using dynamic time warping (DTW). The alignment algorithm, however, needs to be adapted in order to be useful for whispered speech. Initially, we compared two

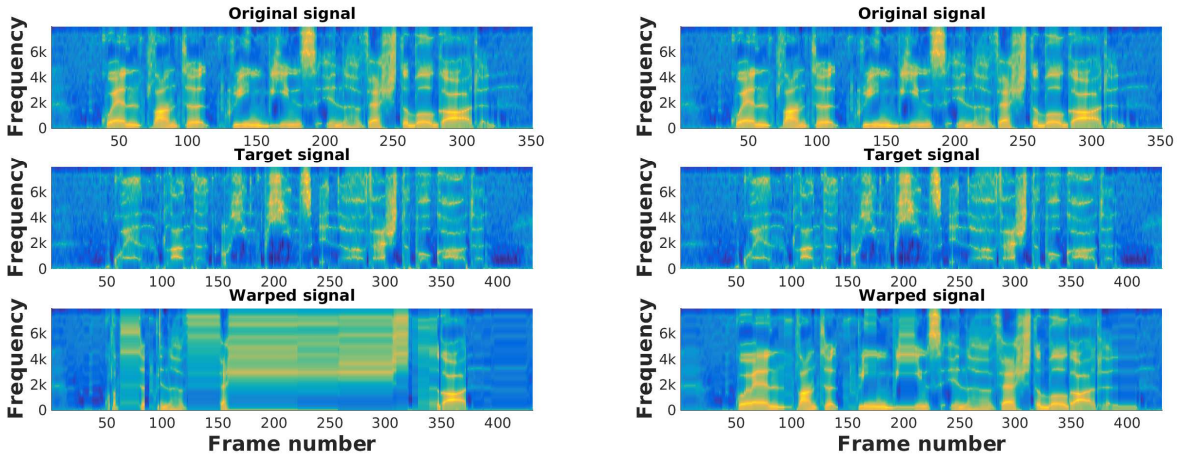
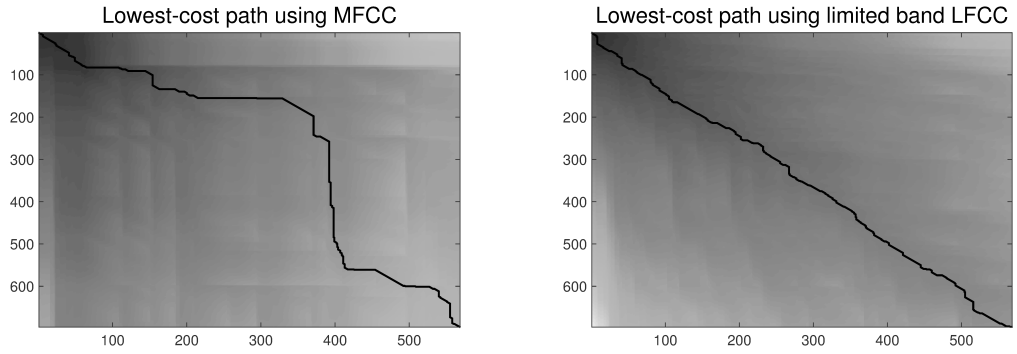


Figure 4.4 – Plots comparing two alignment strategies (a) Using full band MFCC and (b) using limited band LFCC.

feature representations to compute the distance matrix between the two recordings to be aligned, then we selected the best alignment and compare the final performance of the system. For the sake of completeness, here we describe the two approaches and how the best approach was selected. The first feature representation is the standard MFCC previously described in Section 2.2.1. The second approach is based on results presented in previous studies [34] and the analysis performed in Chapter 2 (see Figure 2.3), where it was shown that on average comparing the spectral envelope of both speaking styles, the frequency band where there are less differences is approximately between 1.2 - 4 KHz. As such, a linear spaced filterbank within this acoustic sub-band and 12 linear frequency cepstral coefficients were used to compute the distance matrix. A linear spaced filterbank is preferred for this purpose in order to not emphasize any particular frequency band. Figure 4.4 compares spectrograms before and after warping for the two approaches and illustrate how the choice in feature representation affects considerably the final result. Figure 4.5 on the other hand, compares the two alignment paths; as can be seen, the second approach can be considered as optimal because the lowest-cost paths are always close to the diagonal, which allows that the replicated frames to be evenly distributed along the whole recording and not in a single area, as is the case with the MFCC.



(a) (b)

Figure 4.5 – Plots comparing the lowest cost path computed with two feature representations, namely: (a) standard MFCC, and (b) limited band LFCC (1.2 - 4 kHz).

Evaluation Measures	Norm to Whsp		Whsp to Norm	
	GMM	DNN	GMM	DNN
MCD	13.84	12.78	13.96	12.75
ε_{rms}	0.644	0.596	0.649	0.595

Table 4.5 – Evaluation measures comparison between the two feature mapping techniques. MCD - Mean Cepstral Distance and ε_{rms} - root mean square error

Feature mapping

First, recordings from 14 speakers, seven female and seven male, from the wTIMIT database were used. Each speaker uttered approximately 450 different sentences, each of them in normal-voiced and then in whispered mode; in total 6298 pairs of utterances were included into the analysis. Since phonemes uttered by the same speaker have different duration in time for whispered and normal-voiced speech, we need to ensure that training utterances are phonetically aligned. To guarantee this, we used the alignment algorithm with limited band LFCC to compute the distance matrix. Table 4.5 compares the feature mappings in terms of mean cepstral distance and root mean square error between the original signal and its mapped counterpart. In terms of these measures the DNN performs better than the GMM-based mapping. However, the advantages of using a feature mapping should be decided on the basis of speaker verification performance, as detailed next.

Table 4.6, in turn, reports the equal error rate (EER) results obtained with the standard i-vector/PLDA based system, again using the conventional MFCC features. Four cases are reported to completely illustrate our experiments: *Baseline* illustrates the scenario where only normal speech is available for training and enrollment, no feature mapping is applied and no whispered speech features

Scenario	Normal			Whispered		
	Feature Mapping					
	none	GMM	DNN	none	GMM	DNN
Baseline	**2.81**	–	–	27.31	–	–
Mapping - case a	3.13	8.75	6.25	20.83	24.17	20.00
Mapping - case b	3.13	3.13	3.13	20.83	**17.50**	21.07

Table 4.6 – EER comparison with the baseline system and the two feature mappings in different scenarios. For these results $C = 256$, and $D = 200$.

were used for parameter estimation. *Mapping - case a*): illustrates the case where normal speech features from the enrollment set were mapped to whispered ones using GMM or DNN mapping functions. *Mapping - case b*), in turn, exemplifies the scenario where whispered speech features in the test set were mapped to normal speech ones using the GMM/DNN mapping functions. The latter case assumes an oracle normal/whisper classification system, thus the results for normal speech are unaffected. In both *Mapping case a*) and *b*), whispered speech features from the background speakers set were also included during parameter estimation, i.e., for T-matrix, i-vector post-processing and PLDA training, because by using only the mapped features slight improvements were observed (in the order of 2%), i.e., these two cases complement the *multi-style* model with feature mapping. The three columns in the Table represent no feature mapping (none), GMM or DNN based mapping. For these experiments the model parameters are C=256 and D=400, for the GMM-UBM and T-matrix, respectively.

As can be seen from Table 4.6, both feature mappings add some gains when testing with whispered speech, with relative improvements up to 37%. Despite the results reported in Table 4.5, suggesting that DNN mapping was better than GMM, such gains are not reflected in the EER results. These results also show that the addition of whispered speech during parameter estimation does not suffice to boost performance when testing with this speaking style, as whispered speech data does not contain enough inter-speaker variability.

In summary, feature mapping showed to provide some benefit for the train/test mismatch problem, but still resulted in a large gap between whispered and normal speech performances. As seen previously, addition of whispered recordings from target speakers seems to be the most effective method of shortening this gap [26], but with the disadvantage of hampering normal speech SV accuracy. Next, we explore alternate fusion schemes to investigate how efficiently these approaches can use the limited resources available during parameter estimation.

4.4.2 Fusion schemes

Three fusion schemes are investigated in this chapter, two at the input level and one at the output level, namely: *i) Frame level fusion*, *ii) i-vector concatenation* and *iii) score-level fusion*. Diagrams in Figure 4.6 (a)-(c) depict these fusion schemes, respectively. For frame-level fusion, MFCC and WIF features are concatenated into a final feature vector. Principal component analysis is then performed to remove redundant variables and only the top components are kept as features with 99% of cumulative variance retained. These top components are then used for i-vector computation. With i-vector concatenation, in turn, i-vectors extracted from MFCC and WIF features are concatenated into a final feature vector prior to post-processing, i.e., prior to LDA, whitening, and length normalization. This strategy has shown to be effective in various scenarios such as language recognition and short utterance speaker recognition [109, 110] to combine strengths of i-vectors estimated from different feature representations. This approach does not require training of an additional system thus represents an advantage over score level fusion. Fusion at frame- and i-vector-level (i.e., Figure 4.6 (a) and (b)) are both cases of fusion at the input level when seen in a general scheme as previously depicted by Figure 2.4.

Lastly, for score-level fusion, separate data (different from background and target speakers) is needed to train the fusion system and the systems to be fused (i.e., systems trained on MFCC and WIF feature sets) are evaluated using an unseen evaluation set. A logistic regression function is used as a fusion system and maps evaluation scores into a final decision using the Bosaris toolkit [111]. To estimate the parameters of this fusion system, speech recordings from TIMIT and wTIMIT datasets were used as described in Section 4.3.1.

Next, we present the results of the three fusion schemes, *frame-level fusion*, *i-vector concatenation* and *score-level fusion*. Results are presented in Table 4.7, with the best overall results highlighted in bold per speaking style. By comparing the three fusion schemes we can see significant differences; for example, fusion at frame level seems to be the less efficient way to combine the information from the two feature sets. Results presented for the feature sets separately in Table 4.4 are better than those attained by combining the two feature sets at the frame level. Fusion at higher levels, on the other hand, such as at the i-vector- or score-level, shows to be a better option. While the i-vector concatenation has the advantage of not requiring the training of a separate mapping function, score level fusion resulted in the lowest overall EER. When comparing with

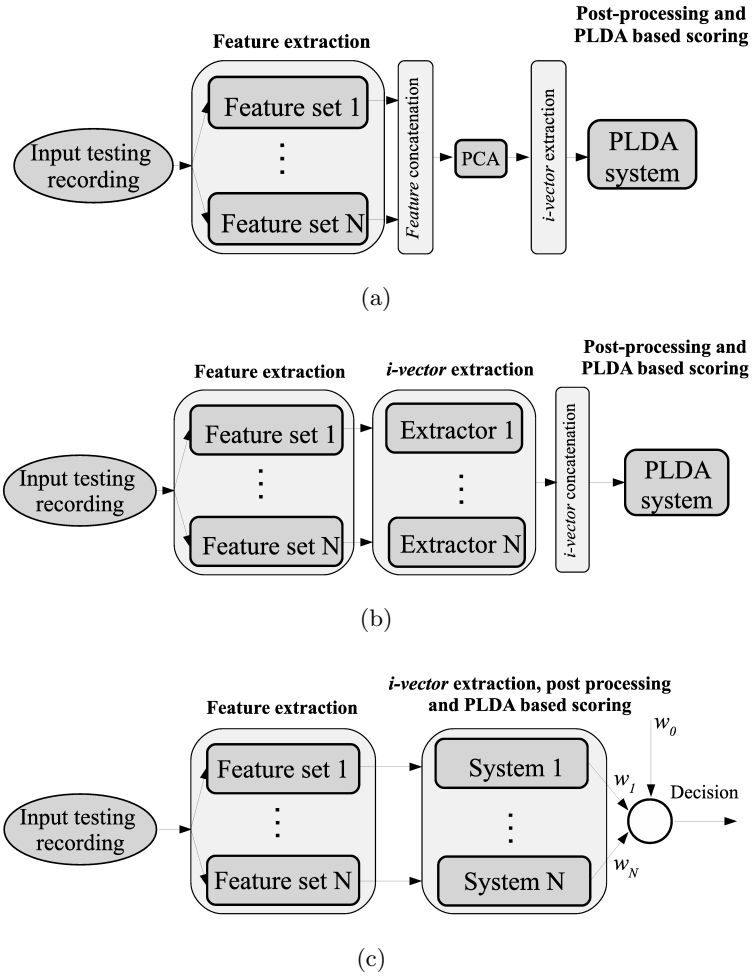


Figure 4.6 – General building blocks of the fusion schemes: (a) Frame level fusion, (b) i-vector concatenation and (c) Score-level fusion.

the baseline system presented in Table 4.3, relative improvements are 44% and 42% for normal and whispered speech, respectively, were obtained relative to MFCC features and 28% and 37%, respectively, relative to WIF features. Figure 4.7 complements EER results with the DET curve of the best configuration per speaking style and per fusion scheme. Solid and dashed lines correspond to testing with normal and whispered speech, respectively. As can be seen, for these feature sets, frame-level fusion performs the poorest. When comparing i-vector concatenation and score-level fusion, there are slight differences and the later is the best choice for the task at hand.

These experimental results show the advantages of using system fusion when using the classical MFCC and features extracted from the AM-FM model as feature vectors. Furthermore, the use AM-FM based features suggested that the phase and envelope of bandpass signals can contain highly discriminative speaker specific information for i-vector extraction purposes and complement

UBM	Normal			Whispered		
	T matrix dimension					
	200	300	400	200	300	400
Frame-level fusion						
128	4.15	3.35	3.44	19.85	22.00	20.83
256	4.04	3.14	3.37	21.67	19.98	20.83
i-vector concatenation						
128	2.19	2.19	2.19	16.54	16.67	17.46
256	1.87	2.03	2.29	21.67	16.67	16.43
Score-level fusion						
128	2.07	2.50	2.38	**15.83**	15.84	16.67
256	**1.56**	1.88	2.30	15.97	19.98	16.67

Table 4.7 – EER comparison using three different fusion schemes, and two feature sets MFCC and WIF.

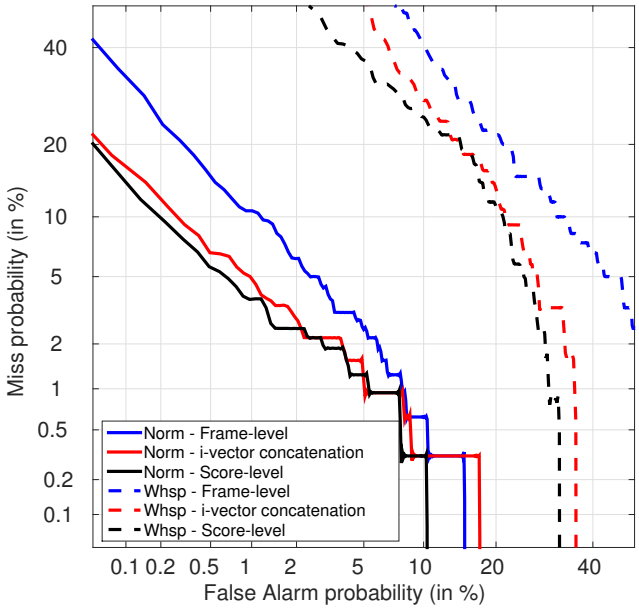


Figure 4.7 – DET curve comparison of the best configuration per fusion scheme. Solid and dashed lines correspond to testing with normal and whispered speech, respectively.

classical MFCC based schemes for both normal and whispered speech. While feature mapping and fusion resulted in improvements over the baseline, the gap between normal and whispered speech still remains, thus suggesting that alternate feature representations may be needed.

4.5 Conclusions

In this chapter we have addressed the issue of speaker verification based on whispered speech in a more realistic scenario. Three databases were pooled together in order to increase the number of speakers and add more flexibility to the experimental evaluation. The addition of whispered data during training, in order to add information about whispered speech variability, combined with feature mapping techniques, to compensate for the lack of whispered speech data from target speakers, was shown to not suffice to boost speaker verification performance for whispered speech. As an alternative, we explored complementary information extracted from WIF and MFCC feature sets via three fusion schemes, namely: *i*) frame level, *ii*) i-vector concatenation, and *iii*) score level. Gains as high as 42% and 44% were obtained for whispered and normal speech, respectively, relative to a baseline system based on i-vectors/PLDA+MFCC with no whispered speech in the training set.

Overall, we observed that existing features (e.g. MFCC) do not convey sufficient reliable speaker identity information across different vocal efforts. Given the lack of sufficient speakers to train independent and dedicated models for whispered speech, techniques such as feature mapping seem to be insufficient to improve performance and fusion schemes seem to be more effective. Nonetheless, the mismatch problem is still present and the gap in performance is still considerable between normal and whispered speech.

These insights suggest that innovative features conveying more speaker-dependent invariant information across different vocal efforts are needed. From the results obtained with WIF features, information extracted from slow varying envelope from bandpass signals or information related to the phase seems to be an alternative to be explored. These new features are described in the next chapter.

