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

SIGNAL PROCESSING AND MACHINE LEARNING FOR ROBUST IN-THE-WILD VOICE APPLICATIONS

By

Anderson Avila

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

Evaluation Commitee

Internal evaluator and committee president:	Jean-Charles Gregoire INRS
External evaluator 1:	Alessandro L. Koerich École de Technologie Supérieure
External evaluator 2:	Andrew Hines University College Dublin
Research Advisors:	Tiago H. Falk (Supervisor) Douglas O'Shaughnessy (Co-Supervisor)

"Ignorance more frequently begets confidence than does knowledge."

Charles Darwin

Acknowledgements

First and foremost, I thank God for paving the way, giving me the strength and ability to complete this endeavour. I would like to express my sincere gratitude towards everyone who contributed to this work. I would like to thank my family, especially my lovely wife, who has supported me every step of the way. I thank my three beautiful children for bearing with me, even when things were tough and I could not be a better father. Special thanks to my grandparents, Julieta e Natalino, and to my parents, Sueli e Darci. There is no words to express my gratitude for your support. I would like to thank Professor Tiago H. Falk for all of his support, patient guidance, encouragement and advice that he has provided with excellence throughout my time as his student. I am very grateful for the time spent at MuSAE Lab, where I learned and grew a lot. I would like to thank my co-supervisor, Professor Douglas O'Shaughnessy, for his generosity and dependability at all times. The support from my peers (Dr. Raymundo Cassani, Dr. João Felipe, João Monteiro, Isabela Albuquerque, Abhishek Tiwari, Shruti Kshirsagar, Belmir Junior, Olivier Rosanne, Stefany Bedoya, Alexandre Drouin, Milton Sarria-Paja, Marília Karla and Diana Tobon) was very important. I would like to thank Microsoft for receiving me as research intern during the summer of 2017. I had a wonderful time there and a rich experience that had a significant impact on my research. Special thanks to Dr. Ivan Tashev for being a fantastic mentor. I would like to show my appreciation to CRIM for hosting me as an intern for almost 18 months. All the facilities and infra-structure available during my internship were impeccable. I would like to thank Dr. Jahangir Alam for sharing his expertise, codes, experience and his time to guide me. I would like to express my gratitude to an old friend, Professor Francisco Fraga, my former supervisor, who helped me to start this beautiful journey.

I also would like to thank the Fonds de Recherche du Québec - Nature et Technologies (FRQNT) and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for their financial support.

Abstract

Quality and reliability are key concepts in today's technological world. Companies strive to perform consistently well in order to surpass users' expectations. Hence, considerable amounts of money are being invested annually worldwide to produce highly reliable and good quality products and services. To remain globally competitive, quality principles and reliability are being applied across different sectors of the economy, such as engineering, robotics, health care, Internet and software in general. In this work, we are interested in the perceived quality of speech and the reliability of speech-based technologies. It is known that such technologies have made their way out of laboratories to be employed in real-world applications. Therefore, their performance beyond laboratory settings has become an increasing concern for the research community.

The first part of this thesis focuses on the estimation of perceived speech quality in noisy and reverberant environments. We propose a new perceptual quality estimator based on the i-vector framework. While widely used across numerous speech applications, the potential of i-vectors to summarize the quality of a speech recording has been overlooked. We conduct a detailed analysis of how the total variability space is capable of capturing ambient factors, such as those related to background noise and reverberation levels. We then propose a full-reference speech quality model based on i-vector similarities. The main motivation behind this lies on the fact that i-vectors are known for carrying out both speaker and channel information. Thus, by considering a full-reference model, we can assume no speaker and speech variability between the reference and degraded representations. That is, speech content will remain mostly the same for the reference and degraded signals and only changes in the channel factors will be present. We also propose a new non-intrusive instrumental quality measure based on the similarity between two i-vector representations. As the reference clean signal is not available in this case, we propose the use of a clean speech Gaussian mixture model to estimate the clean speech spectra from its degraded counterpart, which is then used to attain the reference i-vector representation.

The second part of this thesis is dedicated to the reliability of speech-based technologies, specifically speech emotion recognition (SER) and automatic speaker verification (ASV). We first explore spontaneous SER in-the-wild, where factors such as noise, reverberation and their combined effects compromise SER performance. We show that existing SER systems based on per-frame features (computed from the modulation spectrum), while useful for enacted/posed emotions, perform poorly for spontaneous speech. To overcome such limitation, an environment-robust feature pooling scheme, which combines information from neighbouring frames, is proposed to predict spontaneous arousal and valence emotional primitives. Second, the reliability of speaker verification is also addressed. We propose a new method to minimize the impact of affective speech on ASV performance. For that, a Gaussian mixture model is used to learn a prior probability distribution of the neutral speech for a given speaker (i.e., characterizing his/her source space). This knowledge is then used to minimize the differences between target (affective) and source (neutral) spaces. Besides intra-speaker variability

caused by emotional speech, replay attacks also represent a serious threat to ASV reliability. To mitigate such problems, we propose a front-end based on the use of blind estimation of the channel response magnitude and a residual neural network as back-end. Our hypothesis is that the magnitude response of the channel, obtained by subtracting the log-magnitude spectrum of the observed signal from the estimated log-magnitude spectrum of the observed signal's clean counterpart, will capture the nuances of room ambiences, recordings and playback devices. This can then be used to distinguish bonafide from spoofed speech.

Keywords Quality of experience, Speech quality assessment, i-vector, reliability

Résumé

La qualité et la fiabilité sont des concepts clés dans le monde technologique d'aujourd'hui. Les entreprises s'efforcent de toujours bien performer afin de dépasser les attentes des utilisateurs à leur égard. Par conséquent, des sommes d'argent considérables sont investies chaque année dans le monde pour produire des produits et des services hautement fiables et de bonne qualité. Pour rester compétitif à l'échelle mondiale, les principes de qualité et de fiabilité sont appliqués à différents secteurs de l'économie, tels que l'ingénierie, la robotique, les soins de santé, Internet et les logiciels en général. Dans ce travail, nous nous intéressons à la qualité perçue de la parole et à la fiabilité des technologies basées sur la parole. Il est connu que ces technologies ont fait leur chemin hors des laboratoires pour être utilisées dans des applications réelles. Par conséquent, leur performance au-delà des paramètres de laboratoire est devenue une préoccupation croissante pour la communauté des chercheurs.

La première partie de cette thèse porte sur l'estimation de la qualité perçue de la parole dans des environnements bruyants et réverbérants. Pour ce faire, nous présentons un nouvel estimateur de la qualité perceptuelle basé sur le cadre du i-vector. Bien qu'ils soient largement utilisés dans de nombreuses applications vocales, le potentiel des i-vectors pour résumer la qualité d'un enregistrement vocal a été ignoré. Nous effectuons donc une analyse détaillée de la façon dont l'espace de variabilité totale est capable de capturer des facteurs ambiants, tels que ceux liés au bruit de fond et aux niveaux de réverbération. Nous proposons ensuite un modèle de qualité de la parole à référence complète basé sur les similitudes du i-vector. La principale motivation derrière cette démarche réside dans le fait que les i-vectors sont connus pour fournir à la fois des informations sur les haut-parleurs et les canaux. Ainsi, en considérant un modèle de référence complète, nous ne pouvons supposer aucune variabilité du locuteur et de la parole entre les représentations de référence et dégradées, c'est-à-dire que le contenu de la parole restera essentiellement le même pour le signal de référence et dégradé, et que seuls des changements dans les facteurs de canal seront présents. Nous mettons également de l'avant une nouvelle mesure de qualité instrumentale non intrusive basée sur la similitude entre deux représentations i-vectorielles. Comme le signal propre de référence n'est pas disponible dans ce cas, nous proposons l'utilisation d'un modèle de mélange gaussien de parole propre pour estimer les spectres de parole propre à partir de son homologue dégradé, qui est ensuite utilisé pour atteindre l'i-vector de référence.

La deuxième partie de cette thèse est consacrée à la fiabilité des technologies basées sur la parole, en particulier la reconnaissance automatique des émotions (RAE) de la parole et la vérification automatique du locuteur (VAL). Nous explorons d'abord la RAE de la parole spontané "in-the-wild", où des facteurs tels que le bruit, la réverbération et leurs effets combinés compromettent les performances du RAE de la parole. Nous montrons que les systèmes SER existants basés sur des caractéristiques par trame (calculées à partir du spectre de modulation), bien qu'utiles pour les émotions mises en scène/posées, fonctionnent mal pour la parole spontanée. Pour surmonter cette

limitation, un schéma de mise en commun des fonctionnalités robuste à l'environnement, qui combine des informations provenant de trames voisines, est proposé pour prédire l'excitation spontanée et les primitives émotionnelles de valence. Deuxièmement, la fiabilité de la vérification des locuteurs est également abordée. Nous proposons une nouvelle méthode pour minimiser l'impact de la parole affective sur les performances de la VAL. Pour ce faire, un modèle de mélange gaussien est utilisé pour apprendre une distribution de probabilité antérieure de la parole neutre pour un locuteur donné (c'est-à-dire caractériser son espace source). Ces connaissances sont ensuite utilisées pour minimiser les différences entre les espaces cibles (affectifs) et sources (neutres). Outre la variabilité intra-locuteur causée par le discours émotionnel, les attaques par rejeu représentent également une menace sérieuse pour la fiabilité de la VAL. Pour atténuer ces problèmes, nous mettons de l'avant un frontal basé sur l'utilisation d'une estimation aveugle de l'amplitude de la réponse du canal et d'un réseau neuronal résiduel comme back-end. Notre hypothèse est que la réponse en amplitude du canal, obtenue en soustrayant le spectre de magnitude logarithmique du signal observé du spectre de magnitude logarithmique estimé de la contrepartie propre du signal observé, capturera les nuances des ambiances de la pièce, des enregistrements et des appareils de lecture. Cela peut ensuite être utilisé pour distinguer la bonne foi de la parole usurpée.

Keywords Qualité d'expérience, Évaluation de la qualité de la parole, i-vector, fiabilité

Content

List of Figures

List of Tables

List of abbreviations

ACR	Absolute Category Rating
ADL	Analyse Discriminante Linéaire
AI	Artificial Intelligence
ANN	Artificial Neural Network
ANSI	American National Standards Institute
ASR	Automatic Speech Recognition
ASV	Automatic Speaker Verification
AVEC	Audio/Visual Emotion Challenge
CNN	Convolutional Neural Network
CSD	Cosine Similarity Distance
CV	Conversion Vocale
ECA	Échelle de Classification Absolue
EmotiW	Emotion Recognition in The Wild Challenge
ERB	Equivalent Rectangular Bandwidth
EU	l'Expérience Utilisateur
GMM	Gaussian mixture model
HCI	Human-Computer Interface
IA	Intelligence Artificielle
ICT	Information and Communication Technology
IHM	Interaction Homme-Machine
IRS	Intermediate Reference Systems
ITU-T	International Telecommunication Union
JFA	Joint Factor Analysis
LDA	Linear Discriminant Analysis
LOT	Listening-only Tests
LP	Linear Predictive

LSTM	Long-Short Term Memory
MFCC	Mel-frequency Cepstral Coefficients
MLP	Multi-layer layer Perceptron
MOS	Mean Opinion Score
MSF	Modulation Spectrum Features
MSR	Modulation Spectrum Representation
NCIC	Normalisation de Covariance Intra-Classe
NCM	Normalized Covariance Metric
NOM	Note d'Opinion Moyenne
PESQ	Perceptual Evaluation of Speech Quality
POLQA	Perceptual Objective Listening Quality Assessment
QoE	Quality of Experience
QoS	Quality of Service
RAE	Reconnaissance Automatique des Émotions
RAP	Reconnaissance Automatique de la Parole
ReLU	Rectifier Linear Unit
ResNet	Residual Neural Network
RNN	Recurrent Neural Network
SER	Speech Emotion Recognition
SPL	Sound Pressure Level
SRMR	Speech-to-Reverberation Modulation Ratio
STOI	Short-time Objective Intelligibility
SV	Synthèse Vocale
TIC	Technologies de l'Information et de la Communication
TV	Total Variability
UBM	Universal Background Model
UIT	Union Internationale des Télécommunications
UX	User Experience
VAE	Variational Autoencoder
VAL	Vérification Automatique du Locuteur
WCCN	Within Class Covariance Normalization

Synopsis

0.1 Introduction

Selon la théorie des signes de Charles Sanders Peirce, un signe est tout ce qui signifie autre chose, à savoir l'objet, déterminant ainsi sa signification qui sera ensuite comprise par l'interprète. Dans ses propres mots : *'je définis un signe comme tout ce qui est si déterminé par quelque chose d'autre, appelé son objet, et détermine ainsi un effet sur une personne, effet que j'appelle son interprétant, que ce dernier est ainsi déterminé de façon médiatique par l'ancien'* [7]. On peut considérer le signe comme le signifiant. En ce qui concerne la langue, par exemple, un mot écrit ou un signal vocal acoustique est la forme sous laquelle le signe est présenté. L'objet est tout ce qui est signifié, par exemple ce à quoi le mot écrit ou le signal vocal acoustique fait référence. Le sens est le sens donné au signe par son interprète et qui établit la relation signe/objet [3]. Ceci est représenté par le triangle sémiotique montré dans la figure 1. Dans ce contexte, les télécommunications permettent une interaction entre humains via la transmission de signes (c'est-à-dire des messages, des mots, des écrits, des images et des sons) par fil, radio, optique ou autres systèmes électromagnétiques. Dans une conversation téléphonique, par exemple, les humains communiquent entre eux en échangeant des signes vocaux. Tout en étudiant la qualité de la parole, il est utile d'établir une distinction claire entre la forme acoustique du discours et son contenu (ou sa signification). Dans ce cas, l'objet principal d'intérêt est la forme acoustique, bien que le contenu joue un rôle important sur la perception de la parole et sur la façon dont une certaine qualité est associée à la forme acoustique. En fin de compte, c'est le locuteur et l'auditeur humain qui déterminent la relation entre la forme acoustique et son contenu car aucune relation directe ne peut être établie entre les deux en négligeant l'interprète [1].

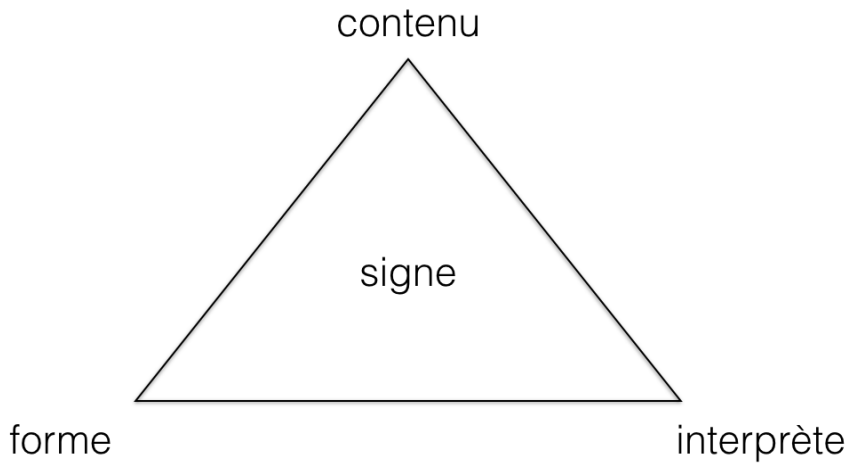


Figure 1 – Triangle sémiotique représentant le signe triadique. Adapté de [1].

Les technologies de l'information et de la communication (TIC) se sont considérablement développées au cours de la dernière décennie. Les fournisseurs de systèmes de communication multimédia sont confrontés à une multitude de défis : de nouvelles méthodes de codage [8], de nouveaux algorithmes de traitement du signal plus sophistiqués [9, 10], ainsi que des approches de transmission plus modernes sont déployés chaque jour [11]. Une préoccupation croissante concerne les systèmes de communication vocale, car les services vocaux basés sur diverses technologies de transmission sont fournis par les réseaux de télécommunication [12, 13]. En tant que principal moyen de communication, le signal vocal, lors de son acquisition, de sa transmission ou de sa distribution, est susceptible de voir sa qualité et son intelligibilité diminuées par différentes altérations. La Figure 2 montre les dégradations typiques présentes dans un canal de communication [2, 14]. Il prend en considération l'interface utilisateur aux deux extrémités du canal de communication (par exemple, les appareils de microphone et les haut-parleurs), et le support de transmission lui-même (par exemple, fil de cuivre, sans fil, fibre optique, etc.) [2]. Un signal vocal, par exemple, peut être affecté par des contraintes de bande passante, une variabilité des conditions de canal et des erreurs d'acquisition ou de stockage [15]. De plus, le bruit de fond et la réverbération peuvent également constituer une source de dégradation supplémentaire, compromettant la qualité globale du signal de parole [16]. En effet, la qualité d'un service de télécommunication vocale est fortement affectée par la qualité du système de transmission. Néanmoins, la qualité globale perçue n'est pas supposée être uniquement liée aux aspects physiques de la transmission. Elle est plutôt déterminée par l'utilisateur. En effet, lorsque l'utilisateur a suffisamment d'expérience dans l'exploitation d'un service de télécommunications déterminé, une idée de qualité est développée et attendue. En ce sens, la qualité peut être comprise

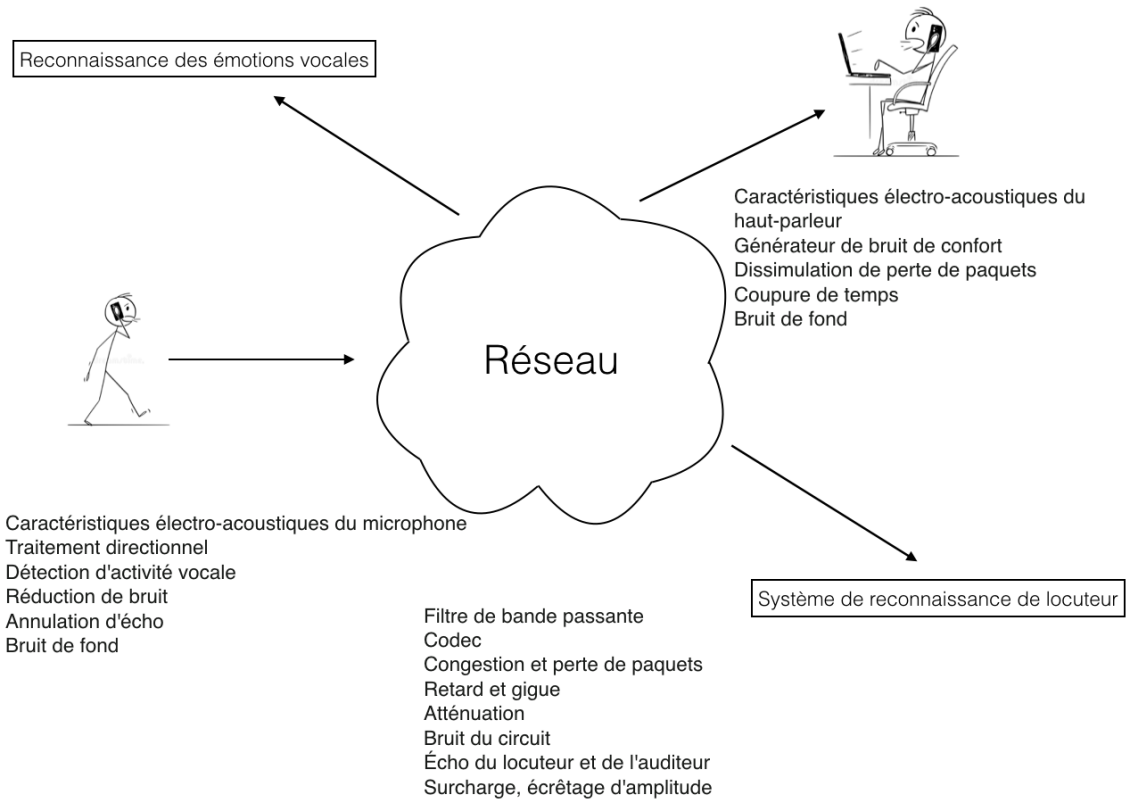


Figure 2 – Principales dégradations d'un canal de communication de bout en bout. Adapté de [2].

comme la réalisation de certaines caractéristiques souhaitées et attendues telles que perçues par l'utilisateur [1].

Traditionnellement, le moyen le plus fiable pour évaluer la qualité d'un système de communication vocale provient de sujets humains. Les tests d'écoute subjectifs ont été largement utilisés à cet égard et une série de recommandations décrivant les méthodes et les procédures pour mener de telles évaluations subjectives ont été mises à disposition par l'Union internationale des télécommunications (UIT-T) dans sa Recommandation UIT-T P.800 [17]. Dans de tels tests subjectifs, les signaux vocaux sont présentés à des auditeurs (naïfs ou experts, selon l'application) qui jugent la qualité du signal sur une échelle de classification absolue (ECA) allant de 1 à 5, représentant « mauvais » et « excellent », respectivement. La note d'opinion moyenne (NOM), qui représente la qualité de la parole perçue après le nivellement des facteurs individuels [4], est atteinte après la moyenne de tous les scores des participants sur une condition spécifique. Cependant, de telles mesures subjectives ne sont pas toujours pratiques, car elles : (1) nécessitent de nombreux auditeurs; (2) peuvent être laborieuses et longues; (3) peuvent être coûteuses; et (4) ne peuvent pas être exécutées en temps réel [18]. Par conséquent, un algorithme pour estimer la qualité vocale perçue pourrait être la meilleure option lors

de la planification du réseau. En fait, des mesures de qualité instrumentales ont été explorées au fil des ans pour surmonter ces limites et elles sont conçues pour être fortement corrélées avec les scores NOM d'écoute subjective, remplaçant ainsi efficacement le panel d'auditeurs par un algorithme de calcul en temps réel [4].

Il est important d'estimer la qualité perçue des services et des applications multimédias existants et émergents, en particulier pour les fournisseurs de communications vocales qui cherchent à optimiser leurs services et à maximiser l'expérience client [19]. La surveillance de la qualité en temps réel, par exemple, peut aider à la conception et au développement du réseau, ainsi qu'à l'adaptation en ligne pour garantir que les attentes des utilisateurs finaux sont satisfaites. De plus, des études ont montré une certaine corrélation entre la qualité de la parole et les performances des technologies vocales telles que la vérification automatique du locuteur (VAL) et la reconnaissance automatique des émotions (RAE) de la parole [20, 21, 22]. Ces systèmes peuvent avoir de bonnes performances dans des paramètres plus contrôlés, mais sont gravement affectés par le décalage entre les énoncés de formation, qui peuvent ne pas tenir compte des déficiences invisibles des canaux, et les énoncés transmis par différents canaux de communication, entraînant une diminution de leurs performances [23]. La mesure instrumentale de la qualité pourrait donc s'avérer utile pour prévoir les performances de telles applications pour une configuration de canal de communication donnée. Par exemple, comme suggéré dans la Figure 2, l'authentification de la voix de l'utilisateur, lors de l'exécution d'une transaction financière par téléphone, peut nécessiter le traitement à distance de la transmission du signal vocal. Cela pourrait entraîner une dégradation de la qualité du signal vocal affectant également la fiabilité de ces applications.

La fiabilité est pourtant un élément important de la qualité qui peut influencer l'expérience utilisateur (EU) [24]. Il consiste à garantir qu'un système fonctionnera de manière cohérente et comme prévu par l'utilisateur final. Ainsi, l'EU des systèmes non fiables est certainement affecté lorsque les utilisateurs n'obtiennent pas les mêmes résultats lors d'essais répétés. Étant donné que de nombreuses technologies basées sur la parole ont quitté les laboratoires pour être utilisées dans des applications réelles, une préoccupation croissante de la communauté des chercheurs concerne la performance de ces technologies dans des scénarios plus réalistes. Ici, nous nous intéressons particulièrement aux systèmes RAE de la parole et VAL. Malgré les progrès récents réalisés dans ce domaine, principalement grâce à de nouvelles incorporations basées sur des modèles de neurones

profonds [25], dans certaines circonstances, la qualité du signal vocal n'est pas assez bonne pour prendre des décisions fiables [26].

Cette thèse traite de la qualité de perception et de l'expérience humaine lors de l'utilisation de systèmes basés sur la voix. Nous nous intéressons en particulier à l'estimation de la qualité perçue de la parole et à l'amélioration de la fiabilité des technologies basées sur la parole. Nous proposons l'utilisation du cadre i-vector comme nouvel outil d'estimation objective de la qualité. Nous présentons ensuite un schéma de mise en commun des fonctionnalités robuste pour augmenter la fiabilité du RAE de la parole "in the wild" [1]. Étant donné que la parole émotionnelle peut nuire aux performances de l'VAL, afin de maintenir la fiabilité de ces systèmes, nous proposons une nouvelle méthode pour apprendre des caractéristiques de parole neutres afin de neutraliser la parole émotionnelle. Enfin, nous mettons de l'avant une interface basée sur l'utilisation de l'estimation aveugle de l'amplitude de réponse du canal et d'un réseau neuronal résiduel pour atténuer la menace d'attaques d'accès physique à l'VAL. Dans la section suivante, nous avons détaillé les problèmes de recherche abordés dans ce travail et les hypothèses respectives à leur égard.

0.1.1 Problème de recherche et hypothèses

Dans la communication vocale, le signal enregistré est corrompu par le bruit ambiant et la réverbération dans la salle de transmission. Celui-ci peut dégrader gravement la qualité et l'intelligibilité de la parole perçues. Les effets néfastes du bruit environnemental peuvent toutefois être atténués par l'utilisation d'algorithmes d'amélioration de la parole tels que la réduction du bruit, l'annulation de l'écho et la dé-réverbération [27]. Bien que ces algorithmes visent à améliorer le signal vocal en réduisant la quantité de bruit et de réverbération, ils peuvent également introduire des artefacts indésirables qui dégradent la qualité de la parole [28]. Dans de tels scénarios, l'évaluation de la qualité de la parole perçue est cruciale [18, 29]. Les tests d'écoute subjective sont généralement considérés comme le moyen le plus fiable d'évaluer la qualité de la parole, mais ils sont coûteux et prennent du temps. À titre d'alternative, plusieurs mesures objectives de qualité instrumentale ont été proposées et normalisées par l'UIT-T. Cependant, des algorithmes objectifs standardisés peuvent devenir obsolètes à mesure que de nouveaux scénarios, tels que la capture de son en champ

1. Ici, "in the wild" signifie au-delà des paramètres de laboratoire, où les conditions acoustiques, présentes dans des scénarios du monde réel, telles que le bruit et la réverbération, conduisent à une inadéquation significative entre les conditions du train et des tests.

lointain, de nouveaux algorithmes de compression audio et de nouveaux modèles d'amélioration de la parole émergent [27]. Pour résoudre ce problème, nous avons exploré l'utilisation de l'i-vector pour la mesure de la qualité instrumentale de la parole bruyante, réverbérante et améliorée. Cela conduit à notre première hypothèse.

- **H1**: les i-vectors ont le potentiel de capturer non seulement des informations dépendantes du locuteur, mais également des informations liées à la distorsion et à la qualité présentes dans le signal vocal et, par conséquent, peuvent être utilisées comme une mesure de qualité instrumentale de référence complète.

La principale motivation derrière cela réside dans le fait que les i-vectors sont connus pour transmettre à la fois des informations sur le canal et sur le locuteur. Néanmoins, la plupart des recherches dans le domaine se sont concentrées sur les caractéristiques du locuteur de la représentation (par exemple, pour la reconnaissance du locuteur). Comme indiqué dans les recherches précédentes [30, 31], les performances des applications basées sur les i-vectors sont gravement affectées par des facteurs environnementaux, tels que le bruit de fond et la réverbération. Pour atténuer ces effets de canal, des techniques de compensation, telles que l'analyse discriminante linéaire (ADL) et la normalisation de covariance intra classe (NCIC) [32], sont couramment appliquées. Étant donné qu'il s'avère peu probable que ces travaux fonctionnent, nous utilisons ces informations comme un corrélat de la qualité vocale perçue.

Les mesures de qualité instrumentale de référence complète nécessitent l'accès au signal de parole propre de référence et à son homologue dégradé [4]. La plupart des mesures intrusives reposent donc sur le calcul de la distance entre une certaine représentation auditive du signal de référence et des signaux dégradés pour estimer la qualité de la parole. Les mesures sans référence, en revanche, ne dépendent que du signal dégradé pour estimer la qualité de la parole. Prédire la qualité de la parole sans le signal de parole propre de référence est une tâche difficile et offre souvent des performances inférieures par rapport aux méthodes intrusives. Cela est dû à la grande variabilité du signal vocal d'entrée, qui est le résultat de différences en tant que locuteurs, voies vocales, caractéristiques de hauteur et contenu de la parole [33].

Notre première hypothèse repose sur la disponibilité du signal propre de référence pour estimer la qualité perçue. Cela n'est pas toujours possible pour certaines applications vocales, car le signal de nettoyage de référence peut ne pas être disponible. Cela met en lumière la question suivante

: l'i-vector peut-elle être utilisée comme mesure de qualité instrumentale sans référence ? Notre deuxième hypothèse tente de répondre à cette question.

- **H2**: il est possible d'utiliser l'i-vector comme mesure de qualité instrumentale sans référence, car le signal propre de référence peut être atteint en utilisant un modèle de mélange gaussien (GMM) de parole propre qui reconstruit la référence permettant de nettoyer les spectres du signal corrompu.

La deuxième partie de cette thèse traite de la fiabilité des technologies basées sur la parole. Nous nous concentrons d'abord sur la robustesse de RAE de la parole. Avec l'essor des applications d'interaction homme-machine (IHM), en particulier sur les appareils mobiles, les paramètres dits « à l'état sauvage » ont constitué une menace sérieuse pour les systèmes de reconnaissance des émotions. Afin d'accélérer les innovations dans le domaine, plusieurs défis ont été organisés au cours des dernières années, notamment INTERSPEECH Emotion Challenge [34, 35], Emotion Recognition in The Wild Challenge (EmotiW) [36, 37], et les défis des émotions audiovisuelles de 2016 et 2017 (AVEC) [38, 39]. Nous sommes particulièrement préoccupés par les effets environnementaux du bruit ambiant et de la réverbération qui peuvent gravement nuire aux performances du RAE de la parole. Cela nous amène à notre troisième hypothèse.

- **H3**: un schéma de mise en commun des fonctionnalités (calculé à partir du spectre de modulation) qui combine les informations des trames voisines peut augmenter les performances du RAE de la parole, favorisant la robustesse vis-à-vis du bruit convolutif et additif.

Bien que la reconnaissance des émotions soit souhaitable dans de nombreuses situations, pour certaines applications informatiques, telles que l'VAL, la parole affective peut avoir un effet néfaste. La variabilité intra-locuteur, par exemple, causée par la parole émotionnelle, représente une menace réelle pour les performances des systèmes de reconnaissance du locuteur. Alors que de nombreux efforts ont été faits pour accroître la robustesse de la vérification automatique des locuteurs (ASV) vis-à-vis des effets de canal ou des attaques d'usurpation, seules quelques études, telles que dans [40] et [41], ont abordé les conséquences néfastes de l'affectif. discours. Cela motive notre quatrième hypothèse.

- **H4**: pour atténuer la variabilité intra-locuteur, causée par différentes émotions, un modèle de mélange gaussien peut être formé pour apprendre une distribution de probabilité antérieure

de la parole neutre pour un locuteur donné (c'est-à-dire caractériser son espace source) et cette connaissance peut ensuite être utilisée pour minimiser les différences entre les espaces cibles (affectifs) et sources (neutres).

L'authentification fait partie de nos vies. Tout au long de la journée, nous devons prouver notre identité généralement à l'aide de ce que nous possédons, ce que nous savons ou ce que nous sommes (par exemple, les clés de voiture).

0.1.2 Objectifs

L'objectif principal de cette thèse est d'augmenter l'expérience humaine tout en utilisant des technologies basées sur la voix. Nous abordons d'abord le problème de l'estimation de la qualité vocale perçue, puis nous examinons la question de l'amélioration de la fiabilité du RAE de la parole et du VAL.

Nos objectifs spécifiques sont :

1. pour développer une nouvelle mesure de qualité instrumentale de référence complète
2. pour développer une nouvelle mesure de qualité instrumentale sans référence
3. pour augmenter la fiabilité de RAE de la parole dans des environnements bruyants et réverbérants
4. pour développer une solution VAL qui augmente la fiabilité en atténuant les effets néfastes de la parole affective
5. pour développer une solution de contre-mesures pour rejouer les attaques sur les systèmes VAL

Par la suite, nous décrivons les principales contributions de cette thèse.