Chapter 5

Exploring Speaker-Dependent Invariant Information Between Normal and Whispered speech

5.1 Preamble

Results presented in this chapter are also detailed in papers #3 and #8 listed in Section 1.3. Paper #3 has been accepted with minor revisions to be published in the journal *Computer Speech & Language* and paper #8 has been submitted to *EUSIPCO 2017* [62, 67]. In this chapter, we focus attention on the extraction of invariant speaker-dependent information from normal and whispered speech, thus allowing for improved multi-vocal effort speaker verification.

5.2 Introduction

In previous chapters we have explored different techniques that have been reported in the literature to be useful in adverse conditions, or to extract important speaker-dependent information useful for speaker recognition tasks. In Chapter 3, for example, we evaluated different feature representations and system configurations. Chapter 4, in turn, evaluated the combination of multi-style models with feature mapping and fusion schemes. Here, to complement the above-mentioned strategies, we

explore the computation of innovative features that extract invariant information embedded within both speaking styles.

Motivated by findings from the previous chapters and from the literature, two classes of features are proposed. First, we propose variants of the classic MFCC feature in order to better extract vocal-effort invariant information. Second, we extract a new feature set based on modulation spectral analysis. Our pilot experiments with AM-FM features have highlighted the advantages of the slowly varying envelope for the task at hand. Moreover, previous studies have shown that the modulation spectral signal representation accurately decouples speech from environment- based components (e.g., noise and reverberation) [43], thus can potentially add robustness to practical speaker recognition systems. Performance of the new features sets are compared to the classical MFCCs and the baseline system described in Section 4.4.

5.3 Towards cross-vocal effort SV: new feature representations

5.3.1 Variants of the MFCCs

According to perceptual and acoustic studies, two of the most salient differences between normal and whispered speech are related to the spectral envelope, i.e. *i*) whispered speech has a lower and flatter power spectral density [25] and *ii*) the formants shift towards higher frequencies [71, 73]. This last observation is more noticeable for the first three formants (F1, F2 and F3), where, F1 shifts can be up to 71% for men and 52% for women; F2 shifts can be up to 24% for men and 20% for women; and F3 shifts can be of 10% and 4.8%, respectively [73]. These differences were discussed in Chapter 2, where it was shown that most of the differences remain below 1.2 kHz. Figure 2.3 illustrates these differences by depicting the average power spectrum of amplitude-normalized and pre-emphasized recordings from 36 speakers (male and female). For normal speech, most of the energy is concentrated below 1 kHz, whereas for whispered speech it is concentrated below 500 Hz, with frequency shifts in the spectral peaks and valleys (F1 shifts are most prominent). In Chapter 2 it was also discussed that there are some similarities between the two speaking styles; for example while it has been documented that characteristics of vowels and voiced consonants are significantly different, unvoiced consonants are relatively similar [33]. Based on these insights, two MFCC variants are explored.

As mentioned in Section 2.2.1, following the source-filter model of speech generation shown in Figure 2.6, it is possible to split the speech signal in two components: an excitation signal and a transfer function which models the vocal tract configuration [68]. The excitation can be visualised as the combination of two different signal generators: one for voiced-speech and another for voiceless (noise-like) speech. The excitation signal is also known as residual. In the past, features extracted from the residual have been shown to contain important speaker-dependent information useful for speaker recognition tasks [112, 113, 114]. This is relevant for whispered speech because by removing the influence of the vocal tract, then differences related to the spectral envelope are no longer a nuisance factor affecting SV performance.

It is widely known that the residual signal of normal speech contains quasi-period pulses corresponding to glottal closure/opening instances during vocal fold vibration of voiced speech segments. Unvoiced segments, in turn, are not caused by a regular vibration (glottal excitation) but rather by turbulent airflow due to a constriction in the vocal tract [68]. Unvoiced sounds have been shown to remain unaffected during whispering mode [33], and also to contain important speaker-dependent information for speaker recognition tasks using whispered speech [26, 115]. As such, it is expected that features extracted from the residual signal will carry some invariant speaker information, particularly from the unvoiced segments. While it is not expected that the residual based feature will perform accurately alone, as most of the speaker-dependent information is typically embedded in spectral envelope associated to the vocal tract configuration, it should carry complementary information that can be fused with other features [114, 116]. These features are termed RMFCC and are used for i-vector computation.

Previously, residuals were also explored for speaker verification of normal narrowband speech (i.e., 8kHz sampling) [114]. In such case, the residual modelled differences in excitation energy and periodicity information amongst speakers. Here, residuals are explored for alternate reasons, as there is no periodicity to be modelled with whispered speech due to the lack of vocal fold vibrations. Our hypothesis is that the resulting spectrally flat signal, even with the harmonic structure for normal speech, has reduced differences when comparing the two speaking styles. Furthermore, unvoiced sounds have more energy concentration at higher frequencies [68] and consonants such as stops, fricatives and affricates have more spectral similarities at frequencies higher than 4 kHz [25]. As such, by analyzing residuals from wideband speech, more information related to the unvoicedness will be captured. Given the similarities between the two vocal efforts for unvoiced speech segments,

it is believed that this information will contain useful speaker-dependent information invariant across the two vocal efforts.

The process to compute the residual or excitation signal is as follows: Given the speech signal, using linear predictive analysis, it is possible to rebuild the vocal tract transfer function by estimating the parameters of a low-order all-pole filter. By definition, linear prediction analysis uses the redundancy in the speech signal to predict the current sample, $\hat{x}(n)$, as a linear combination of past p samples, as shown by Equation 5.1, where $\{a_i\}$ are the linear prediction coefficients and $x(n)$ is the speech sequence. The residual $e(n)$ is the prediction error obtained as the difference between the predicted speech sample and the actual sample [68], as shown in Equation 5.2. Since the excitation signal is spectrally flat, uncorrelated white noise, the transfer function of this all-pole model represents the spectral envelope of the speech signal. Having the transfer function, inverse filtering is used to recover the excitation or residual signal [68], the relation error (output) to speech signal (input) of the inverse filtering process is shown in Equation 5.3.

$$\hat{x}(n) = \sum_{i=1}^{p} a_i x(n-i), \tag{5.1}$$

$$e(n) = x(n) - \hat{x}(n) = x(n) - \sum_{i=1}^{p} a_i x(n-i), \tag{5.2}$$

$$\frac{E(z)}{X(z)} = 1 - \sum_{i=1}^{p} a_i z^{-k}. \tag{5.3}$$

Typically, the use of narrowband (NB) signals for telephone based communications has limited the analysis of speech signals for feature extraction to the range of frequencies in 0.3 - 3.4 kHz. With the use of emerging wideband (WB) communications and advanced digital signal processing technology in the telecommunications infrastructure, this range has been expanded to 8 kHz [117]. This has motivated detailed analyzes to explore the role and relevance of different frequency subbands for speaker recognition tasks. As an example, for NB speech signals in [118] it was shown that the 1.5 - 3.4 kHz frequency sub-band contains more discriminative information than the lower 0.3 - 1.5 kHz frequency sub-band, except for nasals. For WB speech signals, on the other hand, in [119] it was shown that the frequency sub-band 4-8 kHz provides a performance similar to that obtained with

the frequency sub-band 0-4 kHz, thus suggesting the presence of relevant speaker-discriminative information beyond 4 kHz. In a different study, it was shown that for text-dependent speaker identification, higher frequency channels were more relevant for speaker recognition than those located at lower frequencies [120]. It was reported that the lowest identification rates were associated to channels containing information of first and second formants, and that there was a high negative impact in performance when removing channels containing information from the frequency band between 5 kHz to 8 kHz [120]. Moreover, preliminary results presented in Chapter 3 suggested that by using the sub-band from 1.2 kHz to 4 kHz to compute the different feature sets it was possible to improve performance in the mismatch condition, but at the cost of reduced performance in the matched scenario.

By using these insights, we propose an alternate variant of the MFCC, which follows the typical processing pipeline described in Section 2.2.1, but is computed from the 1.2-8 kHz sub-band. By doing this, the sub-band that comprises mostly information from the first formant (F1) is removed which, as mentioned above, can have shifts as high as 71% for men and 52% for women, relative to F1 from their normal speech counterparts. Hence, most of the speaker specific information relevant for speaker recognition tasks is preserved and the performance in normal speech should not be affected. These features are termed LMFCC and are used for i-vector computation.

For the experiments herein, as described for MFCC features in Sections 2.2.1 and 4.3.2, 39-dimensional feature vectors were used, i.e., thirteen LMFCC and RMFCC features were computed including the 0–th order cepstral coefficient using 25 ms windows with 40% overlap. Delta and double delta coefficients were appended to include dynamic or transitional information.

5.3.2 Auditory-inspired amplitude modulation features - AAMF

For the analysis in this section we assume that an observed time-domain signal is the result of multiplying a low-frequency modulator (temporal envelope) by a high-frequency carrier. Hence, the modulation spectrum characterizes the rate of change of long-term speech temporal envelopes [121], and the analysis is carried out by using acoustic subbands. The modulation frequency (modulation domain) represents the frequency content of the subband amplitude envelopes and it potentially contains information about speaking rate and other speaker specific attributes [40]. Auditory-inspired amplitude modulation features have been effectively used in the past to improve automatic

speaker identification in realistic environments, as it was shown that they accurately separate speech from environment-based components [43]. In that case, the technique relied on identifying the modulation frequencies that remained unaffected by environmental noise based on energy levels, and disregarding those that presented significant changes when affected by noise. A similar idea can be applied for the task at hand. However, identification of channels or variables containing invariant information cannot be based on energy levels given their inherent differences between normal-voiced and whispered speech. For this reason, mutual information (MI) is chosen to compare pairs of variables coming from the two speaking styles and determine whether a specific modulation frequency channel contains shared information that can be useful for speaker recognition purposes.

The proposed auditory-inspired amplitude modulation features are computed following the approach described in [122] with some adaptations required fit our needs. More specifically, the speech signal $x(n)$ is first processed by an N-point short-time discrete Fourier transform (STDFT) to generate $X(nL_a, f_a)$ given by:

$$X(nL_a, f_a) = \sum_{m=-\infty}^{\infty} x(m) w_a(nL_a - m) e^{-i\frac{2\pi k}{N}m}, \tag{5.4}$$

where $w_a(n)$ is an acoustic frequency analysis window and L_a denotes the frame shifts, the subscript a stands for acoustic domain. Acoustic frequency components (termed f_a) are aligned in time to form the conventional time-frequency representation. In order to emulate human cochlear processing, the squared magnitudes of the obtained acoustic frequency components are grouped into 27 subbands ($|X_j(\cdot)|, j = 1, \ldots, 27$), spaced according to the perceptual mel scale as depicted by Figure 5.1 (top plot). A second transform is then performed across time for each of the 27 subband magnitude signals to yield:

$$X_j(mL_m, f_m) = \sum_{n=-\infty}^{\infty} |X_j(n)| w_m(mL_m - n) e^{-j\frac{2\pi k}{N}n}, \tag{5.5}$$

where $w_m(m)$ is a modulation frequency analysis window, L_m the frame shift, the subscript m stands for modulation domain, j indexes the acoustic frequency bands, and f_m represents modulation frequency bins. Following recent physiological evidence of an auditory filterbank structure in the modulation domain [123], we further group squared modulation frequency bins into eight subbands using logarithmically-spaced triangular bandpass filters distributed between $0.01 - 80$ Hz modulation frequency as depicted by Figure 5.1 (bottom plot). The speech modulation spectrum results in a high-dimensional feature representation (e.g., 27 acoustic bands $\times 8$ modulation

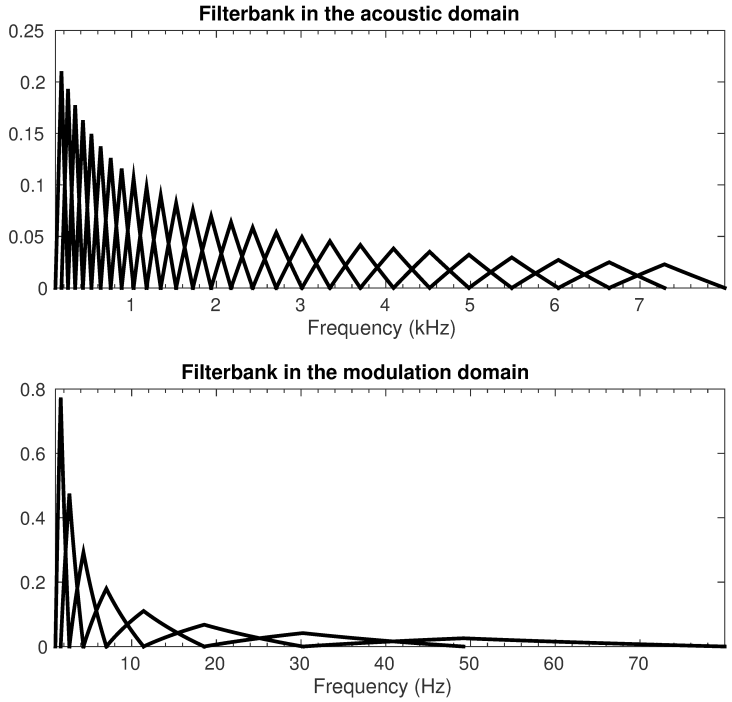


Figure 5.1 – Plots of frequency response of the 27- (top) and 8-channel (bottom) filterbanks used in the experiments herein.

bands= 216 dimensions), finally $\log_{10}$ compression is applied. Figure 5.2 summarizes the above described process. As can be seen, each recording is represented as a 3-dimensional array with dimensions being acoustic frequency channel, modulation frequency channel and modulation frame index. For a given modulation frame, which for this work is 100 ms, a two dimensional array represents the energy distribution across the different channels in both frequency domains. The evolution through time of a particular point with acoustic frequency j and modulation frequency i represents the variable $\xi_{(i,j)}$, as highlighted in dark gray in Figure 5.2.

Each modulation frame (a matrix with $27 \times 8 = 216$ elements) can be collapsed into a vector and used as standard features. However, given the high dimensionality of the resulting space and correlation among different dimensions, each feature vector is projected to a lower dimensional space using principal component analysis (PCA) with 40 components retaining 98.7% of cumulative variance, which according to our experiments showed to be an optimal value. These 40 components are then used for i-vector calculation. The main differences between our approach and the one presented in [122] are in the way the modulation bins are grouped in logarithmic distributed bands, thus better simulating the human auditory system; the log compression, and finally the dimension-

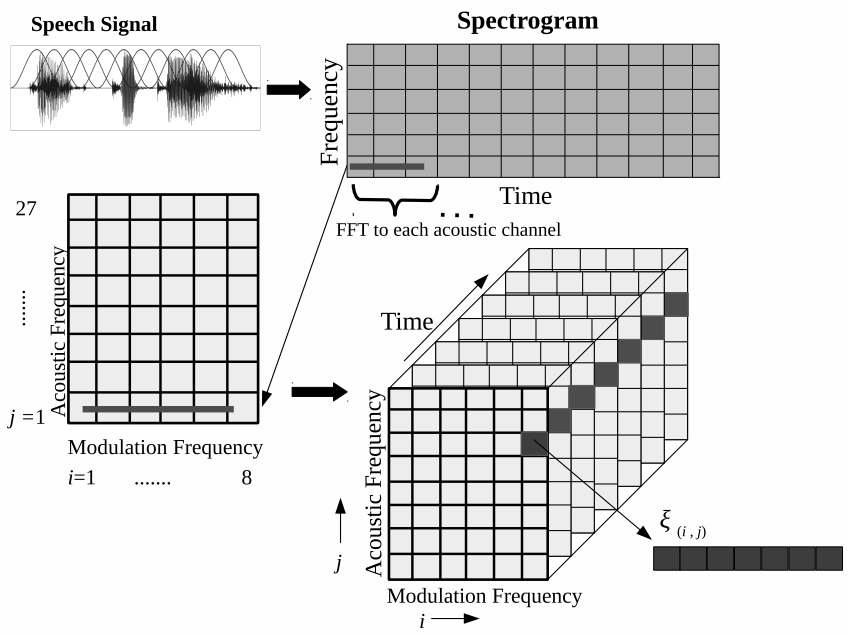


Figure 5.2 – Decomposition of a speech recording in terms of acoustic and modulation frequency components in a short time basis.

		Normal			Whispered		
		T matrix dimension					
Feature set	UBM	200	300	400	200	300	400
MFCC	128	3.23	3.44	3.38	30.00	29.17	28.56
	256	3.05	2.92	**2.81**	29.43	**28.52**	27.31
AAMF	128	1.25	1.07	1.25	**22.85**	25.71	23.33
	256	**0.94**	0.94	1.04	24.17	26.85	25.00

Table 5.1 – EER comparison between MFCC and AAMF using different values for the number of Gaussian components in the UBM and T matrix dimension. No whispered speech recordings were used during parameter estimation.

ality reduction, which according to our experiments result in a more informative feature vector for the task at hand.

As a first step, in order to validate the discriminative capabilities of this feature representation, we carried out an experiment by comparing with the standard MFCC feature vectors using different configurations of the SV system as described in Section 4.4. Results are presented in Table 5.1 with best results highlighted per feature set and per speaking style. As can be seen, AAMF not only performed better in the matched condition for all cases but also helps to reduce error rates in the mismatched condition. Since these results do not rely on whispered speech being used during training, they suggest that the proposed AAMF features are more discriminative than standard

MFCCs for both normal and whispered speech. The gap remaining between the two vocal efforts is still high, however, thus indicating that further processing is needed. As in [43], further analysis is needed in order to investigate which acoustic/modulation channels carry invariant information across the two vocal efforts. This analysis is described next.

Mutual information (MI) based feature selection: In order to verify which acoustic/modulation channels contained invariant information across vocal efforts, we relied on the mutual information (MI), as it conveys both linear and non-linear statistical dependencies between the two efforts. MI has been shown to be an effective tool to measure relevance and redundancy among different modalities (e.g., [44, 45, 46]). By definition given two random variables X and Y with probability mass functions $p(x)$ and $p(y)$ respectively, and joint distribution $p(x, y)$, then the mutual information between X and Y is given by [124]:

$$MI(X, Y) = \sum_{x,y} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}. \tag{5.6}$$

To derive $p(x)$, $p(y)$ and $p(x, y)$ the variables X and Y were partitioned in N uniform intervals and then the observed values for x and y were discretized. Having the number of data pairs, the number of intervals and the discretized values, then the probability mass functions are represented by a histogram [124].

Here, in order to compute MI, recordings from 14 speakers, seven female and seven male, from the wTIMIT database were used. Each speaker uttered approximately 450 different sentences, each of them in normal-voiced and then in whispered mode; in total 6298 pairs of utterances were included into the analysis. Since phonemes uttered by the same speaker have different duration in time for whispered and normal-voiced speech, we need to ensure that training utterances are phonetically aligned such that during MI analysis all sentences have the same duration and the analysis can be performed between two equivalent frames. To guarantee this, we used the dynamic time warping (DTW) approach described in Section 4.4.1. However, since we are exploring the MI of AAMFs between normal and whispered speech, using the LFCCs for time alignment may result in unnatural temporal dynamics that may affect AAMF computation. As such, here we also explore time alignment based on the AAMFs themselves. Figure 5.3 compares the two alignment paths; as

can be seen, both achieve lowest-cost paths close to the diagonal, which allows the replicated frames to be evenly distributed along the whole recording and not in a single area.

Finally, having the time series aligned, MI values were computed per variable per speaker (i.e., acoustic/modulation pair); thus resulting in 216 MI values per speaker. Each value is normalized using the sum of entropies as:

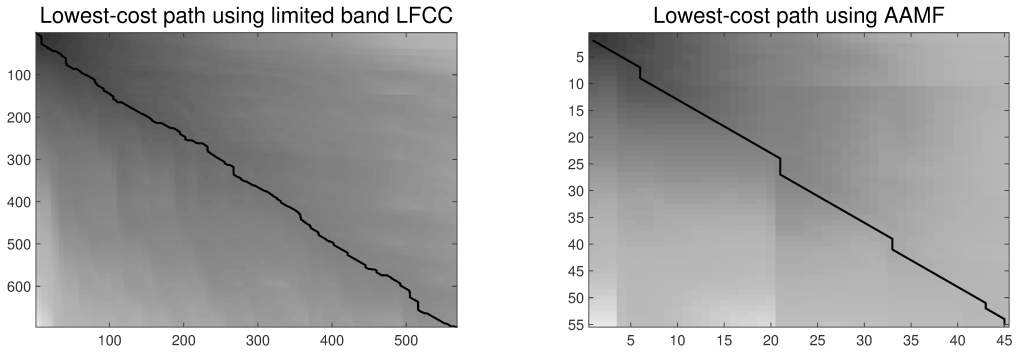
$$\hat{MI} = \frac{2 \cdot MI}{H_1 + H_2}, \tag{5.7}$$

where $\hat{MI}$ is the normalized MI value, H_1 and H_2 are the entropy values of the two variables being compared. Next, having all MI values for a given speaker, they are re-scaled to the range [0-1] by using the transformation :

$$\hat{x} = \left(\frac{x - min_x}{max_x - min_x} \right), \tag{5.8}$$

where x is the original value of a given variable to be re-scaled and $\hat{x}$ the scaled value. Finally, all MI values were averaged over the 14 speakers. Results from this analysis allow us to identify which acoustic/modulation channels have a high degree of shared information between normal-voiced and whispered speech by thresholding, we can then create a binary mask to be used to select which channels to keep for SV system. training. For the experiments herein, the threshold was set to 0.4, which resulted in the selection of 141 and 157 variables depending on which alignment approach is used, i.e., LFCC or AAMF, respectively.

Figures 5.4 depicts the processing steps used in the creation of the MI-based binary mask. Figure 5.5, depicts the obtained binary masks using the LFCC- (left) and AAMF-based (right) alignment algorithms. As can be seen, both approaches eliminate the lower acoustic bands, with the LFCC based alignment method resulting in more suppression in the 2-5 kHz acoustic bands and modulation bands greater than 20 Hz. Finally, principal components analysis (PCA) was used to reduce the high-dimensional feature set to 40 dimensions, accounting for 99.3% and 99.1% of accumulated variance when using LFCC and AAMF for alignment, respectively. Figure 5.6 depicts this final dimensionality reduction step prior to SV system training. Even though there were no big differences in terms of error rate when (pilot) testing with normal speech, the alignment based on AAMF showed to have better performance when testing with whispered speech. As such, results reported henceforth will be based on AAMF alignment. The features resultant from this MI-



(a) (b)

Figure 5.3 – Plots comparing the lowest cost path computed with three feature representations. (a) Using limited band LFCC (1.2 - 4 kHz) as in Section 4.4.1, (b) Using AAMF.

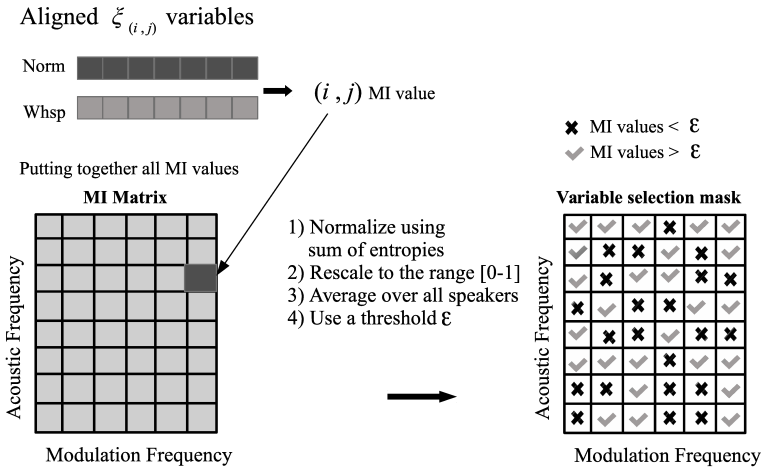


Figure 5.4 – Identification of relevant variables (acoustic and modulation channels) using MI.

based feature reduction method will be referred to as AAMF(FS), which are then used for i-vector calculation.

To test the effectiveness of the proposed feature sets, SV systems as described in Section 4.4 were used (i.e., whispered speech recordings used only for T matrix estimation); results are presented in Table 5.2. For completeness and comparison purposes, the MFCC results from Table 4.4 are included as well.

As can be seen, by using the standard MFCC, there is a gap in performance between normal and whispered speech around 17%, in the best case. Next, by using the RMFCC feature set, it is

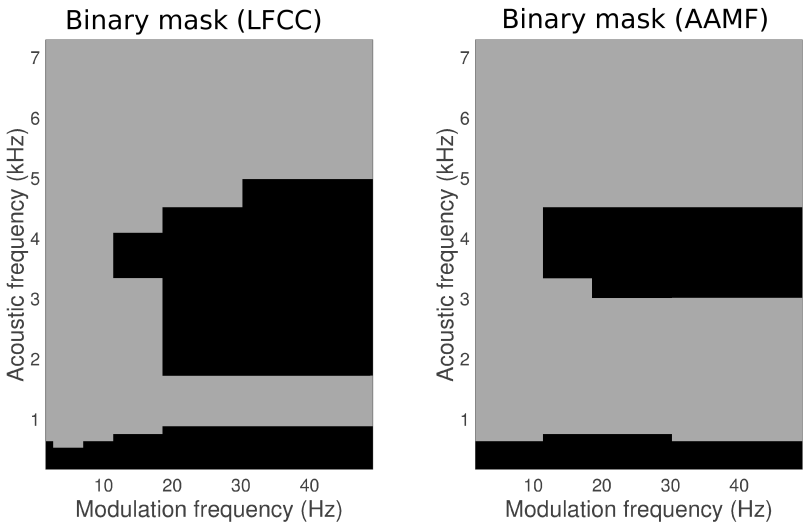


Figure 5.5 – Acoustic and modulation bands selected, these bands contain high degree of information that is common for both, normal-voiced and whispered speech. Grey areas correspond to selected channels, while the black ones to the disregarded channels.

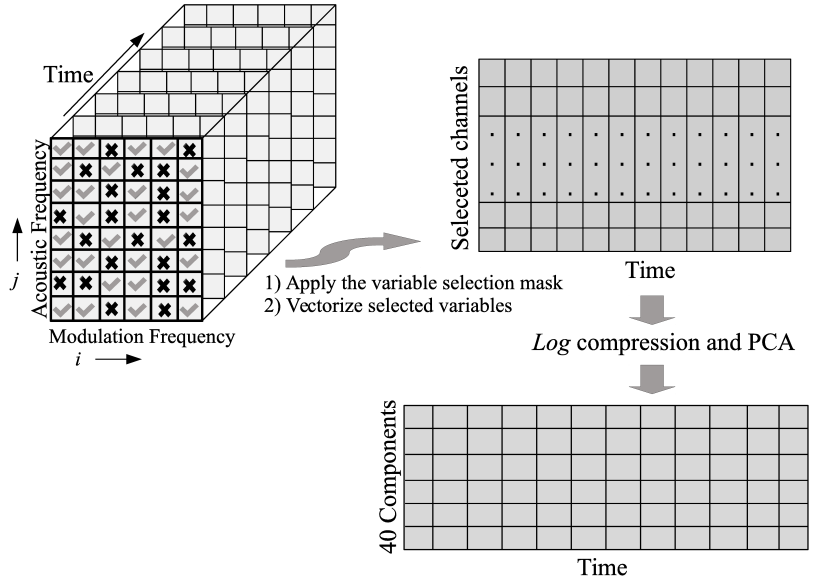


Figure 5.6 – Process to compute modulation spectrum based features using the MI-based binary mask and decorrelation using PCA.

clear that by removing the information related to the spectral envelope important speaker specific information is also removed. Systems based only on this feature set are not expected to perform at the same level as standard MFCCs, but may provide complementary information. And finally, LMFCC are shown to perform equally well to MFCCs for normal speech, but to lower whispered speaker verification error rate from 20% to 16.67%. Moreover, the AAMF(FS) feature set, results in superior performance when testing with normal speech, regardless of the alignment method used

UBM (C)	Normal			Whispered		
	T matrix dimension					
	200	300	400	200	300	400
S1: MFCC						
128	3.78	3.69	3.36	20.99	21.11	20.83
256	3.18	3.44	**3.13**	**20.00**	21.91	20.83
S2: RMFCC						
128	8.44	8.14	7.50	25.83	24.42	25.95
256	8.44	7.50	**6.70**	24.20	25.00	**22.95**
S3: LMFCC						
128	3.13	3.13	3.44	18.13	17.13	17.81
256	3.15	3.44	**3.13**	17.97	18.29	**16.67**
AAMF(FS) - LFCC alignment						
128	1.56	1.58	1.55	21.22	19.82	20.86
256	1.60	1.36	**1.26**	20.51	**19.17**	20.83
S4: AAMF(FS) - AAMF alignment						
128	1.00	1.25	1.18	20.00	20.00	20.57
256	1.56	1.12	**0.94**	18.44	18.29	**14.80**

Table 5.2 – EER comparison using three different feature sets: MFCC, RMFCC, LMFCC and AAMF(FS).

to generate the binary mask. Regarding the configuration of the SV system, i.e., the number of components in the GMM-UBM and the dimensionality of the T-matrix, we can see that for almost all compared feature sets the configuration that offered the best results is C=256 and D=400, hence, these settings will be used henceforth. In the Table, each feature set has received a label "*Si*". These labels will be used in Section 5.5 to indicate which feature sets were used during fusion.

5.4 Score-domain feature complementarity analysis

In order to better understand the contributions and complementarity of each newly proposed feature set, we perform an analysis on the output scores of the systems trained on them. A comparison in the score domain is more feasible and easier to interpret than a comparison in the feature space, given that they are encoding different characteristics of the speech signals. For the analysis we use the Lawley-Hotelling statistic [47], a commonly used measure in MANOVA (multivariate analysis of variance) to compare the mean vectors of k groups of samples for significant differences. In this case, we want to test whether or not the mean of impostor scores equals the mean of target speakers.

The hypotheses are, therefore: $\mathcal{H}_0 : \mu_i = \mu_t$, vs. $\mathcal{H}_1 : \mu_i \neq \mu_t$, where μ_i and μ_t stand for impostors and target speakers mean, respectively. The Lawley-Hotelling statistic is defined as [47]:

$$U^{(s)} = \text{tr}(E^{-1}H) = \sum_{i=1}^{s} \lambda_i, \tag{5.9}$$

where E and H are the "between" and "within" matrices respectively, λ_i are the eigenvalues of $E^{-1}H$, and $s = \min\{p, k\}$, being p the number of variables and k the number of groups or classes. The main advantage of this test is that the multivariate information in E and H about separation of mean vectors is summarized into a single scale, on which we can determine if the separation of mean vectors is significant. We reject $\mathcal{H}_0$ for large values of $U^{(s)}$. This test is carried out by combining different systems in an incremental way, and separating the scores from normal and whispered speech. This allows us to better understand the effect that the addition of a particular system has in the separability of impostors and target speakers scores for each speaking style. The analysis is also carried out per gender and the results are summarized in Figure 5.7.

In Figure 5.7 (a), bars represent the $U^{(s)}$ measure for the combined systems, normalized by the max value because we are interested in the relative improvements from the baseline and not in the absolute value per se. In the plots, normal scores are in dark grey, whispered speech scores, in turn, are in light grey. Dashed lines represent the same measure for the baseline system, black for normal and grey for whispered speech. As can be seen, for all cases the addition of a new system should increase the separability for normal speech scores, but the same effect is not observed for whispered speech. When combining with the baseline MFCC system, the feature set that seems to be most beneficial for whispered speech is RMFCC; whereas AAMF(FS) seems to add more separability to normal speech scores. When the three proposed sets are combined, benefits occur for both vocal efforts. Lastly, maximal separation is seen to occur once the proposed features are combined with MFFCs. It is important to emphasize, however, that these results only give an idea of how each system contributes to the separation of impostors and target speakers scores, and the actual gains in SV accuracy still need to be calculated.