0.1.3 Principales contributions

Cette thèse vise à augmenter l'expérience humaine tout en utilisant des technologies basées sur la parole et ses principales contributions sont :

- Proposition d'une nouvelle mesure de qualité instrumentale de référence complète :

  L'utilisation de l'i-vector pour la mesure de la qualité instrumentale de la parole bruyante, réverbérante et améliorée est explorée. Nous montrons comment l'espace de variabilité total est capable de capturer les facteurs ambiants et une mesure complète et trois mesures sans référence sont proposées. Les résultats expérimentaux sur deux ensembles de données ont montré que la méthode de référence complète obtenait des résultats conformes à deux repères standard et contournait le besoin d'alignement temporel entre les signaux de référence et traités. Sur les mêmes ensembles de données, les trois mesures sans référence présentaient des corrélations plus élevées avec les scores de qualité subjective par rapport à deux références sans référence, montrant ainsi leur efficacité dans le suivi de la qualité de la parole mains libres et améliorée.

- Proposition d'une nouvelle mesure de qualité instrumentale sans référence :

  L'i-vector est proposé comme mesure de qualité de la parole non intrusive. La méthode présentée repose sur un modèle de mélange gaussien (GMM) pour reconstruire le spectre de référence propre à partir du signal vocal dégradé. Les i-vectors sont ensuite calculées pour les spectres propres et dégradés estimés et les corrélats de qualité sont obtenus au moyen de deux méthodes de notation différentes : la distance euclidienne et la distance cosinusoïdale. Les résultats expérimentaux ont montré que la méthode utilisée surpassait plusieurs algorithmes de référence non intrusifs et atteignait une précision alignée sur des algorithmes intrusifs, sans avoir besoin d'un signal de référence propre. Plus important encore, ladite méthode a montré une précision stable dans des conditions dégradées et améliorées, suggérant ainsi une meilleure applicabilité aux conditions émergentes.

- Proposition d'un système de reconnaissance des émotions vocales respectueux de l'environnement :

  Le RAE de la parole spontané "in the wild" est étudié. Des facteurs tels que le bruit, la réverbération et leurs effets combinés ont été explorés. Nous avons montré que les systèmes RAE de la parole existants basés sur des caractéristiques par trame (calculées à partir du spectre de modulation), bien qu'utiles pour les émotions mises en scène/posées, fonctionnent mal pour la parole spontanée. En tant que tel, un schéma de mise en commun des fonctionnalités qui combinent les informations des trames voisines est proposé. Cette mise en commun a considérablement contribué à améliorer les performances du RAE de la parole et s'est également révélée extrêmement importante pour la prédiction de la valence.

En effectuant la mise en commun des fonctionnalités, nous avons également observé une robustesse accrue contre le bruit environnemental. Par rapport à un algorithme de référence de l'AVEC 2016, le schéma de mise en commun des fonctionnalités présenté a fait mieux en présence de bruit et de réverbération bruit plus. Les gains ont été plus importants à mesure que les niveaux de bruit augmentaient, montrant ainsi les avantages des schémas proposés pour le RAE de la parole à l'état sauvage.

- Proposition d'un système VAL robuste à la parole affective :

  Une nouvelle méthode pour compenser les effets néfastes de la parole affective sur un système ASV est proposée. Nous adoptons notamment une méthode basée sur GMM pour « neutraliser » la parole affective, atténuant ainsi les effets néfastes de l'émotion sur la tâche de vérification du locuteur. La méthode utilisée, lorsqu'elle est couplée à un VAL à base de l'i-vector conventionnel, s'avère supérieure de 15% à une ligne de base. Les expériences sont effectuées sur quatre ensembles de données multilingues distincts, ainsi qu'avec un ensemble de données plus grand combiné, et les résultats obtenus montrent systématiquement que la méthode adoptée surpasse la ligne de base jusqu'à huit états émotionnels différents. Plus important encore, l'approche proposée ne compromet pas la performance VAL de la parole neutre, un problème couramment observé dans d'autres approches rapportées dans la littérature.

- Proposition d'une contre-mesure pour rejouer les attaques des systèmes VAL :

  Un front-end basé sur l'estimation de réponse de canal aveugle comme nouvelle approche pour la détection d'attaque de rejeu est proposé. Notre hypothèse est que les nuances de l'ambiance acoustique, des microphones et des appareils de lecture présents dans le spectre contiennent suffisamment d'informations pour établir la distinction entre une attaque de bonne foi et une attaque usurpée. Nous avons exploré un back-end de base basé sur des modèles de mélange gaussiens, ainsi qu'un classificateur de réseau neuronal résiduel profond. Les expériences sur les ensembles de données ASVspoof 2017 et ASVspoof 2019 Challenge montrent les méthodes présentées qui surpassent plusieurs systèmes de base de défi et offrent une robustesse améliorée contre l'inadéquation des ensembles entraînement/évaluation.

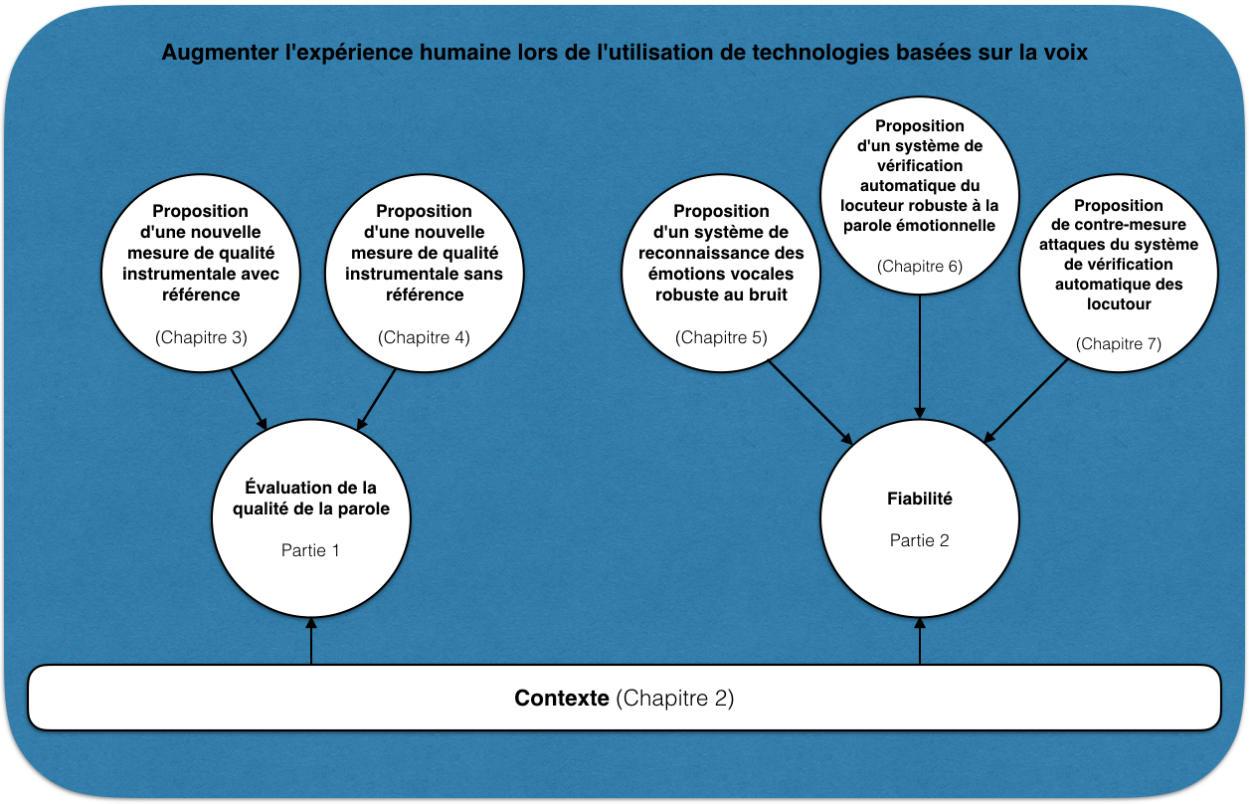


Figure 3 – Organisation de la thèse et relation entre les chapitres.

0.1.4 Organisation de la thèse

Cette thèse est organisée en deux parties principales. Un aperçu des chapitres et de leur relation peut être observé à la Figure 3. Dans le chapitre 2, nous couvrons le contexte lié aux principaux sujets abordés dans cette thèse. Le chapitre 3 présente la mesure de qualité instrumentale de référence complète proposée basée sur les i-vectors. Au chapitre 4, la mesure de la qualité instrumentale sans référence, également basée sur des i-vectors, est discutée. Ces deux chapitres appartiennent à la première partie de cette thèse où l'évaluation de la qualité de la parole est le sujet principal. La deuxième partie de cette thèse comprend les chapitres 5, 6 et 7, et est consacrée à l'amélioration de la fiabilité des technologies basées sur la parole. Au chapitre 5, nous présentons un système RAE de la parole robuste à l'environnement. Le chapitre 6 présente le système VAL proposé pour les environnements émotionnels. Au chapitre 7, une contre-mesure aux attaques d'accès physique aux systèmes VAL est exposée. Le chapitre 8 conclut la thèse en fournissant une discussion finale sur la limitation, l'orientation de la recherche et les travaux futurs.

0.2 Résumé

0.2.1 Chapitre 2: Contexte

Les systèmes qui permettent la communication interhumaine, tels que la voix sur IP, peuvent être évalués en termes de qualité de conversation. En d'autres termes, l'expérience utilisateur est liée à la qualité de l'échange d'informations rendu possible par ces systèmes. Des tests conversationnels, par exemple, ont été couramment appliqués pour effectuer de telles évaluations [4]. Dans l'interaction homme-machine, d'autre part, les systèmes peuvent être évalués en fonction de leur utilisabilité, c'est-à-dire si ces systèmes permettent une interaction efficace avec la machine. Par conséquent, les performances élevées des systèmes de transmission qui ont conduit à une bonne qualité de conversation ou à une bonne convivialité étaient supposées entraîner une grande satisfaction, augmentant ainsi le nombre de clients [42]. Néanmoins, certains services, tels que les premiers systèmes SMS, ont réussi malgré leur mauvaise qualité. Cela a mis en évidence la nécessité de comprendre la qualité du point de vue de l'utilisateur final. La perception de l'utilisateur devient désormais pertinente pour évaluer la qualité des services de communication. Cela a conduit à la formulation du terme «Qualité d'expérience» (QoE), qui peut être décrit succinctement comme le degré de satisfaction ou d'agacement de l'utilisateur envers une application ou un service [43].

Depuis de nombreuses années, la qualité des systèmes de communication est associée à la notion de « Qualité de service » (QoS), qui est définie dans [43] comme la « totalité des caractéristiques d'un service de télécommunication qui portent sur sa capacité pour satisfaire les besoins déclarés et implicites de l'utilisateur du service' ». Cette notion étant limitée aux performances du réseau et aux aspects physiques des systèmes de transmission, elle ne peut pas exprimer pleinement tout ce qui est compris dans un service de communication [44]. Ainsi, la QoE s'est épanouie en tant que concept pertinent non seulement pour les services multimédias, mais aussi pour IHM et l'esthétique [44]. Cela a motivé un changement d'orientation sur les fournisseurs qui se préoccupent désormais davantage de fournir des services avec une QoE élevée plutôt que des services avec uniquement une QoS élevée. Des tendances similaires ont également été observées pour les interfaces homme-machine interactives, « Expérience utilisateur » (EU), celles-ci devenant aussi pertinentes que l'utilisabilité classique (c'est-à-dire l'efficacité et l'efficience des services fournis) [24]. Par conséquent, l'aspect

humain devient le thermomètre de qualité des systèmes et services [45], et le processus de formation des jugements de qualité sonore par l'utilisateur doit être compris et pris en considération.

Ce chapitre se concentre sur les connaissances de base impliquant l'évaluation de la qualité de la parole ainsi qu'une brève description du matériel et des méthodes principaux utilisés tout au long de cette thèse. Le reste de ce chapitre est ensuite organisé comme suit. La section 2.2 fournit une discussion concernant l'évaluation de la qualité de la parole et la section 2.3 traite de la fiabilité des technologies basées sur la parole. Le paramétrage de la parole est présenté dans la section 2.4 avec les éléments fondamentaux du framework i-vector. Dans la section 2.5, nous présentons les principaux modèles de réseaux de neurones profonds utilisés dans cette thèse.

0.2.2 Chapitre 3: i-Vector pour la mesure de la qualité instrumentale de la parole traitée

Le cadre i-vector a été largement utilisé pour résumer les informations dépendantes du locuteur présentes dans un signal de parole. Le cadre peut être vu comme une procédure d'extraction de caractéristiques qui dépend essentiellement du signal vocal observé, d'un modèle du monde (UBM) et de la matrice de variabilité totale (ou matrice T), déjà discutée dans la section 2.4.3. Considéré comme l'état de l'art en matière de vérification du locuteur pendant de nombreuses années, son potentiel d'estimation de la distorsion / qualité de l'enregistrement vocal a été ignoré. Ce chapitre tente de combler cette lacune. Nous effectuons une analyse détaillée de la façon dont les distorsions sont représentées dans l'espace de variabilité totale. Nous présentons ensuite un estimateur intrusif de la qualité de la parole basé sur les similitudes de l'i-vector et trois approches non intrusives. Le premier utilise un seul i-vector de référence basé sur la moyenne des i-vectors extraits de signaux propres. Une deuxième approche repose sur un livre de codes à quantificateur vectoriel (VQ) de i-vectors représentatifs de la parole propre. Enfin, les i-vectors et NOM ont été utilisés pour former un modèle de réseau de neurones profond pour l'évaluation de la qualité de la parole non intrusive. Il est démontré à travers plusieurs expériences que la plupart des méthodes proposées sont bien adaptées pour évaluer la qualité de la parole et surpassent les mesures instrumentales bien établies telles que les algorithmes PESQ et POLQA, mais avec l'avantage supplémentaire de ne pas nécessiter d'alignement temporel avec une référence signal ou, dans le cas des méthodes non intrusives, un signal de référence tout à fait.

La principale motivation derrière ce travail réside dans le fait que les i-vectors sont connus pour transmettre à la fois des informations sur le canal et le locuteur. Néanmoins, la plupart des recherches dans le domaine se sont concentrées sur les caractéristiques du locuteur de la représentation (par exemple, pour la reconnaissance du locuteur) et les effets de canal ont été supprimés. Comme indiqué dans les recherches précédentes [30, 31], les performances des applications basées sur i-vector sont gravement affectées par des facteurs environnementaux, tels que le bruit de fond et la réverbération. Pour atténuer ces effets de canal, des techniques de compensation, telles que ADL et NCIC [32], sont couramment appliquées. Ici, contrairement aux travaux précédents, nous utilisons ces informations comme corrélat de la qualité de la parole perçue. De plus, comme les i-vectors sont mappés à un vecteur d'entité de longueur fixe, quelle que soit la longueur du signal d'origine, l'évaluation de la qualité de référence complète peut contourner l'alignement temporel, ce qui représente une étape cruciale et sujette aux erreurs pour PESQ et POLQA [4].

Conclusions

Dans ce chapitre, nous avons exploré l'utilisation de l'i-vector pour mesurer la qualité instrumentale de la parole bruyante, réverbérante et améliorée. Pour ce faire, nous avons montré comment l'espace de variabilité total est capable de capturer les facteurs ambiants et une mesure de référence complète, et trois mesures de non-référence ont été proposées. Les résultats expérimentaux sur deux ensembles de données ont révélé que la méthode de référence complète obtenait des résultats conformes à deux repères standard et contournait le besoin d'alignement temporel entre les signaux de référence et traités. Sur les mêmes ensembles de données, les trois mesures sans référence ont surpassé deux références sans référence, prouvant ainsi leur efficacité dans le suivi de la qualité de la parole mains libres et améliorée. En fait, une approche sans référence basée sur des réseaux de neurones profonds a surpassé les deux repères de référence complète, sans avoir besoin d'un signal de référence propre.

0.2.3 Chapitre 4: Mesure de la qualité de la parole basée sur une estimation du i-vector propre de la parole propre

L'évaluation de la qualité de la parole instrumentale non intrusive ne repose que sur le signal reçu (traité) pour prédire la qualité. Ces méthodes sont appelées non intrusives et s'avèrent cruciales

dans les applications vocales où les signaux propres de référence ne sont pas accessibles. Prédire la qualité de la parole sans le signal de parole propre de référence est une tâche difficile et offre souvent des performances inférieures par rapport aux méthodes intrusives. Cela est dû à la grande variabilité du signal vocal d'entrée, qui est le résultat de différents locuteurs, voies vocales, caractéristiques de hauteur et contenu de la parole [33]. Une façon de surmonter ce problème consiste à estimer le signal de parole propre à partir du signal de parole corrompu. La Recommandation UIT-T P.563 [46], par exemple, est basée sur une approche similaire. Le bloc principal de la mesure instrumentale non intrusive tente de séparer la parole du contenu non vocal. Il utilise ensuite une analyse statistique de haut niveau pour obtenir des informations supplémentaires sur le caractère naturel de la parole [46].

Le modèle comporte certains inconvénients. Premièrement, il est spécialement conçu pour la prédiction de la qualité de la parole dans les réseaux téléphoniques publics [46]. Il ne convient également qu'aux signaux vocaux à bande étroite échantillonnés à une fréquence d'échantillonnage de 8 kHz [47]. Par conséquent, il ne couvre que les types et l'ampleur des distorsions présentes dans la gamme d'occurrences courantes dans ces réseaux. Cela entraîne de mauvaises prévisions pour des scénarios plus récents et réalistes impliquant du bruit de fond, de la réverbération et une amélioration de la parole [46].

Le rapport de modulation parole-réverbération (SRMR) a été développé comme alternative. La méthode tente de séparer les composants propres et de réverbération/bruit du signal dégradé. En fait, il repose sur le principe selon lequel l'énergie de modulation d'une parole claire est généralement concentrée dans des fréquences de modulation plus faibles (inférieures à 20 Hz) tandis que les artefacts acoustiques de la pièce surviennent généralement dans des fréquences de modulation plus élevées supérieures à 20 Hz [47]. Bien que la métrique se révèle très efficace pour évaluer la qualité de la parole réverbérante, elle n'a montré que des performances modérées avec des types de distorsions plus récents [48].

Des travaux plus récents ont proposé le réseau neuronal profond (DNN) comme mesures instrumentales non intrusives. Dans [48] et [27], par exemple, les auteurs présentent une étude de l'applicabilité d'approches de réseaux de neurones profonds pour estimer le NOM sans le signal de référence. Dans [49], une architecture de réseau de neurones profonds, basée sur le système auditif humain, est proposée pour extraire des caractéristiques à utiliser pour une évaluation objective de la qualité non intrusive. Dans [28], les auteurs présentent une nouvelle mesure non intrusive basée sur

un réseau neuronal récurrent utilisant des cellules de mémoire à court terme et des caractéristiques d'énergie de modulation. Bien que ces résultats soient prometteurs, les modèles basés sur un réseau de neurones nécessitent souvent une énorme quantité de données pour surpasser de manière significative les algorithmes d'apprentissage machine plus traditionnels ainsi que les approches basées sur l'ingénierie des fonctionnalités. Par conséquent, il est important d'étudier des modèles non intrusifs qui peuvent bien fonctionner dans des scénarios où une grande quantité de données étiquetées n'est pas disponible. Dans ce chapitre, nous entendons combler cette lacune en proposant une nouvelle mesure de qualité instrumentale non intrusive basée sur la similitude entre deux i-vectors.

Nous mettons notamment de l'avant le cadre i-vector comme mesure de qualité de la parole non intrusive. Afin de surmonter le problème de l'indisponibilité du signal de parole propre de référence, nous proposons de reconstruire les spectres propres à partir du signal dégradé lui-même. Dans notre solution exposée, un GMM de discours propre doit être formé avec des coefficients cepstraux de fréquence de mel filtrés par RASTA (RASTA-MFCC), extraits de plusieurs fichiers de discours clair. Cela nous permet d'atteindre un modèle de caractéristiques de spectre propres. Ainsi, le discours propre GMM nous permet d'extraire la représentation i-vector proposée en utilisant les spectres propres de référence estimés à partir du signal corrompu.

Conclusions

Dans ce chapitre, nous présentons le cadre i-vector pour la mesure non intrusive de la qualité de la parole. La méthode utilisée repose sur un modèle de mélange gaussien (GMM) pour estimer un spectre de référence propre à partir du signal vocal dégradé. Les i-vectors sont ensuite calculées pour les spectres propres et dégradés estimés et les corrélats de qualité sont obtenus au moyen de deux méthodes de notation différentes. Les résultats expérimentaux ont montré que la méthode proposée surpassait plusieurs algorithmes de référence non intrusifs et atteignait une précision alignée sur des algorithmes intrusifs, sans avoir besoin d'un signal de référence propre. Plus important encore, la méthode adoptée a montré une précision stable dans des conditions dégradées et améliorées, suggérant ainsi une meilleure applicabilité aux conditions émergentes.

0.2.4 Chapitre 5: Mise en commun des caractéristiques pour une meilleure reconnaissance des émotions vocales « in-the-wild »

De nouveaux algorithmes d'apprentissage automatique basés sur des techniques d'apprentissage approfondi ont permis aux ordinateurs de résoudre des problèmes complexes en assimilant les informations directement à partir des données brutes de la même manière que les humains [50]. Malgré les percées dans de nombreux domaines importants tels que la reconnaissance automatique de la parole (RAP) [51] et la reconnaissance d'objets [52], les tâches impliquant la compréhension automatisée des états émotionnels humains restent non résolues. Cela est devenu une préoccupation croissante pour la communauté de l'intelligence artificielle (IA) car, dans un avenir proche, les machines intelligentes devraient posséder un certain niveau de compréhension des émotions humaines afin de garantir que des décisions sûres, impliquant des tâches liées aux cognitions, seront prises [53]. Pour atténuer cette limitation, l'informatique affective vise à permettre aux machines de reconnaître, d'analyser et de synthétiser les états affectifs humains [54].

En fait, l'intérêt pour l'informatique affective est en plein essor, en grande partie en raison de son rôle dans les interfaces affectives homme-ordinateur émergentes. À ce jour, la majorité des recherches existantes sur l'analyse automatisée des émotions se sont appuyées sur des données collectées dans des environnements contrôlés. Cependant, avec l'essor des applications IHM sur les appareils mobiles, les paramètres dits « à l'état sauvage » ont constitué une menace sérieuse pour les systèmes de reconnaissance des émotions, en particulier ceux basés sur la voix. Dans ce cas, des facteurs environnementaux tels que le bruit ambiant et la réverbération entravent gravement les performances du système. Dans ce chapitre, nous quantifions les effets néfastes de l'environnement sur la reconnaissance des émotions et explorons les avantages réalisables avec l'amélioration de la parole. De plus, nous présentons un système de mise en commun des caractéristiques spectrales de modulation qui s'avère supérieur à un système de référence de pointe pour la prédiction robuste à l'environnement des éveils spontanés et des primitives émotionnelles de valence. Deux DNN sont également explorés, à savoir un réseau de perceptrons multicouches (MLP) et le réseau neuronal récurrent basé sur la mémoire à long terme et à court terme (LSTM). Leurs performances sont comparées à celles du benchmark basé sur la machine à vecteurs de support (SVM) pour quantifier les avantages d'un système de machine learning par rapport à un autre pour un RAE de la parole « in-the-wild ». Des expériences sur une version corrompue de l'environnement du jeu de données RECOLA

d'interactions spontanées montrent le schéma de mise en commun des fonctionnalités proposé, combiné à l'amélioration de la parole, surpassant la référence dans différentes conditions de bruit uniquement, de réverbération uniquement et de bruit plus réverbération. Des tests supplémentaires avec la base de données SEWA montrent les avantages de la méthode proposée pour les applications "in the wild".

Conclusion

Ce chapitre a exploré la reconnaissance spontanée des émotions de la parole "in the wild", où des facteurs tels que le bruit, la réverbération et leurs effets combinés, ont été explorés. Nous montrons que les systèmes RAE de la parole existants basés sur des caractéristiques par trame (calculées à partir du spectre de modulation), bien qu'utiles pour les émotions mises en scène/posées, fonctionnent mal pour la parole spontanée. En tant que tel, un schéma de mise en commun des fonctionnalités est proposé, et celui-ci combine les informations des trames voisines. Cette mise en commun a largement contribué à booster les performances du RAE de la parole; il s'est également révélé extrêmement important pour la prévision de la valence, même dans des conditions de propreté. En effectuant la mise en commun des fonctionnalités, nous avons également observé une robustesse accrue contre le bruit environnemental. Par rapport à un algorithme de référence du Audio/Video Emotion Challenge 2016, le schéma de mise en commun des fonctionnalités proposé a fait mieux en présence de bruit et de réverbération bruit-plus. Les gains se sont avérés plus importants à mesure que les niveaux de bruit augmentaient, montrant ainsi les avantages des schémas proposés pour le RAE de la parole à l'état sauvage. Nos résultats ont été reproduits dans un sous-ensemble de l'ensemble de données SEWA, qui est considéré comme sauvage. Les expériences de cette base de données corroborent la principale contribution de ce chapitre, c'est-à-dire que l'application du regroupement de fonctionnalités est essentielle pour augmenter les performances des fonctionnalités basées sur le spectre de modulation. Nous avons également constaté que pour les prédictions d'excitation, notre méthode a surpassé le système de référence utilisé pour l'Audio/Video Emotion Challenge 2017. Dans cette étude, nous avons exploré l'inadéquation train/test, où la parole propre était utilisée pour la formation et la parole dégradée pour les tests.

0.2.5 Chapitre 6: Vérification automatique du locuteur à partir d'un discours émotionnel

Les avancées récentes dans les techniques de compensation des canaux [55, 56, 57, 58] et l'utilisation des intégrations d'apprentissage en profondeur, telles que les x-vectors [25], ont fait passer la biométrie vocale au niveau supérieur. Néanmoins, l'atténuation des effets de la parole affective sur la VAL est restée une question ouverte et la robustesse à la variabilité intra-locuteur causée par la parole affective demeure un défi. Des études antérieures ont porté sur les effets néfastes de l'effort vocal [59], de l'inadéquation du langage [60] et de la variation du style de parole [61], mais seules quelques études ont spécifiquement présenté des méthodes pour atténuer la VAL dégradation des performances due à la parole expressive.

Les auteurs de [62], par exemple, ont proposé de modifier les paramètres prosodiques tels que la durée, la hauteur et l'amplitude de la parole affective pour atténuer les effets de l'inadéquation entre l'inscription et les énoncés de test. Le système a été formé avec des caractéristiques extraites de la parole neutre et des caractéristiques modifiées extraites de la parole émotionnelle. Bien que les trois paramètres prosodiques utilisés aient été largement explorés pour synthétiser la parole expressive, il n'y a pas d'accord sur le fait qu'ils peuvent être universellement utilisés pour toutes les émotions [63]. De plus, l'étude semble négliger le fait que chaque locuteur comporte des caractéristiques uniques dans la façon dont l'émotion est exprimée, ce qui rend une solution de VAL indépendante du locuteur dans des environnements émotionnels assez difficile.

Dans [64], les auteurs ont eu recours à un cadre comprenant trois étapes de classification : (1) l'identification du sexe, (2) la reconnaissance des émotions et (3) une étape de vérification du locuteur. Bien que les auteurs aient affirmé que la méthode proposée se révélait supérieure aux systèmes de vérification des locuteurs basés uniquement sur le sexe ou l'émotion uniquement, les performances du système global se sont avérées sensibles aux erreurs dans ces classificateurs, augmentant ainsi la demande d'une reconnaissance émotionnelle très précise. Dans des paramètres réalistes, cela peut être difficile [23]. Les auteurs de [40] ont utilisé le cadre i-vector et ont proposé d'estimer la fiabilité des systèmes de VAL en tenant compte des dimensions émotionnelles telles que l'excitation, la valence et la dominance. Pour ce faire, les performances du système de reconnaissance de locuteur sont cartographiées en fonction des primitives d'excitation et de valence. Un module de reconnaissance des émotions vocales est ensuite formé pour prédire si le contenu émotionnel tombe sur une région

fiable. Dans une étude de suivi, les auteurs ont ensuite analysé les performances de vérification des locuteurs en termes d'éveil (calme/actif), de valence (négatif/positif) et de dominance (faible/fort) [41].

Dans une étude plus récente [65], les auteurs explorent le transfert de l'apprentissage des connaissances acquises pour la tâche de reconnaissance du locuteur à la RAE de la parole. Ils ont donc utilisé un modèle pré-formé pour extraire des x-vectors, qui ont également été utilisés pour étudier l'impact de l'émotion sur la vérification du locuteur. Cette recherche met en lumière deux aspects importants de la vérification du locuteur dans des environnements émotionnels. Premièrement, elle montre que les plongements extraits pour la reconnaissance du locuteur exécutent des signaux émotionnels pertinents qui pourraient être caractérisés par la façon dont chaque locuteur exprime particulièrement ses émotions. En d'autres termes, cela suggère que la reconnaissance du locuteur peut tirer parti des traits émotionnels spécifiques du locuteur, utiles également pour la RAE de la parole. Deuxièmement, le travail rapporte un taux d'erreur égal élevé (EER), même lorsque le même discours affectif (par exemple, en colère) est utilisé à la fois pour l'inscription et le test, ce qui suggère que les modèles de locuteur dépendant des émotions peuvent ne pas être suffisants pour atténuer les effets néfastes du discours affectif sur ASV. Comme leur système est basé sur un x-vector pré-formé, ce problème peut être surmonté en entraînant directement le x-vector avec des données émotionnelles. Cependant, comme indiqué dans [25], les incorporations basées sur des x-vectors nécessitent une grande quantité de données étiquetées avec l'ID du locuteur, qui se trouve rarement dans les ensembles de données émotionnelles.

La plupart des travaux mentionnés ci-dessus se sont appuyés sur un seul ensemble de données de discours émotionnel avec peu de locuteurs pour mener leurs expériences. En tant que tel, il n'est pas clair si ces résultats persistent dans tous les ensembles de données,

Conclusions

Dans ce chapitre, nous avons mis de l'avant une nouvelle méthode pour compenser les effets néfastes que la parole émotionnelle a sur un système de VAL. Nous présentons plus particulièrement une nouvelle méthode basée sur un GMM pour « neutraliser » la parole affective, atténuant ainsi les effets de non-concordance des conditions néfastes sur la tâche de vérification du locuteur. La méthode proposée, lorsqu'elle est couplée à un VAL à base de i-vector conventionnel, s'avère supérieure de 15

% à une ligne de base. Les expériences sont effectuées sur quatre ensembles de données multilingues distincts, ainsi qu'avec un ensemble de données plus grand combiné, et les résultats obtenus montrent de manière cohérente la méthode proposée surpassant la ligne de base jusqu'à huit états émotionnels différents. Plus important encore, l'approche proposée ne compromet pas la performance VAL de la parole neutre, un problème couramment observé dans d'autres approches rapportées dans la littérature. À mesure que de nouveaux ensembles de données émotionnelles deviennent disponibles, d'autres systèmes VAL basés sur des architectures d'apprentissage en profondeur plus récentes peuvent devenir une alternative viable.

0.2.6 Chapitre 7: Estimation de la réponse des canaux et réseau neuronal résiduel pour détecter les attaques physiques à la vérification automatique des locuteurs

La VAL a considérablement mûri au cours des dernières années [66]. Les avancées réalisées dans les techniques de compensation des canaux [55] [56] et l'utilisation des intégrations d'apprentissage en profondeur, telles que les x-vector [25] [67], ont porté la vérification automatique des haut-parleurs à un niveau supérieur. Le déploiement de produits commerciaux de reconnaissance vocale mobile est déjà devenu une réalité [68]. Pour améliorer les mécanismes d'authentification par mot de passe, par exemple, un certain nombre d'institutions financières investissent dans des solutions d'authentification vocale [69]. Cela s'explique principalement par l'utilisation accrue des appareils mobiles, ainsi que par la commodité et la non-intrusion offertes par ces technologies. En fait, des rapports récents prédisent une croissance continue du secteur de la biométrie mobile en raison de la demande accrue de sécurité des consommateurs, en particulier lors de l'utilisation d'appareils mobiles pour les transactions bancaires et le commerce électronique [69].