Lastly, in order to explore possible gender biases within these feature sets, Figures 5.7 (b) and (c) depict the same analysis, but separately for female and male speakers only, respectively. As can be seen, the overall behaviour seems to be independent of gender, but the gains over the baseline for

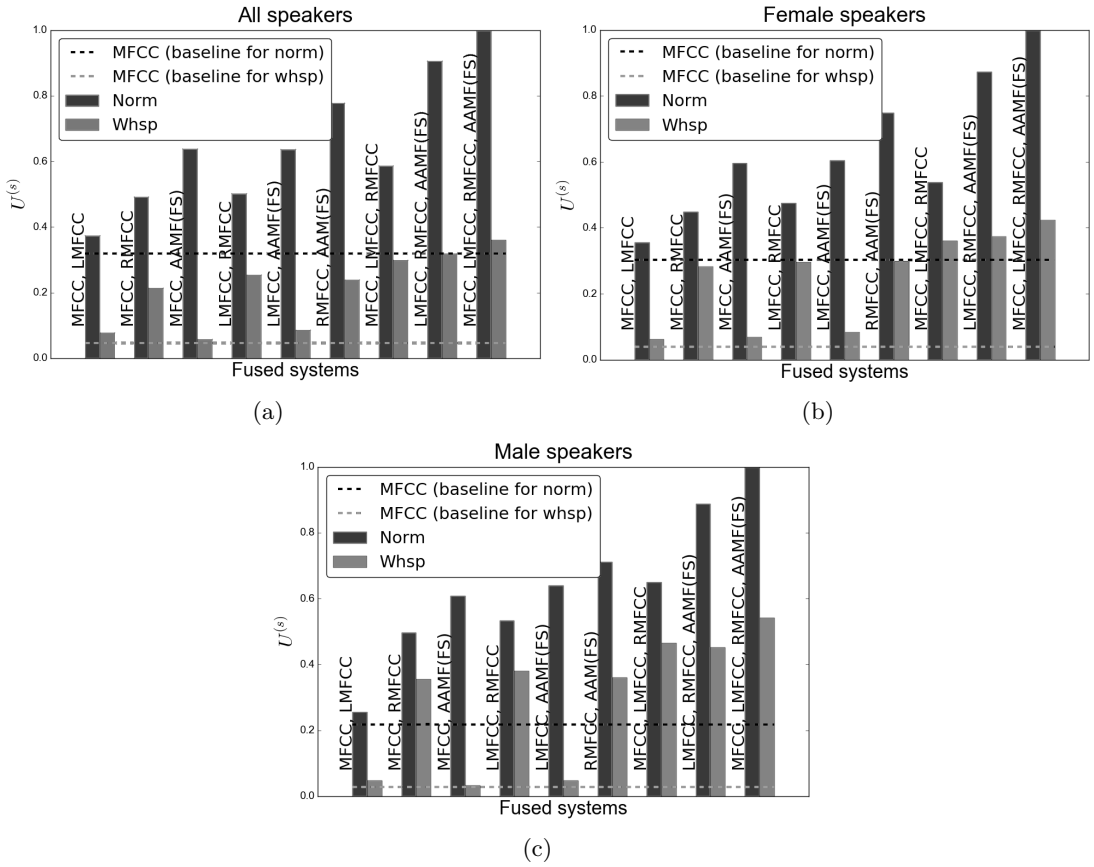


Figure 5.7 – Lawley-Hotelling statistic analysis using combination of different systems to explore contributions of individual feature sets. (a) Gender independent, (b) Female speakers and (c) Male speakers.

whispered speech were shown to be higher for males. This corroborates the that male speech is highly affected in whispered mode [73], thus most recent whispered speech speaker verification studies have relied solely on female speakers [26, 37, 125]. These observations suggest that the proposed features can improve whispered speech separability for target and impostor speakers particularly for male speech recordings. Notwithstanding, given the limitations in our available whispered speech datasets (e.g., gender imbalance with roughly twice as many male data points as female), gender-specific models are not explored herein and are left for a future study. As such, results in the following section refer only to gender independent systems.

Fusion level	**S5: S1 + S2 + S3**		**S6: S1+S2**		**S7: S1+S3**		**S8: S2+S3**	
	Norm	Whsp	Norm	Whsp	Norm	Whsp	Norm	Whsp
Score	2.50	13.49	2.81	15.63	2.81	15.83	2.50	**11.67**
i-vector	2.19	14.71	2.19	16.46	2.53	15.70	2.19	13.26
Frame	**1.88**	15.63	2.19	16.84	2.34	16.33	1.95	15.40

Table 5.3 – EER comparison for different fusion systems. For these experiments $C = 256$ and $T = 400$. Si: Sj+Sk represents the combination of feature set Sj and Sk according to the label in Table 5.2

5.5 Multi-style models trained with proposed feature sets

The results reported in Section 5.4 suggest that the proposed features encode information that is invariant across vocal efforts and that they carry information that is complementary to each other and to MFCCs. In Chapter 4, it was observed that system fusion was a reliable way of combining information from complementary feature sets, thus the same approach is explored here as well. More specifically, three fusion schemes are explored: *i*) frame level, *ii*) i-vector concatenation, and *iii*) score level, as depicted by Figures 4.6 (a), (b), and (c), respectively.

Table 5.3 reports EER values for different combinations of MFCC and proposed MFCC variants, for each of the three fusion strategies; lowest EER values are highlighted in bold per speaking style. As can be seen, fusion of conventional MFCC with either of the proposed feature sets (i.e., set **S6** and **S7** in the Table) showed improvements for both normal and whispered speech, relative to the **S1** results reported in Table 5.2, thus highlighting the complementarity of the proposed features to conventional ones. Notwithstanding, fusion of only the proposed MFCC variants (i.e., set **S8**) resulted in further gains, particularly for whispered speech, with score-level fusion achieving the lowest EER for whispered speaking mode. These findings corroborate those from the score analysis shown in Figure 5.7 (a). Lastly, fusion of all features (i.e., set **S5**), while it did not improve the performance for whispered speech, it did slightly lower the EER for normal speech when using the frame-level fusion strategy.

Relative EER improvements, when comparing best results with the standard MFCC/PLDA system presented in Table 5.2 (**S1**), are 19% and 43% for normal and whispered speech, respectively, with score-level fusion. With i-vector concatenation, in turn, 30% and 36% gains for normal and whispered speech are seen, respectively. Lastly, frame-level fusion resulted in relative improvements of 39% for normal speech and 26% for whispered speech. Comparing to the baseline results presented

Feature sets	Fusion level			
	Score		i-vector	
	Norm.	Whsp.	Norm.	Whsp.
S9: S2 + S4	1.44	12.64	1.28	14.17
S10: S3 + S4	0.94	13.40	0.94	15.69
S11: S2 + S3 + S4	**0.94**	**10.04**	1.20	12.14
S12: S1 + S2 + S3 + S4	1.25	12.43	0.94	13.22

Table 5.4 – Equal Error Rate (EER) comparison for different feature sets and fusion schemes under two testing conditions. For these results $C = 256$ and $T = 400$.

in Table 4.4, the relative improvement achieved with the fusion of the proposed features in the mismatch condition was of 57%.

By comparing the fusion schemes, it can be seen the one with best performance for whispered speech is score-level fusion, while the best for normal speech is frame-level fusion. For normal speech, these results are in agreement with most recent reports showing that feature concatenation to be the an effective strategy to improve speaker verification accuracy [22, 23]. i-vector concatenation, on the other hand, showed to be the scheme with a tradeoff between performance and computational burden, as additional fusion scheme training is not needed, as was the case with score level fusion. A major drawback of frame-level fusion is the need for synchronization of frame size and frame rate of the features being concatenated. This poses a challenge, for example, when exploring fusion of AAMFs with the MFCC variants. As a consequence, in the subsequent analyzes, score- and i-vector level fusion only are explored.

Table 5.4 reports EER values of the fusion of the two classes of proposed features. As can be seen, fusion of the AAMF set with either RMFCC (**S9**) or LMFCC (**S10**) resulted in gains in normal speech, relative to results reported in Table 5.3, but not for whispered speech. Gains were seen irrespective of the fusion strategy. Improvements for whispered speech were only seen when all three proposed feature sets were combined (set **S11**) and score-level fusion was used. Interestingly, the best results were achieved with set **S11**, i.e., without the inclusion of the baseline MFCCs. These findings contradict those of the theoretical score-level analysis in Figure 5.7 where fusion of all four feature sets indicated the best separability. This is likely due to the fact that the limited amount of data available to train the linear function for score fusion did not model the boundary found with the Lawlel-Hotelling analysis. Notwithstanding, the fusion analysis results from Table 5.4 follow the general tendencies observed in Figure 5.7 and show the proposed features extracting

complementary information from speech recordings, thus helping not only to reduce error rates when testing with whispered speech, but to also improve system performance for normal speech. These observations apply also for i-vector concatenation, where we can see that attained results are slightly better for normal speech than those achieved with fusion at score-level, but error rates for whispered speech are on the contrary slightly higher.

Overall, fusion of the systems using the proposed feature sets (**S11**) achieved relative improvements of 66% and 63% for normal and whispered speech, respectively, when comparing to the MFCC/PLDA system without whispered speech in the background set (Table 4.4), and 69% and 51% when comparing to **S1**, respectively, using fusion at score-level (Table 5.2). Overall, the proposed feature sets were capable of reducing the gap between normal and whispered speech from 17% (Table 4.4) to 9%. While this improvement is substantial, an EER of 10% for whispered speech is the equivalent EER of a state-of-the-art PLDA based system using multi-condition training to handle noisy and reverberant conditions [126]. Next, we will explore the use of these innovative features as input to BNF systems.

5.6 Conclusions

In this chapter we have described three innovative feature sets shown to provide invariant information across vocal efforts and complementary information to existing features for an SV task. The proposed features were built on insights obtained from previous chapters, as well as from those reported in the literature. Two variants of the MFCC were proposed, one focused on just the LP residual, thus emphasizing the similarities in unvoiced speech segments between the two vocal efforts, and the other on the 1.2-8 kHz subband shown to be less affected by whispering. Both MFCC variants were shown to provide complementary information to the classic MFCC and to provide gains as high as 39% and 41% for normal and whispered speech, respectively, relative to using just MFCCs. A third feature set was built on evidence from Chapter 4 showing that slowly varying subband envelopes conveyed useful information for cross vocal effort SV. By using the mutual information criterion, a binary mask was developed to select acoustic/modulation channels invariant to vocal effort changes. When all three features sets were combined, improvements of 66% and 63% over an MFCC-based baseline were achieved for normal and whispered speech, respectively. While the gap between normal and whispered speech EER was substantially reduced, the levels attained

for whispered speech can still be considered high at around 10% EER. As such, the next chapter explores the use of these newly proposed features as input to state-of-the-art deep neural network approaches.

Chapter 6

Deep Learning Approaches for Multi-Vocal Effort Speaker Verification

6.1 Preamble

Results presented in this chapter are also detailed in the paper #4 listed in Section 1.3 and is under preparation, to be submitted to to the journal *Speech communications* [63]. In this chapter, we focus attention on the extraction of invariant speaker-dependent information from normal and whispered speech using deep learning approaches and exploring the complementarity of these approaches with features proposed in Chapter 5.

6.2 Introduction

Existing state-of-the-art systems rely on the extraction of i-vectors [19], and most recent techniques have replaced the classical MFCC as acoustic features by approaches based on deep learning to extract the so-called bottleneck features (BNF). The robustness of these approaches, however, has not been tested under varying speaking styles such as whispered speech. In this chapter we aim to fill this gap. First, we explore a standard bottleneck neural network configuration with input consisting

of the classical log Mel-scale filterbank outputs. Next, we explore the use of the newly proposed features from Chapter 5 as alternate input modalities to the DNN. We evaluate how efficiently the proposed approaches handle the addition of whispered speech data from target speakers. Overall, it is found that different strategies result in optimal results for normal speech and whispered speech separately. Such findings suggest the need for a multi-model approach [25, 27], as depicted by Figure 2.12.

In order for these systems to work, whispered speech needs to be detected, such that the correct vocal-effort speaker models can be used for authentication. Whispered speech detection in *silent* environments has been proposed in the past. As examples, the ratio between the spectral energy in high- ($\geq$ 2.5 kHz) and low-frequency bands ($\leq$ 1 kHz) was explored by [127]. Alternately, spectral tilt, spectral flatness, and linear prediction analysis have been shown to be useful indicators of whispered speech [127, 128]. These measures, however, can be severely affected by ambient noise, as well as by pre-processing stages present in existing ASR and ASV systems, such as pre-emphasis filtering and/or power normalization. To overcome these limitations, more recent work has explored the use of entropy-based speech features [129], linear prediction analysis based on minimum variance distortionless response modelling of speech [128], and mel-frequency cepstral coefficients (MFCC) [27, 35]. Here, we explore the use of the auditory-inspired modulation features for detection of whispered speech. Lastly, to explore the noise robustness of the proposed features and developed systems, we explore the accuracy of such a multi-model approach in realistic settings involving different levels of ambient noise.

6.3 Exploring bottleneck feature representations

As described in Section 2.2.1, bottleneck features (BNF) are the current state-of-the-art paradigm for feature extraction in speaker verification systems. It uses a DNN trained to classify sub-phonetic units, known as "senones" which are generated by an ASR system. In our experiments, the targets for the DNN were obtained using a CD-GMM-HMM (context dependent - hidden Markov model using Gaussian mixture models to model observations) ASR trained with kaldi [48]. Training data corresponds to 460 hours extracted from the LibriSpeech dataset [49] (see Section 2.3). The DNN input features are concatenated time contexts of 15 frames, each frame is represented by 27 log Mel-scale filterbank outputs, using the same setting as in MFCC feature computation, which results in

a d=405 dimensional vector (27×15). In our case, K, the number of target labels defined by the output transcription file given by the ASR system, is set to 4121 senones. For the baseline experiments a DNN with five hidden layers is used as depicted by Figure 2.8, and we fix as the bottleneck layer the third hidden layer, and its number of units is $bnf = 80$; all these are typical values used in previous reports [21, 22, 23]. The DNN was trained using Theano [130].

Even though there is little evidence regarding how exactly the non linear operations in the DNN simulate the auditory system, the above described DNN architecture is chosen on the basis of recently presented evidence pointing towards the idea that the position of the bottleneck layer has to do with the task at hand and how similar the data used for training the DNN is to the evaluation set. For practical applications it is suggested to use the bottleneck layer close to the DNN output layer when DNN training data is matched to the evaluation conditions, and a layer more central to the DNN otherwise [50]; this latter setting represents the task at hand. Due to the limited amount of whispered speech available for DNN training, it is assumed that since humans can still recognize speakers in whispered mode [30], a DNN trained on normal speech, with a fairly central bottleneck layer, will capture vocal-effort invariant information useful for SV tasks. We will refer to the feature set based on mel filterbank outputs as filterbank bottleneck features (FBBNF).

In addition to this, in Section 5.5, we have shown that it is possible to extract invariant speaker-dependent information from normal and whispered speech using variants of the mel-frequency cepstral coefficients (MFCC). Using these insights, we propose to use a DNN architecture to extract bottleneck features using as input information related to these MFCC variants. In a similar setting as described above for the FBBNF feature set, we concatenated features from thirteen consecutive frames: *i*) Thirteen MFCC, these features are aimed at the original task to train the DNN, sub-phonetic units classification, *ii*) 27 log Mel-scale filterbank outputs, the triangular filters are spaced between 1.2kHz and 8kHz, and *iii*) 27 log Mel-scale filterbank outputs, extracted from the LP residual. The time context of thirteen frames was defined after an exploratory analysis, where this number was found to best tradeoff the time needed to train the DNN and overall system performance. The hypothesis in this case is that while the MFCCs contain useful phonetic information together with speaker dependent information important for normal speech related tasks, limited band and residual log-filterbank outputs contain mostly information related to speaker identity invariant across vocal efforts. The second and third feature sets, however, disregard important information useful to discriminate among phonetic sounds. Hence, the feature vectors used as input to

Feature set	UBM	Normal			Whispered		
		T matrix dimension					
		200	300	400	200	300	400
MFCC	128	3.78	3.69	3.36	20.99	21.11	20.83
	256	3.18	3.44	**3.13**	**20.00**	21.91	20.83
FBBNF	128	8.13	6.25	6.56	12.50	12.50	13.33
	256	6.88	6.25	**5.94**	14.19	12.14	**12.33**
LRBNF2	128	2.98	2.50	2.19	14.17	14.17	13.80
	256	2.35	2.37	**2.19**	13.74	12.89	**12.50**
LRBNF3	128	5.68	5.31	**4.38**	13.73	14.59	12.68
	256	5.31	5.00	4.39	12.97	13.33	**11.41**
LRBNF4	128	4.02	3.96	3.69	20.14	22.50	20.69
	256	3.44	3.44	**2.75**	**19.59**	21.96	20.00

Table 6.1 – Equal Error Rate (EER) comparison between MFCC, BNF and AAMF using different values for the number of Gaussian components in the UBM and T matrix dimension.

DNN are complementary to each other, and we expect the resulting feature vector in the bottleneck layer to be more informative than the FBBNF feature set described above. With this, the input to the DNN is a d=871 dimensional vector ($(13 + 27 + 27) \times 13$). In addition to this, motivated by [50], we also vary the location of the bottleneck layer, from layer two to layer four and will refer to these features as LRBNFi, with $2 \leq i \leq 4$, where i stands for the layer where the bottleneck is located, and LR stands for limited band and residual information.

We compared the performance of the four bottleneck SV systems using whispered speech from background speakers during T-matrix parameter estimation as described in Section 4.4. Results are reported in Table 6.1, where, for the sake of comparison, results using the standard MFCC/PLDA system have been included as well. For the UBM and T matrix, different number of Gaussians and dimensions were tested, i.e., $C = \{128, 256\}$ and $D = \{200, 300, 400\}$, respectively. Best results are highlighted in bold letters in the table per feature representation and per speaking style.

In a similar way as done for the standard MFCC/PLDA based system, these experiments help to quantify the effects of whispered speech on a multi-style SV system using bottleneck features, and their limitations in our specific task using short length utterances. First we compare the FBBNF feature set with the baseline system. As can be seen, when testing with normal speech, the standard MFCC based system outperforms the FBBNF based one. FBBNF features, on the other hand, are more robust against changes in vocal effort. Next, by evaluating the system with the proposed input to the DNN, and varying the bottleneck layer i.e., the LRBNFi feature sets,

we can observe that the two feature sets using the bottleneck layer closer to the input seems to perform better than the one closer to the output, which corroborates the observations in [50]; and this applies for both speaking styles. If we compare the feature sets extracted in the third layer, i.e., the FBBNF and LRBNF3 feature sets, it is clear that the architecture of the DNN is important, but the input used for parameter estimation plays an equally important role. In this case, the LRBNF3 feature set shows a tradeoff in performance between the two speaking styles, which also corroborates our hypothesis. When comparing the different LRBNFi feature sets, we can see that the scheme using the bottleneck in the second layer is the best configuration for normal speech. For whispered speech, on the other hand, the best results are achieved with the bottleneck in the third layer (EER = 11.41%), but with an absolute difference in EER values of just 1% relative to LRBNF2. Next, we explore the potential gains that fusion schemes can bring to systems based on DNNs and the proposed input features.

6.4 Fusion schemes using bottleneck features

In Chapter 5, it was shown that optimal results could be achieved by combining systems independently trained with AAMF(FS), LMFCC and RMFCC feature sets, for both score level fusion and i-vector concatenation. In Table 6.2, we show the results obtained with different fusion schemes (score-level and i-vector concatenation) based on the different proposed BNF features and the top features from Chapter 5. In the Table, best results have been highlighted in bold letters per speaking style and per fusion scheme.

As can be seen, for normal speech, fusion of BNF-based features and the proposed features from Chapter 5 resulted in the lowest EER, irrespective of the fusion strategy. Interestingly, the fusion of the LRBNF4 feature set with AAMF(FS) and with AAMF resulted in the lowest EER for i-vector concatenation and score-level fusion, respectively. Such findings show the complementarity of the bottleneck DNN setup with slowly varying envelope features. Overall, for normal speech, i-vector concatenation showed to be the best fusion strategy and achieved the lowest error rates (EER=0.63%). It is worth emphasizing that i-vector fusion does not require the training of a separate score fusion mapping, thus exhibits an interesting practical advantage.

Fusion Level	BNF features	Feature sets from Chapter 5							
		AAMF(FS)		AAMF		**S11**		**S13**	
		Norm	Whsp	Norm	Whsp	Norm	Whsp	Norm	Whsp
Score	FBBNF	1.88	10.83	2.19	**7.73**	1.65	10.00	1.88	8.33
	LRBNF2	1.25	12.50	1.00	11.02	1.02	11.20	0.98	9.97
	LRBNF3	1.07	11.54	1.19	10.51	0.88	10.00	0.94	9.17
	LRBNF4	0.96	16.39	**0.85**	10.66	0.94	11.94	0.94	9.17
i-vector	FBBNF	1.20	11.43	1.56	**8.74**	0.65	9.63	1.11	9.03
	LRBNF2	**0.63**	15.00	1.20	11.67	0.94	11.63	0.94	10.00
	LRBNF3	0.71	12.43	0.94	11.50	**0.63**	10.83	1.06	10.83
	LRBNF4	**0.63**	17.50	0.73	10.46	1.25	14.17	0.90	11.72

Table 6.2 – Equal Error Rate (EER) comparison between two fusion schemes for systems trained with FBBNF, LRBNFi, AAMF, AAMF(FS), LMFCC and RMFCC features. Columns labelled as Si represent fusion of systems trained with S11: AAMF(FS), LMFCC and RMFCC and S13: AAMF, LMFCC and RMFCC. For these results C=256 and D=400.

With whispered speech, fusion of LRBNFi with AAMF(FS), LMFCC and RMFCC did not result in improvements, relative to EER values reported in Table 5.4, thus suggesting that bottleneck features were not adding complementary information. As such, we explored the fusion with the entire AAMF feature set, i.e., prior to mutual-information based feature selection, as it has already been shown to be highly discriminative and to be more robust than standard MFCC (see Table 5.1). The hypothesis in this case is that by not eliminating channels before dimensionality reduction, we expect that additional and non redundant information to be included into the fusion schemes with bottleneck features. As a result, the fusion of FBBNF and AAMF features resulted in the lowest EER for whispered speech.

Overall, when comparing with previous results, we can see that for normal speech a relative gain of 79% was achieved when comparing with the standard MFCC based system presented in Table 6.1. This is achieved by concatenating i-vectors extracted from AAMF(FS) features with LRBNF2 or LRBNF4 feature sets. When comparing with fusion at score level of systems trained with AAMF(FS) and LMFCC feature sets, for whispered speech, on the other hand, a relative gain of 61% was achieved relative to the MFCC based system presented in Table 6.1. In this case, the best performance was attained by combining two systems trained with AAMF and FBBNF at the score level.

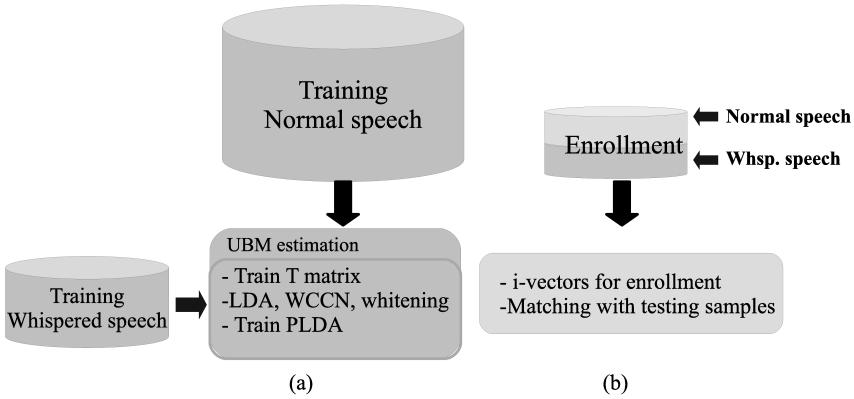


Figure 6.1 – Use of whispered speech data in the different stages of the speaker verification system. (a) Depicts the use of whispered speech recordings from a limited set background speakers, (b) depicts the combination of enrollment utterances from target speakers using both speaking styles.

6.5 Multi-style models with whispered during enrollment

Results presented up to now have only considered small amounts of whispered speech from background speakers for T-matrix estimation (as depicted by Figure 6.1(a)). Notwithstanding, whispered speech from target speakers could also be used during enrollment, as shown in Figure 6.1(b). Here, we explore the addition of small amounts of whispered speech from target speakers using the **S11** and **S12** feature sets described in Section 5.5.

Results are presented in Table 6.3. The column labeled *Number of Wshp. utterances in enrollment* represents the number of whispered speech utterances, each in average 4.5 seconds, from target speakers that were added during enrollment. The columns labeled as *Baseline*, represent the performance of the standard i-vector/PLDA based system with MFCC as feature vectors. As can be seen, for the two fusion schemes and using the **S11** feature set, only by adding one utterance the performance for whispered speech is already inline with the performance of the baseline system when using eight utterances. In addition to this, for each new utterance in the enrollment set there is a relative difference of about 50% between the baseline system and the proposed schemes. This shows that less data is needed to improve performance when testing with whispered speech, which clearly represents an advantage. The other aspect to highlight is the degradation in performance for normal speech as more utterances of whispered speech are present during enrollment. This is a problem that affects both the baseline and the proposed fusion schemes, but is more noticeable in the baseline system. Overall, the proposed fusion schemes keep the error rate below 2% for normal

Number of Whsp. utterances in enrollment	Baseline		Fusion level							
			Score				i-vector			
	S1		S11		S12		S11		S12	
	Norm.	Whsp.	Norm.	Whsp.	Norm.	Whsp.	Norm.	Whsp.	Norm.	Whsp.
0	3.13	20.83	0.94	10.04	1.25	12.43	1.20	12.14	0.94	13.22
1	3.03	17.14	0.97	8.33	1.25	10.40	1.25	8.95	1.01	9.76
2	3.44	14.17	1.02	7.50	1.25	9.07	1.25	6.97	1.25	8.33
3	3.75	13.07	1.16	6.51	1.56	7.50	1.23	5.83	1.25	7.37
4	4.23	11.56	1.14	5.72	1.64	6.59	1.25	5.21	1.25	5.99
5	4.69	10.80	1.25	4.61	1.56	5.45	1.53	4.98	1.25	5.00
6	4.87	9.58	1.56	4.47	1.56	4.86	1.56	4.17	1.25	4.24
7	5.31	8.92	1.56	3.63	1.80	3.71	1.88	4.17	1.47	4.17
8	5.31	8.25	1.61	3.33	**1.56**	3.33	1.88	4.32	1.61	**3.13**

Table 6.3 – Equal Error Rate (EER) comparison for different feature sets and the fusion systems under two *Training/Testing* conditions with varying amounts of whispered speech during enrollment. For these results $C = 256$ and $T = 400$. S1: MFCC, S11: AAMF(FS), RMFCC and LMFCC feature sets, S12: AAMF(FS), RMFCC, LMFCC and MFCC

speech, which is in fact better than the performance achieved by the initial MFCC/PLDA system without whispered speech in the background set (see Table 4.4). An additional and important aspect is that the final error rate achieved for *whispered* speech is closer to the performance achieved by the baseline system with normal speech, thus supporting the idea that within whispered speech there is as much discriminative information as in normal speech, as was suggested by the preliminary experiments performed in Chapter 3.

Lastly, by comparing the two fusion schemes, we can see that there are no significant differences between the two. Notwithstanding, the performance achieved with feature set **S12** is somewhat lower than the performance with **S11** for both speaking styles with fusion at score level, except when adding eight whispered speech utterances. With i-vector concatenation, on the contrary, best performance is achieved by using the feature set **S12**, i.e., concatenating i-vectors from the four feature sets, which coincides with the separability analysis shown in Figure 5.7. With score-level fusion, it is not necessary to include MFCC features into the fusion scheme, and the proposed feature sets are capable of handling both speaking styles. With i-vector concatenation, it is necessary to include four feature sets with the advantage that no fusion scheme is needed to be trained.

Next, we perform similar experiments using fusion schemes with bottleneck features. Two feature sets were used: *i*) Systems trained with FBBNF and AAMF feature sets, as this combination showed to be the best for whispered speech according to Table 6.2; we will refer to this set as **S14**, and *ii*) Systems trained with LRBNF3, LMFCC, RMFCC and AAMF(FS), which is a combination that not only shows the best performance for normal speech but also shows a competitive performance for

Number of Whsp. utterances in enrollment	**S12** from Table 6.3 with i-vector concatenation		Fusion level							
			Score				i-vector			
			S14		**S15**		**S14**		**S15**	
	Norm.	Whsp.	Norm.	Whsp.	Norm.	Whsp.	Norm.	Whsp.	Norm.	Whsp.
0	0.94	13.22	2.19	7.73	0.88	10.00	1.56	8.74	0.63	10.83
1	1.01	9.76	2.19	6.67	0.76	8.02	1.52	8.88	0.94	9.41
2	1.25	8.33	1.95	5.83	0.73	7.50	1.25	7.19	1.04	6.67
3	1.25	7.37	1.89	5.00	1.01	5.83	1.25	5.83	1.05	5.40
4	1.25	5.99	1.88	4.61	1.25	4.74	1.30	5.00	1.02	4.32
5	1.25	5.00	1.88	4.69	1.42	4.17	1.88	4.75	1.14	3.53
6	1.25	4.24	1.90	4.17	1.59	3.54	1.90	4.17	1.41	3.33
7	1.47	4.17	1.90	3.95	1.86	3.33	2.41	4.17	1.56	2.78
8	**1.61**	3.13	2.01	3.87	2.19	3.33	2.50	3.24	1.66	**2.35**

Table 6.4 – Equal Error Rate (EER) comparison between two fusion schemes for systems trained with different feature combinations. S12: MFCC, LMFCC, RMFCC and AAMF(FS), S14: FBBNF and AAMF, and S15: LRBNF3, LMFCC, RMFCC and AAMF(FS). With varying amounts of whispered speech during enrollment. For these results $C = 256$ and $T = 400$.

whispered speech independent of the fusion approach; we will refer to this setting as **S15**. Results are reported in Table 6.4.

As can be seen, addition of whispered speech enrollment utterances induce similar negative effects for normal speech speaker verification as reported in Table 6.3. At the same time, systems involving feature sets **S12** and **S15** better handle the addition of whispered speech data, and final error rates are below 2% for normal speech using i-vector concatenation. In particular, fusion schemes using systems trained with FBBNF and AAMF feature sets (**S14**) present competitive performance for both speaking styles, but are less efficient using the new information included by whispered speech utterances from target speakers. This contrasts with results in Table 6.2, where it was shown that in absence of whispered speech recordings from target speakers, these were the best feature sets to use in a fusion scheme.

It is important also to note that with five utterances of whispered speech in the erollment set, fusion schemes using **S15** have already better performance than **S12** and **S14** for both normal and whispered speech. These results show that the proposed feature sets, using insights from acoustic studies and mutual information, combined with a DNN architecture using also as input a feature vector aimed to carry invariant speaker-dependent information across vocal efforts, can handle both speaking styles in a multi vocal effort speaker verification task. However, when comparing best results achieved for normal speech in Table 6.2 and best results for whispered speech in Table 6.4, we can see that different strategies can offer different benefits for each speaking style. For example, best error rates when testing with normal speech were achieved when no whispered speech was

Number of Whsp. utterances in enrollment	Baseline System MFCC	Fusion level					
		Score			i-vector		
		S12	**S14**	**S15**	**S12**	**S14**	**S15**
3	11.51	3.31	5.40	3.33	5.83	9.17	5.16
4	11.67	2.71	5.71	2.98	5.66	8.60	5.79
5	10.00	2.50	5.00	2.50	5.83	8.33	5.00
6	10.71	2.70	5.04	2.50	5.72	8.33	4.31
7	9.84	2.85	4.76	2.84	5.01	8.09	4.18
8	10.83	2.50	5.00	2.50	4.27	7.43	3.33

Table 6.5 – EER comparison only for whispered speech among different systems using two fusion schemes with varying amounts of whispered speech during enrollment, normal speech recordings are not included for enrollment. S12: MFCC, RMFCC, LMFCC and AAMF(FS), S14: FBBNF and AAMF, and S15: LRBNF3, LMFCC, RMFCC and AAMF(FS). For these results $C = 256$ and $T = 400$.

added during the enrollment stage. Lowest error rates for whispered speech are achieved with a different scheme. This opens the possibility to implement dedicated systems for each speaking style, and combined with a normal/whisper speech classification system, to implement a multi-style type system, such as that shown in Figure 2.12.

Before exploring dedicated systems, we explore the performance of the same systems but enrolling and testing only with whispered speech, normal speech recordings were not included for enrollment and testing in this experiment. Results are presented in Table 6.5. In the Table, we have included the performance of the standard MFCC/PLDA based system as a baseline. According to these results, for some systems the performance of whispered speech is dependent on having normal speech recordings in the enrollment set as well. That is the case for the baseline system, and also for the fusion schemes using the **S13** feature sets. We refer for example to Table 6.3, where the baseline system achieved an EER=8.25% by combining eight normal and whispered speech utterances during enrollment. Similar behaviour is observed with **S13**; when combining both speaking styles during enrollment, the achieved error rates were 3.24% and 3.87% for i-vector concatenation and score level fusion, respectively. Regarding **S12** and **S15** feature sets, even though they are less dependent on having normal speech as well during enrollment, it can be seen in Table 6.4 that the lowest error rate is achieved with **S15** feature sets, using i-vector concatenation and including normal speech during enrollment; the differences are minimal nevertheless.

Results in Table 6.5 can be explained by the fact that the features that better perform in this scenario, are features aimed to extract invariant information from both speaking styles. By not enrolling normal speech, we are not including important variability that is shared by normal and whispered speech, which is better modelled in the T-matrix, post-processing stage of i-vectors

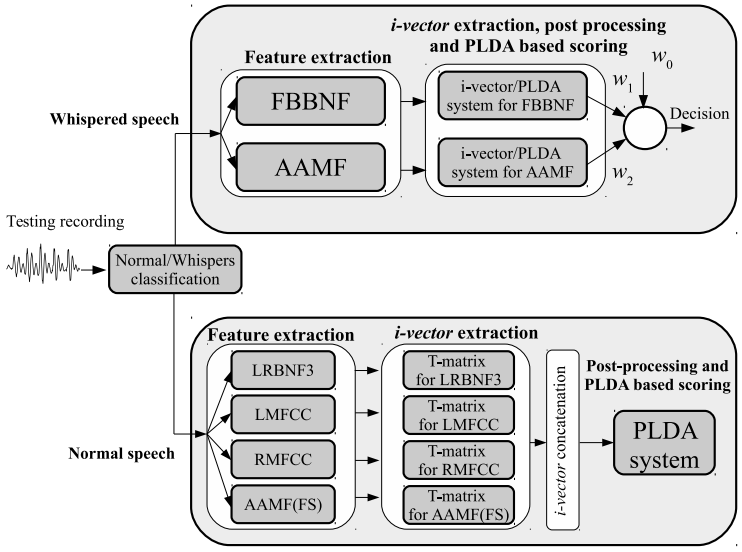


Figure 6.2 – Building blocks for a speaker verification system using a normal/whispers classification system in the input to select the best system. In this case, it is assumed that there are not whispered speech recordings from target speakers.

and the PLDA model and has been learned from the background speakers. With this, the optimum system for multi vocal effort speaker verification involving normal and whispered speech is as shown in Figure 6.2, when there are no whispered speech recordings for enrollment. Figure 6.2, on the other hand, illustrates the case when there are whispered speech recordings from target speakers. In this later case, to achieve optimal performance for both speaking styles, the best scoring strategy must be selected according to the testing recording, i.e., if it is normal speech, then, only normal speech is used for enrollment, if it is whispered speech, then, a combination of normal and whispered speech is used for enrollment. In order for such systems to work, however, accurate normal/whispered speech detection is needed. One such proposed system is described next.