Malgré toutes ces avancées, les attaques malveillantes d'usurpation d'identité ont été reconnues comme une menace sérieuse pour la VAL [66]. Caractérisée par une tentative d'une personne ou d'un programme de contourner illégalement la sécurité en masquant son identité, la vulnérabilité de la VAL face à des attaques d'usurpation, telles que l'emprunt d'identité, les attaques par rejeu, la synthèse vocale et la conversion vocale, suscite de plus en plus d'inquiétudes. wu2015spoofing. En tant que tel, une poignée d'initiatives pour développer des contre-mesures d'usurpation d'identité ont été prises récemment [70] [71]. De nombreux efforts dans ce sens ont été axés sur le développement

de techniques anti-usurpation pour protéger les systèmes VAL contre la synthèse vocale (SV) et la conversion vocale (CV) [70]. Dans cette étude, nous nous intéressons particulièrement aux contre-mesures pour rejouer les attaques, qui consistent à tenter de tromper un système de la VAL en lisant un échantillon de parole préenregistré. Dans de telles circonstances, la détection préalable de l'attaque de relecture s'avère cruciale pour maintenir la fiabilité de la VAL.

Compte tenu de l'intérêt croissant pour le sujet, le défi de vérification automatique de l'usurpation et des contre-mesures a été créé, les dernières versions datant de 2017 [71] et 2019 [72]. Désormais appelés ASVspoof 2017/2019, ces défis ont fourni des bases de données, des protocoles et des métriques communs pour évaluer différentes solutions de contre-mesures. Les deux compétitions contenaient des ensembles de données avec diverses formes d'attaques de rejeu, l'ensemble 2019 offrant un plus grand nombre de configurations d'attaque de rejeu. Jusque-là, les attaques par rejeu, également appelées attaques physiques, avaient reçu peu d'attention de la part de la communauté de recherche par rapport à d'autres modalités d'usurpation (par exemple, la synthèse vocale).

Dans [73], par exemple, les auteurs proposent une solution de contre-mesure de bout en bout utilisant la forme d'onde brute. Alors que de nombreuses méthodes de contre-mesures sont basées sur une combinaison d'extraction de fonctionnalités et d'un classificateur principal, cette méthode rejette la nécessité de tout prétraitement sur la forme d'onde de la parole. Bien que l'ingénierie des fonctionnalités puisse faciliter l'interprétation des fonctionnalités extraites, elle peut négliger les informations vocales riches qui pourraient être trouvées dans la forme d'onde brute par une approche de bout en bout. Dans [74], les auteurs ont utilisé une architecture de réseau neuronal résiduel profond (ResNet-18), avec un mécanisme d'attention visuelle sur les représentations temps-fréquence basées sur des caractéristiques de retard de groupe, comme contre-mesure pour les attaques par rejeu. Les résultats obtenus en termes de taux d'erreur égal (EER) étaient assez faibles, mais uniquement rapportés dans l'ensemble de données ASVspoof 2017. Dans [75], les auteurs ont suggéré de compléter les caractéristiques spectrales à court terme par deux nouvelles fonctionnalités basées sur le spectre de modulation. Ce dernier capture les caractéristiques statiques et dynamiques du signal de parole du spectre de modulation, qui complètent les caractéristiques spectrales à court terme pour une utilisation dans la détection de relecture. Les auteurs de [76], à leur tour, se sont appuyés sur des bitmaps spectraux ou des pics spectraux, qui sont des points temps-fréquence supérieurs à un seuil prédéfini. Le score de similitude a été atteint en calculant un produit élément par élément entre le bitmap spectral de l'échantillon de vérification et les modèles de bitmap spectral stockés. Plus

récemment, les performances de plusieurs fonctionnalités et classificateurs ont été décrites dans [77]. Les auteurs ont rapporté les résultats de six fonctionnalités basées sur le spectre d'amplitude et trois fonctionnalités basées sur le spectre de phase sur le défi de détection d'attaque de relecture ASVspoof 2017, avec des expériences révélant la supériorité des fonctionnalités du spectre d'amplitude sur les fonctionnalités basées sur la phase pour les quatre classificateurs testés. Un réseau de filtrage attentif combiné à un classificateur basé sur ResNet est proposé dans [78], générant ainsi de meilleures caractéristiques discriminantes à la fois dans les domaines temporel et fréquentiel.

Malgré les récents progrès dans ce domaine, il est toujours nécessaire d'étudier de nouvelles solutions de contre-mesures applicables aux scénarios émergents et plus difficiles. Dans ce travail, nous mettons de l'avant l'utilisation d'estimation de spectre de canal aveugle en combinaison avec une classification basée sur un réseau de neurones profond pour détecter les attaques de rejeu. Étant donné que dans une attaque de relecture, l'énoncé sera affecté acoustiquement par des facteurs tels que l'environnement de la pièce, l'enregistrement et les appareils de lecture, il est prévu que ces effets génèrent une « signature » unique dans le spectre de magnitude logarithmique du signal. Par conséquent, nous faisons le choix de détecter de telles signatures spectrales en estimant la réponse en amplitude du canal.

Ici, cela est obtenu en entraînant d'abord un GMM en langage prope. Le modèle est formé à l'aide de coefficients cepstraux de fréquence de mel filtrés par RASTA (RASTA-MFCC) extraits de plusieurs fichiers de discours propres, ce qui nous permet d'atteindre un modèle de caractéristiques de spectre propres. Le spectre de réponse de canal est ensuite estimé en calculant la moyenne du spectre de magnitude logarithmique des signaux propres et en le soustrayant ensuite du spectre de magnitude logarithmique du signal observé. En tant que classificateur, nous avons adopté le GMM de référence pour faire la distinction entre les énoncés vrais et ceux qui sont usurpés. Ensuite, motivés par les résultats récents obtenus avec ResNets [79, 80], nous explorons également l'utilisation de tels réseaux.

Les résultats expérimentaux montrent que la méthode proposée surpasse les repères sur les ensembles de développement et d'évaluation pour les ensembles de données ASVspoof 2017 et 2019. À notre connaissance, seules quelques études ont abordé l'utilisation de l'estimation de canal comme solution de contre-mesure. Dans [81], par exemple, les auteurs ont proposé l'utilisation de deux descripteurs de bas niveau, les coefficients cepstraux à Q constant (CQCC) et les coefficients

cepstraux à haute fréquence (HFCC), comme entrée dans un réseau neuronal convolutionnel (CNN). Les auteurs affirment que le modèle CNN estime les conditions du canal, bien qu'aucune explication claire ne soit donnée concernant la façon dont le canal est estimé. Le présent travail fournit une étude plus complète de l'utilisation de l'estimation de canal, ce qui représente une contribution importante à l'atténuation du problème des attaques par répétition d'usurpation. De plus, par rapport à notre travail précédent [82], cette étude (1) présente des résultats supplémentaires et améliorés sur un ensemble de données étendu; (2) évalue l'impact de la résolution de l'approche d'estimation de canal sur les performances de détection d'usurpation; (3) effectue une analyse de la qualité des deux ensembles de données testés et discute de l'impact de la qualité du signal et de la précision de la détection d'usurpation; et (4) présente des améliorations de performances avec ResNet, tout en comparant les résultats avec un algorithme de pointe.

Conclusions

Dans ce chapitre, nous avons mis de l'avant l'utilisation de l'estimation de la réponse de canal aveugle comme nouvelle approche pour la détection d'attaque par rejeu. Notre hypothèse est que les nuances de l'ambiance acoustique, des microphones et des appareils de lecture présents dans le spectre contiennent suffisamment d'informations pour faire la distinction entre une attaque authentique et une usurpation d'identité. Nous avons exploré un back-end de base basé sur des modèles de mélange gaussiens, ainsi qu'un classificateur de réseau neuronal résiduel profond. Les expériences sur les ensembles de données ASVspoof 2017 et ASVspoof 2019 Challenge montrent que les méthodes utilisées surpassent plusieurs systèmes de base de défi et offrent une robustesse améliorée contre l'inadéquation des ensembles train/évaluation. Une discussion sur les effets de la qualité du signal sur les performances de détection d'usurpation est également rapportée, fournissant ainsi quelques informations préliminaires sur la meilleure façon de former des modèles pour la tâche à accomplir.

Introduction

According to the Theory of Signs of Charles Sanders Peirce, a sign is anything which signifies something else, namely an object, thus determining its meaning that is later understood by the interpreter. In his own words: *'I define a sign as anything which is so determined by something else, called its Object, and so determines an effect upon a person, which effect I call its interpretant, that the later is thereby immediately determined by the former'* [7]. One can think of the sign as the signifier. In respect to language, for example, a written word or an acoustic speech signal is the form in which the sign is presented. The object is whatever is signified, for example, what the written word or the acoustic speech signal refers to. The meaning is the sense made of the sign by its interpreter and that establishes the sign/object relation [3]. This is represented by the semiotic triangle shown in Figure 1.1. In this context, telecommunication enables human-human interaction via transmission of signs (i.e., messages, words, writings, images and sounds) by wire, radio, optical or other electromagnetic systems. In a telephone conversation, for example, humans communicate with each other by exchanging speech signs. While investigating speech quality, it is useful to make a clear distinction between the acoustic form of the speech and its content (or meaning) [1]. In such cases, the main object of interest is the acoustic form, although the content plays an important role in the perception of speech, and on how a certain quality is associated with the acoustic form. Ultimately, it is the human talker and listener who determine the relationship between the acoustic form and its content as no direct relationship can be established between the two by neglecting the interpreter [1].

Information and communication technology (ICT) has significantly developed over the past decade. Providers of multimedia communication systems are facing a multitude of challenges as

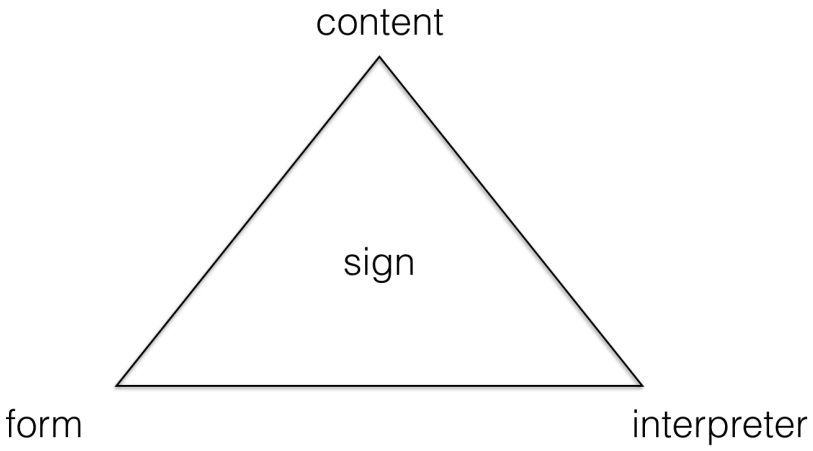


Figure 1.1 – Semiotic triangle representing the triadic sign. Adapted from [1].

novel coding methods [8], new and more sophisticated signal processing algorithms [9, 10], as well as more modern transmission approaches are being deployed every day [11]. A growing concern regards speech communication systems, as voice services based on a variety of transmission technologies are being provided by telecommunication networks [12, 13]. As our main way of communication, the speech signal, during its acquisition, transmission or distribution, is likely to have its quality and intelligibility reduced by different impairments. Figure 1.2 shows the typical impairments present in a communication channel [2, 14]. It accounts for the user interface at both ends of the communication channel (e.g., microphone devices and loudspeakers), and the transmission medium itself (e.g., copper wire, wireless, fibre optic, etc.) [2]. A speech signal, for instance, may be affected by bandwidth constraints, variability in channel conditions and errors in acquisition or storage [15]. Moreover, background noise and reverberation may also be a source of additional degradation, compromising the overall quality of the speech signal [16]. In fact, the quality of a speech telecommunication service is greatly affected by the quality of the transmission system. Nevertheless, the overall perceived quality cannot be assumed to be solely related to physical aspects of the transmission. It is instead determined as well by the user. When the user has enough experience operating a determined telecommunication service, a notion of quality is developed and then expected. In this sense, quality can be understood as the fulfillment of certain desired and expected characteristics as perceived by the user [1].

Traditionally, the most reliable way to assess the quality of a speech communication system comes from human subjects. Subjective listening tests have been widely used in that regard and a series of recommendations describing methods and procedures for conducting such subjective

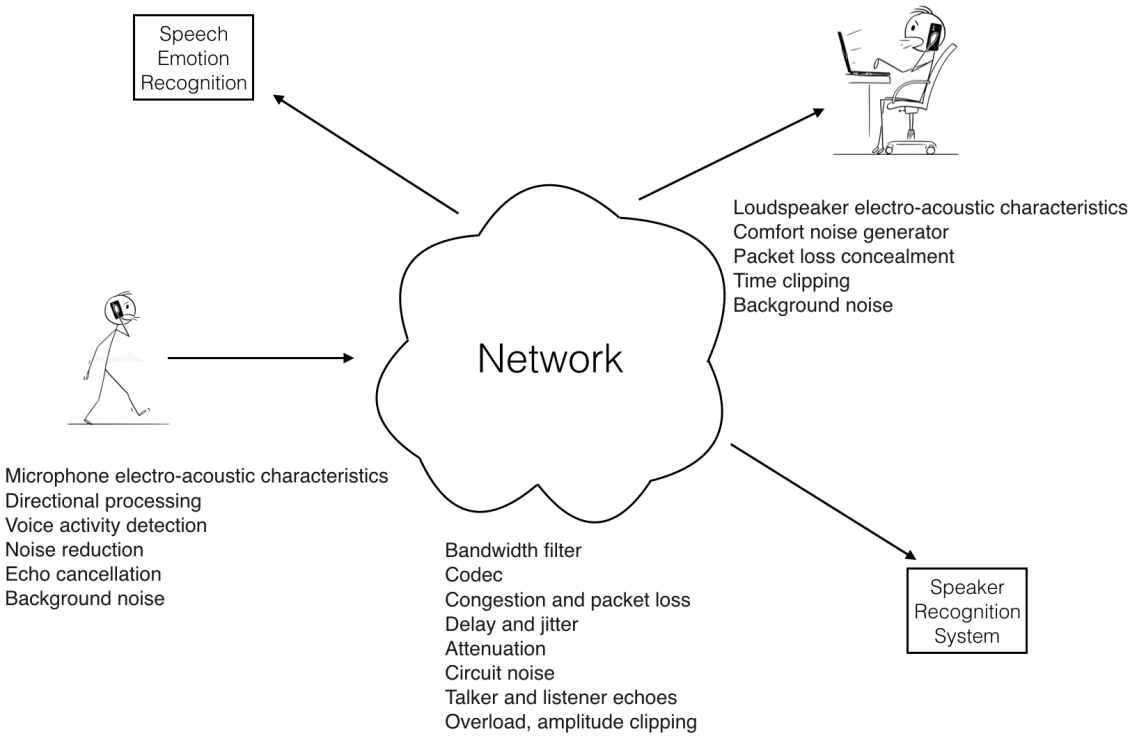


Figure 1.2 – Main impairments of an end-to-end communication channel. Adapted from [2].

evaluations has been made available by the International Telecommunication Union (ITU-T) in its ITU-T Recommendation P.800 [17]. In such subjective tests, speech signals are presented to listeners (either naive or experts, depending on the application) who judge the signal quality on an Absolute Category Rating (ACR) scale ranging from 1 to 5, representing "bad" and "excellent", respectively. The mean opinion score (MOS), which represents the perceived speech quality after leveling out individual factors [4], is attained by averaging all participant scores over a specific condition. Such subjective measurements, however, are not always practical as they: (1) require many listeners; (2) can be laborious and time-consuming; (3) can be expensive; and (4) cannot be performed in real-time [18]. Therefore, an algorithm for estimating perceived speech quality might be the best option during network planning. Instrumental quality measures, in fact, have been explored over the years to overcome these limitations and they are built to be highly correlated with subjective listening MOS scores, thus effectively replacing the listener panel by a real-time computational algorithm [4].

Estimating the perceived quality of existing and emerging multimedia services and applications is important, especially for speech communication providers seeking to optimize their services and maximize customer experience [19]. Real-time quality monitoring, for example, can help with

network design and development, as well as with online adaptation to assure that the end users' expectations are met. Moreover, studies have shown some correlation between speech quality and the performance of voice-based technologies such as automatic speaker verification (ASV) and speech emotion recognition (SER) [20, 21, 22]. These systems may have good performance in more controlled settings, but are severely affected by mismatch between training utterances, which may not account for unseen channel impairments and by utterances transmitted through different communication channels, causing a decrease in their performance [23]. Instrumental quality measurement, therefore, could be useful for predicting the performance of such applications for a given communication channel configuration. For example, as suggested in Figure 1.2, the authentication of the user's voice, while performing a financial transaction over the telephone, might require the transmission of the speech signal to be processed remotely. This could cause degradation of the quality of the speech signal, affecting as well the reliability of those applications.

Reliability is yet an important component of quality that may influence the user experience (UX) [24]. It consists of guaranteeing that a system will perform consistently and as expected by the end user. Thus, the UX of non-reliable systems is certainly affected when the users are not experiencing the same results on repeated trials. Because many speech-based technologies have made their way out of laboratory settings to be employed in real-world applications, a rising concern by the research community regards the performance of such technologies in more realistic scenarios. Here, we are specifically interested in SER and ASV systems. Despite the recent advances made in the field, mainly driven by new embeddings based on deep neural network models [25], in some circumstances the quality of the speech signal is not good enough to make reliable decisions [26].

This thesis deals with the perceptual quality and human experience while using voice-based systems. We are, particularly, concerned with the estimation of perceived speech quality and with improving the reliability of speech-based technologies. We propose the use of the i-vector framework as a new objective quality estimation tool. Following, we propose an environment-robust feature pooling scheme to increase SER reliability "in the wild" [1]. As emotional speech can be detrimental to ASV performance, in order to maintain reliability of these systems, we propose a new method to learn neutral speech characteristics to neutralize emotional speech. Lastly, we propose a front-end based on the use of blind estimation of the channel response magnitude and a residual neural network

1. Here, "in the wild" means beyond laboratory settings, where acoustic conditions, present in real-world scenarios, such as noise and reverberation, lead to a significant mismatch between train and test conditions.

to mitigate the threat of physical access attacks on ASV. In the next section, we detail the research problems addressed in this work and the respective hypotheses towards them.

1.1 Research Problem and Hypotheses

In speech communication the recorded signal is often corrupted by ambient noise and reverberation in the transmitting location. This may severely degrade the perceived speech quality and intelligibility. Detrimental effects of environmental noise can be mitigated by the use of speech enhancement algorithms such as noise reduction, echo cancellation, and de-reverberation [27]. Although such algorithms aim at enhancing the speech signal by reducing the amount of noise and reverberation, they may also introduce unwanted artifacts that degrade the speech quality [28]. In such scenarios, assessing the perceived speech quality is crucial [18, 29]. Subjective listening tests are usually considered the most reliable way to evaluate the quality of speech, but are costly and time-consuming. As an alternative, several objective instrumental quality measures have been proposed and standardized by the ITU-T. Standardized objective algorithms, however, may become obsolete as new scenarios, such as far field sound capture, new audio compression algorithms, and new speech enhancement models emerge [27]. To address this issue, we explored the use of i-vector speech representations for instrumental quality measurement of noisy, reverberant and enhanced speech. This leads to our first hypothesis.

- **H1**: i-vectors have the potential to capture not only speaker-dependent information, but also information related to distortion and quality present in the speech signal and, therefore, can be used as a full-reference instrumental quality measure.

The main motivation behind this lies in the fact that i-vectors are known to convey both channel and speaker information. Nevertheless, most research in the field has focused on the speaker characteristics of the representation (e.g., for speaker recognition). As shown in previous research [30, 31], the performance of i-vector based applications is severely affected by environmental factors, such as background noise and reverberation. Thus, indicating that such environmental factors are also capture by the i-vectors. To mitigate these channel effects, compensation techniques, such as linear discriminant analysis (LDA) and within class covariance normalization (WCCN) [32], are

commonly applied. Unlike these works, we utilize this information as a correlate of perceived speech quality.

Full-reference instrumental quality measures require access to the reference clean speech signal and its degraded counterpart [4]. Most intrusive measures, therefore, rely on computing the distance between some auditory representation of the reference and degraded signals to estimate the speech quality. No-reference measures, on the other hand, depend only on the degraded signal to estimate the speech quality. Predicting the speech quality without the reference clean speech signal is a challenging task and often offers lower performance compared to intrusive methods. This is due to the wide variability of the input speech signal, which is the result of different speakers, vocal tracts, pitch characteristics and speech content [33].

Our first hypothesis relies on the availability of the reference clean signal to estimate the perceived quality. This is not always possible for certain speech applications as the reference clean signal may not be available. This brings to light the following question: can the i-vector speech representation be used as a no-reference instrumental quality measure? Our second hypothesis attempts to answer this question.

- **H2**: it is feasible to use the i-vector representation as a no-reference instrumental quality measure, as the reference clean signal can be attained by using a clean speech Gaussian mixture model (GMM) that reconstructs the reference clean spectra characteristics from the corrupted signal.

The second part of this thesis deals with reliability of speech-based technologies. We first focus on the robustness of SER. With the rise of human-computer interaction (HCI) applications, especially on mobile devices, the so-called “in the wild” settings have posed a serious threat for emotion recognition systems. To accelerate innovations in the field, several challenges have been organized over the last few years, most notably the INTERSPEECH Emotion Challenge [34, 35], the Emotion Recognition in The Wild Challenge (EmotiW) [36, 37], and the 2016 and 2017 Audio/Visual Emotion Challenges (AVEC) [38, 39]. We are, particularly, concerned with the environmental effects of ambient noise and reverberation that can severely hamper SER performance. This leads us to our third hypothesis.

- **H3**: a feature pooling scheme (computed from the modulation spectrum) that combines information from neighbouring frames can boost SER performance, promoting robustness towards convolutive and additive noise.

Although recognizing emotion is desirable in many situations, for some computer applications, such as ASV, affective speech can have a detrimental effect. Intra-speaker variability, for example, caused by emotional speech, represents a real threat to the performance of speaker recognition systems. While many efforts have been made to increase automatic speaker verification (ASV) robustness towards channel effects or spoofing attacks, only a handful of studies, such as in [40] and [41], have addressed the detrimental consequences of affective speech. This motivates our fourth hypothesis.

- **H4**: to mitigate intra-speaker variability, caused by different emotions, a Gaussian mixture model can be trained to learn a prior probability distribution of neutral speech for a given speaker (i.e., characterizing his/her source space) and this knowledge can later be used to minimize the differences between target (affective) and source (neutral) spaces.

Authentication is a common part of our lives. Throughout the day, we must prove our identity usually by what we possess, know, or what we are (e.g., car keys, password, and fingerprint, respectively). Whether getting into a restricted area, making a bank transaction, accessing our cellphones or even our cars, we must authenticate ourselves. Because personal authentication based on what we possess or know can be ultimately stolen or forgotten, an ever-increasing interest on biometric authentication has been seen. Moreover, taking into account the increasing amount of sensitive information (e.g., e-mails, financial transactions) available on our personal handhelds, additional security strategies are needed to assure reliable access control to personal information carried out on these gadgets. In such a scenario, automatic voice-based authentication has emerged as a popular biometric modality [83]. Despite all these advances in the field, malicious spoofing attacks have been recognized as a serious issue to ASV reliability [66]. Our main concern, here, is on countermeasures to replay attacks. This leads to our next hypothesis.

- **H5**: a front-end based on the blind estimation of the channel response magnitude is able to capture nuances of room ambiences, recordings and playback devices, and combined with a deep neural network as back-end can mitigate the problem of replay attacks to ASV systems.

In this section, we have detailed the research problems addressed in this thesis and the respective hypotheses towards them. The first two hypotheses are related to prediction of perceived speech quality and the last three hypotheses are related to the reliability of voice-based technologies, specifically for SER and ASV systems. Next, we introduce the reader to the main objectives of this thesis.

1.2 Objectives

The overarching goal of this research is to build tools that improve the human experience while using voice-based technologies. To achieve this goal, two objectives exist: First, we address the problem of estimating perceived speech quality from noisy and enhanced speech. Second, we address the problem of improving the reliability of speech applications, namely SER and ASV, for in the wild conditions.

To achieve these two objectives, the following sub-objectives are tackled:

1. to develop a new full-reference instrumental speech quality measure
2. to develop a new no-reference instrumental speech quality measure
3. to increase the reliability of SER in noisy and reverberant environments
4. to develop an ASV solution that increases the reliability by mitigating the detrimental effects of affective speech
5. to develop a countermeasure solution for replay attacks on ASV systems

In the following, we describe the main contributions of this thesis.

1.3 Main Contributions

This thesis aims at increasing human experience while using speech-based technologies and its main contributions are:

- Proposal of a new full-reference instrumental quality measure:

  The use of i-vector speech representations for instrumental quality measurement of noisy, reverberant and enhanced speech is explored. We show how the total variability space is

capable of capturing ambient factors and a full-reference and three no-reference measures are proposed. Experimental results on two datasets showed the full-reference method achieving results in line with two standard benchmarks and bypassing the need for time alignment between reference and processed signals. On the same datasets, the three no-reference measures presented higher correlations with subjective quality scores compared to two no-reference benchmarks, thus showing their effectiveness in tracking the quality of hands-free and enhanced speech.

- Proposal of a new no-reference instrumental quality measure:

  The i-vector framework is proposed as a non-intrusive speech quality measure. The proposed method relies on a Gaussian mixture model (GMM) to reconstruct the clean reference spectrum from the degraded speech signal. The i-vector representations are then computed for both the estimated clean and degraded spectra, and quality correlates are obtained by means of two different scoring methods: euclidean and cosine distance. Experimental results showed the proposed method outperforming several non-intrusive benchmark algorithms, and achieving accuracy aligned with intrusive algorithms, without the need for a clean reference signal. More importantly, the proposed method showed stable accuracy across degraded and enhanced conditions, thus suggesting better applicability to emerging conditions.

- Proposal of an environment-robust speech emotion recognition system:

  Spontaneous SER in the wild is investigated. Factors such as noise, reverberation and their combined effects were explored. We showed that existing SER systems based on per-frame features (computed from the modulation spectrum), while useful for enacted/posed emotions, perform poorly for spontaneous speech. As such, a feature pooling scheme that combines information from neighbouring frames is proposed. This pooling has significantly contributed to boost SER performance and it has also been shown to be extremely important for valence prediction. By performing feature pooling, we also showed increased robustness against environmental noise. Compared to a benchmark algorithm from the AVEC 2016, the proposed feature pooling scheme did better when noise and noise-plus reverberation were present. The gains were more substantial as noise levels increased, thus showing the advantages of the proposed schemes for in the wild SER.

- Proposal of an ASV system robust to affective speech:

A new method to compensate for the detrimental effects that affective speech has on an ASV system is proposed. In particular, we adopt a GMM based method to "neutralize" affective speech, thus mitigating the detrimental effects of emotion on the speaker verification task. The proposed method, when coupled with a conventional i-vector based ASV, is shown to outperform a baseline by as much as 15%. Experiments are performed across four separate multi-lingual datasets, as well as with a combined larger dataset, and the results obtained consistently show the proposed method outperforming the baseline across up to eight different emotional states. More importantly, the proposed approach does not compromise the ASV performance of neutral speech, an issue commonly observed in other approaches reported in the literature.

- Proposal of a countermeasure to replay attacks for ASV systems:

  A front-end based on the blind channel response estimation as a new approach for replay attack detection is proposed. Our assumption is that the nuances of the acoustic ambience, microphones and playback devices present in the spectrum contain enough information to distinguish between bonafide and spoofed attacks. We explored a baseline back-end based on Gaussian mixture models, as well as a deep residual neural network classifier. Experiments on the ASVspoof 2017 and the ASVspoof 2019 Challenge datasets show the proposed methods outperforming several challenge baseline systems as well as providing improved robustness against train/evaluation set mismatch.

1.4 List of Publications

This thesis proposes two new instrumental speech quality estimation tools based on the i-vector framework and three approaches to increase the reliability of speech-based technologies. The main publications resulting from these innovations include:

[J1] A. Avila, D. O'Shaughnessy, T. Falk, *Non-intrusive Speech Quality Prediction Based on the Blind Estimation of Clean Speech and the i-vector Framework*, J. Quality and User Experience, accepted, in press, 2020.

[**J2**] A. Avila, J. Alam, D. O'Shaughnessy, T. Falk, *On the Use of the I-vector Speech Representation for Instrumental Quality Measurement*, J. Quality and User Experience, 2020, vol. 5, no 1, pp. 1-14, DOI: https://doi.org/10.1007/s41233-020-00036-z.

[**J3**] A. Avila, J. Alam, F. Prado, D. O'Shaughnessy, T. Falk, *On the Use of Blind Channel Response Estimation and a Residual Neural Network to Detect Physical Access Attacks to Speaker Verification Systems*, J. Computer Speech & Language, accepted, in press, 2020.

[**J4**] A. Avila, O'Shaughnessy, T. Falk, *Automatic Speaker Verification from Affective Speech Using Gaussian Mixture Model Based Estimation of Neutral Speech Characteristics*, J. Speech Communication, submitted, under review, 2020.

[**J5**] B. Sadou, A. Lahoulou, T. Bouden, A. Avila, T. Falk, Z. Akhtar, *Free-Reference Image Quality Assessment Framework using Metrics Fusion and Dimensionality Reduction*, Signal & Image Processing, Vol. 10, No. 5, Oct. 2019, pp. 1-14, DOI: 10.5121/sipij.2019.10501.

[**J6**] A. Avila, Z. Akhtar, J. Santos, D. O'Shaughnessy, T. Falk, *Feature Pooling for Spontaneous Speech-Based Emotion Recognition in the wild*, in IEEE Transactions on Affective Computing, pp. 1-12, DOI: 10.1109/TAFFC.2018.2858255.

[**C1**] A. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, J. Gehrke, *Non-intrusive speech quality assessment using neural networks*, ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 631-635, DOI: 10.1109/ICASSP.2019.8683175.

[**C2**] A. Avila, J. Alam, D. O'Shaughnessy, T. Falk, *Blind Channel Response Estimation for Replay Attack Detection*, Interspeech 2019, pp. 2893-2897, DOI: 10.21437/Interspeech.2019-2956.

[**C3**] A. Avila, S. Kshirsagar, A. Tiwari, D. Lafond, D. O'Shaughnessy, and T. Falk, *Speech-Based Stress and Emotion Classification Based on Modulation Spectral Features and Convolutional Neural Networks*, 27th European Signal Processing Conference (EUSIPCO) 2019, pp. 1-5, DOI: 10.23919/EUSIPCO.2019.8903014.

[**C4**] A. Avila, J. Alam, D. O'Shaughnessy, T. Falk, *Intrusive Quality Measurement of Noisy and Enhanced Speech based on i-Vector Similarity*, QoMEX 2019, 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX), pp. 1-5, DOI: 10.1109/QoMEX.2019.8743285. **Nominated for Best Paper Award.**

[**C5**] B. Sadou, A. Lahoulou, T. Bouden, A. Avila, T. Falk, Z. Akhtar, *Blind Image Quality Assessment using SVD based Dominant Eigenvectors for Feature Selection*, SIPRO 2019, pp. 233-242, DOI: 10.5121/csit.2019.90919.

[**C6**] A. Avila, J. Alam, D. O'Shaughnessy, T. Falk, *Investigating Speech Enhancement and Perceptual Quality for Speech Emotion Recognition*, Interspeech 2018, pp. 3663-3667, DOI: 10.21437/Interspeech.2018-2350.

[**C7**] R. Gupta, A. Avila, and T. Falk, Towards a neuro-inspired no-reference instrumental quality measure for text-to-speech systems, QoMEX 2018, 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX), pp. 1-6, DOI: 10.1109/QoMEX.2018.8463392.

[**C8**] A. Avila, J. Monteiro, D. O'Shaughnessy, T. Falk, *Speech Emotion Recognition on Mobile Devices Using a Modulation Spectrum Pooling and Deep Neural Networks*, ISSPIT 2017, 2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), pp. 360-365, DOI: 10.1109/ISSPIT.2017.8388669.

[**C9**] A. Avila, B. Cauchi, S. Goetze, S. Doclo, T. Falk, Performance Comparison of Intrusive and Non-instrusive Instrumental Quality Measures for Enhanced Speech, IWAENC 2016, 2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 1-5, DOI: 10.1109/IWAENC.2016.7602907.