6.6 Normal/Whispered speech classification

According to results presented in previous the section, it is evident the need for a normal/whispered speech classification system. Even though Figure 6.2 does not make it explicit, Figure 6.3 shows that such a classification system may need different features specially designed for the task at hand. This section describes the processing steps involved for normal/whisper speech classification. First, we explore a classification system in clean conditions using *i-vectors* extracted from all feature sets described in previous sections. This is done using a standard classification system depicted by

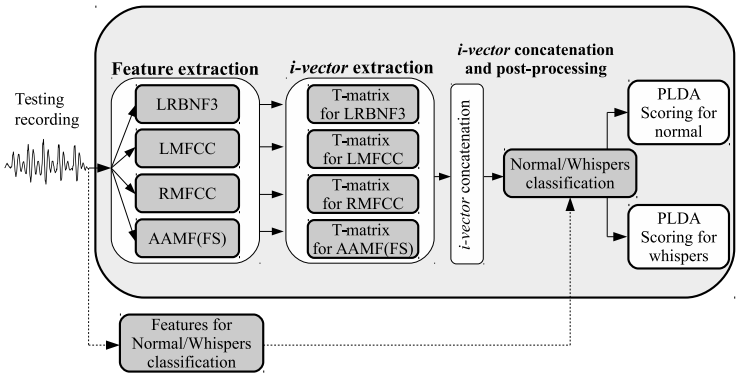


Figure 6.3 – Building blocks for a speaker verification system using a normal/whispers classification system to select the best scoring strategy. In this case, it is assumed that there are whispered speech recordings from target speakers.

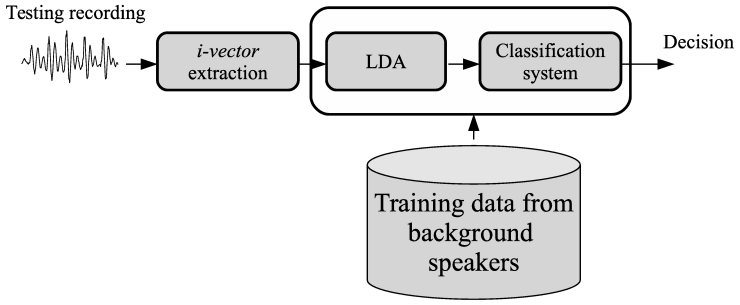


Figure 6.4 – Building blocks for a normal/whispered speech classification system using *i-vectors*.

Figure 6.4. First, i-vectors are pre-processed by using linear discriminant analysis (LDA), then labels are assigned to the testing recordings using a linear classification system. The LDA and the classification system use data from background speakers for parameter estimation and results are presented in the top row of Table 6.6.

As can be seen, the proposed auditory-inspired amplitude modulation features with or without feature selection can offer perfect discrimination between the two speaking styles. Hence, the systems illustrated in Figures 6.2 and 6.3 are completely feasible. However, in practical systems noisy settings are expected, thus the robustness against noise also needs to be evaluated. To this end, testing recordings were contaminated using three different types of noise to simulate possible testing scenarios, namely babble [131], office, and subway [132], at four different signal to noise

SNR	MFCC	RMFCC	LMFCC	AAMF	AAMF(FS)	FBBNF	LRBNF2	LRBNF3	LRBNF4
∞	97.50	99.55	85.45	100	100	97.95	99.77	99.77	99.32
Babble									
15	75.23	80.45	73.64	74.55	88.18	98.64	74.09	85.45	77.73
10	58.18	74.09	68.64	35.45	72.50	94.77	46.82	57.95	65.91
5	40.45	63.18	58.41	27.27	40.45	74.55	29.77	29.32	50.68
0	30.68	47.50	48.41	27.27	28.41	37.95	27.50	27.27	37.73
Office									
15	99.09	84.32	85.45	100	100	97.95	99.55	99.77	99.09
10	99.55	79.77	82.50	100	99.77	98.86	99.55	99.77	98.86
5	99.55	75.68	82.27	100	99.55	98.86	99.32	99.55	97.50
0	99.55	70.68	73.86	99.09	99.32	98.64	99.09	99.55	96.36
Subway									
15	95.68	88.64	83.64	100	98.64	99.09	95.91	99.32	93.86
10	92.95	86.36	76.59	94.32	93.18	97.50	88.41	93.64	87.95
5	84.09	72.50	65.91	69.09	80.91	94.09	70.68	80.68	79.09
0	72.95	60.00	56.82	35.68	53.86	82.95	53.18	57.27	65.68

Table 6.6 – Accuracy (%) comparison among different i-vector based normal/whispered speech classification. Testing recordings have been contaminated with three different kinds of noise at four different signal to noise ratio (SNR) levels.

ratio (SNR) levels: 15, 10, 5 and 0 dB. As can be seen, from the remainder of Table 6.6, babble and subway noise can heavily affect the performance of the classification system, while office noise only affects certain feature sets. Overall, no feature set stood out as being robust against all types and levels of noise. Babble noise, in fact, showed to be the most detrimental to normal/whispered speech detection, thus the remainder of this section will focus on this type of noise.

We explored this task in an early publication. A more detailed set of experiments were carried out using only the CHAINS corpus, as it contains continuous speech recordings long enough that allows to block the speech recording in several short duration segments and validate the proposed scheme. We used the speech stimuli generated from reading the first paragraph of the *Rainbow Text* for training of the classifiers (average duration: 30 seconds; minimum duration: 23 seconds), and kept the stimuli generated from reading the *Cinderella story* for testing (average duration: 55 seconds, minimum duration: 48 seconds). Lastly, in order to generate the noisy speech stimuli, babble noise was added at three different signal-to-noise ratios (SNR): 0, 5, and 10 dB.

6.6.1 Robust features for Normal/Whispered speech classification

Prior to feature extraction, each speech recording was down-sampled to 16 kHz, normalized to -26 dBov (dB overload) and pre-emphasized using a first order finite impulse response filter with

constant $a = 0.97$. Feature vectors were extracted per speech frame and three feature representations were explored, we will refer to these feature sets as 'NWi' as they are used for Normal/Whispered speech classification:

RASTA-PLP features: In our experiments, 19 relative spectral perceptual linear prediction (RASTA-PLP) coefficients were computed using 24 Bark-scale triangular bandpass filters. Delta and double-delta coefficients were also included to convey temporal dynamics information. Preliminary experiments showed RASTA-PLP coefficients to be more robust to noise than conventional MFCCs. A detailed description of RASTA-PLP is beyond the scope of this thesis. This feature set is termed 'NW1'.

Entropy-based features: In order to compute entropy-based features, the methodology described in [129] was used. The feature vector is composed by three spectral information entropy based features, namely he high-to-low entropy ratio (HLER), the low-frequency band entropy within the band $B = [300, 4150]Hz$ and the high-frequency band entropy within the band $B = [4150, 8000]Hz$. A fourth feature related to spectral tilt was also included. This feature set is termed 'NW2'. When using feature combination for 'NW1' and NW2', the resulting feature set is termed 'NW3', thus yielding a 61-dimensional feature set

Auditory-inspired amplitude modulation features: In a similar setting as the one described in Section 5.3.2, the AAMF features were used in these experiments. More specifically, 24 acoustic subbands and eight modulation subbands were used. Features were extracted from the first six modulation frequency bands, thus corresponding to the $0.1 - 24$ Hz modulation frequency range. The speech modulation spectrum results in a high-dimensional feature representation (e.g., 24 acoustic bands $\times 6$ modulation bands= 144 dimensions). Next, principal components analysis (PCA) was used to reduce the 144-dimensional feature set to 55 dimensions. Moreover, since speech-like noise (babble) has also been shown to affect the $0.2 - 20$ Hz modulation frequency band [133], we further calculate a so-called modulation spectral tilt parameter, per modulation frequency band, based on a minimum mean-squared error linear fitting across the 24 acoustic frequency bands. Since the majority of the reported spectral differences between naturally-phonated and whispered speech occur above 1 kHz acoustic frequency [129], the tilt parameter is computed only within this range. The the modulation spectral tilt parameters alone are termed 'NW4', and 'NW5' to

Classifier Type	**Feature set**				
	NW1	NW2	NW3	NW4	**NW5**
Linear	70.9	74.8	75.9	75.2	**99.2**
Quadratic	72.8	73.5	76.8	76.1	**99.1**
GMM (C=2)	79.8	91.2	92.7	80.3	**97.8**
GMM (C=5)	86.2	96.8	97.1	86.5	**99.7**
GMM (C=10)	95.5	98.3	98.2	89.1	**99.9**
GMM (C=20)	95.9	98.7	98.6	89.3	**99.9**
SNR level	Experiments in Noisy Environment				
10 dB	62.0	73.2	67.4	81.3	**99.4**
5 dB	57.5	69.8	65.6	76.3	**88.4**
0 dB	54.3	62.5	63.6	76.1	**84.8**

Table 6.7 – Accuracy results and performance comparison in clean conditions for normal/whispered speech classification among different classification algorithms (top) and noisy environment with different SNR levels and using a GMM based classifier (bottom).

their combination with the PCA-reduced modulation frequency features, thus also resulting in a 61-dimensional feature set.

In order to investigate the performance of each individual feature, our experimental evaluation used a fixed duration of 23 seconds for *clean* normal and whispered speech training data; test data was fixed to one second duration segments. The top part of Table 6.7 presents the obtained results for the linear and quadratic discriminant function-based classifiers, as well as GMM classifiers with varying number of components C. As can be seen, for all feature sets the GMM based classifier outperforms the linear and quadratic classifiers. Relative to only the GMM-based classifiers, it can also be observed that the performance increases as C increases, with $C = 10$ (referred to as GMM-10) showing a tradeoff between accuracy and complexity. Regarding the benchmark features, RASTA-PLP outperformed entropy-based features across all tested conditions. Moreover, no significant gains in performance were achieved with NW3, suggesting potential redundancy of information in RASTA-PLP and entropy features for clean whispered speech. Regarding the proposed features, the modulation spectrum tilt parameter (NW4) was shown to achieve performance levels in line with the entropy measures (NW1); when combined with the modulation spectrum based features (NW5), improved performance was obtained across all tested conditions. It is also important to emphasize that for NW5, a linear discriminant function classifier for 1-second duration test utterances achieved comparable performance with a more complex 20-component GMM classifier. This finding suggests that the obtained performance figures are attributed mostly to the information-bearing advantages of the modulation spectrum features and not the complexity of the classifier.

As mentioned previously, the GMM-10 classifier showed to strike a balance between performance and complexity, thus it is used in our experiments involving ambient noise conditions. In Table 6.7 these results are labeled as "Experiments in Noisy Environment". As can be seen, detection performance in mismatch conditions decreased for all feature representations as the SNR decreased. The two benchmark feature representations achieved similar performance figures and no improvements in performance were observed with the combined feature set NW3, thus suggesting the sensitivity of the benchmark features to ambient noise. The proposed modulation spectral based features, on the other hand, were shown to be more robust to environment noise. The modulation tilt parameters alone (NW4) consistently outperformed the more complex RASTA-PLP-based features, as well as the combined RASTA-PLP-entropy feature set NW3. When combined with the PCA-based modulation spectral features, a gain of approximately 33% could be achieved at an SNR of 0 dB over the benchmark NW3 feature set, thus showing the noise-robustness properties of the proposed features.

Detailed results of these experiments were published in [60] using test segments of different duration.

6.7 Speaker verification in noisy conditions

Though the focus of this dissertation is not robustness of speaker verification systems in noisy environments, in the following experiments we evaluate the robustness of the different feature sets against environmental noise, thus exploring the potential of the developed tools for everyday usage. In this case, we compare the fusion schemes by concatenating i-vectors and at the score level of three feature sets: *i*) **S13** and **S14**, which were the feature sets that showed to perform better when only whispered speech recordings from a limited set of background speakers is available and when recordings from target speakers are included during erollment, respectively, and *ii*) **S9**, LMFCC and AAMF(FS), which, according to our experiments, was the set that showed to be the best choice in presence of environmental noise. Results are presented in Tables 6.8 and 6.9 for i-vector concatenation and score level fusion, respectively.

In the Tables, the best results have been highlighted in bold per speaking style, SNR level, and noise type. As can be seen, when using i-vector concatenation, the fusion of systems based on LMFCC and AAMF(FS) (**S10**) seems to be the most robust to varying types and levels of noise,

SNR Level	Feature set					
	S10		**S14**		**S15**	
	Norm	Whsp	Norm	Whsp	Norm	Whsp
Babble						
15	**4.71**	**6.37**	15.96	11.81	12.19	8.83
10	**9.06**	**9.43**	26.45	15.31	19.68	12.65
5	**17.73**	**12.50**	36.40	19.59	28.51	17.76
0	**29.06**	**15.00**	42.04	21.21	36.49	20.83
Office						
15	**2.19**	3.30	3.44	5.19	2.23	**2.59**
10	2.50	**3.23**	4.36	5.55	**2.44**	3.33
5	3.14	**3.18**	5.69	5.00	**2.26**	4.17
0	4.44	**5.00**	7.81	6.03	**4.38**	5.83
Subway						
15	**3.04**	**5.14**	6.56	7.50	4.79	6.67
10	**5.00**	**5.83**	10.80	8.46	8.75	7.74
5	**9.42**	**8.33**	17.81	10.83	14.38	10.00
0	**16.88**	10.69	26.92	13.88	22.67	**10.61**

Table 6.8 – EER comparison among three i-vector based speaker verification systems using i-vector concatenation. Testing recordings have been contaminated with three different kinds of noise at four different signal to noise ratio (SNR) levels. In the table S10: LMFCC and AAMF(FS), S14: FBBNF and AAMF, and S15: LRBNF3, LMFCC, RMFCC and AAMF(FS).

SNR Level	Feature set					
	S10		**S14**		**S15**	
	Norm	Whsp	Norm	Whsp	Norm	Whsp
Babble						
15	**4.76**	**5.26**	8.52	8.29	5.94	5.37
10	**7.65**	**7.40**	18.03	11.54	11.95	8.65
5	**13.98**	**11.67**	29.24	16.54	20.31	15.29
0	**25.00**	**14.17**	38.35	19.17	29.28	17.63
Office						
15	2.42	**3.33**	2.50	3.92	**2.19**	4.17
10	2.64	**3.20**	3.13	4.06	**2.26**	4.17
5	3.13	3.98	3.75	**3.72**	**2.84**	5.00
0	4.69	4.87	5.00	**3.97**	**3.75**	5.35
Subway						
15	3.31	4.17	4.69	5.58	**3.13**	**4.10**
10	**4.99**	5.83	6.74	6.67	5.09	**5.83**
5	**7.54**	**7.50**	12.50	9.17	9.69	7.78
0	**13.36**	**10.00**	20.06	13.33	16.41	10.15

Table 6.9 – EER comparison among three i-vector based speaker verification systems using score level fusion. Testing recordings have been contaminated with three different kinds of noise at four different signal to noise ratio (SNR) levels. In the table S108: LMFCC + AAMF(FS), S14: FBBNF and AAMF, and S15: LRBNF3, LMFCC, RMFCC and AAMF(FS)

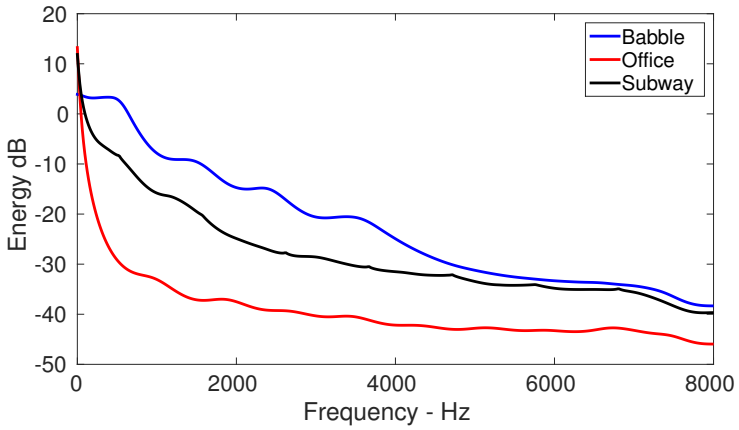


Figure 6.5 – Average power spectrum of the three different types of noise added during testing stage.

thus highlighting their importance for multi vocal effort speaker verification in realistic settings. **S15**, in turn, only showed robustness agains office noise in the score-level fusion scheme. To better explain these results, Figure 6.5 depicts the average power spectrum of the three types of noise. As can be seen, for all three cases, most of the energy is concentrated at low frequencies, typically below 1 kHz. Recall from Chapter 5 that the LMFCC and AAMF(FS) feature sets suppress this frequency band, thus not only introducing cross vocal effort robustness for speaker verification, but also noise robustness for realistic applications.