1.5 Thesis Organization

This thesis is organized in two main parts. An overview on the chapters and their relationship can be observed in Figure 1.3. In Chapter 2, we cover the background related to the main topics discussed in this thesis. Chapter 3 presents the proposed full-reference instrumental quality measure based on i-vectors. In Chapter 4, the no-reference instrumental quality measure, also based on

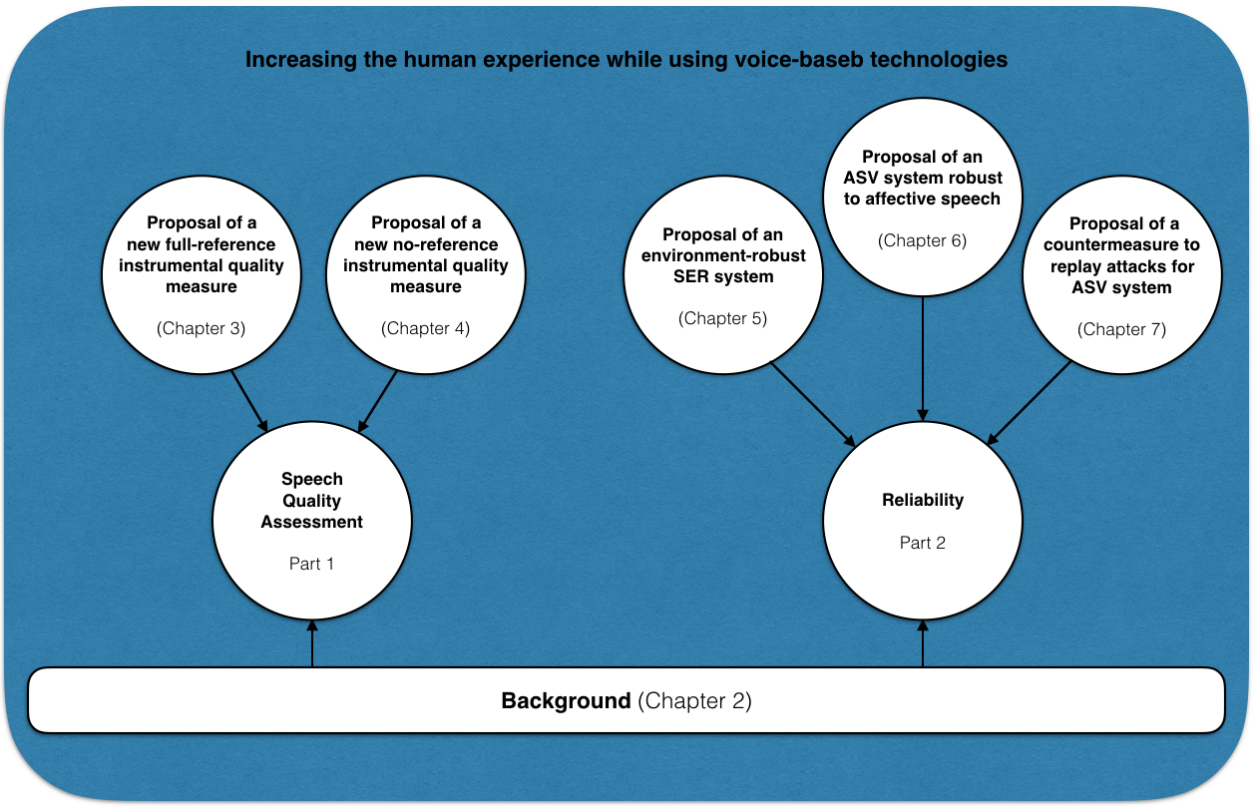


Figure 1.3 – Thesis organization and the relationship among chapters.

i-vectors, is discussed. These two chapters belong to the first part of this thesis where speech quality assessment is the main topic. The second part of this thesis comprises Chapters 5, 6 and 7, and is dedicated to improving the reliability of speech-based technologies. In Chapter 5, we propose an environment-robust SER system. Chapter 6 introduces the proposed ASV system for emotional environments. In Chapter 7, a countermeasure to physical access attacks to ASV systems is proposed. Chapter 8 concludes the thesis providing a final discussion about limitations, research directions and future work.

Background

2.1 Introduction

Systems that enable human-to-human communication, such as Voice-over-IP, may be assessed in terms of conversation quality. In other words, the user experience is related to how good the exchange of information enabled by these systems is. Conversational tests, for instance, were commonly applied to perform such assessments [4]. In human-machine interaction, on the other hand, systems can be evaluated with respect to their usability, that is, whether these systems allow for an effective interaction with the machine. Hence, high performance of transmission systems that led to good conversation quality or good usability used to be assumed to result in high satisfaction, thus increasing the number of customers [42]. Nevertheless, some services, such as the early SMS systems, achieved success despite having low quality. This brought to light the necessity of understanding quality from the perspective of the end user. The user's perception now becomes relevant to assessing the quality of communication services. This led to the formulation of the term 'Quality-of-Experience' (QoE), which can be succinctly described as the user's degree of delight or annoyance towards an application or service [43].

For many years, the quality of communication systems has been associated with the notion of 'Quality of Service' (QoS), which is defined in [43] as the "Totality of characteristics of a telecommunication service that bear on its ability to satisfy stated and implied needs of the user of the service". Because this notion is limited to network performance and the physical aspects of transmission systems, it cannot fully express everything comprised in a communication service [44]. Thus, QoE has flourished as a relevant concept not just for multimedia services but also for

HCI and aesthetics [44]. This motivated a shift on the focus of providers that now become more concerned with delivering services with high QoE rather than services with only high QoS. Similar trends were also seen for interactive human-machine interfaces, with the so-called "User Experience" (UX) becoming as relevant as classical usability (i.e., effectiveness and efficiency of provided services) [24]. Therefore, the human aspect becomes the quality thermometer of systems and services [45], and the user's formation process of sound-quality judgments must be understood and taken into consideration.

This chapter focuses on addressing the background knowledge involving speech quality assessment along with a short description of the main material and methods used throughout this thesis. The remainder of this chapter is then organized as follows. Section 2.2 provides a discussion regarding speech quality assessment and Section 2.3 discusses reliability of speech-based technologies. Speech parameterization is presented in Section 2.4 with the fundamentals of the i-vector framework. In Section 2.5 we present the main deep neural network models used in this thesis.

2.2 Speech Quality Assessment

Speech quality assessment aims at estimating what listeners perceive from the acoustic form. In speech quality tests, participants must describe their quality perception, that is, they are asked to assess the quality of a particular speech sample or system used for its transmission [3]. Typically, they assign numbers that categorize the perceived quality. As a non-deterministic process, it may lead to different conclusions even when the same acoustic form and environmental conditions are presented. That is because the process of speech perception and assessment depends on how the judgment is made by human listeners. Figure 2.1 depicts the assessment process as performed by humans according to [3]. Based on *modifying factors*, which consist of a set of features based on individual expectations, relevant demands and/or social requirements, the listener is able to anticipate the percept at some level. The *modifying factors* define the context and situation in which the physical event (or sound) occurs. The context is influenced by personal factors such as the mood and motivation and prior experience with the nature of the sound. The *desired characteristics* are stored schemes related to previous percepts of similar context of communication [3]. During the reflection process, desired features are identified after the listener decomposes the desired characteristics. Similarly, the listener decomposes the perceived characteristics into a set of perceived features. In

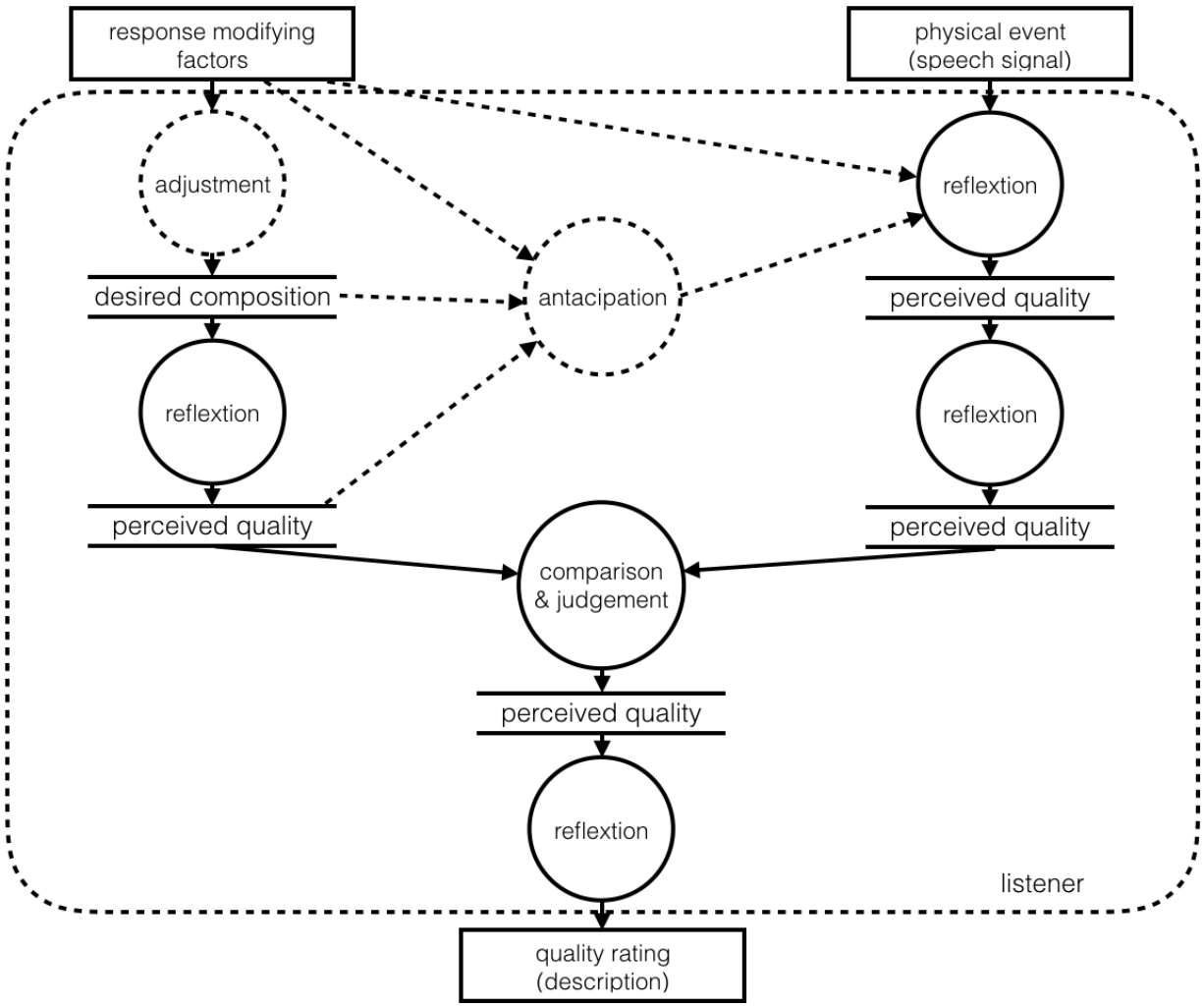


Figure 2.1 – Diagram describing the perception composition of the quality event. Adapted from [3].

the last stage, the perceived quality is attained from the judgment on the comparison from the desired and the perceived features [3].

2.2.1 Subjective Quality Assessment

In order to qualify or quantify the aforementioned perceptive effects, auditory assessment methods are needed. Because the choice of an adequate test method is an important factor, the ITU-T Recommendation 800 [17] provides a number of suggestions on how to administer subjective tests of transmission quality. Traditionally, subjective listening-only tests (LOT) have been used and shown to be a reliable method for assessing the quality of speech [17]. In such scenarios, speech signals are presented to listeners who judge the signal quality, typically, on a 5-point ACR scale.

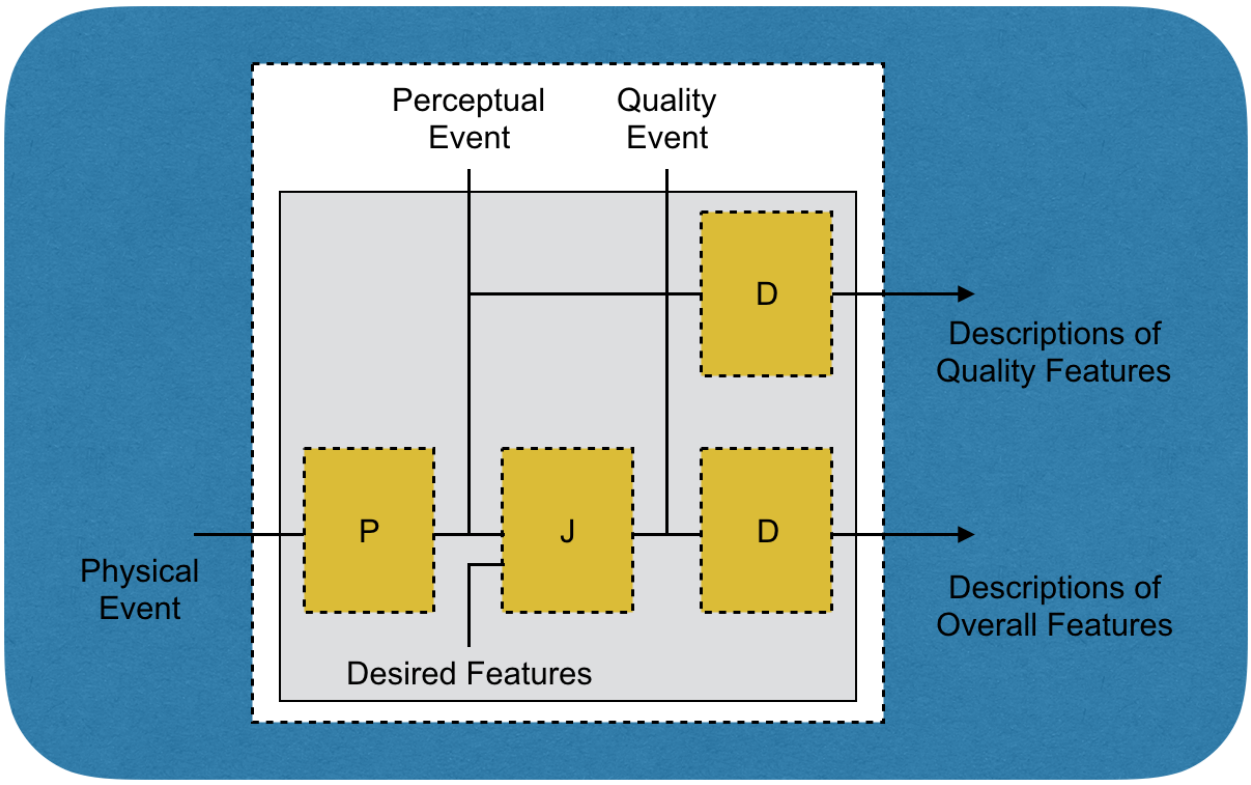
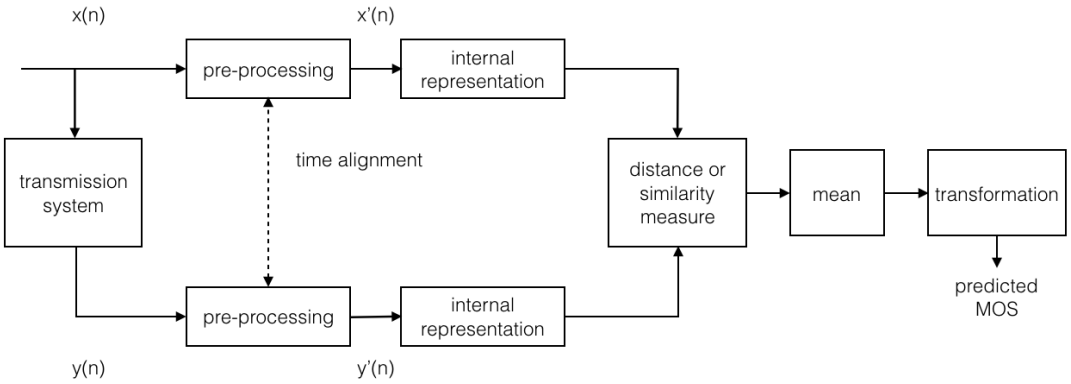


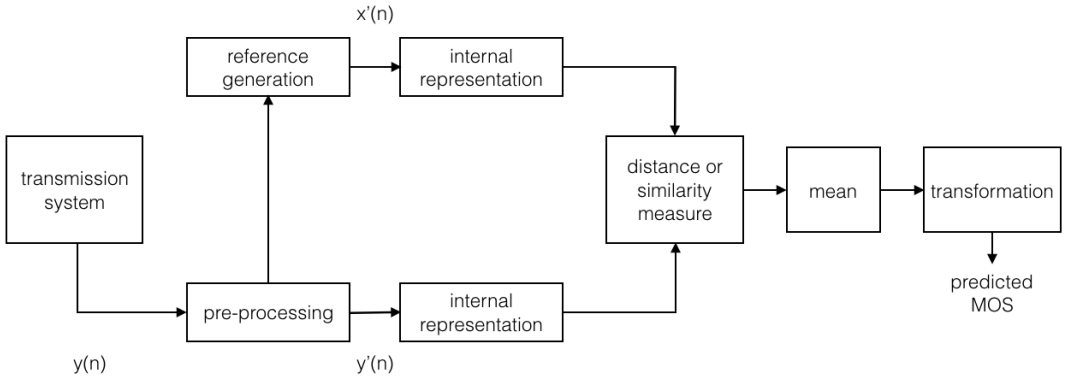
Figure 2.2 – Diagram describing a test subject in a listening quality test. Adapted from [4].

The categorical labels are chosen in order to provide some kind of 'absolute' level information. In such tests, the results are strongly affected by the choice of conditions, as the settings are quite artificial and the subjects are mainly focused on the form of the speech sound and not necessarily on its content [1].

Figure 2.2 describes a listener in a quality test situation. The participant is expected to provide a description of the features perceived in a sound sample. Rather than focusing on judgments to attain the perceived quality, such tests are looking for a description of the perceived (quality) features and are regarded as an analytical type of speech quality test. As depicted in Figure 2.2, speech quality is a multidimensional construction that comprises perception (P), judgment (J), and description (D) [4]. The perceptual event gets triggered by the physical event, that is, as the sound waves reach out the human ears. The "perceptual event" will result in quality features experienced by the listener. These features are then compared to the desired features based on the individual internal reference as discussed previously. A "quality event" is then formed based on the result of this comparison and can be quantitatively described as a judgment of the "overall quality". The MOS, which represents



(a) Double-ended model.



(b) Single-ended model.

Figure 2.3 – Diagram representing a typical objective quality assessment approach for signal-based measures. Adapted from [3].

the perceived speech quality after leveling out individual factors [4], is attained after averaging all participant scores over a specific condition.

2.2.2 Objective Quality Assessment

Because subjective listening tests are time-consuming, expensive and for the most part cannot be done in real-time, much effort has been made for the development of instrumental quality methods, also referred to as objective quality models. It is important to mention that such models are usually designed for a specific domain of application and for a certain range of network impairments. Therefore, it is rarely the case that they can be universally applied to all circumstances [84]. For instance, regarding the application domains, the models can be classified as (1) signal-based models; (2) network planning models; and (3) monitoring models [84]. This thesis focuses on the first category, i.e., signal-based ones.

Signal-based models use speech signals transmitted or modified by speech processing algorithms to predict quality [4]. Such models express the quality of a speech signal according to the ACR listening quality scale. There are two types of signal-based models: double-ended (also referred to as "intrusive" or "full-reference"), and single-ended (also known as "non-intrusive" or "no-reference"). Double-ended models rely on a reference speech signal (system input) and its corresponding degraded version (system output). Single-ended models, on the other hand, depend only on the degraded speech signal. Figure 2.3 shows a general diagram of signal-based measures for double-ended models, described in Figure 2.3-a, as well as for single-ended ones, detailed in Figure 2.3-b.

2.2.3 Full-reference instrumental measures

Most full-reference measures estimate transmission quality by computing a perceptually weighted distance between the channel input and output signals, which is assumed to be a good indicator of quality. In this section, the main full-reference instrumental measures are presented.

Perceptual evaluation of speech quality, PESQ

The Perceptual Evaluation of Speech Quality, or ITU-T Recommendation P.862, is the most widely-used intrusive instrumental measure available for narrowband and wideband telephone speech [85]. In order to directly compare the reference and processed signals, PESQ relies on a time alignment algorithm. The signals are then transformed to an internal representation that is analogous to the psychophysical representation of audio signals in the human auditory system. This is achieved by means of perceptual frequencies and compressive loudness scaling. The audible difference between the two signals is then calculated based on the difference between the two internal representations. Audible errors are lastly evaluated by a cognitive model that separates asymmetrical and symmetrical disturbances and maps them into a predicted MOS score using a weighted linear combination of the two disturbances. More details about the metric can be found in [86, 85].

Perceptual objective listening quality Assessment, POLQA

ITU-T Recommendation P.863 [87], also known as Perceptual Objective Listening Quality Assessment (POLQA), is the successor of PESQ to allow for more recent distortions to be modeled,

such as speech enhancement and internet protocol networks [88]. Time alignment, for instance, was improved due to more recent packet loss concealment strategies. Moreover, in addition to the symmetric and asymmetric disturbances used within PESQ, POLQA includes a noise analysis and a reverberation analysis module, which provide input to a cognitive mapping module that estimates the MOS. Unlike PESQ, POLQA has been validated from narrowband to super-wideband speech conditions [87].

Normalized covariance metric, NCM

The normalized covariance metric (NCM) is based on the covariance between auditory-inspired envelopes of the clean and processed speech signals [89]. Such envelopes are attained via the Hilbert transform, from outputs of a gammatone filter bank used to emulate cochlear processing. The NCM metric was recently shown to be useful in estimating the quality and intelligibility of reverberant speech for impaired listeners with and without hearing devices [47]. The final NCM value is given by

$$NCM = \frac{\sum_{k=1}^{23} W(k) SNR(k)}{\sum_{k=1}^{23} W(k)}, \tag{2.1}$$

where the SNR is computed from the envelopes of the reference clean and degraded signal. These values are weighted in each frequency channel, according to the so-called articulation index weights $W(k)$, recommended in the American National Standards Institute (ANSI) S3.5 Standard [90].

Short-time objective intelligibility, STOI

The STOI metric is used to estimate intelligibility. As with the NCM, the STOI metric relies on the covariance of the temporal envelopes of the clean and processed speech and assumes that both signals are time-aligned. The main difference is in the fact that the STOI algorithm computes this covariance over short time segments and then aggregates the per-frame disturbances into a final quality rating [91]. Unlike NCM, the speech signals are decomposed by a one-third octave filterbank, followed by level normalization and clipping. The latter processing is performed in order to lower the bound for the signal-to-distortion ratio and is used to avoid changes in intelligibility prediction. More details can be found in [91].

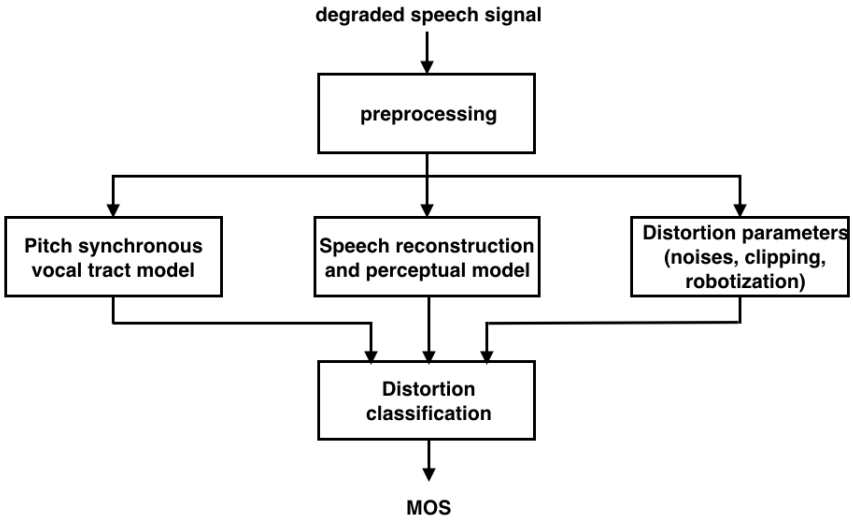


Figure 2.4 – P.563 components.

2.2.4 Non-reference instrumental measures

Although more difficult, reference-free models may be the most suitable solution for live network monitoring, as the reference clean speech signal is not promptly available in such scenarios [4]. Predicting the speech quality without the reference clean speech signal is a challenging task, due to the wide variability of the input speech signal, e.g., different speakers, vocal tract dimensions, pitch characteristics and speech content [33]. One way to overcome this is by estimating the clean speech signal from the corrupted speech signal as depicted in Figure 2.3-b. In this section, the main non-reference instrumental measures are introduced.

ITU-T Recommendation P.563

The non-intrusive ITU-T Recommendation P.563 is a standard algorithm for telephone-band speech. To evaluate distortions, five components are combined as described in Figure 2.4. The first is the pre-processing component. At this stage, the model considers that the signal level may differ from the listening level. Therefore, the model assumes that the signal is presented with a sound level pressure of 79 dB SPL at the ear reference point. Moreover, to avoid discrepancies, the input signal level is also normalized to - 26 dBov (dB overload) [92]. Another important step in the pre-processing block is the generation of two additional versions of the degraded signal. The first version is filtered as to simulate the properties of the modified intermediate reference system

(IRS), with the frequency characteristics described in the ITU-T Recommendation P.830 [93]. The second component is the pitch synchronous vocal tract model, which extracts pitch marks using a hybrid temporal/spectral method [92]. The pitch length is determined with the calculation of the maximum normalized autocorrelation for 64-ms frames with 50% overlap. When the maximum autocorrelation value exceeds 0.5, frames are marked voiced and unvoiced otherwise. The third component aims at reconstructing the clean speech, which is used as input for the perceptual model (or full-reference model) as a quasi-reference. This is attained by constraining LP (Linear Predictive Coding) coefficients to fit the vocal-tract model of a typical human talker [92]. The perceptual model used in this step is a modified version of PESQ [85]. The fourth component performs the detection and quantification of specific distortions encountered in voice channels, such as temporal clipping, robotization, and noise. The distortion levels are estimated based on noise occurring in non-speech segments, and the measurement is used by the distortion classifier for identifying very noisy signals. The different distortion types are then ranked and a distortion-dependent weighted linear mapping is applied to estimate the final MOS scores rating. For complete details on this algorithm, the interested reader is referred to [92, 46].

Speech-to-reverberation modulation ratio, SRMR

The SRMR computes the ratio of low to high modulation energy after an auditory model is applied based on a 23-channel gammatone cochlear filterbank and an 8-channel modulation filterbank to emulate the human hearing system. The measure relies on the principle that the modulation energy of clean speech is generally concentrated in lower modulation frequencies (below 20 Hz), while room acoustic artefacts typically arise in higher modulation frequencies beyond 20 Hz. As such, the model has been shown to accurately characterize room acoustics, as well as the quality and reverberation level of reverberant speech and speech processed by enhancement algorithms [94, 95].

$\mathbf{SRMR}_{norm}$

An extended version of the SRMR metric was proposed in [95]. The main motivation for this was to reduce the variability caused by the effects of pitch and speech content. The frequency range of the modulation filters was reduced from 4-128 Hz in the original SRMR implementation to 4-40 Hz in order to reduce pitch effects. Moreover, to reduce the sensitivity to spoken content, a per-frame

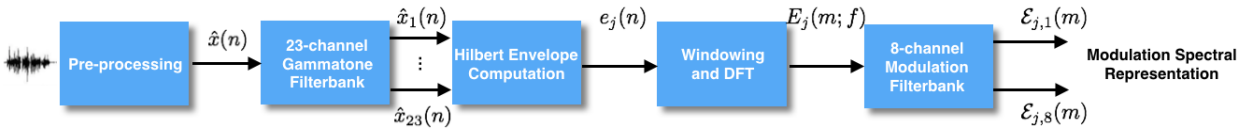


Figure 2.5 – Block diagram describing steps for computing the modulation spectrum representation.

energy thresholding scheme was implemented where only frames below 30 dB were used. The so-called SRMR_{norm} metric was shown to reduce intra- and inter-speaker variability and to better estimate the intelligibility level of speech under reverberation, noise, and reverberation-plus-noise conditions [95].

2.3 Reliability and user experience (UX)

The reliability of speech-based technologies has always been an important issue. Particular interest lies in real-world scenarios where mismatch between training and testing conditions are commonly encountered. It is not uncommon to have a recognition system offering robustness in one environment but failing in another. This is due to the fact that training is often performed in laboratory settings, in a noise-free tranquil type of environment [96]. Thus, these systems are expected to significantly degrade in more realistic scenarios. For example, when distortions such as convolutive and additive noise are present or in an emotional/stress type of situation. Thus, several studies have attempted to either improve or predict the reliability of such systems. In [96], for example, the authors propose an environmental-robust feature to mitigate performance degradation caused by additive background noise. The study also formulates a quantitative measure of recognition performance based on the estimation of the speech quality. In a more recent work [26], both convolutive and additive noise are addressed. The authors investigated several instrumental quality measures and their capability to predict speaker verification reliability. The authors in [40] proposed to map the performance of an ASV system as a function of arousal and valence primitives. With that they were then able to estimate the reliability of the recognition system by considering emotional dimensions such as arousal, valence and dominance. In a similar study, the speaker verification performance is analyzed in terms of arousal (calm/active), valence (negative/positive) and dominance (weak/strong) [41].

As can be seen, reliability is an important issue for speech-based systems. Moreover, it is a component of usability that contributes to the overall UX.

2.4 Feature Extraction

In this section, we describe the main features used in this work. We start with the modulation spectrum representation, followed by the cepstrum parameterization and conclude by presenting the i-vector framework.

2.4.1 Modulation spectrum signal representation

The modulation spectrum corresponds to an auditory spectro-temporal representation that captures long-term dynamics of the speech signal, which has been shown to carry relevant speech and speech emotion information [97]. Here, we follow the processing pipeline proposed by [98] to compute the modulation spectral representation. The processing pipeline is depicted in Figure 2.5. During the pre-processing step, the speech activity level is normalized to -26 dBov, thus eliminating unwanted energy variations caused by different loudness levels in the speech signal. Next, the pre-processed speech signal $\hat{x}(n)$ is filtered by a 23-channel gammatone filterbank simulating the cochlear processing [99]. The first filter of the filterbank is centered at 125 Hz and the last one at half of the sampling rate [98]. Each filter bandwidth follows the equivalent rectangular bandwidth (ERB) [99] given by:

$$ERB_j = \frac{f_j}{Q_{\text{ear}}} + B_{\min}, \tag{2.2}$$

where f_j represents the center frequency of the j-th filter. Q_{ear} represents the asymptotic filter quality at large frequencies and $B_{\min}$ is the minimum bandwidth for low frequencies. They are set, respectively, to 9.265 and 24.7.

The temporal envelope $e_j(n)$ is then computed from $\hat{x}_j(n)$, the output of the j-th acoustic filter, via the Hilbert transform:

$$e_j(n) = \sqrt{\hat{x}_j(n)^2 + \mathcal{H}\{\hat{x}_j(n)\}^2}, \tag{2.3}$$

where $\mathcal{H}\{\cdot\}$ denotes the Hilbert Transform. Temporal envelopes $e_j(n), j = 1, \ldots, 23$ are then windowed with a 256-ms Hamming window and shifts of 40 ms. The discrete Fourier transform $\mathcal{F}\{\cdot\}$ of the temporal envelope $e_j(m; n)$ (m indexes the frame) is then computed in order to obtain the

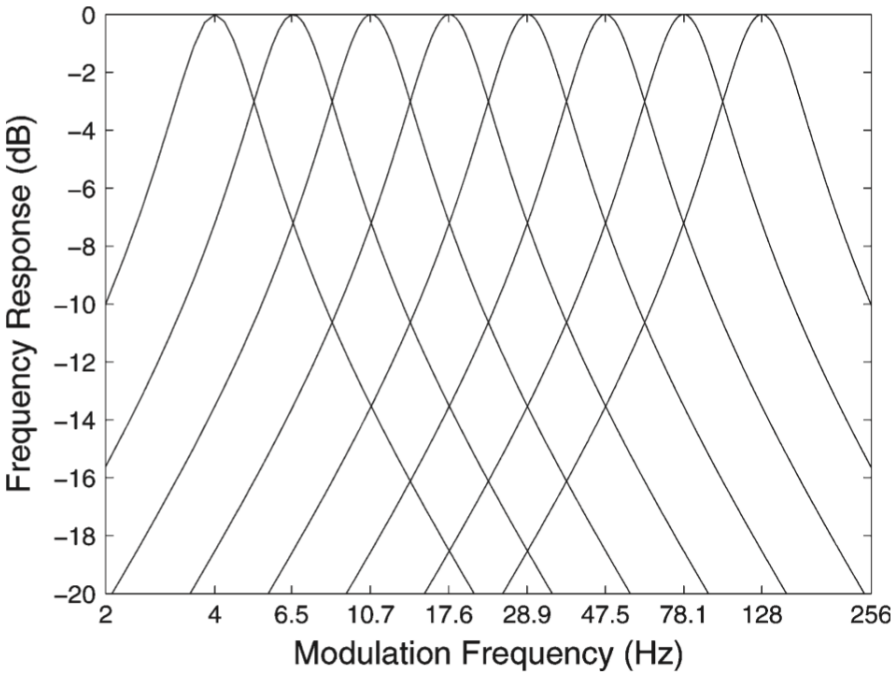


Figure 2.6 – Frequency responses of the 8-channel modulation filterbank. Adapted from [5].

modulation spectrum $E_j(m, f_m)$, i.e.,

$$E_j(m; f_m) = |\mathcal{F}(e_j(m; n))|, \tag{2.4}$$

where m represents the m-th frame obtained after every Hamming window multiplication and f_m designates modulation frequency. The time variable n is dropped for convenience. Lastly, following recent physiological evidence of a modulation filterbank structure in the human auditory system [100], an auditory-inspired modulation filterbank is further used to group modulation frequencies into eight bands. These are denoted as $\mathcal{E}_{j,k}(m)$, $k = 1, ..., 8$, where j indexes the gammatone filter and k the modulation filter. Figure 2.6 depicts the frequency response for the 8-channel modulation filterbank used in our experiments. Note that the filter center frequencies are equally spaced in the logarithmic scale from 4 to 128 Hz.