6.8 Conclusions

In this chapter, we have addressed the problem of finding invariant speaker dependent information across vocal efforts using deep neural networks. In addition to this, we continued exploring the benefits of two fusion schemes (score-level and i-vector level) to overcome existing challenges namely: *i*) Short duration utterances (4.5 seconds average), *ii*) No whispered speech data available during enrollment from target speakers, and *iii*) the negative effects seen when adding whispered speech recordings during enrollment.

In previous works it has been shown that the performance of speaker verification systems are strongly dependent on the condition of the speech material provided as input [51]. While characterizing the baseline systems (Table 6.1) for normal speech, it became evident that MFCC and standard BNF features achieved EER figures higher than what is typically reported in the literature [134, 22, 23]. This is likely due to the short speech duration which limits the phonetic variability

present in the training set [52, 53]. Notwithstanding, the proposed LRBNFi features seem to reduce such negative effect, in particular when the bottleneck layer is closer to the input. For whispered speech, on the other hand, the BNF features outperform all previously studied individual feature sets by a substantial amount. It is hypothesized that this was due to the superior capabilities of the DNN to model the invariant information when comparing normal and whispered speech. Overall, if only a baseline system was to be used, the BNF based one, using as input the combination of MFCC, residual and limited band log-filterbank outputs, and a bottleneck layer closer to the input (LRBNF2), resulted in the best overall multi vocal effort performance.

Fusion strategies were shown to clearly have many advantages, as they not only reduced error rates when there were no whispered speech recordings from target speakers for enrollment, but also helped to reduce the observed negative impact of adding whispered speech during parameter estimation. Overall, the proposed AAMF feature set was shown to be the most informative for both normal speech and whispered speech, to be used in a fusion scheme with bottleneck features when there is no whispered speech data from target speakers. This configuration, however, does not show the best results when whispered speech recordings from target speakers were included, and alternative features such as LRBNF3, LMFCC, RMFCC and AAMF(FS), were shown to be a better choice in a multi vocal effort speaker verification task. When comparing the fusion schemes, the i-vector concatenation scheme shows to be the best fusion strategy to be used. This is not only justified by the attained results, but also by the fact that it is not necessary to train any additional fusion system, as is the case in score level fusion. Overall, with the proposed fusion scheme, it is shown that only 4.5 seconds (aprox.) of whispered enrollment data is needed to achieve the same performance as the baseline system, which in turn, required 22.5 seconds (aprox.) of whispered enrollment data. Hence, the proposed features are well posed to handle vocal effort variations and low resource speaker verification tasks.

As was observed, different strategies offered different benefits for each speaking style. In such cases, dedicated systems per vocal effort offer a promising solution. To this end, a normal/whispered speech classification system needs to be implemented. This was explored in clean and noisy conditions. It was found that while the features used for cross-vocal effort speaker verification could provide accurate whispered speech detection in clean conditions, alternate feature representations were needed for noisy settings. Overall, a system based on FBBNF features was able to achieve clas-

sification accuracy as high as 98% at an SNR of 15 dB for babble noise, but performance degrades quickly as SNR diminishes. Similar behavior was observed for AAMF(FS) features.

Chapter 7

Conclusions and Future Research Directions

In this chapter, a general discussion for this doctoral thesis is presented, moreover some suggestions for future research directions are also proposed.

7.1 Conclusions

This doctoral thesis has addressed the important, yet not sufficiently explored, problem of speaker verification in multi vocal effort testing scenarios. In particular, we have centered the attention on two speaking styles, i.e., normal and whispered speech. We found that whispered speech can contain as much speaker specific information as normal speech, but standard approaches designed for normal speech tend to fail for whispered speech. In this regard, we have developed strategies to incorporate this speaking style into the possible testing scenarios of standard speaker verification systems. These strategies allow to efficiently use the limited resources available to overcome existing challenges namely: *i*) Short duration utterances (4.5 seconds average), *ii*) No whispered speech data available during enrollment from target speakers, and *iii*) the negative effects seen when adding whispered speech recordings during enrollment. By addressing these problems, the proposed SV system configurations were shown to achieve high performance levels for whispered speech inline with the performance obtained for normal speech.

In Chapter 3 the speaker verification (SV) task based on whispered speech recordings was addressed in an ideal scenario, using a limited number of speakers, and a closed-set scheme for speaker verification by using speech recordings from target speakers also for parameter estimation. For the experimental setup and given the limitations in number of speakers, the performance bounds of a standard GMM–UBM SV system using MAP adaptation were explored. To this end, the effectiveness of several strategies, such as frequency warping, sub-band analysis, alternate feature representations, feature combination, as well as class-dependent modeling (i.e., speaking-style) were evaluated. According to these preliminary experiments mismatch *train/test* conditions can highly affect the performance of a SV system, independent of the feature representation used. As in previous studies in adjacent areas, it was shown that in order for a SV system to handle both normal and whispered speech for practical applications, speaker model training had to involve data of both vocal efforts. Such approach, however, resulted in poorer verification performance for normal speech. Overall, feature representations evaluated here have been mainly proposed for normal-voiced speech applications, thus suggesting that alternate feature representations, tuned for whispered speech speaker verification, are still needed.

In Chapter 4 the issue of speaker verification based on whispered speech was addressed in a more realistic scenario. Three databases were pooled together in order to increase the number of speakers and add more flexibility to the experimental evaluation. In addition to this, and following results presented in Chapter 3, experiments in this chapter were carried out using specific feature representations such as the classical MFCC and WIF as they showed a good tradeoff between performance in matched and mismatched conditions for both speaking styles. Overall, we observed that existing features (e.g. MFCC) do not convey sufficient reliable speaker identity information across different vocal efforts. Given the lack of sufficient speakers to train independent and dedicated models for whispered speech, the multi-style modelling approach was explored. The addition of whispered data during training was shown to not suffice to boost speaker verification performance for whispered speech. As an alternative, we explored techniques such as feature mapping and complementary information extracted from WIF and MFCC feature sets via three fusion schemes, namely: *i*) frame level, *ii*) i-vector concatenation, and *iii*) score level. As a result, feature mapping approaches seem to be insufficient to improve performance and fusion schemes seem to be more effective. In this latter scenario, gains as high as 42% and 44% were obtained for whispered and normal speech, respectively, relative to a baseline system based on i-vectors/PLDA+MFCC with no whispered

speech in the training set. These findings suggest that innovative features conveying more speaker-dependent invariant information across different vocal efforts are needed as the mismatch problem is still present and the gap in performance is still considerable between normal and whispered speech.

In Chapter 5 we centered attention on the extraction of invariant speaker-dependent information from normal and whispered speech, thus allowing for improved multi-vocal effort speaker verification. To this end, three innovative feature sets were described. These feature sets were shown to provide invariant information across vocal efforts and complementary information to existing features for an SV task. More specifically, two variants of the MFCC were proposed, one focused on just the LP residual, and the other on the 1.2-8 kHz subband shown to be less affected by whispering. A third feature set was built on evidence from Chapter 4 showing that slowly varying subband envelopes conveyed useful information for cross vocal effort SV. By using the mutual information criterion, a binary mask was developed to select acoustic/modulation channels invariant to vocal effort changes. Prior to SV system evaluation, complementarity of the proposed feature sets was evaluated by means of the Lawley-Hotelling statistic in the score domain, thus allowing to better understand the contributions and complementarity of each newly proposed feature set and which combinations could achieve better performance when testing the SV system. Final results showed improvements of 66% and 63% over an MFCC-based baseline for normal and whispered speech, respectively, when all three proposed features sets were combined. While the gap between normal and whispered speech EER was substantially reduced, the levels attained for whispered speech can still be considered high at around 10% EER.

In Chapter 6, we continued exploring speaker-dependent invariant information across vocal efforts using deep learning approaches together with the benefits of fusion schemes. Bottleneck features were shown to add robustness to the SV system not only when facing vocal effort variations, but also to reduce previously mentioned negative effects when combining data from two speaking styles. This is likely due to the superior capabilities of the DNN to model the invariant information when comparing normal and whispered speech. Overall, if only a baseline system was to be used, the BNF based one, using as input the combination of MFCC, residual and limited band log-filterbank outputs, and a bottleneck layer closer to the input, resulted in the best overall multi vocal effort performance. When combining bottleneck features with feature sets proposed in Chapter 5, additional gains could be seen, as not only reduced error rates were achieved for both speaking styles, but also helped to reduce the observed negative impact of adding whispered speech during parameter

estimation. Overall, the proposed auditory-inspired amplitude modulation features were shown to be the most informative for both normal speech and whispered speech, to be used in a fusion scheme with bottleneck features when there is no whispered speech data from target speakers. On the other hand, fusion of MFCC variants, bottleneck and auditory-inspired amplitude modulation features was shown to be the best choice when combining normal and whispered speech recordings from target speakers during enrollment. Finally, when comparing the fusion schemes, the i-vector concatenation scheme shows to be the best fusion strategy to be used. This is not only justified by the attained results, but also by the fact that it is not necessary to train any additional fusion system.

7.2 Future Research Directions

1. *Incorporate modulation spectrum based features to DNN approaches*: Given the promising results obtained with modulation spectrum based features, one step forward to take advantage of this signal representation is to use the capabilities that deep neural networks have shown to extract highly discriminative features. This is motivated by results obtained with variants of MFCC extracted using limited band and residual log filterbank outputs, which achieved improved performance in fusion schemes and it was possible to achieve additional improvements when incorporating bottleneck features extracted with the same information. This however is not a trivial task, as it is necessary to devise a scheme to synchronize the long time contexts used for modulation spectrum signal representation with the transcriptions given by standard automatic speech recognition systems.
2. *Extend research in fusion schemes*: Fusion schemes have shown to be an efficient and effective way to incorporate the strengths of different feature representations into one single system. Fusion schemes explored in this thesis can be considered simple but effective solutions for the task at hand, that allow us to better interpret our results knowing that the load of the improvements were not in the machine learning algorithms but in the feature extraction process. Once the understanding on how we can extract invariant information across vocal efforts advances, we can also advance in more elaborate techniques to explore how to better combine complementary information extracted from different feature representations. As future research direction, more complex data-driven or machine learning based fusion systems

may be explored, including the use of DNNs to fuse information from multiple features. However, more whispered speech data will be needed before this can be accomplished.

3. *Explore the potential of modulation spectral features in noisy environments*: Not only in this thesis, but also in previous work, it has been shown that modulation spectral based features can add robustness to speaker recognition systems [43]. In particular, in this work we have explored a technique to find speaker dependent information invariant across vocal efforts using mutual information analysis. Resulting features not only showed to be highly discriminative, but also to make fusion schemes more robust against coloured noise. New analysis techniques can be incorporated in order to extract information that is not affected by noise and still highly important for speaker recognition tasks, such as the work reported in [43] for normal speech speaker identification in reverberant environments.
4. *Explore gender dependencies specifically for whispered speech*: In our experiments we found a strong gender dependency. For example, for normal speech the feature representations that performed best for male speech did not perform at the same level for female speech, thus corroborating previous findings [56, 57]. When exploring gender and speaking dependent models, we found that whispered speech speaker verification performance was higher for female speakers. This suggests that female whispered speech carries more speaker-specific information that is captured by the investigated features. In fact, most of the recent published research in the field has been done only with females [26, 37], thus making the improvements seem more noticeable. Nevertheless, given the lack of sufficient data, we could not perform additional and more detailed experiments. It is necessary to collect more whispered speech data from female speakers to allow further advances in this direction.
5. *Explore applications to other vocal efforts*: Here in, we have focus only on the problem of speaker verification using whispered speech. However it is important also to explore how to extract invariant information between normal speech and other vocal efforts such as *shouted speech*. There are some preliminary experiments reported in the literature using shouted speech [13] for speaker identification, but just as whispered speech, there are not enough studies and additional research into this field is still needed.

Bibliography

[1] "Speech technology: A global strategic business report," Global Industry Analysts, Inc., Tech. Rep., 2012.

[2] J. Bonneau, C. Herley, P. Oorschot, and F. Stajano, "The quest to replace passwords: A framework for comparative evaluation of web authentication schemes," in *Proc. IEEE Symposium on Security and Privacy*, May 2012, pp. 553–567.

[3] A. Slomovic, "Privacy issues in identity verification," *IEEE Security Privacy*, vol. 12, no. 3, pp. 71–73, May 2014.

[4] F. Towhidi, A. Azizah, M. Salwani, and L. Habibi, "The knowledge based authentication attack," in *Proc. International Conference On Security & Management*, 2011, pp. 1–5.

[5] S. Sahu and A. Singh, "Survey on various techniques of user authentication and graphical password," *International Journal of Computer Trends and Technology*, vol. 16, no. 3, pp. 553 – 560, 2014.

[6] J. Unar, W. Chaw Seng, and A. Abbasi, "A review of biometric technology along with trends and prospects," *Pattern Recognition*, vol. 47, no. 8, pp. 2673 – 2688, 2014.

[7] K. Saeed, "A note on problems with biometrics methodologies," in *Proc. ICBAKE*, Sept 2011, pp. 20–22.

[8] R. O'Neil King, "Speech and voice recognition white paper," Biometrics Research Group, Inc., Tech. Rep., May 2014.

[9] Research and Markets, "Global mobile biometrics market 2015-2019," Research and Markets, Tech. Rep., 2015.

[10] J. Ming, T. Hazen, J. Glass, and D. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Audio, Speech, Language Process.*, vol. 15, no. 5, pp. 1711–1723, July 2007.

[11] K. Rao and S. Sarkar, *Robust Speaker Recognition in Noisy Environments*, ser. SpringerBriefs in Speech Technology. Springer International Publishing, 2014.

[12] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 16, no. 6, pp. 1097–1111, August 2008.

[13] C. Hanilci, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, and F. Ertas, "Speaker identification from shouted speech: Analysis and compensation," in *Proc. ICASSP*, May 2013, pp. 8027–8031.

[14] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, no. 2-3, pp. 225–254, June 2000.

[15] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, January 2000.

[16] N. Dehak and G. Chollet, "Support vector GMMs for speaker verification," in *Proc. The IEEE Odyssey Speaker and Language Recognition Workshop*, 2006, pp. 1–4.

[17] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[18] N. Dehak, P. Kenny, R. Dehak, O. Glembek, P. Dumouchel, L. Burget, V. Hubeika, and F. Castaldo, "Support vector machines and joint factor analysis for speaker verification," in *Proc. ICASSP*, 2009, pp. 4237–4240.

[19] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[20] A. Sizov, K. Lee, and T. Kinnunen, "Unifying probabilistic linear discriminant analysis variants in biometric authentication," in *Proc. S+SSPR*, 2014, pp. 464–475.

[21] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. Springer Publishing Company, Incorporated, 2014.

[22] F. Richardson, D. Reynolds, and N. Dehak, "A unified deep neural network for speaker and language recognition," *arXiv:1504.00923*, 2015.

[23] P. Matejka, O. Glembek, O. Novotny, O. Plchot, F. Grézl, L. Burget, and J. Cernocky, "Analysis of DNN approaches to speaker identification," in *Proc. ICASSP*, March 2016, pp. 5100–5104.

[24] H. Traunmuller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," *Journal of the Acoustical Society of America*, vol. 107, no. 6, pp. 3438–3451, 2000.

[25] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, no. 2, pp. 139–152, February 2005.

[26] X. Fan and J. H. L. Hansen, "Speaker identification within whispered speech audio streams," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1408–1421, July 2011.

[27] P. Zelinka, M. Sigmund, and J. Schimmel, "Impact of vocal effort variability on automatic speech recognition," *Speech Communication*, vol. 54, no. 6, pp. 732–742, July 2012.

[28] N. Lass, L. Waters, and V. Tyson, "Speaker sex identification from voiced, whispered, and filtered isolated vowels," *Journal of the Acoustical Society of America*, vol. 59, no. 3, pp. 975–678, 1976.

[29] V. Tartter, "What's in a whisper?" *Journal of the Acoustical Society of America*, vol. 86, no. 5, pp. 1678–1683, 1989.

[30] V. Tartter, "Identifiability of vowels and speakers from whispered syllables." *Perception & Psychophysics*, vol. 49, no. 4, pp. 365–372, 1991.

[31] G. Chenghui, Z. Heming, Z. Wei, W. Yanlei, and W. Min, "A preliminary study on emotions of chinese whispered speech," in *International Forum on Computer Science-Technology and Applications*, vol. 2, December 2009, pp. 429–433.

[32] K. Tsunoda, S. Sekimoto, and T. Baer, "Brain activity in aphonia after a coughing episode: Different brain activity in healthy whispering and pathological aphonic conditions," *Journal of Voice*, vol. 26, no. 5, pp. 668.e11 – 668.e13, September 2012.