2.4.2 Cepstrum Parametrization

Mel-frequency cepstral coefficients (MFCC) have been widely used in speech applications for many years as they simulate the mel-scale present in the human cochlea. Prior to their extraction, input speech signals are normalized and segmented into frames. The signals also undergo a pre-emphasis

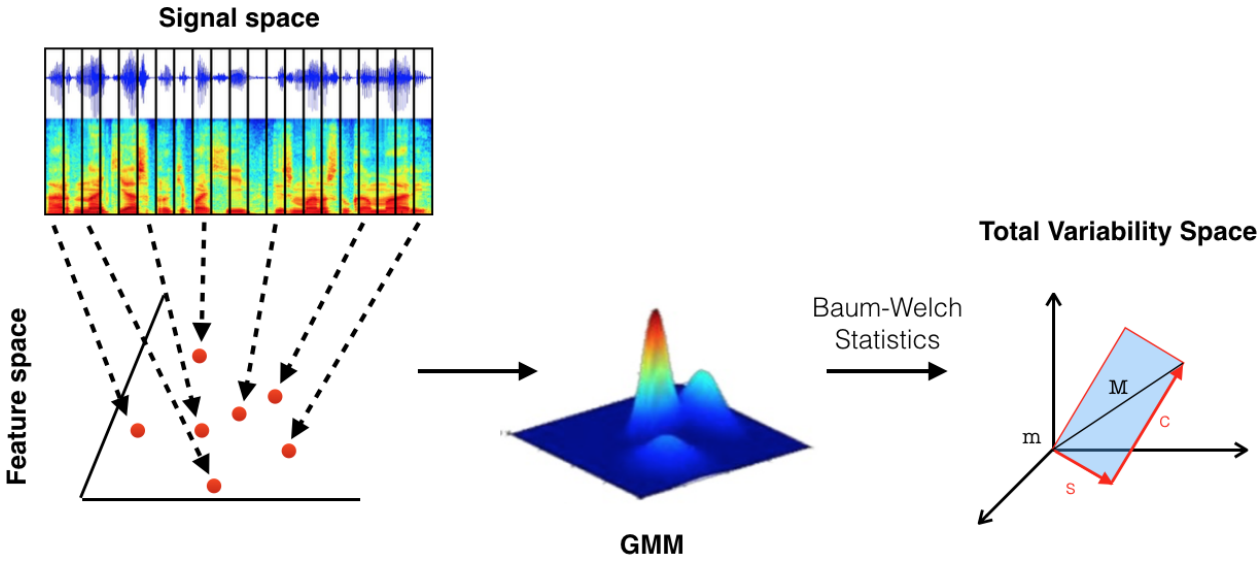


Figure 2.7 – Block diagram describing the steps for i-vector extraction.

filter, which is meant to balance low and high frequency magnitudes. MFCCs are extracted from short-time speech frames, typically between 20 and 30 ms with a 50% hop-size, according to:

$$c_n = \sum_{m=1}^{M} [Y_m] cos \left[\frac{\pi n}{M} \left(m - \frac{1}{2} \right) \right], n = 1, 2, 3, ..., N, \tag{2.5}$$

where c_n is the n^{th} mel-cepstral coefficient and Y_m refers to the log-energy of the m^{th} filter. The set of N coefficients forms the feature vector from each frame. As can be seen in (2.5), the MFCC representation is based on a short-term log-power spectrum and a cosine transformation on the nonlinear mel scale of frequency.

2.4.3 i-vector

The i-vector framework was developed inspired by joint factor analysis (JFA) [101]. Both frameworks consider speaker and channel variability to lie in a low-rank subspace. Notwithstanding, while the JFA approach models speaker and channel variability in separated subspaces (i.e., the eigenvoice V and eigenchannel U), the i-vector framework considers only one subspace. This subspace is defined as total variability (TV) and comprises both speaker effects and channel effects [30]. Similar to JFA, the TV space is defined by GMM supervectors, which contain the mean values of a GMM universal background model (UBM) [83]. The supervector in the TV space can be represented as

follows:

$$M = m + Tw, \tag{2.6}$$

where M is the speaker- and channel-dependent supervector extracted from a specific recording, m is the independent supervector from the UBM, T corresponds to the total variability matrix trained with multiple recordings using the same procedure for learning the eigenvoice matrix [102], and w is a random vector with normal distribution, $N(0, I)$. This vector is referred to either as identity vector or i-vector, and conveys the total factors.

Figure 2.7 depicts the steps involving the extraction of i-vectors. Note that the framework ultimately maps a list of feature vectors into a fixed-length vector, $w \in R^D$. These feature vectors, denoted here as $O = \{o_t\}_{t=1}^{N}$, where $o_t \in R^F$, are extracted during the speech parameterization phase. In order to obtain w, a GMM model, $\lambda = (\{p_k\}, \{m_k\}, \{\sigma_k\})$, must be trained using multiple utterances. When such utterances come from different speakers, the model is referred to as an UBM. As depicted in Figure 2.7, after training the GMM-UBM model, Baum-Welch statistics are extracted from each utterance u [103]. Note that the total factor w is the posterior distribution conditioned on the Baum-Welch statistics [30][102], which are computed as follows:

$$N_k = \sum_{l=1}^{L} P(k|y_k, \lambda), \tag{2.7}$$

$$F_k = \sum_{l=1}^{L} P(k|y_k, \lambda) y_k, \tag{2.8}$$

$$\tilde{F}_k = \sum_{l=1}^{L} P(k|y_k, \lambda)(y_k - m_k), \tag{2.9}$$

where the k-th frame is represented by y_k and L denotes the total number of frames extracted from a given utterance, and λ is the UBM. The mean of the k-th mixture component is represented by m_k. The posterior probability that the vector y_k is generated from the mixture component k is given by $P(k|y_k, \lambda)$. Note that Eq. (2.7) and Eq. (2.8) represent the zero-th and first-order Baum-Welch

statistics respectively. Eq. (2.9) is the centralized version of Eq. (2.8) [30]. The i-vector is then attained by

$$w = (I + T^t\Sigma^{-1}N(u)T)^{-1}.T^t\Sigma^{-1}\tilde{F}(u), \tag{2.10}$$

with $N(u)$ being a diagonal matrix of $KF \times KF$ dimension and $\tilde{F}(u)$ is a supervector of dimension $KF \times 1$ obtained by the concatenation of first-order Baum-Welch statistics $\tilde{F}_k$ for a given utterance u. A diagonal covariance matrix of $KF \times KF$ dimension is defined by Σ.

2.5 Deep Neural Networks

Motivated by recent advances in artificial intelligence, we explore deep neural network architectures as tools to map relevant speech features to a desired outcome. In this thesis, the outcomes can range from perceived quality to user emotional state, to a speaker, to name a few. Next, we will introduce some of the fundamental concepts behind the deep neural network models used herein.

2.5.1 Multi-layer Perceptron

Multiple layer Perceptron (MLP) architectures learn to map the input space into a space where the output is linearly separable [50]. This is achieved by performing a set of transformations given by a linear combination of the input variables as follows:

$$\mathbf{z}_i = \mathbf{w}_i\mathbf{x}_i + \mathbf{b}_i \tag{2.11}$$

where $\mathbf{W}_i$ and $\mathbf{b}_i$ are the weights and biases, respectively, while $\mathbf{x}_i$ represents the input and $\mathbf{z}_i$ the activation of each hidden unit in a given hidden layer. This computation is followed by a non-linear

function g, also referred to as the activation function [104][105], which provides the layer's output: $\mathbf{y}_i = g(\mathbf{z}_i)$. Model parameters $\mathbf{W}_i$ and $\mathbf{b}_i$ are defined such that:

$$\arg\min_{\mathbf{w},\mathbf{b}} L(\mathbf{x}, \mathbf{y}), \tag{2.12}$$

where $L(\mathbf{x}, \mathbf{y})$ is some loss function (e.g., mean square error) to be minimized over the data set $(\mathbf{x}, \mathbf{y})$, where $\mathbf{x}$ and $\mathbf{y}$ are the example features and corresponding outputs, respectively.

2.5.2 Convolutional Neural Network

Convolutional neural networks have successfully been applied to 2D image classification, as well as video, speech and audio processing applications [50]. CNNs are typically comprised of two layers: convolution and pooling. Convolutional layers are in charge of mapping, into their units, detected features from local connections in previous layers. Known as feature maps, this is the result of a weighted sum of the input features (or feature maps from previous convolutional layers) passed through a non-linearity such as ReLU [50]. A pooling layer will typically take the maximum or average of a set of neighboring feature maps, reducing dimensionality (i.e., subsampling) by merging semantically similar features.

2.5.3 Recurrent Neural Network

Recurrent neural networks (RNNs), in turn, have been designed to process sequential data. They can be seen as feed-forward neural networks with a parameter-sharing scheme that allows the output corresponding to a specific input to be dependent on previously seen examples [105]. This aspect makes recurrent neural networks a good alternative to model sequential data. Here, we use the Long-short Term Memory (LSTM) network, which has been widely employed mainly due to the fact that its cells have shown to be able to avoid the vanishing/exploding gradients problem [106, 107]. LSTM cells substitute hidden layers of vanilla recurrent neural networks and include a learnable

gating process that allows long-term dependencies. In a given time step t, the input i_t, forget f_t and output o_t gates are given by:

$$f_t = \sigma(\mathbf{b}_f + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{W}_f \mathbf{x}_t), \tag{2.13}$$

$$i_t = \sigma(\mathbf{b}_i + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{W}_i \mathbf{x}_t), \tag{2.14}$$

$$o_t = \sigma(\mathbf{b}_o + \mathbf{U}_o \mathbf{h}_{t-1} + \mathbf{W}_o \mathbf{x}_t), \tag{2.15}$$

where $\mathbf{x}_t$ is the input example at time step t, $\mathbf{h}_{t-1}$ is the intermediate layer representation of $\mathbf{x}_t$ and $\mathbf{b}$, $\mathbf{U}$ and $\mathbf{W}$ are the gate parameters and σ is the sigmoid function. The hidden state $\mathbf{h}_t$ is given according to:

$$\mathbf{h}_t = o_t \odot \tanh(c_t), \tag{2.16}$$

where c_t is the LSTM cell state defined as:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t. \tag{2.17}$$

The operation $\odot$ is an element-wise multiplication and $\tilde{c}_t$ is:

$$\tilde{c}_t = \sigma(\mathbf{b}_c + \mathbf{U}_c \mathbf{h}_{t-1} + \mathbf{W}_c \mathbf{x}_t). \tag{2.18}$$

2.5.4 Residual Neural Network

Deep residual networks (ResNets) were proposed initially in [80] as a strategy to mitigate the problem of vanishing gradients, encountered while optimizing a deep neural network. Since its introduction, the training of considerably deep models (e.g., over 50 layers) has become more feasible. It has been shown, for example, that training and testing errors increase with deeper networks [80],

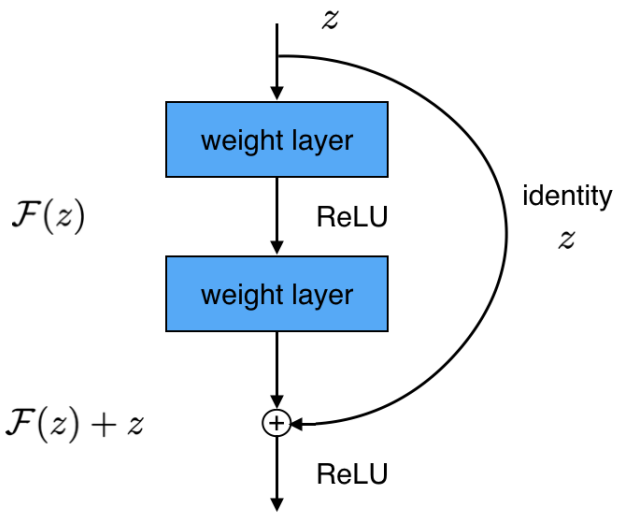


Figure 2.8 – ResNet building block.

not necessarily due to overfitting [80], but also possibly due to optimization challenges with deeper models. The problem has been mitigated with the introduction of an identity mapping [80][108] depicted in Figure 2.8. The identity mapping is represented by the skip (or shortcut) connection that allows the simple element-wise addition between the input, z, and the residual function, $\mathcal{F}(z)$. In such a scenario, it is expected that the input and the residual function would have the same dimension. The residual function is the difference between the underlying mapping, $\mathcal{T}(z)$, represented by few stacked layers and its actual input, x. Therefore, rather than approximating the actual mapping function, $\mathcal{T}(z)$, the deep residual learning approach attempts to approximate the residual function given by $\mathcal{F}(z) = \mathcal{T}(z) - z$. The residual block can be defined as:

$$y = \mathcal{F}(z, W_i) + z \tag{2.19}$$

where the input and the output are represented by z and z, respectively. The residual mapping (or function) is given by $\mathcal{F}(z, W_i)$ [80]. The main idea behind the residual block is that it allows the network to retain the information learned in previous layers. This is achieved by using the identity mapping weight function, thus preventing vanishing/exploding gradients [80].

Part I

Speech Quality Assessment

i-vector speech representations for instrumental quality measurement of unprocessed and enhanced speech

3.1 Preamble

This chapter is compiled from material extracted from the manuscript published in the Journal of Quality and User Experience [J2], and its earlier version appeared in the Proceedings of the 11^{th} International Conference on Quality of Multimedia Experience (QoMEX 2019) [C4].

3.2 Introduction

The i-vector framework has been widely used to summarize speaker-dependent information present in a speech signal. The framework can be seen as a feature extraction procedure that depends on the observed speech signal, the UBM and the total variability matrix (or T matrix), already discussed in Section 2.4.3. Considered the state-of-the-art in speaker verification for many years, its potential to estimate speech recording distortion/quality has been overlooked. This chapter is an attempt to fill this gap. We conduct a detailed analysis of how distortions are represented in the total variability space. We then propose an intrusive speech quality estimator based on i-vector similarities and three non-intrusive approaches. The first makes use of a single reference i-vector

based on the average of i-vectors extracted from clean signals. A second approach relies on a vector quantizer (VQ) codebook of representative clean speech i-vectors. Lastly, i-vectors and MOS were used to train a deep neural network model for non-intrusive speech quality assessment. It is shown through several experiments that many of the proposed methods are well-suited for assessing speech quality, and outperform well-established instrumental measures such as the PESQ and the POLQA algorithms, but with the added advantages of not requiring time alignment with a reference signal or, in the case of the non-intrusive methods, a reference signal altogether.

The main motivation behind this work lies in the fact that i-vectors are known to convey both channel and speaker information. Nevertheless, most research in the field has focused on the speaker characteristics of the representation (e.g., for speaker recognition) and channel effects have been suppressed. As shown in previous research [30, 31], the performance of i-vector based applications is severely affected by environmental factors, such as background noise and reverberation. To mitigate these channel effects, compensation techniques, such as LDA and WCCN [32], are commonly applied. Here, unlike previous work, we utilize this information as a correlate of perceived speech quality. Moreover, as i-vectors are mapped to a fixed length feature vector, regardless of the originating signal length, full-reference quality assessment can bypass time-alignment, which is a crucial and error-prone step for PESQ and POLQA [4].

The remainder of this chapter is organized as follows. Section 3.3 presents the proposed method. Section 3.4 describes the experimental setup and Section 3.5 presents the results and discussion. Lastly, conclusions are presented in Section 3.6.

3.3 Background and proposed method

In this section, the proposed full-reference instrumental quality measure based on the i-vector representation is presented. We start introducing the measure used for estimating the distortion between reference and degraded signals. In the following, we discuss the effects of distortions in the total variability space.

3.3.1 Cosine distance for similarity scoring

The cosine similarity measure has been widely used to compare two supervectors in the total variability space [109]. It represents the angle between two total factor vectors, generated by (2.10) via the projection of two supervectors in the total variability space. The measure can be computed as follows [110]:

$$\cos(\theta) = \frac{w_{ref} \cdot w_{deg}}{||w_{ref}|| \cdot ||w_{deg}||}, \tag{3.1}$$

where w_{ref} is the i-vector extracted from the reference speech recording and w_{deg} is the i-vector representation for the degraded speech recording. Note that, according to [110], the cosine distance, on the other hand, is defined as:

$$cosine\_distance = 1 - \cos(\theta). \tag{3.2}$$

Figure 3.1 represents the total variability space, comprising the speaker and channel factors for two speech recordings of the same speaker. Considering that the proposed model is full-reference, we can assume no speaker and speech variability in the two representations. That is, speech content will remain the same for the reference and degraded signal and only changes in the channel factors will be present, as depicted in Figure 3.1. Note that when significant alterations occur in the channel factors, the angle between w_{ref} and w_{deg} is expected to increase as well as the values of the cosine similarity distance. As such, the computation of the cosine similarity distance provides values close to 0 for high similarities and low distortions, and values close to 1 for low similarities and high distortions. Therefore, the similarities being captured are directly related to levels of distortions in the speech signal, as we show in the next section, and, thus, inversely proportional to speech quality.

3.3.2 Effects of distortions on the total variability subspace

In this section, we illustrate the impact of ambient noise on the i-vector representation to motivate our findings and hypotheses. For illustration purposes, we focus here only on reverberation and noise, as well as on MFCC features.

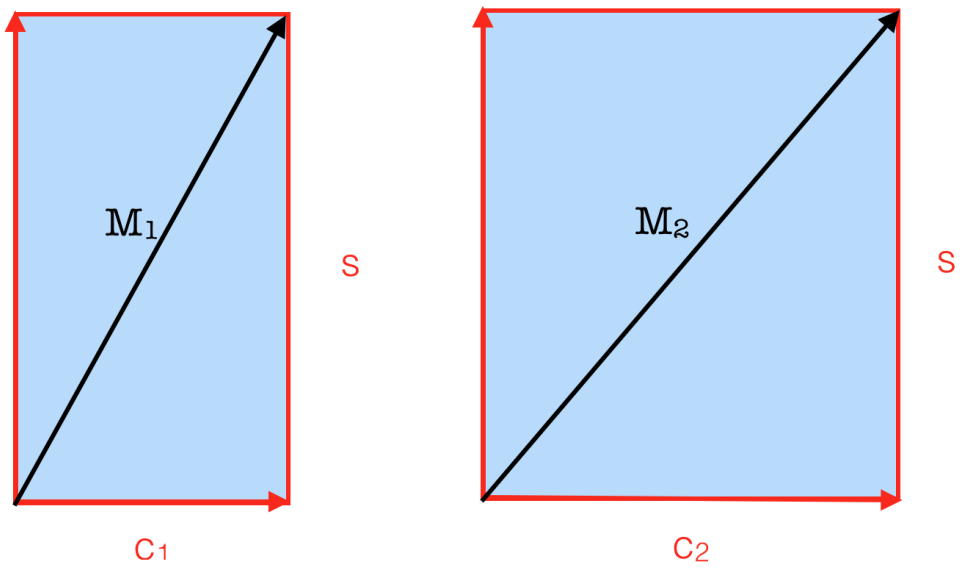


Figure 3.1 – Representation of speaker- and channel-dependent supervectors of two recordings from the same speaker where only the channel factors are affected.

t-distributed stochastic neighbor embedding

In order to visualize the similarities between high-dimensional data points, it is convenient to map them into a two- or three-dimensional space. Such a projection must preserve the distances between data points, maintaining the structure of the high-dimensional data as much as possible. Note that in this process the interpretation of the coordinates becomes less important whereas the distances between data points and their clustering carry out much more meaning. Here, we adopted a tool commonly used in machine learning, namely t-distributed stochastic neighbor embedding (t-SNE).

Different from reduction techniques, such as principal components analysis (PCA) that attempts to keep the low-dimensional representations of dissimilar data points far apart, the t-SNE method keeps the low-dimensional representations of very similar data points close together. According to [111], the t-SNE technique can capture local structure of the high-dimensional data and at the same time keep global structure such as the presence of clusters at several scales. To achieve this, the method embeds high-dimensional data into a lower-dimension space, usually two or three dimensions, by minimizing the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data. The joint probability for the high-dimensional data can

be expressed as:

$$p(x_j, x_i) = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq l} \exp(-||x_k - x_l||^2/2\sigma_i^2)}, \tag{3.3}$$

where x_i and x_j are data points and σ_i is the variance of the Gaussian distribution centered at x_i. Note that the conditional probability, $p(x_j|x_i)$, is assumed to be high for nearby data points, x_i and x_j, and low when x_i and x_j are far apart [111]. For the low-dimensional data, the joint probability takes the form of:

$$q(x_j, x_i) = \frac{\exp(-||y_i - y_j||^2)}{\sum_{k \neq l} \exp(-||y_k - y_l||^2)}, \tag{3.4}$$

where y_i and y_j are data points for data where we assume that the conditional probability, $q(x_j|x_i)$, is high for nearby data points, y_i and y_j, and low when y_i and y_j are far apart [111].

The cost function of the t-SNE is given by:

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p(x_j, x_i) log \frac{p(x_j, x_i)}{q(x_j, x_i)}, \tag{3.5}$$

where $p(x_i, x_i)$ and $p(y_i, y_i)$ are set to zero as modeling interests lie on pairwise similarities. It is important to mention that the t-SNE is an improved version of the Stochastic Neighbor Embedding (SNE), which attempts to mitigate the so-called "crowd problem" encountered during optimization [111].

Effects of background noise

Background noise plays an important role in the perception of speech quality. Therefore, it is expected that an instrumental quality measure will be sensitive to changes in signal-to-noise ratios (SNR's). To summarize how the proposed model captures these changes, Figure 3.2 depicts the effects of ambient noise on the TV subspace. For this, we added background noise at different levels (0, 5, 10, 15 dB) to clean speech files. In order to visualize the similarities between data points (i.e., between i-vectors), we use t-SNE to embed high-dimensional i-vectors into two-dimensional space. Hence, each dot indicates an i-vector extracted from a speech recording and projected onto a two-dimensional space.

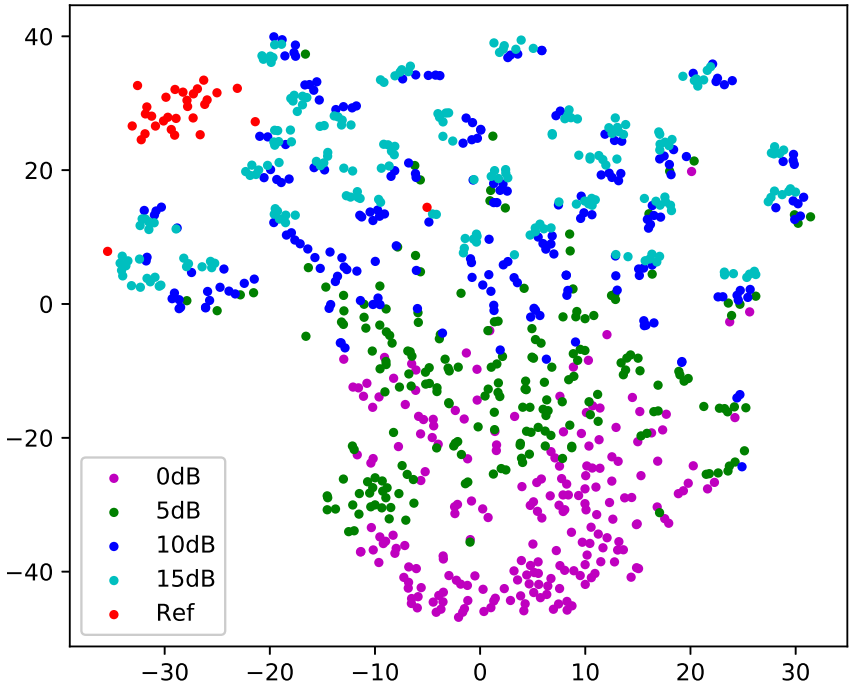


Figure 3.2 – i-Vector projection onto a 2-D space using t-SNE in the TV subspace at different levels of SNR.

In Figure 3.2, the recordings are labelled by SNR levels in the range of 0-15 dB (see different colors). Note that the speech recordings with the same distortion levels are closely clustered. Moreover, as the SNR decreases, the clusters deviate from the clean speech cluster, with larger "distances" being seen for noisier cases. It is expected that the cosine similarity distance will be able to capture this information. To give the reader more insights into the expected behaviour of the cosine similarity distance index, Figure 3.4-a provides the distribution of MOS as a function of SNR and Figure 3.4-b gives the cosine similarity distance as a function of SNR. As can be seen, cosine similarity distance is inversely proportional to SNR levels, which in turn are directly related to MOS.

Effects of reverberation

Reverberation is characterized by the reflections of the speech signal on surfaces (e.g., walls) and objects present in an enclosed environment [94]. This directly changes the frequency response of the speech signal [112], which can have either positive or negative effects on the perceived quality of

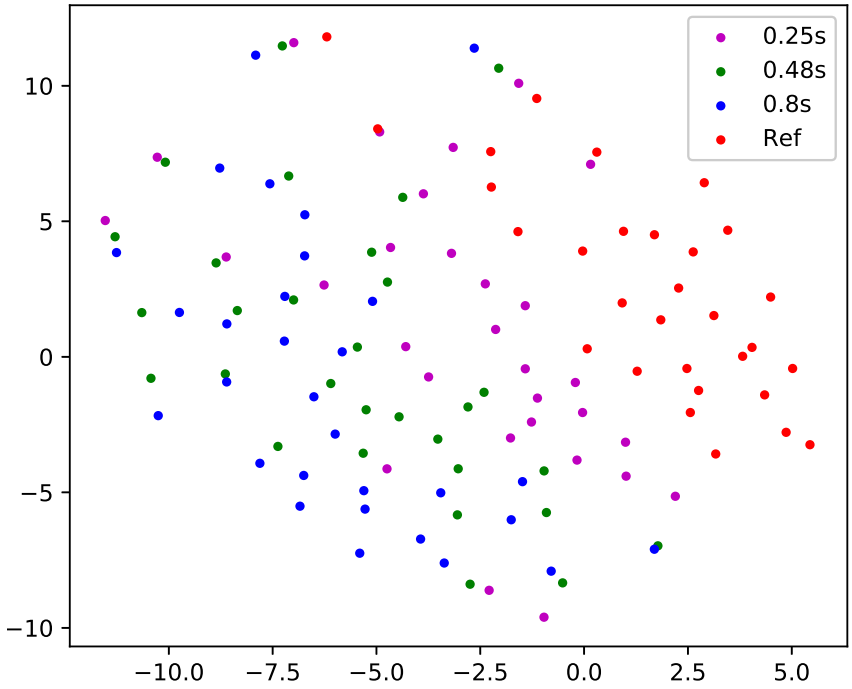
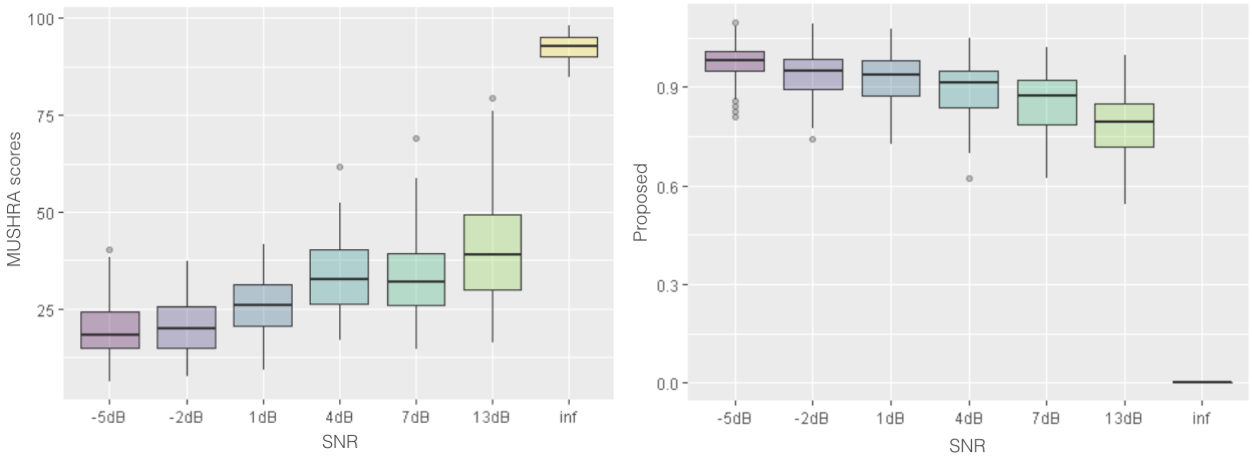


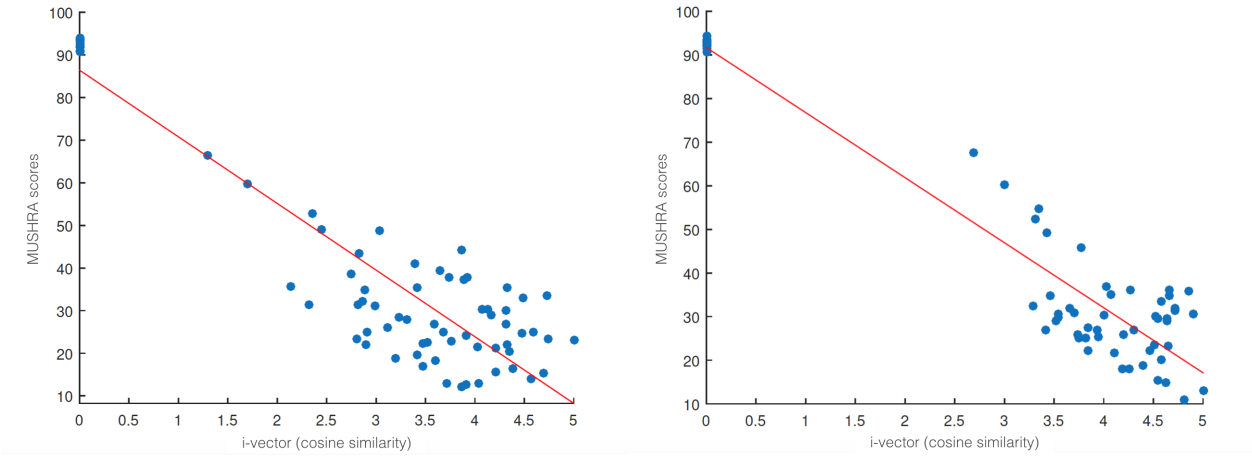
Figure 3.3 – i-vector disposition in the TV subspace at different levels of reverberation time (RT).

the speech. Early reflections, for instance, are desired as they cause changes in the signal timbre or coloration [112]. Late reflections, however, provide unwanted distortions represented by temporal smearing of the speech signal. As reverberation may affect the perceived quality of the speech [94], it is also expected that an instrumental quality measure will be able to rank different levels of reverberation, i.e., the time required for a signal to decay by 60 dB (also referred to as reverberation time or T_{60}) [113]. In Figure 3.3, we show how different levels of reverberation are captured and represented in the TV space. For this, a small sample of clean speech files, each with different speech content, was convolved with an impulse response (IR) representing the following reverberation time: 0.25 s, 0.48 s and 0.8 s. In the figure, "*Ref*" stands for the corresponding reference clean speech signals. As previously, as reverberation levels increase, greater "distances" from the clean speech recordings are observed. Note, for instance, that the blue dots indicating recordings with $T_{60} = 0.80$ s are farther away from the red dots, which represent i-vector from clean recordings, than the purple dots, representing i-vectors from recordings with $T_{60} = 0.25$ s. This is expected as blue dots are the recordings with the highest amount of reverberation.



(a) (b)

Figure 3.4 – Box-plot of (a) MUSHRA scores vs SNR and (b) cosine similarity distance vs SNR.



(a) (b)

Figure 3.5 – Scatter-plots of MUSHRA scores vs cosine similarity distance metric for speech corrupted with (a) only noise and (b) only reverberation.

3.3.3 Label distribution vs SNR levels

To give the reader more insights of the expected behaviour of the proposed instrumental measure, Figure 3.4 provides the distribution of MUSHRA scores and cosine similarity distance according to values of SNR's, which ranges from -5 dB to 13 dB, including clean signal (see "inf"). For this, we used the INRS dataset, which will be described in section 3.4.1. Note that, as mentioned before, cosine similarity distance is inversely proportional to perceived quality while the MUSHRA score is directly proportional to it. That is, while the MUSHRA score increases with the SNR the proposed instrumental measure decreases. The best similarity is achieved with "inf", when no noise is present

in the speech signal. This leads to the maximum MUSHRA scores of 100 and minimum cosine similarity distance of 0, as can be seen in Figure 3.4-b. Moreover, we observe a close trend in both distributions. For example, for the two lowest SNR's (i.e., -5 dB and -2 dB) we also have the two lowest MOS, followed by a slight improvement on the perceived quality. We see a similar pattern with the proposed method, but with maximum values of cosine similarity distance for low SNR's. Of course, we cannot assume by this that MUSHRA scores and our method is then correlated. However, we show in Figure 3.5 that there is a clear trend between our predictions and the MUSHRA scores for the same database. Note that Figure 3.5-a is the scatterplot for different levels of SNR while Figure 3.5-b is the scatterplot for different reverberation times.

3.3.4 Proposed methods

Given the insights mentioned above, four new instrumental measures are proposed, one full-reference and three no-reference, as detailed next.

Full-reference

This approach relies on the cosine similarity distance between i-vectors extracted from the reference (clean) signal and i-vectors extracted from their degraded speech signal counterparts. This similarity index is used as a correlate of speech quality.

No-reference - average reference model

As no-reference models do not have access to a reference signal, models of clean speech are required. The first approach proposed here models clean speech as an average i-vector computed from clean speech. More specifically, i-vectors are extracted from clean speech data (see Section 3.4.1) and averaged to obtain one reference i-vector to always be used in the computation of the cosine distance metric. As discussed in Section 3.3.2, larger distances from this average i-vector should indicate lower quality signals.

No-reference - vector quantizer codebook reference model

Inspired by earlier works on instrumental speech quality measurement [33], the second proposed approach relies on a vector quantizer (VQ) codebook of reference i-vectors obtained from clean speech. This builds upon the previous "average i-vector" method in that a different reference i-vector is used for each processed signal. In particular, the i-vector that most closely resembles the degraded signal i-vector is used for computation of the cosine similarity distance. Here, the k-means algorithm is used to build the vector quantization codebook. We tested different values of k in the range of 5-500 and found that the optimal number usually represented the number of distortions in the datasets used in our experiments.

No-reference - deep neural network reference model

Lastly, we are inspired by recent innovations in no-reference methods based on deep neural networks [48][28]. Here, a DNN is trained to estimate MOS. In particular, a fully-connected model with 400 input units (i.e., it receives i-vectors with 400 factors) and three hidden layers is used. The first hidden layer has 200 units, followed by 100 and 50 units. We adopted ReLU as the activation function and the output unit is a simple linear function. We used dropout with 0.2 rate and Adadelta as the optimization method. Dropout function is used as a regularizer to avoid over-fitting, and the Adadelta optimizer is used to dynamically update the learning rate and is suited to sparse data [114].

3.4 Experimental setup

In this section, the databases used in our experiments are presented, details about feature extraction are given and a description of the figure-of-merit used is presented.

3.4.1 Database description

Three main datasets are used herein: (1) the noise speech database developed by [115], (2) the INRS audio quality dataset [116], and (3) the open-source noise speech corpus NOIZEUS [117]. The

noise speech database is a clean and noisy parallel speech dataset, developed for the purpose of training speech enhancement algorithms, such as the speech enhancement generative adversarial network (SEGAN) [9]. It contains pairs of clean and noisy speech samples from 28 speakers (14 males and 14 females), all from the same accent region (England), taken from the larger Voice Bank corpus [118]. The dataset is sampled at 48 kHz. For the purpose of our experiments, only clean utterances were used from this dataset and solely to train the i-vector framework (i.e., the GMM-UBM and total variability matrix), as well as the two of our proposed no-reference approaches: (1) the one based on the reference i-vector, attained from the average of the extracted clean signals, and (2) the one based on the reference VQ codebook, also attained from the extracted clean signals.

The INRS dataset, in turn, contains speech files sampled at 16 kHz and degraded by noise and reverberation. To this end, clean signals from the TIMIT database were corrupted with babble and factory noises at SNRs of -2 dB, -5 dB, 1 dB, 4 dB, 7 dB and 13 dB. Noise signals were obtained from the NOISEX-92 dataset [119]. Reverberant utterances, in turn, were generated by convolving the clean utterances with 740 room impulse responses (RIR) with the following reverberation times (T_{60}): 0.3 s, 0.6 s, 0.9 s, 1.2 s and 1.5 s. For each T_{60} value, twenty different simulated RIRs (with different room geometry, source microphone positioning and absorption characteristics) were used. The RIRs were generated using an image-source method tool for simulating sound fields in virtual reverberant environments [120]. From the noisy signals, three speech DNN-based enhancement models were used. The first uses a feed-forward neural network to estimate the spectral data. The second model proposes the use of a feed-forward neural network in combination with arbitrary features to estimate a spectral mask. The final investigated enhancement model is based on spectral estimation through a context-aware recurrent neural network model. Details about these algorithms can be found in [116].

The listening tests followed the MUSHRA methodology. We performed two different online listening tests, one for the dereverberation (112 participants and 10 conditions) and one for the noise suppression (245 participants and 12 conditions). Both tests were MUSHRA-style tests where the output of all models (i.e., dereverberation or noise suppression), a hidden reference, a corrupted anchor, and the corrupted signals were presented to each participant. A slider with their positions quantized as integers ranging from 0 to 100 was used by the listeners to rate the signal quality. For each noise type, an anchor was the same stimulus corrupted with a 5 dB lower SNR. For the dereverberation conditions, in turn, the anchor was a signal convolved with an RIR with a T_{60} of 2 s. More details can be found in [116].

Lastly, the NOIZEUS database contains 30 IEEE sentences recorded by three male and three female speakers in a sound-proof booth. The speech signals, sampled at 8 kHz, are contaminated with eight different noise types taken from the AURORA database [121] and include car, train, babble, exhibition, restaurant, street, airport and station noises. Noise is added to the clean speech signals at SNRs of 0, 5, 10 and 15 dB. In our experiments, a subset of NOIZEUS is considered and includes 4 noise types (i.e., babble, car, street and train) and two SNR levels (i.e., 5 and 10 dB). Thirteen speech noise-suppression algorithms are also applied to the corrupted samples; a complete list is available in [121].

The subjective test conducted was based on the ITU-T Recommendation P.835, which aims at reducing the listeners' uncertainty to which component (i.e., the speech signal, the background noise, or both) to take into account when rating the signal quality. To this end, listeners are instructed to first evaluate the speech signal alone using a five-point scale of signal distortion (SIG). Next, they attend to the background noise alone using a five-point background intrusiveness (BAK) scale, and lastly, they are instructed to focus on the overall effect using the five-point mean opinion score (OVRL): [1 = bad, 2 = poor, 3 = fair, 4 =good, 5 = excellent]. A total of 32 participants, between the age of 18 and 50, took part in the test. The interested reader is referred to [121] for more details about the subjective listening test, including the details regarding the SIG and BAK scales.

3.4.2 Feature extraction

Prior to the extraction of MSF and MFCC features, all utterances were re-sampled to 16 kHz in case they were in a different sampling frequency. All utterances were normalized to -26 dBOV. In the case of MFCCs, a pre-emphasis filter of coefficient 0.95 was also applied and 30-ms hamming windows with 50% overlap were used. From each frame, a total of 19 coefficients together with log energy, delta and delta-delta coefficients were attained, thus leading to a final 60-dimensional feature vector for each frame. General details about these features can be found in Section 2.4.

Once speech parameterization was performed, an universal background model (GMM-UBM) with 1024 Gaussians was trained using 824 clean utterances from the noise speech database [118]. Five different total variability matrix T sizes were explored, containing 100, 200, 300, 400 or 500 total factors. The motivation behind testing different TV subspaces was to obtain insight on what

number of factors is optimal for speech quality assessment. After training, i-vectors were extracted from all recordings from the INRS and NOIZEUS datasets.

3.4.3 Figures-of-merit and benchmark algorithms

Here, the Pearson correlation coefficient between the ground truth subjective ratings (e.g., the MOS) and the cosine similarity distance between two i-vectors is used. Therefore, high correlations with subjective ratings will be considered a measure of good performance. As this is a first attempt at exploring the usefulness of the measure and its generalizability across datasets, aside from the DNN based results discussed in Section 3.5.3, we do not attempt to map the i-vector similarity score to an estimated MOS, as this would require the use of one of the two available datasets for training of this mapping. As such, the results reported herein are expected to be a lower bound on what performance could be achieved with the proposed method. Improved results should be expected once a similarity score to estimated MOS mapping is devised; this, however, is left for future work.

Correlations are reported on a per-condition basis, where all files under the same acoustic condition are first averaged prior to correlation calculation. Table 3.1 describes the acoustic conditions present in the INRS dataset excluding references and anchors. Two noise types (babble and factory) are considered at six different SNR levels. Moreover, three enhancement algorithms were applied, referred herein as Santos2018 [122], Williamson2017 [123] and Wu2016 [124]. Thus, for the noisy samples, an acoustic condition is defined by the noise type, processing status and SNR level, while for the reverberant samples, an acoustic condition is configured by the processing status and the reverberation time (T_{60}). The NOIZEUS dataset, on the other hand, presents four noise types: babble, car, street and train at two different SNR levels. For noise suppression, two algorithms based on the Wiener filter are tested, referred herein as Wavthre [125] and Tsoukalas [126]. Table 3.2 summarizes the conditions in the NOIZEUS database.

Lastly, to gauge the advantages of the proposed methods, two standard full-reference methods are used as benchmarks, i.e., ITU-T P.862 (PESQ) [127] and ITU-T P.863 (POLQA) [87]. For the no-reference measures, ITU-T P.563 [46] and SRMR [94] are also used as benchmarks.

Table 3.1 – Overview of the INRS speech quality dataset where acoustic conditions are presented for denoising and dereverberation processes, excluding reference files and anchors.

	Noise type	Enhancement algorithms	SNR		Enhancement algorithms	T_{60}
Denoising	Babble, Factory	Unprocessed, Santos2018, Williamson2017, Wu2016	-5 dB -2 dB 1 dB 4 dB 7 dB 13 dB	Dereverberation	Unprocessed, Santos2018, Williamson2017, Wu2016	0.3 s 0.6 s 0.9 s 1.2 s 1.5 s

Table 3.2 – Overview of the NOIZEUS speech quality dataset where acoustic conditions are presented for denoising, excluding references.

Noise type	Enhancement algorithms	SNR
Babble, Car, Street, Train	Unprocessed, Wavthre, Tsoukalas	5 dB 10 dB

3.5 Experimental results and discussion

In this section, we describe our experiments and provide a discussion of the achieved results.

3.5.1 Experiment I: Full-reference measurement

Table 3.3 presents the per-condition performances on the INRS database for the noise-only and reverberation-only settings (each condition also includes the enhanced counterpart), as well as the performances on the NOIZEUS dataset for noisy and enhanced speech. In the table, the performances of the two benchmark full-reference algorithms are also presented for comparison purposes. As can be seen, i-vectors extracted from MSFs are able to better correlate with subjective ratings for the noise, reverberation, and enhanced conditions, whereas MFCC-based ones are not as effective for reverberation and enhancement cases. For noise and reverberation conditions, a larger number of factors (400-500) resulted in the best results, whereas for enhancement alone, a smaller number of factors sufficed (200). Overall, MFCC-based i-vectors showed to be more sensitive to the number of factors relative to MSF-based ones. Correlation values were in line with those obtained with the benchmarks, but with the added benefit of not requiring temporal alignment. In the case of the

Table 3.3 – Per-condition performance of the proposed full-reference approach on the INRS and NOIZEUS databases. Numbers in subscript indicate the number of factors in the total variability space.

	INRS		NOIZEUS	
Metrics	Noise	Reverb	Noise	Enhanced
PESQ	**0.95**	0.92	0.96	0.88
POLQA	**0.95**	0.89	0.92	0.89
$MFCC_{100}$	-0.73	-0.45	-0.95	-0.61
$MFCC_{200}$	-0.73	-0.52	-0.97	-0.47
$MFCC_{300}$	-0.77	-0.54	-0.97	-0.44
$MFCC_{400}$	-0.78	-0.58	-0.96	-0.41
$MFCC_{500}$	-0.89	-0.64	-0.97	-0.64
MSF_{100}	-0.93	-0.86	-0.95	-0.89
MSF_{200}	-0.92	-0.88	-0.94	**-0.90**
MSF_{300}	-0.94	-0.90	-0.95	-0.85
MSF_{400}	**-0.95**	-0.92	-0.97	-0.83
MSF_{500}	**-0.95**	**-0.93**	**-0.98**	-0.88

Table 3.4 – Per-condition performance of the no-reference approach based on average of i-vectors for the INRS and NOIZEUS databases. Numbers in subscript indicate the number of factors in the Total Variability Space.

	INRS		NOIZEUS	
Metrics	Noise	Reverb	Noise	Enhanced
SRMR	0.71	0.48	0.93	**0.76**
P563	0.39	0.35	0.86	0.12
$MFCC_{100}$	-0.14	-0.18	**-0.94**	-0.51
$MFCC_{200}$	-0.01	-0.06	-0.33	-0.08
$MFCC_{300}$	-0.58	-0.22	-0.43	-0.13
$MFCC_{400}$	-0.05	-0.04	-0.45	-0.30
$MFCC_{500}$	-0.35	-0.20	-0.79	-0.13
MSF_{100}	**-0.87**	**-0.68**	-0.39	-0.29
MSF_{200}	-0.81	-0.53	-0.55	-0.56
MSF_{300}	-0.78	-0.58	-0.09	-0.51
MSF_{400}	-0.55	-0.44	-0.08	-0.59
MSF_{500}	-0.52	-0.44	-0.44	-0.03

NOIZEUS dataset, the proposed method achieved slightly higher correlations, i.e., 0.98 for the noisy and 0.90 for the enhanced conditions, when compared to the benchmarks.

Table 3.5 – Per-condition performance of the no-reference approach based on VQ codebook for the INRS and NOIZEUS databases.

	INRS		NOIZEUS	
Metrics	Noise	Reverb	Noise	Enhanced
SRMR	0.71	0.48	0.93	**0.76**
P.563	0.39	0.35	0.86	0.12
$MFCC_{k10}$	-0.09	-0.32	**-0.95**	-0.63
$MFCC_{k30}$	-0.23	-0.10	-0.93	-0.34
$MFCC_{k40}$	-0.01	**-0.58**	-0.92	-0.44
MSF_{k10}	**-0.72**	-0.42	-0.93	-0.48
MSF_{k30}	-0.31	-0.08	-0.01	-0.13
MSF_{k40}	-0.47	-0.32	-0.91	-0.19

3.5.2 Experiment II: No-reference measurement based on average model

Table 3.4 shows the results attained using the simplest no-reference measure, as well as the two no-reference benchmarks. As expected, results are much lower than what can be achieved with a full-reference measure. Notwithstanding, the simple average i-vector model extracted from MSFs outperformed both benchmarks on the INRS database when 100, 200 and 300 factors were adopted. The simple approach, however, was not capable of accurately tracking the quality of the enhanced signals in the NOIZEUS database, despite outperforming ITU-T P.563. I-vectors extracted from MFCCs, however, were able to quantify the distortions in the noise-only case for the NOIZEUS database and results in line with SRMR were achieved.

Overall, reverberation showed to be a harder problem. This was expected based on insights from Figures 3.2 and Figure 3.3. In fact, the projection of i-vectors is less clustered in the case of reverberation, thus making it harder to model using such a simple approach. Furthermore, it is interesting to note that, in this case (MSF features), performance is inversely proportional to the number of factors.

3.5.3 Experiment III: No-reference measurement based on the VQ codebook

Table 3.5 provides results obtained with the VQ codebook based approach. We tested different numbers of clusters (i.e., 10, 30 and 40) and we adopted 10 clusters, which seemed to be the optimal value. The performances are presented for both MFCCs and MSF-based systems. We can note

Table 3.6 – Per-condition performance of a no-reference DNN-based model trained with i-vector features from speech samples from the INRS database.

Metrics	All
PESQ	0.90
POLQA	0.88
MFCC_{100}	0.77
MFCC_{200}	0.86
MFCC_{300}	0.80
MFCC_{400}	**0.94**
MFCC_{500}	0.89
MSF_{100}	0.84
MSF_{200}	0.86
MSF_{300}	0.88
MSF_{400}	0.86
MSF_{500}	0.85

that the proposed metric based on MSF-k10 provides more stable results throughout the tested conditions. Moreover, it presents competitive performance compared to the two benchmarks, SRMR and P.563, outperforming the latter for all tested conditions and SRMR in two situations. See the first and last columns with the respective correlations equal to -0.72 and -0.80. We can verify that distortions caused by reverberation and enhanced speech are more challenging to the proposed metric. In fact, the results were more reliable for distortions caused by noise for all the metrics, including the proposed MSF-k10.

3.5.4 Experiment IV: No-reference measurement based on DNNs

In this experiment, the model is trained considering both noisy and reverberant samples. We randomly sampled 70% of the examples in the INRS database to train. During training, 20% of these examples are used for validation. The remainder of the dataset is kept for testing. The results presented in Table 3.6 are based on the average value after running the experiment 10 times, randomly picking samples for the training and test sets. As can be seen, the best results are achieved by the MFCC-based i-vector with 400 factors, outperforming both PESQ and POLQA. Models based on MSFs and 300 factors achieved similar results to PESQ and POLQA.

3.6 Conclusions

In this Chapter, we explored the use of i-vector speech representations for instrumental quality measurement of noisy, reverberant and enhanced speech. We show how the total variability space is capable of capturing ambient factors and one full-reference and three no-reference measures are proposed. Experimental results on two datasets showed the full-reference method achieving results in line with two standard benchmarks and bypassing the need for time alignment between reference and processed signals. On the same datasets, the three no-reference measures outperformed two no-reference benchmarks, thus showing their effectiveness in tracking the quality of hands-free and enhanced speech. In fact, one no-reference approach based on deep neural networks outperformed the two full-reference benchmarks, without the need for a clean reference signal.

Non-intrusive speech quality measurement based on clean speech i-vector estimation and similarity scoring

4.1 Preamble

This chapter is compiled from material extracted from the manuscript accepted to the Journal of Quality and User Experience [J1].

4.2 Introduction

Non-intrusive instrumental speech quality assessment relies only on the received (processed) signal to predict quality. Such methods are called non-intrusive and are crucial in speech applications where reference clean signals are not accessible. Predicting the speech quality without the reference clean speech signal is a challenging task and often offers lower performance compared to intrusive methods. This is due to the wide variability of the input speech signal, which is the result of different speakers, vocal tracts, pitch characteristics and speech content [33]. One way to overcome this is by estimating the clean speech signal from the corrupted speech signal. The ITU-T Recommendation P.563 [46], for instance, is based on a similar approach. The main block of the non-intrusive instrumental measure

attempts to separate speech from non-speech content. It then uses high-order statistical analysis to attain additional information about how natural the speech is [46].

The model has some drawbacks. First, it is specifically designed for the prediction of speech quality in public telephone networks [46]. It is also only appropriate for narrow-band speech signals sampled at an 8-kHz sampling rate [47]. Therefore, it only covers the types and the amount of the distortions present in the range of common occurrences in such networks. This results in poor predictions for more recent, realistic scenarios involving background noise, reverberation, and enhanced speech [46].

The speech-to-reverberation modulation ratio (SRMR) was developed as an alternative. The method attempts to separate clean and reverberation/noise components from the degraded signal. In fact, it relies on the principle that the modulation energy of clean speech is generally concentrated in lower modulation frequencies (below 20 Hz) while room acoustic artefacts typically arise in higher modulation frequencies above 20 Hz [47]. While the metric is found to be very efficient to assess the quality of reverberant speech, it has only shown moderate performance with more recent types of distortions [48].

More recent works have proposed deep neural networks (DNN) as non-intrusive instrumental measures. In [48] and [27], for instance, the authors present an investigation of the applicability of deep neural network approaches to estimate the MOS without the reference signal. In [49], a deep neural network architecture, based on the Human Auditory System, is proposed to extract features to be used for non-intrusive objective quality assessment. In [28], a novel non-intrusive measure based on a recurrent neural network using long short-term memory cell and modulation energy features is proposed by the authors. Although these findings are promising, neural network-based models often require a massive amount of data to significantly outperform more traditional machine learning algorithms as well as approaches based on feature engineering. Therefore, it is important to investigate non-intrusive models that can perform well in scenarios where a large amount of labelled data is not available. In this Chapter, we intend to fill this gap by proposing a new non-intrusive instrumental quality measure based on the similarity between two i-vectors.

In particular, we propose the i-vector framework as a non-intrusive speech quality measure. In order to overcome the problem of the unavailability of the reference clean speech signal, we propose to reconstruct the clean spectra from the degraded signal itself. In our proposed solution, a clean speech Gaussian mixture model (GMM) must be trained with RASTA-filtered mel-frequency cepstral

coefficients (RASTA-MFCCs), extracted from several clean speech files. This allows us to attain a model of clean spectrum characteristics. Thus, the clean speech GMM allows us to extract the proposed i-vector representation using the reference clean spectra estimated from the corrupted signal.

The remainder of this Chapter is organized as follows. Section 4.3 presents the proposed method and background on the i-vector framework. Section 4.4 discusses how the i-vector similarity relates to speech quality. Section 4.5 describes the experimental setup and Section 4.6 presents the results and discussion. Lastly, conclusions are presented in Section 4.7.

4.3 Proposed non-intrusive speech quality method

In this section, our proposed instrumental quality measure is explained. We start with the speech parameterization step, followed by a description on how to train the clean speech model, as well as on how to estimate the clean spectrum from the degraded speech signal. Lastly, the i-vector framework used to extract the i-vector representation is also discussed.

4.3.1 Cepstrum parameterization and RASTA filtering

In this work, MFCCs are used with two different purposes: first, to train the clean speech Gaussian Mixture Model (see Sections 4.3.2 and 4.3.3), and second, to train the i-vector framework. Prior to the extraction of i-vector representation, all utterances were re-sampled to 16 kHz [1] in case they were in a different sampling frequency. All utterances were normalized to -26 dBov and a pre-emphasis filter of coefficient 0.97 was also applied. A 32-ms Hanning window with 50% overlap was used. From each frame, a total of 13-dimensional MFCC coefficients were attained.

For the clean speech modeling part of the proposed method, in turn, we also apply the relative spectral (RASTA) filtering technique. This is particularly important to remove irrelevant information that might be "embedded" in the speech signal, such as those characterized by the communication channel [128]. The RASTA filter technique relies on the fact that linguistic content is coded based on the movements of the vocal tract at rates-of-change different from other non-linguistic content.

1. In the case of P.563, speech samples were resampled to 8 kHz.

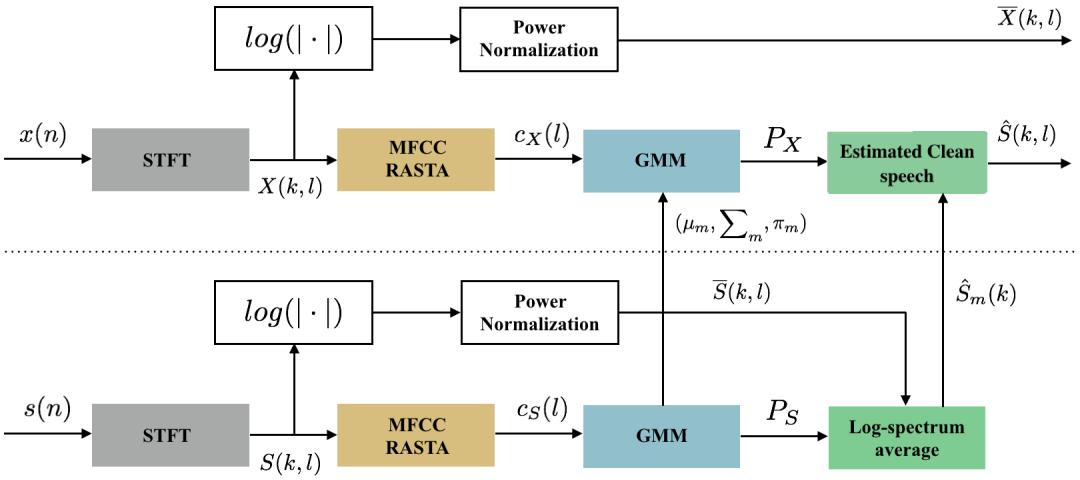


Figure 4.1 – Diagram representing the steps to estimating the reference clean log-magnitude spectrum.

In the past, it has been shown to be efficient in mitigating the negative effects of convolution and additive noise. More details about the steps to apply the RASTA technique can be found in [128].

4.3.2 Clean speech Gaussian Mixture Model

Inspired by earlier work on GMM-based non-intrusive quality measurement [129], as well as by our recent work on blind channel estimation [82], we propose to train a clean speech Gaussian mixture model to blindly estimate the reference clean spectrum from the corrupted speech signal.

The procedure to estimate the reference clean spectrum magnitude is depicted in Figure 4.1. The clean speech, $s(n)$, is used to train the clean speech GMM. This is depicted in the lower half of Figure 4.1. Prior to the training process, some pre-processing steps are taken. The speech signals are first segmented into frames of 512 samples (32-ms length for speech signal sampled at 16 kHz), with 50% hop-size (i.e., 256 samples). Pre-emphasis is also adopted with a filter of coefficient 0.97. A Hanning window is applied on each frame, followed by the computation of the STFT. The attained log-spectrum, $\underline{S}(k,l) = \log(|S(k,l)|)$, is then normalized by subtracting the log-spectrum mean as follows:

$$\overline{S}(k,l) = \underline{S}(k,l) - \frac{1}{K}\sum_{k=1}^{K} \underline{S}(k,l), \tag{4.1}$$

where K is the number of STFT points. Note that a total of 12 MFCC coefficients plus the log energy are attained, leading to a 13-dimensional vector for each frame. Lastly, a RASTA filter is also applied to mitigate channel effects [130].

The Gaussian mixture model is trained with the MFCC-RASTA coefficients, $c_S(l)$, obtained after the speech parameterization is performed. The GMM adopted here contains $M = 1024$ Gaussians and the mixture probabilities given by:

$$p_{l,m}(c_s(l)) = \frac{\pi_m \mathcal{N}(c_s(l)|\mu_m, \sum_m)}{\sum_{j=1}^{M} \pi_j \mathcal{N}(c_s(l)|\mu_j, \sum_j)}, \tag{4.2}$$

where $\lambda = \{\mu_m, \sum_m, \pi_m\}$ are the parameters of a multivariate Gaussian distribution represented by $\mathcal{N}(c_s(l)|\mu_m, \sum_m)$. To attain the average of the short-term log-spectra, $p_{l,m}(c_s(l))$ and $\overline{S}(k,l)$ are combined over all available frames of the training data. This leads to M average clean speech log-spectra:

$$\hat{S}_m(k) = \frac{\sum_{l=1}^{L} p_{l,m}(c_s(l))\overline{S}(k,l)}{\sum_{l=1}^{L} p_{l,m}(c_s(l))}, \forall k, m = 1, \dots M, \tag{4.3}$$

Note that each mixture, m, is associated with a clean speech spectrum, attained from the weighted average of multiple clean speech spectra.

4.3.3 Clean spectrum estimation

The upper part of Figure 4.1 provides a description of how to estimate the clean spectrum from the degraded signal. For instance, the speech signal, $x(n)$, is segmented into overlapping frames and the same pre-processing steps, discussed previously, are taken prior to extracting the STFT, $X(k,l)$, and the RASTA-MFCC coefficients, $c_x(l)$. For a given speech signal, the clean log-spectrum is then obtained using the feature vectors (i.e., $c_x(l)$) and the GMM parameters ($\mu_m, \sum_m$ and π_m) computed during the training phase. Then, the likelihood that a feature vector $c_x(l)$ belongs to the m-th mixture can be computed as in (4.2), which leads to a probability, $0 < p_{l,m} < 1$, for each mixture $m = 1,...,M$. This probability can be used to estimate the clean speech spectra of the l-th frame using the weighted average of the clean speech spectra, $\hat{S}(k,l)$, as described below:

$$\hat{S}(k,l) = \sum_{m=1}^{M} p_{l,m}(c_x(l))\hat{S}_m(k), \forall k. \tag{4.4}$$

Considering $\hat{S}(k,l) \approx S(k,l)$, an estimation of the clean speech spectrum is attained from the degraded speech spectrum, $X(k,l)$. This can be used as reference clean speech to estimate the speech

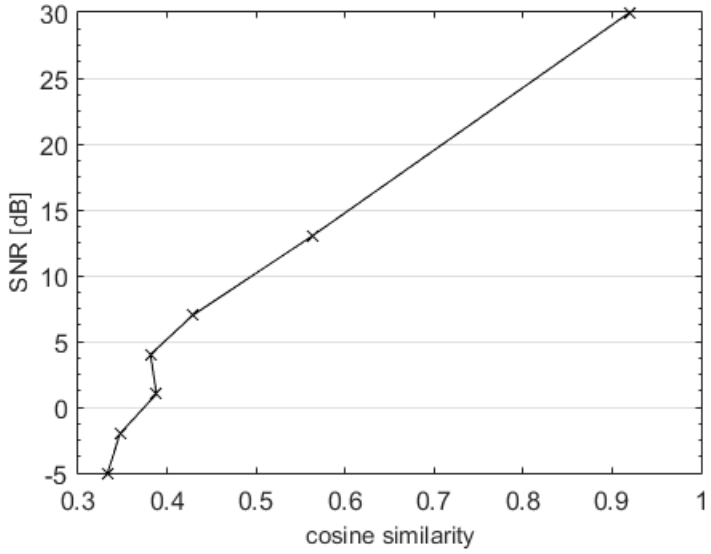


Figure 4.2 – Cosine similarity versus SNR where each point represents the average value over 10 speech files.

quality. Next, we discuss the i-vector framework, used here to extract the i-vector representation from the degraded speech signal, $\overline{X}(k,l)$, as well as from the estimated reference clean signal, $\hat{S}(k,l)$.

4.4 Speech quality assessment based on i-vector similarity

A similarity measure compares two vectors and computes a number that represents their similarity. In this section, we present the cosine distance and the euclidean distance as two alternatives to estimate the similarity between two i-vectors. We also discuss how distortions are captured by the i-vector framework and how it is expected to correlate with speech quality.