[33] S. Jovicic and Z. Saric, "Acoustic analysis of consonants in whispered speech," *Journal of Voice*, vol. 22, no. 3, pp. 263–274, May 2008.

[34] X. Fan and J. H. L. Hansen, "Speaker identification with whispered speech based on modified LFCC parameters and feature mapping," in *Proc. ICASSP*, April 2009, pp. 4553–4556.

[35] Q. Jin, S.-C. Jou, and T. Schultz, "Whispering speaker identification," in *IEEE International Conference on Multimedia and Expo*, July 2007, pp. 1027–1030.

[36] X. Fan and J. H. L. Hansen, "Speaker identification for whispered speech based on frequency warping and score competition," in *Proc. INTERSPEECH*, 2008, pp. 1313–1316.

[37] X. Fan and J. H. L. Hansen, "Acoustic analysis and feature transformation from neutral to whisper for speaker identification within whispered speech audio streams," *Speech Communication*, vol. 55, no. 1, pp. 119–134, January 2013.

[38] J. S. Garofolo, L. D. Consortium *et al.*, "TIMIT: acoustic-phonetic continuous speech corpus," 1993.

[39] B. P. Lim, "Computational differences between whispered and non-whispered speech," Ph.D. dissertation, University of Illinois, 2011.

[40] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, January 2010.

[41] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, vol. 1, May 1998, pp. 285–288.

[42] S. Desai, A. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *IEEE Audio, Speech, Language Process.*, vol. 18, no. 5, pp. 954–964, July 2010.

[43] T. Falk and W.-Y. Chan, "Modulation spectral features for robust far-field speaker identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 90–100, January 2010.

[44] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug 2005.

[45] P. Estevez, M. Tesmer, C. Perez, and J. Zurada, "Normalized mutual information feature selection," *IEEE Transactions on Neural Networks*, vol. 20, no. 2, pp. 189–201, Feb 2009.

[46] A. Clerico, R. Gupta, and T. Falk, "Mutual information between inter-hemispheric EEG spectro-temporal patterns: A new feature for automated affect recognition," in *Proc. IEEE/EMBS NER*. IEEE/EMBS, 2015, pp. 2106–2109.

[47] A. Rencher, *Methods of multivariate analysis*. John Wiley & Sons, 2003, vol. 492.

[48] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *Proc. ASRU*, December 2011.

[49] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, April 2015, pp. 5206–5210.

[50] M. McLaren, L. Ferrer, and A. Lawson, "Exploring the role of phonetic bottleneck features for speaker and language recognition," in *Proc. ICASPP*, March 2016, pp. 5575–5579.

[51] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech," in *Proc. Odyssey Speaker and Language Recognition Workshop*, June 2010.

[52] R. J. Vogt, C. J. Lustri, and S. Sridharan, "Factor analysis modelling for speaker verification with short utterances," in *Proc. The IEEE Odyssey Speaker and Language Recognition Workshop*. IEEE, 2008.

[53] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "I-vector based speaker recognition on short utterances," in *Proc. INTERSPEECH*, 2011, pp. 2341–2344.

[54] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74–99, Nov 2015.

[55] J. H. L. Hansen, C. Swail, A. South, R. Moore, H. Steeneken, E. Cupples, T. Anderson, C. R. A. Vloeberghs, I. Trancoso, and P. Verlinde, "The impact of speech under "stress"

on military speech technology," NATO Research & Technology Organization RTO-TR-10, AC/323(IST)TP/5 IST/TG-01, Tech. Rep., 2000.

[56] M. Senoussaoui, P. Kenny, N. Brummer, E. de Villiers, and P. Dumouchel, "Mixture of plda models in i-vector space for Gender-Independent speaker recognition," in *Proc. INTERSPEECH*. ISCA, 2011, pp. 25–28.

[57] J. Alam, P. Kenny, and D. O'Shaughnessy, "Low-variance multitaper mel-frequency cepstral coefficient features for speech and speaker recognition systems," *Cognitive Computation*, vol. 5, no. 4, pp. 533–544, 2013.

[58] S. Ghaffarzadegan, H. Bosil, and J. H. L. Hansen, "Generative modeling of pseudo-target domain adaptation samples for whispered speech recognition," in *Proc. ICASSP*, April 2015, pp. 5024–5028.

[59] S. Ghaffarzadegan, H. Boril, and J. H. L. Hansen, "Generative modeling of pseudo-whisper for robust whispered speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 10, pp. 1705–1720, Oct 2016.

[60] M. Sarria-Paja and T. Falk, "Whispered speech detection in noise using auditory-inspired modulation spectrum features," *IEEE Signal Processing Letters*, vol. 20, no. 8, pp. 783–786, August 2013.

[61] M. Sarria-Paja and T. Falk, "Strategies to enhance whispered speech speaker verification: A comparative analysis," *Journal of the Canadian Acoustical Association*, vol. 43, no. 4, pp. 31–45, 2015.

[62] M. Sarria-Paja and T. Falk, "Fusion of auditory inspired amplitude modulation spectrum and cepstral features for whispered and normal speech speaker verification," *Computer Speech & Language*, Accepted with minor revisions.

[63] M. Sarria-Paja and T. Falk, "Bottleneck and amplitude modulation features for improved whispered speech speaker verification in train/test mismatch conditions," *Under preparation – to be submitted to Speech Communications*, (11 pages).

[64] M. Sarria-Paja, T. Falk, and D. O'Shaughnessy, "Whispered speaker verification and gender detection using weighted instantaneous frequencies," in *Proc. ICASSP*, May 2013, pp. 7209–7213.

[65] M. Sarria-Paja, M. Senoussaoui, and T. H. Falk, "The effects of whispered speech on state-of-the-art voice based biometrics systems," in *Proc. IEEE CCECE*, May 2015, pp. 1254–1259.

[66] M. Sarria-Paja, M. Senoussaoui, D. O'Shaughnessy, and T. Falk, "Feature mapping, score-, and feature-level fusion for improved normal and whispered speech speaker verification," in *Proc. ICASSP*, March 2016, pp. 5480–5484.

[67] M. Sarria-Paja and T. Falk, "Variants of mel-frequency cepstral coefficients for improved whispered speech speaker verification in mismatched conditions," *EUSIPCO 2017*, Under review (5 pages).

[68] D. O'Shaughnessy, *Speech communications - human and machine (2. ed.)*. IEEE, 2000.

[69] M. Matsuda and H. Kasuya, "Acoustic nature of the whisper," in *Proc. EUROSPEECH*, 1999, pp. 133–136.

[70] I. Thomas, "Perceived pitch of whispered vowels," *Journal of the Acoustical Society of America*, vol. 46, no. 2B, pp. 468–470, 1969.

[71] M. Higashikawa, K. Nakai, A. Sakakura, and H. Takahashi, "Perceived pitch of whispered vowels-relationship with formant frequencies: A preliminary study," *Journal of Voice*, vol. 10, no. 2, pp. 155–158, 1996.

[72] M. Schwartz and H. Rine, "Identification of speaker sex from isolated, whispered vowels," *Journal of the Acoustical Society of America*, vol. 44, no. 6, pp. 1736–1737, 1968.

[73] H. Sharifzadeh, I. McLoughlin, and M. Russell, "A comprehensive vowel space for whispered speech," *Journal of Voice*, vol. 26, no. 2, pp. 49–56, March 2012.

[74] T. Tran, S. Marioryad, and C. Busso, "Audiovisual corpus to analyze whisper speech," in *Proc. ICASSP*, May 2013, pp. 8101–8105.

[75] R. Morris and M. Clements, "Reconstruction of speech from whispers," *Medical Engineering & Physics*, vol. 24, no. 7-8, pp. 515–520, October 2002.

[76] H. Sharifzadeh, I. McLoughlin, and F. Ahmadi, "Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 10, pp. 2448–2458, October 2010.

[77] F. Ahmadi, I. McLoughlin, and H. Sharifzadeh, "Analysis-by-synthesis method for whisper-speech reconstruction," in *Proc. IEEE Asia Pacific Conference on Circuits and Systems*, December 2008, pp. 1280–1283.

[78] L. Rabiner and R. Schafer, *Digital processing of speech signals*, ser. Prentice-Hall signal processing series. Englewood Cliffs, N.J. Prentice-Hall, 1978.

[79] S. Stevens, J. Volkmann, and E. Newman, "A scale for the measurement of the psychological magnitude pitch," *JASA*, vol. 8, no. 3, pp. 185–190, 1937.

[80] J. Proakis and D. Manolakis, *Digital signal processing : principles, algorithms, and applications.* Upper Saddle River, N.J. Prentice Hall, 1996.

[81] D. O'Shaughnessy, "Invited paper: Automatic speech recognition: History, methods and challenges," *Pattern Recognition*, vol. 41, no. 10, pp. 2965 – 2979, 2008.

[82] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of MFCC," *Journal of Computer Science and Technology*, vol. 16, no. 6, pp. 582–589, 2001.

[83] Z. Tao, X.-J. Zhang, H.-M. Zhao, and W. Kulesza, "Noise reduction in whisper speech based on the auditory masking model," in *Proc. International Conference on Information Networking and Automation*, vol. 2, October 2010, pp. 272–277.

[84] C. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag New York, Inc., 2006.

[85] Q. Jin and T. F. Zheng, "Overview of front-end features for robust speaker recognition," 2011.

[86] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE transactions on signal processing*, vol. 41, no. 10, pp. 3024–3051, 1993.

[87] S. Sadjadi and J. H. L. Hansen, "Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions," in *Proc. In INTERSPEECH*, 2010, pp. 2138–2141.

[88] P. Clark and L. Atlas, "Time-frequency coherent modulation filtering of nonstationary signals," *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4323 –4332, November 2009.

[89] V. Mitra, M. McLaren, H. Franco, M. Graciarena, and N. Scheffer, "Modulation features for noise robust speaker identification," in *Proc. INTERSPEECH*, 2013, pp. 3703–3707.

[90] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 201–216, Mar 2001.

[91] V. Mitra, J. VanHout, W. Wang, C. Bartels, H. Franco, D. Vergyri, A. Alwan, A. Janin, J. H. L. Hansen, R. Stern *et al.*, "Fusion strategies for robust speech recognition and keyword spotting for channel-and noise-degraded speech," in *Proc. INTERSPEECH*, vol. 2016, 2016.

[92] M. J. Alam, P. Ouellet, P. Kenny, and D. O'Shaughnessy, "Comparative evaluation of feature normalization techniques for speaker verification," in *International Conference on Nonlinear Speech Processing*. Springer, 2011, pp. 246–253.

[93] R. Duda, P. Hart, and D. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[94] D. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, Jan 1995.

[95] P.-M. Bousquet, D. Matrouf, and J.-F. Bonastre, "Intersession compensation and scoring methods in the i-vectors space for speaker recognition," in *Proc. INTERSPEECH*, 2011, pp. 485–488.

[96] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE 11th ICCV*, Oct 2007, pp. 1–8.

[97] E. Khoury, L. El Shafey, and S. Marcel, "SPEAR: An open source toolbox for speaker recognition based on Bob," in *Proc. ICASSP*, 2014, pp. 1655–1659.

[98] S. Ghaffarzadegan, H. Boril, and J. H. L. Hansen, "UT-VOCAL EFFORT II: Analysis and constrained-lexicon recognition of whispered speech," in *Proc. ICASSP*, 2014, pp. 2544–2548.

[99] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The CHAINS corpus: Characterizing individual speakers," in *Proc of SPECOM*, vol. 6, 2006, pp. 431–435.

[100] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, July 2008.

[101] ITU-T P.56, *Objective measurement of active speech level*, International Telecommunication Union, 1993.

[102] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Odyssey Speaker and Language Recognition Workshop*, June 2001.

[103] K. McDougall, "Speaker-specific formant dynamics: An experiment on Australian English /aI/," *Speech, Language and the Law.*, vol. 11, no. 1, pp. 103–130, June 2004.

[104] K. McDougall and F. Nolan, "Discrimination of speakers using the formant dynamics of /u:/ in British English," in *Proc. 16th International Congress of Phonetic Sciences*, August 2007, pp. 1825–1828.

[105] R. Lyon, A. Katsiamis, and E. Drakakis, "History and future of auditory filter models," in *Proc. IEEE International Symposium on Circuits and Systems*, June 2010, pp. 3809–3812.

[106] Z. Tao, J.-H. Gu, X.-D. Tan, Y.-S. Xu, T. Han, and H.-M. Zhao, "Reconstruction of normal speech from whispered speech based on RBF neural network," in *Proc. IITSI*, April 2010, pp. 374–377.

[107] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1859–1872, Dec. 2014.

[108] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. ICML*, July 2008, pp. 1096–1103.

[109] S. Irtza, H. Bavattichalil, V. Sethu, and E. Ambikairajah, "Scalable i-vector concatenation for PLDA based language identification system," in *Proc. APSIPA*, Dec 2015, pp. 1182–1185.

[110] Z.-Y. Li, W.-Q. Zhang, and J. Liu, "Multi-resolution time frequency feature and complementary combination for short utterance speaker recognition," *Multimedia Tools and Applications*, vol. 74, no. 3, pp. 937–953, 2015.

[111] N. Brummer and E. de Villiers, "The BOSARIS Toolkit User Guide: Theory, algorithms and code for binary classifier score processing," CAGNITIO Research, South Africa, Tech. Rep., 2011.

[112] T. Drugman and T. Dutoit, "On the potential of glottal signatures for speaker recognition," in *Proc. INTERSPEECH*. ISCA, 2010, pp. 2106–2109.

[113] M. Chetouani, M. Faundez-Zanuy, B. Gas, and J. Zarader, "Investigation on LP-residual representations for speaker identification," *Pattern Recognition*, vol. 42, no. 3, pp. 487 – 494, 2009.

[114] P. Debadatta and M. Prasanna, "Processing of linear prediction residual in spectral and cepstral domains for speaker information," *International Journal of Speech Technology*, vol. 18, no. 3, pp. 333–350, 2015.

[115] J. Xu and H. Zhao, "Speaker identification with whispered speech using unvoiced-consonant phonemes," in *Proc. IASP*, Nov 2012, pp. 1–4.

[116] M. Sahidullah, S. Chakroborty, and G. Saha, "Improving performance of speaker identification system using complementary information fusion," *CoRR*, vol. abs/1105.2770, 2011.

[117] J. Rodman, "The effect of bandwidth on speech intelligibility - white paper," POLYCOM Inc., USA, Tech. Rep., September 2006.

[118] H. Lei and E. Lopez-Gonzalo, "Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition," in *Proc. INTERSPEECH*, 2009.

[119] L. Gallardo, M. Wagner, and S. Möller, "Advantages of wideband over narrowband channels for speaker verification employing MFCCs and LFCCs." in *Proc. INTERSPEECH*, 2014.

[120] L. Besacier and J.-F. Bonastre, "Subband approach for automatic speaker recognition: Optimal division of the frequency domain," in *Proc. AVBPA*, ser. Lecture Notes in Computer Science, March 1997, pp. 195–202.

[121] T. Falk, W.-Y. Chan, and F. Shein, "Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility," *Speech Communication*, vol. 54, no. 5, pp. 622–631, 2012.

[122] T. Kinnunen, K. Lee, and H. Li, "Dimension reduction of the modulation spectrogram for speaker verification," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2008)*, 2008.

[123] J. Xiang, D. Poeppel, and J. Simon, "Physiological evidence for auditory modulation filterbanks: Cortical responses to concurrent modulations," *JASA-EL*, vol. 133, no. 1, pp. EL7–El12, 2013.

[124] R. Moddemeijer, "On estimation of entropy and mutual information of continuous distributions," *Signal Processing*, vol. 16, no. 3, pp. 233–248, 1989.

[125] X. Fan and J. H. L. Hansen, "Acoustic analysis for speaker identification of whispered speech," in *Proc. ICASSP*, March 2010, pp. 5046–5049.

[126] D. Ribas, E. Vincent, and J. Calvo, "Full multicondition training for robust i-vector based speaker recognition," in *Proc. INTERSPEECH*, 2015.

[127] M. Carlin, B. Smolenski, and S. Wenndt, "Unsupervised detection of whispered speech in the presence of normal phonation," in *Proc. INTERSPEECH*, 2006.

[128] A. Mathur, S. Reddy, and R. Hegde, "Significance of parametric spectral ratio methods in detection and recognition of whispered speech," *EURASIP Adv Signal Process*, vol. 2012, no. 157, 2012.

[129] C. Zhang and J. H. L. Hansen, "Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing," *IEEE Audio, Speech, Language Process*, vol. 19, no. 4, pp. 883–894, 2011.

[130] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, may 2016.

[131] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[132] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. 21st International Congress on Acoustics*. Acoustical Society of America, Jun. 2013.

[133] T. Falk and W.-Y. Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 4, pp. 978–989, 2010.

[134] S. O. Sadjadi, M. Slaney, and L. P. Heck, "MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker Recognition Research," in *IEEE Speech and Language Processing Technical Committee Newsletter*, November 2013.