4.4.1 Cosine similarity

The cosine similarity, $\cos(\theta)$, is computed as described in (3.1). In our previous work [29], we showed that by having access to the reference signal no speaker and speech variability is found between the two i-vector representations. That is, speaker and speech content will remain the same for the reference and degraded signal and only changes in the channel factors will be present. Therefore, significant alterations in the channel factors will decrease the values of the cosine similarity as the angle between w_{ref} and w_{deg} is expected to increase. Hence, the cosine similarity provides

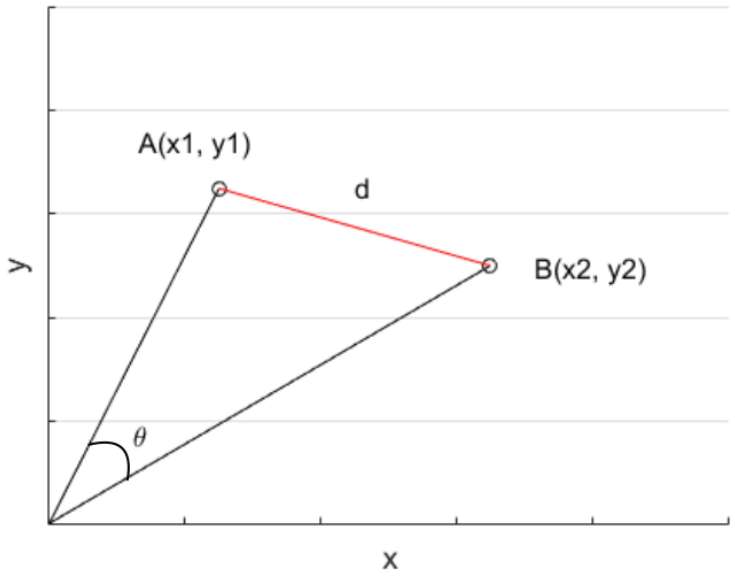


Figure 4.3 – Geometric interpretation of the cosine and euclidean similarities.

values close to 0 for orthogonal i-vectors (i.e., approximately 90°), representing low similarities and high distortions, and values close to 1 for i-vectors in the same direction (i.e., approximately 0°), representing high similarities and low distortions.

In the present work, we propose to use a reference i-vector representation based on the estimation of the clean speech spectrum obtained from the degraded signal. Our hypothesis is that the same behaviour will be observed while using the reference signal. That is, the cosine similarity will be lower for large distortions and higher for small distortions, as shown in Figure 4.2. Each point in this figure represents the average value of the cosine similarity over 10 speech files with the same condition, that is, same noise type, i.e., factory noise, and same SNR. Note that for low SNRs (e.g., -5 and -2 dB), the cosine similarity is also low, indicating that the i-vectors are close to being orthogonal. As the SNR increases, on the other hand, the cosine similarity also increases, which implies that the i-vectors are more similar with minimal amount of distortion. We can conclude that the cosine similarity is, therefore, directly proportional to the SNR and is expected to correlate to speech quality.

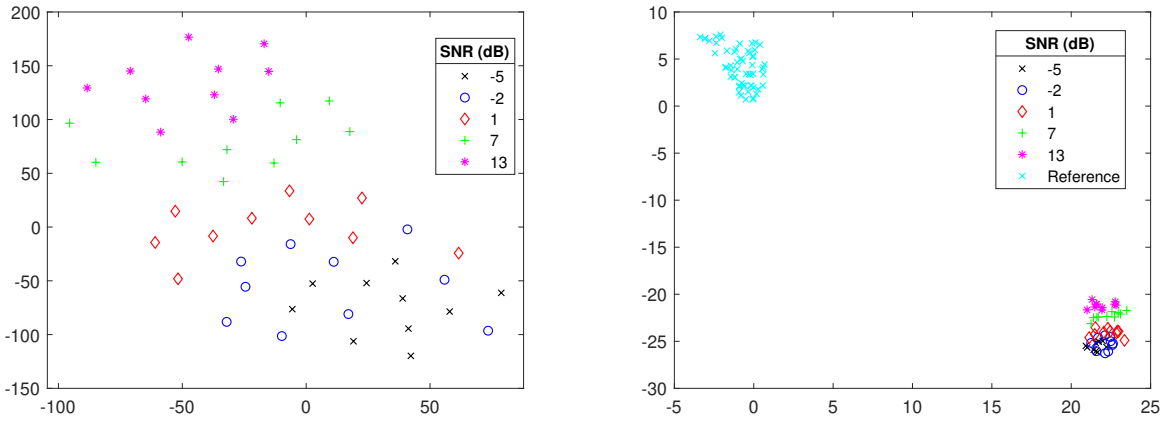


Figure 4.4 – I-vector projection onto a 2-D space using t-SNE in the TV subspace with SNRs varying between -5 to 13 dB in (a) and in (b) the i-vector representation obtained from estimated clean spectrum is also provided.

4.4.2 Euclidean distance for similarity scoring

The euclidean distance is the square root of the sum of squared differences between corresponding components of two vectors and is defined as follows:

$$d_{(A,B)} = \sqrt{\sum_{i=1}^{N} (x_n - y_n)^2}, \tag{4.5}$$

where $A = (x_1, x_2, ..., x_n)$, $B = (y_1, y_2, ..., y_n)$ and $d_{(A,B)}$ represents the distance between the vectors A and B. Figure 4.3 depicts an example of the euclidean distance in two dimensions. This distance can be used as a measure of similarity between two i-vectors, with smaller distances linked to higher similarities whereas larger distances relate to lower similarities.

4.4.3 Effects of distortions on the total variability subspace

An instrumental quality measure is expected to be sensitive to various distortions occurring on the speech signal. The level of background noise and reverberation, for instance, plays an important role in the perception of speech quality. To give the reader some insight on how the proposed model captures changes in SNR, Figure 4.4 depicts the effects of ambient noise on the TV subspace. The t-distributed stochastic neighbor embedding (t-SNE) is used to visualize the similarities between

data points (i.e., between i-vector representations). The method embeds high-dimensional data into a lower-dimension space, usually two or three dimensions, by minimizing the divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data [111]. An i-vector extracted from a speech recording is projected onto a two-dimensional space and is represented by the dots in Figure 4.4.

In Figure 4.4-a, the recordings are labelled by SNR levels ranging from -5 to 13 dB (see different colors). Note that i-vector representations extracted from speech recordings with the same distortion levels are clustered together. In Figure 4.4-b, an i-vector representation obtained from the estimation of the reference clean spectrum is also provided and labelled as "Reference. " Our hypothesis is that by estimating the clean counterpart of the degraded i-vector representation, we will be able to mitigate variability caused by speaker and speech content. Therefore, the main source of variability between i-vectors for a given degraded and estimated reference clean spectra pair will come from distortions such as background noise and reverberation.

It can be observed in Figure 4.4-b that the projections of the reference i-vectors are clustered away from the distorted samples. As each representation is extracted from different recordings, it is expected to encounter some level of variability among them. The variability here is due to different speakers and speech content. However, considering the i-vector representation of a degraded recording and its respective estimated reference counterpart, the speaker and content information is expected to remain the same or with minimum variation. That is, while computing the cosine and Euclidean distances, the main factors of variability should be alterations that occurred in the channel information.

To give the reader more insights into the expected behaviour of the cosine distance index, we use the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA [131]) scores available in the INRS audio quality dataset [116], described in Section 4.5.1. Note that the MUSHRA is a listening test method to evaluate audio systems that introduce intermediate impairments. It is based on a multi-stimulus where listeners are presented with the stimuli processed by a system under test, a hidden reference, and an anchor stimuli. The scale used ranges from 0 to 100 and is divided into five equal intervals categorized as "bad", "poor", "fair", "good", and "excellent" [132]. Figure 4.5 provides the distribution of the MUSHRA scores and the cosine distance (referred to as "Proposed" in the figure) as a function of SNR and reverberation time (T_{60}). For better interpretation of the

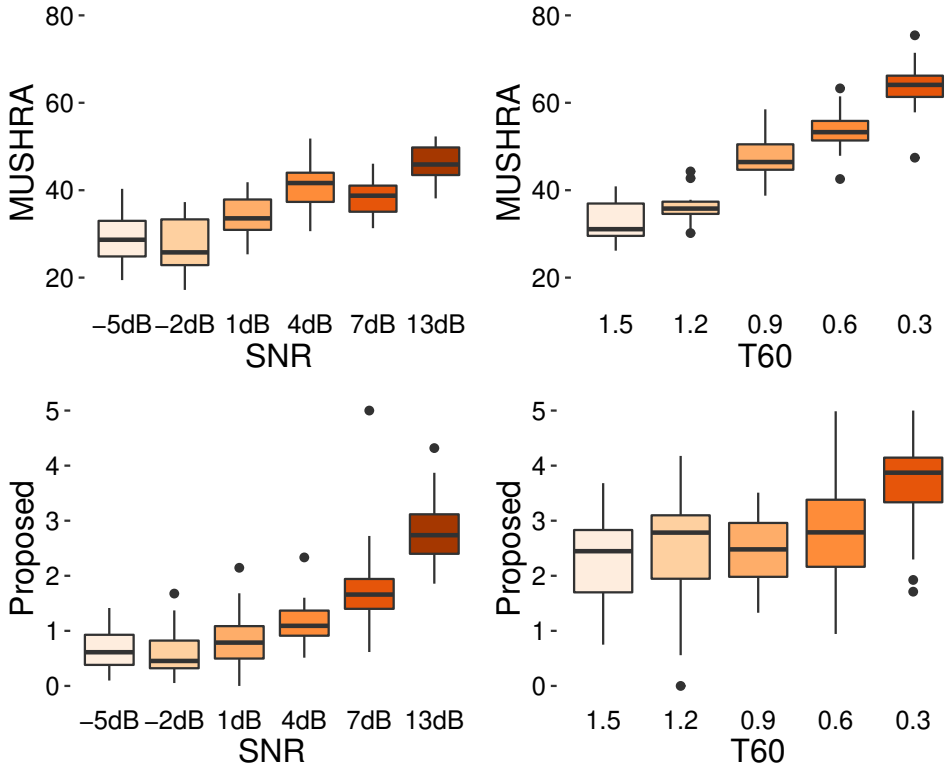


Figure 4.5 – Box-plot of the cosine similarity based on i-vectors versus SNR (first column) and versus reverberation time, $T60$ (second column).

scores, the proposed measure was re-scaled to have its values between 0 and 5. We can observe a similar trend between the MUSHRA scores and the proposed measure as the SNR increases. Note that as the SNR increases both scores increases as well. Moreover, for -5 dB and -2 dB SNR, the scores are interchangeable for both measures. This effect can be confirmed in Figure 4.4-a, where the points for -2 and -5 dBs are mixed together, which makes it more difficult for the model to distinguish them. Lastly, the effects of reverberation are also explored. We can verify that both the MUSHRA scores and the cosine similarity increase as the T_{60} decreases. Interestingly, for low T_{60} MUSHRA scores and cosine distance scores were significantly high.

4.5 Experimental setup

In this section, we give a short description of the databases adopted in this work, as well as the instrumental measures used as benchmarks. Details about feature extraction and figures-of-merit are also addressed.

4.5.1 Database description

Two main datasets are used herein: (1) the noise speech database developed by [115] and (2) the INRS audio quality dataset [116]. The noise speech database was developed for the purpose of training speech enhancement algorithms, such as the speech enhancement generative adversarial network (SEGAN) [9]. It contains pairs of clean and noisy speech samples from 28 speakers (14 males and 14 females), all from the same accent region (England), taken from the larger Voice Bank corpus [118]. Since the purpose of this dataset is to solely train our clean speech model discussed in 4.3.2, only clean utterances (824 in total) were used from this dataset. Note that no audio quality information is available from this data.

The INRS dataset, in turn, contains speech files degraded by noise and reverberation. To this end, babble and factory noises at SNRs of -2 dB, -5 dB, 1 dB, 4 dB, 7 dB and 13 dB were used to corrupt clean signals from the TIMIT database. Noise signals were obtained from the NOISEX-92 dataset [119]. To attain reverberant utterances, 740 room impulse responses (RIR) with reverberation time (T_{60}) values ranging from 0.2 s to 2.0 s in 0.05-s steps were convolved with clean utterances. Twenty different RIRs (with different room geometry, source microphone positioning and absorption characteristics) were used. From the noisy signals, three speech DNN-based enhancement models were used, referred herein as Santos2018 [122], Williamson2017 [123] and Wu2016 [124]. The first uses a feed-forward neural network to estimate the spectra. The second model proposes the use of a feed-forward neural network in combination with arbitrary features to estimate the spectral mask. The final investigated enhancement model is based on spectral estimation through a context-aware recurrent neural network model. Details about these algorithms can be found in [116].

The listening test was conducted online and consisted of the speech quality being evaluated using the MUSHRA methodology, described in Section 4.4.3. Separate tests for the outputs from the dereverberation and noise reduction models were performed by the authors in [116]. The denoising quality tests had 12 conditions and were performed by a total of 245 participants, while the dereverberation quality tests had 112 participants and 10 conditions. For more details on the listening test, the reader is referred to [116].

Table 4.1 – Four configurations based on 2 similarities and 2 cepstral parameterization.

Configuration	MFCC
cosine	13 coef. (with energy)
euclidean	13 coef. (with energy)
cosine_{deriv}	39 coef. (energy + derivatives)
euclidean_{deriv}	39 coef. (energy + derivatives)

4.5.2 i-Vector configuration

Table 4.1 summarizes the four i-vector representation configurations investigated in our experiments. Five different total variability matrix T sizes were also explored, containing 100, 200, 300, 400 or 500 total factors. The motivation behind testing different TV subspaces was to obtain insight on what number of factors is optimal for speech quality assessment. After training the framework with the noise speech database, i-vectors were extracted from all recordings in the INRS audio quality dataset.

4.5.3 Non-intrusive instrumental measures based on deep neural network

In this section, we present two non-intrusive deep neural network approaches to estimate speech quality. Because these methods require a subset of the INRS quality dataset to be trained, they are evaluated in a separated experiment using a reduced number of test examples. The first approach is an end-to-end solution, used as benchmark, namely Quality-NET. The second one is based on the proposed i-vector combined with a simple fully connected neural network.

Quality-Net

Quality-Net is an end-to-end non-intrusive speech quality assessment model based on a bidirectional long short-term memory (BLSTM). In [133], the authors have successfully trained the model to predict PESQ scores. Although highly correlated to PESQ, it had not been trained to predict subjective scores. Here, we attempt to fill this gap by training the Quality-Net model to predict MUSHRA scores. The model takes as input magnitude spectrograms and quality scores are associated to each utterance. The model was designed to consider quality at the frame-level despite the fact that labels are provided at the utterance level. This represents an advantage considering

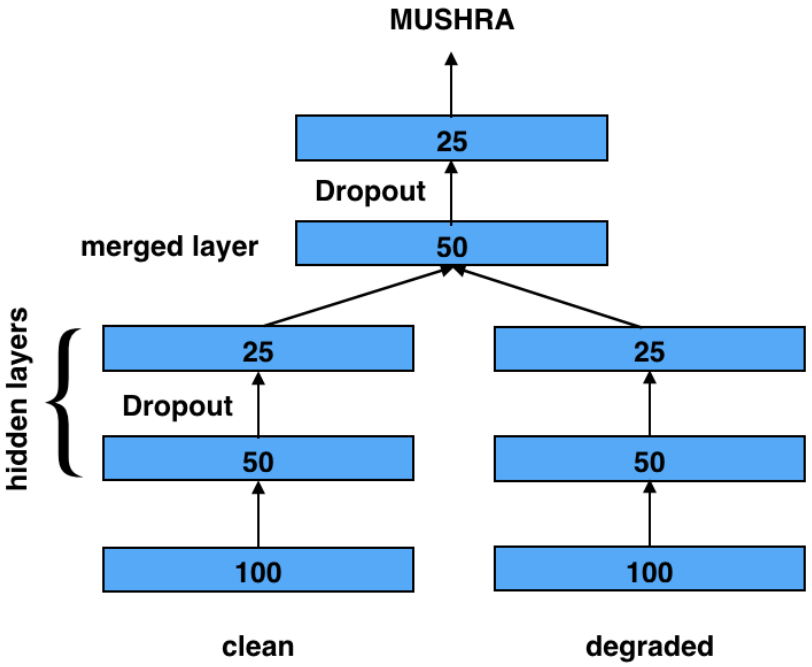


Figure 4.6 – Non-intrusive deep neural network model based on the i-vector representation of the degraded and the estimated clean reference speech signal.

that noise or speech distortions may not occur in all frames, but should still be accounted for while assessing the quality of an utterance. Therefore, the model must learn to estimate the overall quality score based on assessment performed on a per-frame basis. The overall quality, estimated at the utterance-level, is then attained by combining frame-wise scores via global averaging [133].

i-vector

Since the i-vector extractors (i.e., the UBM and the T matrix) are trained in an unsupervised fashion, no prior information about quality is captured by the representation other than what is already intrinsically present in the speech signal itself. Therefore, further supervised training is likely to lead to improved results. Thus, we propose to train a fully connected neural network that receives as input the degraded i-vector representation and its estimated clean reference counterpart. Such a DNN learns a better feature representation by mapping the input features into a linearly separable feature space [50]. This is achieved by successive linear combinations of the input variables, $z_i = w_i x_i + b_i$, where w_i and b_i are weights and biases, followed by a non-linear activation function. As can be seen in Figure 4.6, the DNN architecture adopted here has 100 input units for each i-vector representation (i.e., 100 for the degraded version and 100 for its clean counterpart). Following the

input layer there are, respectively, 50 and 25 units in the first and second hidden layers. The outputs of the second hidden layers are concatenated into a 50 units merged layer that is followed by another hidden layer with 25 units. We used ReLU as activation function. The dropout rate adopted was 0.3. The authors had explored similar approach in [48], but using as input only the degraded i-vector representation of the speech signal. Hence, this new approach is expected to achieve improved results.

4.5.4 Comparing subjective and objective ratings: scale adjustments

Subjective listening tests commonly use the five-point absolute category rating (ACR) listening quality (LQ) scale, defined in ITU-T Recommendation P.800 [17]. The scale, from 1 to 5, refers to categories bad, poor, fair, good, and excellent, respectively. Other tests, such as MUSHRA, have listeners rate quality based on a 0-100 scale. While the ranking of scores between tests remain similar (i.e., low score relates to poor quality), direct comparison between different subjective scales has to be done carefully. The same is true when comparing objective metrics, which may have been calibrated using a specific scale), with subjective ratings. Here, we adopt the use of a 3rd order monotonic mapping function in order to map objective and subjective scores to the same scale [134]. Monotonic 3rd order polynomials have been widely used in ITU-T evaluations and are used here for all comparisons.

4.5.5 Figures-of-merits

Three figures-of-merit are used to assess the performance of the proposed and the benchmark instrumental measures. The linear relationship between estimated quality and subjective ratings is computed using the Pearson correlation coefficient, ρ_{Person}. The Spearman rank correlation coefficient, $\rho_{Spearman}$, which computes the ranking capability of each measure is also considered. Lastly, the epsilon-insensitive root mean square error (ϵ-$RMSE$) is used to measure the difference between values predicted by a model and subjective ratings [135]. The ϵ-$RMSE$ measure is similar to $RMSE$, but takes into account the uncertainty of the subjective test by using a 95% confidence interval defined as,

$$ci_{95}(c) = t(0.05, M)\frac{\sigma(c)}{\sqrt{M}}, \tag{4.6}$$

where a condition type is denoted by c and the total number of conditions by M, and σ is the standard deviation of subjective scores per condition, $t(0.05, M)$ is the t-value computed at a 0.05 significance level. The per-condition ϵ-$RMSE(c)$ is then defined as

$$\epsilon\text{-}RMSE(c) = max(0, |S(c) - O(c) - ci_{95}(c)), \tag{4.7}$$

where $S(c)$ corresponds to the average subjective quality score for a specific condition c and $O(c)$ is the corresponding average objective score. The ϵ-$RMSE$ is hence given by,

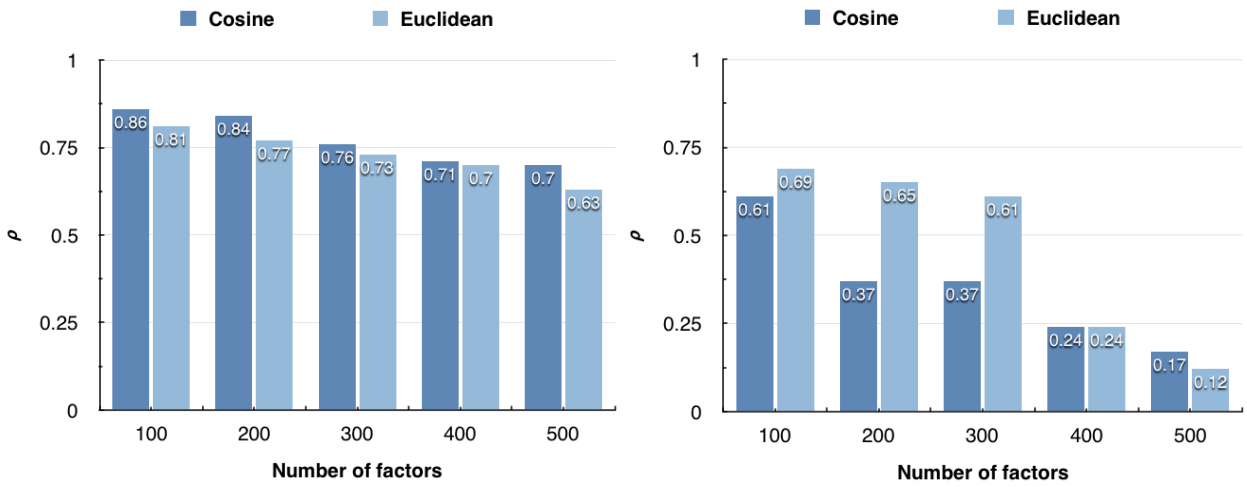
$$\epsilon\text{-}RMSE = \sqrt{\frac{1}{M-d}\sum_{c=1}^{M} \epsilon\text{-}RMSE(c)} \tag{4.8}$$

with d representing the degree of freedom.

Correlations are reported on a per-condition basis, where all files under the same acoustic condition are first averaged prior to correlation calculation. Table 3.1 describes the acoustic conditions present in the INRS dataset, excluding references and anchors. Thus, for the noisy samples, an acoustic condition is defined by the noise type, processing status, and SNR level, while for the reverberant samples, an acoustic condition is configured by the processing status and the reverberation time (T_{60}).

The wideband mode ITU-T Recommendation P.862.2 (PESQ) [136] and the fullscale mode ITU-T Recommendation P.863 (POLQA) [87] are used as benchmarks [2]. For the non-intrusive measures, the narrow-band ITU-T Recommendation P.563 [92] and wide-band SRMR [5] are used as benchmarks. Note that no standardized wideband non-intrusive measure is available. Therefore, the results reported herein for the ITU-T Recommendation P.563 are in somewhat of a disadvantage, considering that the proposed method considers wide-band signals. Performance comparisons with wide-band POLQA, PESQ and SRMR, nevertheless, provide a more fair comparison to gauge the benefits of the proposed method. All results for both proposed and benchmark algorithms are reported after a 3rd order monotonic mapping to map objective and subjective ratings into the MUSHRA scale.

2. Out-of-scope usage of POLQA as input and reference signals are wideband whereas reference signals are expected to be superwideband [87].



(a) Noise+enhanced

(b) Reverberation+enhanced

Figure 4.7 – Pearson correlation per-sample between the proposed systems and human rating (i.e., MUSHRA scores). Performance is given in respect to different numbers of total factors (MFCCs with no derivatives) for the cosine and euclidean similarities. Results for noise speech samples and their enhanced counterparts are presented in (a), whereas the results for reverberant speech samples and their enhanced counterparts are presented in (b).

4.6 Experimental results

In this section, experimental results are provided. First, we verify the impact of the number of total factors on the cosine and euclidean similarities. Next, we compare our solution to the benchmark measures for noise, enhanced, and reverberant speech conditions. Lastly, we investigate deep neural network models as non-intrusive speech quality assessment methods.

4.6.1 Impact of number of total factors on quality prediction

Here, we evaluate the performance of our method with respect to the dimension adopted for the i-vector representation. Total factors of size 100, 200, 300, 400 or 500 were explored along with the two similarity approaches. Figure 4.7-a presents the results in terms of Pearson correlation for noisy speech and their enhanced counterpart. All correlations were computed on the per-sample basis. We can observe that the cosine similarity, providing the highest correlation, 0.91, and being less affected by the number of total factors than the euclidean distance. Note that there is a clear decline in performance for both cosine and euclidean distances as the number of total factors is increased.

As the i-vector also contains speaker-dependent information, increasing the number of factors could also increase the amount of speaker-correlated information, thus reducing the strength of the proposed metric. This may explain the performance decay discussed above. In a study of speaker inter-variability, the authors in [137] found that the optimal number of factors for speaker recognition is around 300, while the number of channel factors is 50. In our experiments, 100 factors were found to be the optimal value for quality assessment as experiments performed with 50 provided lower correlations.

In Figure 4.7-b the results are reported in terms of Pearson correlation for reverberant speech and their enhanced counterpart. Correlations are now lower than the ones achieved with noisy data. We can observe from Figure 4.5 that the scores of the proposed method seem to be less sensitive to changes of T60. Another possible cause might be that during enhancement artifacts might be introduced by the enhancement algorithms.

As previously, the correlations for the cosine and euclidean similarities decay as the number of factors increases. For the distortions caused by reverberation, the euclidean distance metric outperformed the cosine similarity one across most scenarios presented in Figure 4.7-b. It was also less affected by the number of factors when compared to the experiments involving noisy speech. In the next sections, more detailed analyses are performed and compared to several benchmarks.

Table 4.2 – Performance comparison on per-condition basis for noisy and enhanced speech. Results are after a 3rd order monotonic polynomial mapping.

Metric	Noise+Enhanced			Noise			Enhanced		
	$\rho_{Pearson}$	$\rho_{Spearman}$	ϵ-$RMSE$	$\rho_{Pearson}$	$\rho_{Spearman}$	ϵ-$RMSE$	$\rho_{Pearson}$	$\rho_{Spearman}$	ϵ-$RMSE$
PESQ	0.97	0.79	6.24	0.99	**0.83**	5.98	**0.98**	0.85	**5.14**
POLQA	**0.99**	0.86	**5.64**	**0.99**	0.65	**4.27**	0.98	0.88	5.20
NCM	0.96	**0.92**	7.24	0.94	0.83	8.72	0.98	**0.91**	5.36
STOI	0.96	0.84	7.06	0.97	0.83	6.75	0.97	0.87	6.37
P563	0.54	0.55	27.92	0.66	0.53	39.09	0.55	0.57	30.12
SRMR	0.73	0.75	24.51	0.76	0.67	34.71	0.74	0.79	26.44
SRMR_{norm}	0.82	**0.77**	18.79	0.88	0.61	24.85	0.83	**0.79**	20.01
Cosine	0.96	0.69	7.70	0.98	**0.75**	7.73	0.97	0.76	6.87
Euclidean	0.95	0.67	8.70	0.98	0.62	5.88	0.96	0.73	8.40
Cosine_{deriv}	**0.96**	0.72	**7.06**	0.98	0.65	6.71	**0.98**	0.78	**6.03**
Euclidean_{deriv}	0.96	0.71	7.77	**0.98**	0.65	**6.05**	0.97	0.77	7.15

4.6.2 Experiment A: Noise and enhancement only conditions

In this experiment, we compare the performance of our proposed instrumental measure to the instrumental benchmark measures. Results are presented in Table 4.2 and are separated into noisy-plus-enhanced, noise-only, and enhanced-only conditions, thus allowing us to verify the effects that enhancement may have on the proposed measure. The first four lines on the top refers to the full-reference metrics while the seven ones on the bottom are the no-reference metrics. The best results are marked in bold.

As can be seen, the method that relies on i-vectors computed from MFCCs and their derivatives resulted in the largest correlations and lowest ϵ-$RMSE$. The same trend is found for noisy and enhanced samples, thus suggesting that spectral dynamics plays an important role in non-intrusive quality assessment. The proposed method outperforms the benchmark non-intrusive measures in terms of Pearson correlation and ϵ-$RMSEs$, achieving values inline with those achieved with intrusive measures. Moreover, Spearman correlation values similar to SRMR and SRMR$_{norm}$ are achieved, but lower compared to intrusive measures. It is important to emphasize that the proposed solution offers consistent results for both unprocessed and processed speech.

Table 4.3 – Performance comparison on per-condition basis for reverberant and enhanced speech. Results are after a 3rd order monotonic polynomial mapping.

	Reverberation+Enhanced			Reverberation			Enhanced		
Metric	$\rho_{Pearson}$	$\rho_{Spearman}$	ϵ-$RMSE$	$\rho_{Pearson}$	$\rho_{Spearman}$	ϵ-$RMSE$	$\rho_{Pearson}$	$\rho_{Spearman}$	ϵ-$RMSE$
PESQ	**0.92**	**0.44**	**9.04**	**0.96**	**0.92**	**12.16**	0.97	0.47	**6.09**
POLQA	0.92	0.39	9.17	0.96	0.83	12.14	0.97	0.41	6.34
NCM	0.92	0.41	9.39	0.95	0.85	12.93	**0.97**	**0.51**	6.12
STOI	0.86	0.27	12.04	0.88	0.72	16.18	0.95	0.36	8.47
P563	0.35	0.11	25.12	0.90	0.80	30.67	0.39	0.12	26.99
SRMR	0.80	**0.79**	22.72	0.49	0.22	33.27	0.49	0.11	12.43
SRMR$_{norm}$	0.49	0.11	22.15	0.90	0.82	30.82	0.54	0.16	23.46
Cosine	0.91	0.72	13.75	0.91	0.80	19.92	0.90	0.61	14.76
Euclidean	0.88	0.77	13.28	0.92	0.76	18.04	0.87	**0.71**	14.38
Cosine$_{deriv}$	0.86	0.57	14.77	0.94	**0.88**	19.68	0.86	0.39	15.59
Euclidean$_{deriv}$	**0.91**	0.60	**11.77**	**0.97**	0.85	**14.37**	**0.92**	0.52	**12.43**

4.6.3 Experiment B: Reverberation and enhancement only conditions

Here, only (de)reverberation as the main source of distortions is considered. Results are given in Table 4.3 and presented separately for reverberation-plus-enhanced, reverberation-only, and enhanced-only. As expected from the plots in Figure 4.7, speech quality assessment of reverberant

speech is a more challenging task as compared to noise conditions. In this case, addition of MFCC delta coefficients still provides some gains compared to the configurations where spectral dynamics are not considered. The proposed method outperformed all non-intrusive benchmarks, specially the approach based on euclidean distance. The SRMR_{norm} provided 0.90 correlation for the reverberation-only condition. This benchmark measure, however, showed to be severely affected by the DNN enhancement procedures, whereas the proposed method remained stable across all tested conditions. Overall, the measure provided results aligned with POLQA and PESQ, but without the need for a clean reference signal.

4.6.4 Experiment C: Non-intrusive assessment based on deep neural networks

Table 4.4 – Per-sample performance comparison with speech samples from the INRS database.

Metric	$\rho_{Pearson}$	$\rho_{Spearman}$	ϵ-$RMSE$
PESQ	0.92	0.62	9.45
POLQA	0.92	0.66	9.61
P563	0.32	0.25	23.23
SRMR	0.46	0.42	21.75
i-vector	**0.93**	**0.76**	**8.79**
Quality-net	0.91	0.72	9.77

In this experiment, we explore two non-intrusive deep neural network approaches presented in Section 4.5.3. The two models are trained considering both noisy and reverberant samples. 70% of the samples in the INRS database are randomly selected to train. The remainder 15% of the data is used for validation and the other 15% are used for testing. Results are shown in Table 4.4. We present two intrusive benchmarks, PESQ and POLQA, and two non-intrusive ones, P563 and SRMR. The DNN-based instrumental measures are the two on the bottom of Table 4.4. As can be seen, among the non-intrusive metrics the two DNN models achieved the highest performance with the i-vector solution presenting slightly better performance.

It is important to notice that the DNN based methods rely on a subset of the INRS quality database for training, which certainly leverage their performance. The benchmarks, on the other hand, were attained with no prior information regarding the quality dataset. Notwithstanding, the attained results are promising, as they show that if labeled data is available, DNN-based models can ultimately derive objective quality assessment tools that are accurate and adaptive as they may be quickly re-trained.

4.6.5 Study limitations

The use of ITU-T Recommendation P.835 is advised when assessing the quality of noise suppression algorithms [138]. This methodology adopts separate rating scales to independently estimate the subjective quality of the speech signal alone, the background noise alone, and the overall quality [138], thus helping reduce any subjective biases with noise suppression artifacts. The dataset used herein, in turn, was collected using the MUSHRA methodology, thus may have some biases. Notwithstanding, despite not being indicated by ITU-T Recommendation P.835 as a listening assessment methodology for noise suppressed speech, several recent works have relied on it (e.g., [139, 116, 140]), thus suggesting its usefulness.

4.7 Conclusions

In this Chapter, we propose the i-vector framework for non-intrusive speech quality measurement. The proposed method relies on a Gaussian mixture model (GMM) to estimate a clean reference spectrum from the degraded speech signal. i-vector representations are then computed for both the estimated clean and the degraded spectra and quality correlates are obtained by means of two different scoring methods. Experimental results showed the proposed method outperforming several non-intrusive benchmark algorithms, and achieving accuracy aligned with intrusive algorithms, without the need for a clean reference signal. More importantly, the proposed method showed stable accuracy across degraded and enhanced conditions, thus suggesting better applicability to emerging conditions.

Part II

Reliability

Feature Pooling for Improved Speech Emotion Recognition in the wild

5.1 Preamble

This chapter is compiled from material extracted from the manuscript published in the IEEE Transactions on Affective Computing [J6], also from material that appeared in the Proceedings of the 17^{th} IEEE International Symposium on Signal Processing and Information Technology (ISSPIT 2017) [C8], Interspeech 2018 [C6], and the 27^{th} European Signal Processing Conference (EUSIPCO 2019) [C3].

5.2 Introduction

New machine-learning algorithms based on deep learning techniques have enabled computers to solve intricate problems by assimilating information directly from raw data in a similar manner as humans do [50]. In spite of breakthroughs in many important areas such as automatic speech recognition (ASR) [51] and object recognition [52], tasks involving automated comprehension of human emotional states still remain unsolved. This has become an increasing concern for the artificial intelligence (AI) community as, in the near future, intelligent machines will be expected to possess some level of human emotion understanding in order to assure that safe decisions, involving cognitive

related tasks, will be made [53]. To mitigate such limitation, affective computing aims at enabling machines to recognize, analyze and synthesize human affective states [54].

In fact, interest in affective computing is burgeoning, in great part due to its role in emerging affective human-computer interfaces (HCI). To date, the majority of existing research on automated emotion analysis has relied on data collected in controlled environments. With the rise of HCI applications on mobile devices, however, so-called "in the wild" settings have posed a serious threat for emotion recognition systems, particularly those based on voice. In this case, environmental factors such as ambient noise and reverberation severely hamper system performance. In this chapter, we quantify the detrimental effects that the environment has on emotion recognition and explore the benefits achievable with speech enhancement. Moreover, we propose a modulation spectral feature pooling scheme that is shown to outperform a state-of-the-art benchmark system for environment-robust prediction of spontaneous arousal and valence emotional primitives. Two DNNs are also explored, namely a multi-layer perceptron (MLP) network and the recurrent neural network based on Long-short Term Memory (LSTM). Their performance are compared to the SVM-based benchmark to quantify the advantages of one machine learning system over another for in the wild SER. Experiments on an environment-corrupted version of the RECOLA dataset of spontaneous interactions show the proposed feature pooling scheme, combined with speech enhancement, outperforming the benchmark across different noise-only, reverberation-only and noise-plus-reverberation conditions. Additional tests with the SEWA database show the benefits of the proposed method for in the wild applications.

The remainder of this chapter is organized as follows. Section 5.3 presents the modulation spectral representation and gives details on the extraction of the proposed features and pooling schemes. Section 5.4 describes the experimental setup and Section 5.5 reports the obtained results. Lastly, conclusions are drawn in Section 5.6.

5.3 Modulation Spectral Features for Robust SER

In this section, we describe six modulation spectral features extracted from the modulation spectral representation discussed in Section 2.4 and we present feature pooling strategy proposed to boost SER performance in the wild.

5.3.1 Modulation spectral features

Previously, six features extracted from the modulation spectral representation discussed in Section 2.4.1 were used for SER [97]. These features were shown to be useful for enacted emotions and were not tested in the wild. Here, we show that such features are indeed affected by spontaneous emotions collected in realistic settings. As such, we propose two new feature pooling strategies to enhance SER performance. For completeness, the six original features, computed on a per-frame basis, are described below.

The first feature, termed $\psi_{1,m}(k)$, represents the mean of the energy samples with respect to the k-th modulation channel and represent the energy distribution of the modulation frequency (see Section 2.4.1), namely:

$$\psi_{1,m}(k) = \frac{\sum_{j=1}^{N} \mathcal{E}_{j,k}(m)}{N}. \tag{5.1}$$

The second feature, $\psi_{2,m}(k)$, is the ratio of the geometric mean of the spectral energy of the k-th modulation channel to its arithmetic mean. This measure represents the spectral flatness, where a spectral value close to 1 is associated with a flat spectrum while a spectral value close to 0 indicates a great variation in terms of spectrum amplitude. This measure is defined as:

$$\psi_{2,m}(k) = \frac{\sqrt[N]{\prod_{j=1}^{N} \mathcal{E}_{j,k}(m)}}{\psi_{1,m}(k)}. \tag{5.2}$$

The third feature, $\psi_{3,m}(k)$, captures the center of mass of each modulation channel, where j represents the index of the critical band. The spectral centroid across the modulation dimension is computed as follows:

$$\psi_{3,m}(k) = \frac{\sum_{j=1}^{N} j\mathcal{E}_{j,k}(m)}{\sum_{j=1}^{N} \mathcal{E}_{j,k}(m)}. \tag{5.3}$$

The last three features correspond to spectral measures and the relationship between adjacent acoustic channels. As such, the 23 acoustic channels are grouped into five groups, namely: $G_1 = \{1-4\}; G_2 = \{5-8\}; G_3 = \{9-12\}; G_4 = \{13-18\}; G_5 = \{19-23\}$. Modulation information from

channels within the same group are summed together as $E_m(l,k) = \sum_{i \in G_l} \mathcal{E}_{i,k}(m), l = 1, \ldots, 5$. As such, the fourth feature, $\psi_{4,m}(l)$, computes the spectral centroid across acoustic frequency group "l" (as opposed to $\psi_{3,m}(k)$, which measures it across the modulation dimension). The feature is computed similarly to (5.3) but with the overall per-group energy, i.e.,

$$\psi_{4,m}(l) = \frac{\sum_{k=1}^{8} kE_m(l,k)}{\sum_{k=1}^{8} E_m(l,k)}. \tag{5.4}$$

The last two spectral measures $\psi_{5,m}(l)$ and $\psi_{6,m}(l), l = 1, \ldots, 5$ are, respectively, the linear regression coefficient (slope) and the regression error (root mean squared error, RMSE) associated with the first-degree polynomial model used to fit $E_m(l)$ across all eight modulation channels. This captures the rate of change of each acoustic frequency group, and thus provides some indication of temporal dynamics [97].

5.3.2 Proposed feature pooling scheme

When measuring emotional primitives continuously, no consensus has been achieved on what is the best temporal window size for automatic affect analysis [141]. Studies have shown that emotional cues can last between $0.5 - 4$ s, depending on the modality and/or emotional primitive. For SER, appropriate window sizes in the range $2 - 6$ s have been reported [142]. The modulation spectral representation described in 2.4, however, as well as the six basic features from [97], described above, rely on information extracted over 256-ms windows. While such an approach was shown to be reliable for enacted emotions, for realistic spontaneous SER that may not be the case. As such, here we explore feature pooling as a method of capturing relevant SER information present in adjacent speech frames. Different window sizes in the range $1 - 6$ s were explored and eight statistical functionals (i.e., mean, standard deviation, variance, kurtosis, skewness, range, minimum, and maximum) are computed per window. Moreover, in allied domains, pooling of modulation spectral features has also shown to improve feature "signal-to-noise" ratio [143]. As such, it is expected that feature pooling will not only improve SER performance for spontaneous speech, but also for SER in realistic (noisy) scenarios.

We assume a dataset with N observations $\mathrm{x}_1, ..., \mathrm{x}_N$ of a D-dimensional vector $\mathrm{x} = (x_1, ..., x_D)$. The pooling operation is then performed on a subset of n adjacent observations, $\mathrm{x}_1, ..., \mathrm{x}_n$, for each dimension, providing the respective pooling feature vector. The procedure is then repeated on the next subset (i.e., $\mathrm{x}_{1+timestep}, ..., \mathrm{x}_{n+timestep}$) [1]. This is defined according to the number of labels in the dataset, in a way that the feature pooling vectors synchronize with the label's frame rate. Figure 5.1 (a)-(d) depicts the four system configurations investigated herein. The first two approaches (subplots (a) and (b)) show benchmark systems that rely on per-frame features (i.e., without feature pooling) that have been vectorized. The former relies on 184 modulation spectrum energy features ($\mathcal{E}_{j,k}, j = 1, \ldots, 23; k = 1, \ldots, 8$), while the latter on these 184 features, plus the 39 features described above ($\psi_1(k)$, $\psi_2(k)$ and $\psi_3(k)$ for $k = 1 \ldots, 8$ and $\psi_4(l)$, $\psi_5(l)$, and $\psi_6(l)$ for $l = 1, \ldots, 5$), thus totalling 223 features. The last two approaches (subplots (c) and (d)) correspond to these same systems, but with feature pooling, respectively. These are henceforth referred to as *feature pooling scheme 1* and *feature pooling scheme 2*, respectively. For fair comparison, principal component analysis (PCA) is used to reduce the feature dimension to 60 prior to inputting to the regression system. This value was decided empirically based on performance on a pilot step and no improvement was found when a higher number of eigenvectors were used. Here, three regressor types are explored; one based on support vector regression and the other two on deep neural network architectures.

5.4 Experimental Setup

This section describes the databases, the DNN-based architectures, figures of merit and the benchmark SER system used in our experiments.

5.4.1 Database Description

In our experiments, we used the anechoic speech files from the REmote COllaborative and Affective interactions (RECOLA) database [144]. This dataset was also used during the recent 2016 Audio/Visual Emotion Challenge (AVEC) [38]. Speech signals were synchronously recorded from 27 French-speaking subjects, 16 females and 11 males, from 3 different nationalities (French,

1. $(n + timestep)$ should be less than $(N + timestep/2)$

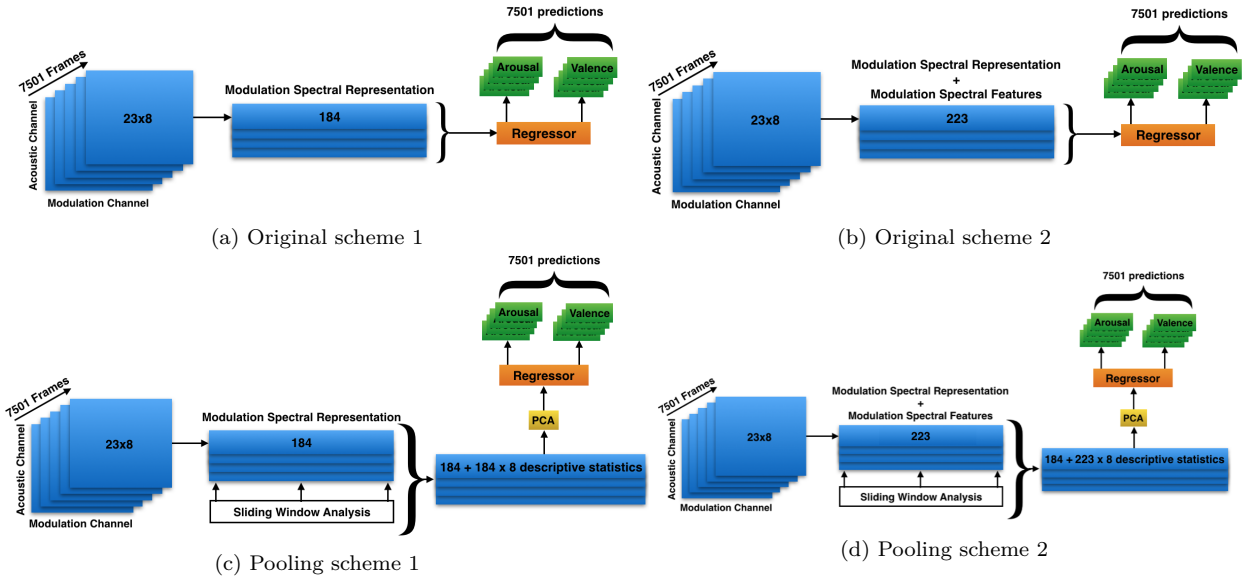


Figure 5.1 – Four configurations used for emotion primitive prediction. Features correspond to: (a) vectorized 184-dimensional $\mathcal{E}_{j,k}, j = 1-23; k = 1-8$ without feature pooling (termed original scheme 1); (b) 184 + 39 features ($\psi_1(k)$, $\psi_2(k)$ and $\psi_3(k)$ for $k = 1-8$ and $\psi_4(l)$, $\psi_5(l)$, and $\psi_6(l)$ for $l = 1-5$) without feature pooling (original scheme 2); (c) same as (a) but with feature pooling (termed pooling scheme 1); and (d) same as (b), but with feature pooling (pooling scheme 2).

Italian and German). Based on spontaneous interactions collected from a collaborative task, six annotators measured emotion continuously using a time-continuous scale for two emotion primitives, namely arousal and valence. For the AVEC Challenge, the dataset was partitioned into training, development and testing subsets, all balanced based on gender and mother tongue. Each subset of the RECOLA database contains 9 speech files of 5 minutes. Here, we use the training and development subsets to train and test the proposed schemes, as access to the testing set labels is not available to non-challenge participants. The annotated data was binned with a frame rate of 40 ms. The ground truth was computed as the mean value of the annotations at every timestep.

In order to simulate in the wild speech, clean signals were corrupted with noise samples from the AURORA and DEMAND databases [121][145]. These databases were designed to evaluate speech recognition systems and offer several real world noise settings. Three noise types were considered: (1) noise multi-talker babble, (2) noise recorded inside a commercial airplane [121] and (3) recorded at an university cafeteria [145]. The noise signals were added at five signal-to-noise ratios (SNRs), namely 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB. Next, in order to simulate SER in reverberant environments, three recorded room impulse responses with reverberation times of $T_{60} = 0.25$ s, 0.48 s, and 0.8 s [146] were convolved with the anechoic signals, thus simulating speech recorded in small-, medium-, and large-sized rooms, respectively. Lastly, to simulate a noise-plus-reverberation condition typically

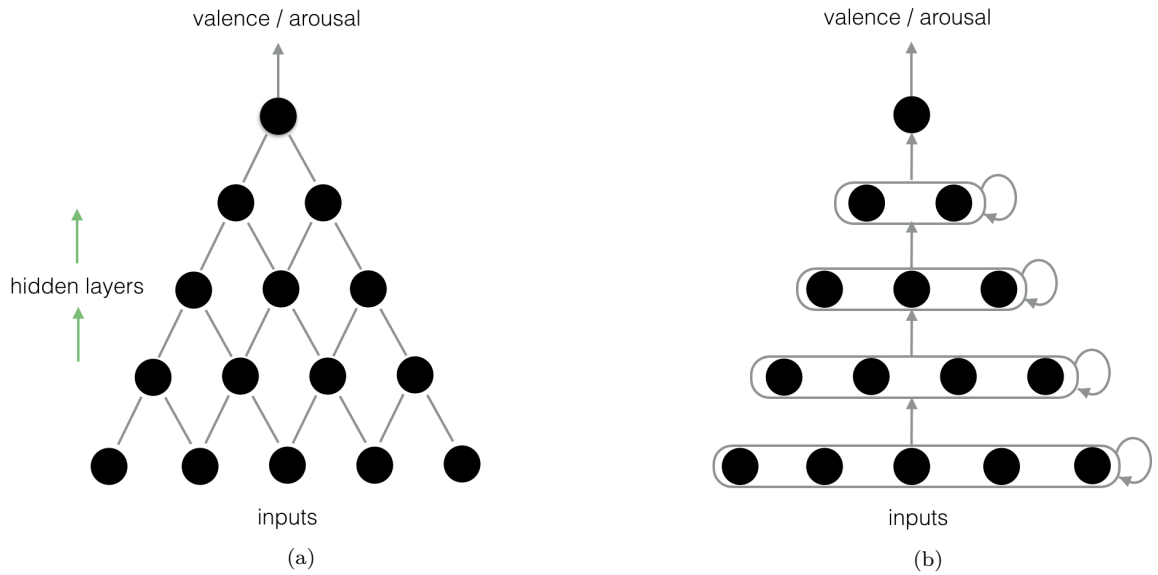


Figure 5.2 – Illustration of the DNN-based architectures adopted in our experiments. In (a) 3 hidden layers with 64, 32 and 16 hidden units for arousal and 32, 16 and 8 for valence and in (b) 3 hidden layers with 64, 32 and 16 hidden units for arousal and 2 hidden layers for valence with 32, 16 hidden units for valence. Input is a 60-dimensional feature vector for all models.

observed in the wild, the three noise types were added to the different reverberant signals at the same SNRs listed above.

An additional dataset is included in our experiments in order to validate our results. The corpus was the same used for the AVEC Challenge 2017 [39], and is a subset of the Sentiment Analysis in the Wild (SEWA) database. The recordings in this dataset were collected 'in the wild', with microphones from the computers used to capture speech in offices or homes. Each recording sample contains approximately 90 seconds of dyadic conversation about a commercial they had watched. There are 32 pairs present in the SEWA subset, thus 64 subjects in total.

5.4.2 Deep Neural Network Models

Figure 5.2 depicts the two architectures used in our experiments (MLP, subplot a; RNN-LSTM, subplot b). Nonlinearities were introduced in the hidden layers by using the activation function tanh() with output unit linear for both models. After exhaustive experiments, these functions showed to be the best option for each model. The numbers of layers and units were also defined empirically after running a considerable number of pilot experiments. The dropout technique was used in the intermediate layers. The technique is a powerful regularization method that significantly reduces

the problem of overfitting by randomly removing a predetermined percentage of units creating sub-networks from an underlying base network [50][147]. The adopted retention of 0.5, which is considered an optimal value for a large range of networks [147], means that each hidden unit has probability of 0.5 to be omitted from the network. In a sense, the technique can be seen as a way to average the predictions of the sub-networks attained [148]. Moreover, by using an RNN, we aim at predicting arousal and valence by capturing the ties found in the 60-dimensional input sequence, as described in Section II. For both models, the loss function was optimized by using stochastic gradient descent (SGD).

5.4.3 Benchmark SER system and figure-of-merit

Since we are using the speech corpus from AVEC, it is desirable to compare the performance of the proposed solution with that obtained from the AVEC benchmark algorithm. The benchmark SER system relies on acoustic low-level descriptor features, which cover spectral, cepstral, prosodic and voice quality information. The features are based on the so-called Geneva Minimalistic Acoustic Parameter Set (GeMAPS) extracted from the open-source OpenSMILE toolkit [149]. A detailed description of all the benchmark features is beyond the scope of this work and the interested reader is referred to [38] and references therein for more details. In order to estimate valence and arousal, the benchmark system relies on a support vector regressor (SVR) [150] with parameters described in [38]. For comparison purposes, an SVR is also used for the proposed features. As in the AVEC, the concordance correlation coefficient (CCC) is used as a figure-of-merit. The method computes the correlation between two variables while also considering reproducibility [151]. As such, it combines Pearson's correlation coefficient (ρ) and the square difference between the mean of the two samples x and y, via:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{5.5}$$

where σ and μ represent the sample standard deviation and mean, respectively.

5.4.4 Test Setup

In order to compare the benchmark and proposed features, the learning process is divided into two phases: training and testing. Mismatch conditions were characterized by using only clean speech

for training and corrupted data was introduced only during the testing phase. No compensation technique was considered in our experiments, as proposed for example in [152]. After training the SVR regressor on the training set, the attained model with its respective parameters are validated on the development set. This is the same approach used for the AVEC Challenge 2016 [38]. These steps are performed several times until the optimal model is obtained. In the testing phase, a total of 9 speech samples per condition were used to test our model. The training procedure was slightly different for the DNN models. In order to avoid overfitting, the development set was used only for testing the models and the training set was segmented with 80% used for training and 20% for evaluation.

5.5 Experimental Results and Discussion

In this section, we provide experimental results under clean, noise-only, reverberation-only, and noise-plus-reverberation conditions. Results involving speech enhancement as a pre-processing step are also presented and discussed.

5.5.1 Experiment I: Clean Speech

In this experiment, we investigate the effects of different pooling window lengths on predicting valence and arousal. Clean speech samples are used during the training and evaluation phases. Figure 5.3 depicts the obtained CCC for arousal and valence with the setup described in Figure 5.1 (a) (i.e., *Original scheme 1*), as well as the proposed setup shown in Figure 5.1 (c) (i.e., *Pooling scheme 1*). Results show that all window sizes being tested improved the predictions of both primitives, demonstrating the effectiveness of the proposed method. The highest performance was achieved by applying feature pooling with a window size of 2 s on *Original scheme 1*. Arousal predictions improved 70%, going from CCC=0.400 to CCC=0.682, while valence had a significant improvement going from CCC=0.080 to CCC=0.395. These findings reveal that the MSR can be useful for predicting emotional primitives once information from adjacent frames is taken into account.

Figure 5.4, in turn, depicts the results attained with the configuration illustrated in Figure 5.1 (b) *Original scheme 2*) as well as the improvements attained with the setup of Figure 5.1 (d) (*Pooling*

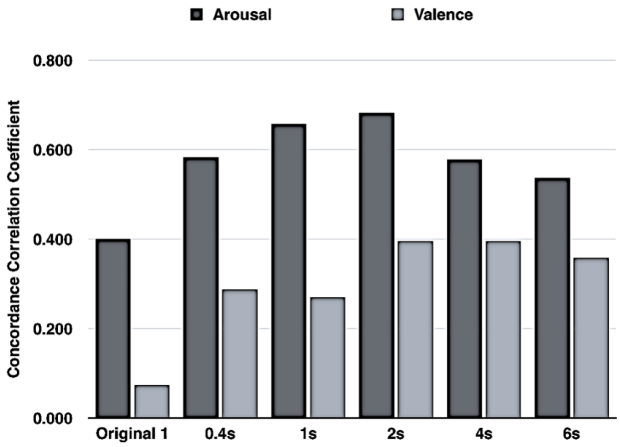


Figure 5.3 – Results comparing the *Original scheme 1* and *Pooling scheme 1* under different sliding window analysis length

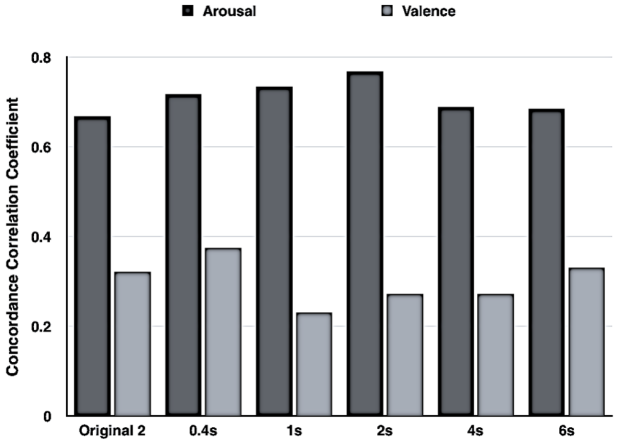


Figure 5.4 – Results comparing the *Original scheme 2* and *Pooling scheme 2* under different sliding window analysis length

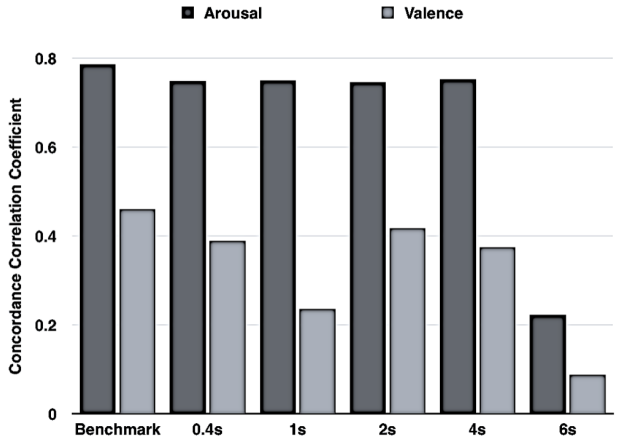


Figure 5.5 – Results of applying *Pooling scheme 2* on the benchmark features under different sliding window analysis length

scheme 2). We can observe a gradual improvement on arousal predictions as the proposed pooling window size increases from 0.4 to 2 s. The best performance for arousal is achieved with window size of 2 s, resulting in a CCC=0.768, thus a 15% improvement relative to the *Original scheme 2*. For valence, the highest result is attained with window size of 0.4 s, which represents a 16% improvement (from CCC=0.322 to 0.375) when compared to the *Original scheme 2*.

Lastly, results in Figure 5.5 explore the effects of feature pooling, but for the benchmark features. Unlike MSR features, pooling did not bring benefits for the benchmark features in the clean conditions. For arousal, feature pooling resulted in a slight decay of 4%, whereas for valence, a 10% decrease in performance was seen. As will be shown in the next sections, however, feature pooling of benchmark features showed to be more beneficial under noisy speech conditions.

Table 5.1 summarizes the best CCC values achieved for all four configurations shown in Figure 5.1, as well as the values provided by the benchmark features. As can be seen, feature pooling was shown to be extremely important to deal with spontaneous speech and outperformed the per-frame features, particularly for valence estimation. Such findings suggest that measurement of spontaneous valence requires information from neighbouring frames. Comparing the *Original schemes 1* and *2*, we observe that the six spectral features (MSF's) extracted from the MSR had a significant impact on arousal and valence predictions. These features have been applied in [97] and although they were successful in predicting discrete emotions, the performance reported by the authors for continuous emotion recognition was limited, especially for valence. In our experiments, the best performance for this particular dimension is achieved with *Pooling scheme 1* while *Pooling scheme 2* showed to be better for arousal. Within the pooling scheme 1, the highest CCC attained for arousal (CCC=0.761) was seen with a window length of 2 s, whereas the highest CCC attained for valence (0.395) was achieved with 4 s.

Moreover, despite the small amount of data available, the performance of two DNN algorithms was also investigated for the task at hand. The LSTM and MLP were not trained in an end-to-end method, and thus relied on features resultant from *Pooling scheme 1*, for valence, and *Pooling scheme 2*, for arousal. The criteria to choose each scheme was based on their performance towards each one of the emotional primitives. The main goal was to investigate if deep neural architectures would work as better discriminators for the proposed features when compared to SVR. The results were especially interesting for arousal. Both DNN architectures outperformed the benchmark regression

Table 5.1 – Results for affective recognition of clean speech in terms of concordance correlation coefficients

Configuration	Arousal	Length	Valence	Length
Benchmark	0.786	-	**0.461**	-
Benchmark (Pooling)	0.753	4 s	0.418	2 s
Original scheme 1	0.400	-	0.080	-
Original scheme 2	0.669	-	0.322	-
Pooling scheme 1	0.682	2 s	0.395	2, 4 s
Pooling scheme 2	0.768	2 s	0.375	0.4 s
LSTM	**0.795**	-	0.265	-
MLP	0.769	-	0.348	-

method. LSTM outperformed also the benchmark system (features + regression method), providing CCC=0.795 for arousal.

5.5.2 Experiment II: Noise-only conditions

Table 5.2 – Performance, in terms of CCC, of benchmark features and the proposed system under different noise levels.

		0dB		5dB		10dB		15dB		20dB	
Noise type	Features (Model)	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence
Airport	Benchmark (SVR)	0.222	0.048	0.326	0.047	0.455	0.033	0.542	0.096	0.594	0.151
	Proposed (SVR)	**0.396**	0.175	**0.495**	**0.226**	0.592	0.235	0.621	0.235	0.737	0.222
	Proposed (MLP)	0.315	**0.193**	0.366	0.225	0.586	**0.238**	0.672	**0.257**	0.746	**0.248**
	Proposed (LSTM)	0.359	0.119	0.406	0.152	**0.628**	0.166	**0.712**	0.172	**0.763**	0.174
Babble	Benchmark (SVR)	0.190	0.095	0.309	0.018	0.453	0.061	0.559	0.105	0.645	0.148
	Proposed (SVR)	0.295	**0.207**	**0.513**	**0.176**	0.536	**0.224**	0.693	0.176	**0.713**	0.225
	Proposed (MLP)	0.198	0.189	0.336	0.133	0.472	0.221	0.672	**0.221**	0.733	**0.258**
	Proposed (LSTM)	**0.298**	0.146	0.434	0.122	**0.626**	0.172	**0.735**	0.161	**0.749**	0.179
Cafeteria	Benchmark (SVR)	0.349	0.031	0.445	0.043	0.544	0.010	0.616	0.145	0.669	0.195
	Proposed (SVR)	0.375	0.126	0.497	0.162	0.514	**0.176**	0.581	0.195	0.622	**0.242**
	Proposed (MLP)	0.314	0.105	0.550	0.172	0.474	0.123	0.643	**0.219**	0.723	0.233
	Proposed (LSTM)	**0.383**	**0.140**	**0.603**	**0.185**	**0.592**	0.124	**0.724**	0.170	**0.756**	0.179

Table 5.2 reports the performance achieved with the benchmark and proposed feature pooling scheme when only noise is present at varying levels. As mentioned previously, *Pooling scheme 2* is used for predicting arousal while *Pooling scheme 1* is used for valence. The highest CCC values are highlighted in bold in the Table. It can be seen that the prediction of valence using the benchmark system is severely affected by the presence of noise, being outperformed by all three proposed approaches under all tested conditions. Although MLP combined with feature pooling provided the best overall performance for predicting valence, SVR seems to be more robust towards more severe SNRs (0 and 5 dB), as can be seen for the airport and babble noise types. For arousal, the proposed systems outperformed the benchmark in all conditions. LSTM combined with the proposed feature

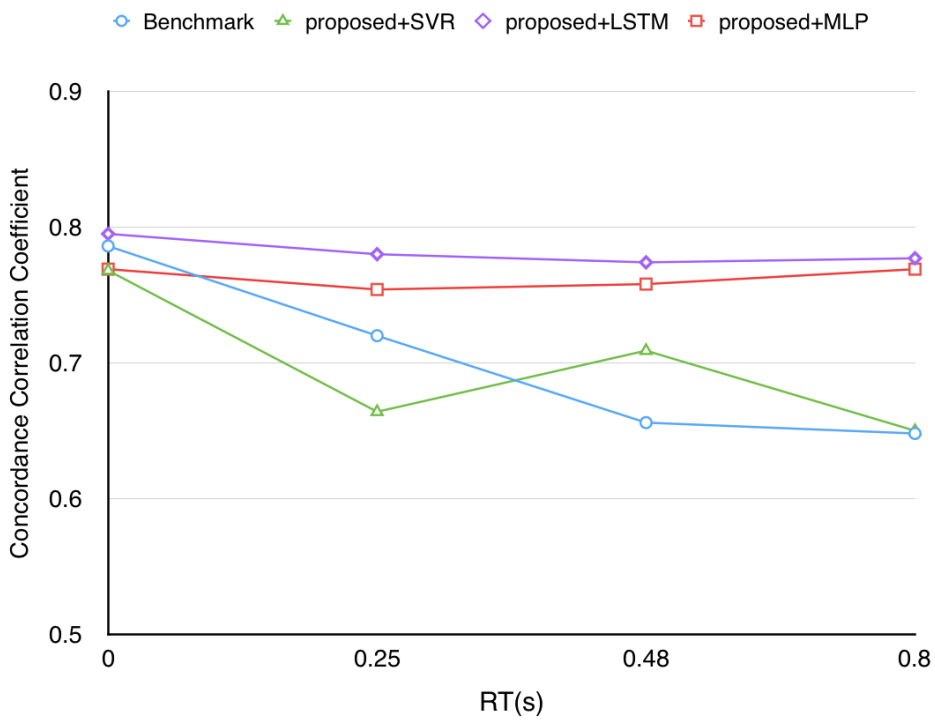


Figure 5.6 – Performance comparison between proposed and benchmark systems as a function of RT for arousal

pooling scheme presented the best performance overall, being outperformed only by the SVR in three conditions. From the last column, we observe that overall LSTM offered the best results for predicting arousal while MLP was the best system for predicting valence. Overall, an improvement of up to 22% and 56% could be observed with the proposed-LSTM system and the proposed-MLP system over the benchmark for arousal and valence estimation, respectively.

5.5.3 Experiment III: Reverberation-only conditions

Reverberation has also shown to be a detrimental factor for in the wild SER [153]. To explore the effects of room reverberation on the benchmark and proposed systems, Figures 5.6 and 5.7 show the CCCs attained for arousal and valence, respectively, as a function of RT. As can be seen, both the benchmark and the proposed feature pooling scheme combined with SVR are affected by reverberation, particularly at higher reverberation levels. The two DNN architectures, on the other hand, combined with the proposed features and pooling scheme showed to be independent of reverberation levels; LSTM achieved a slightly higher performance than the MLP. For valence, on the contrary, the benchmark system outperformed all proposed methods, including those involving

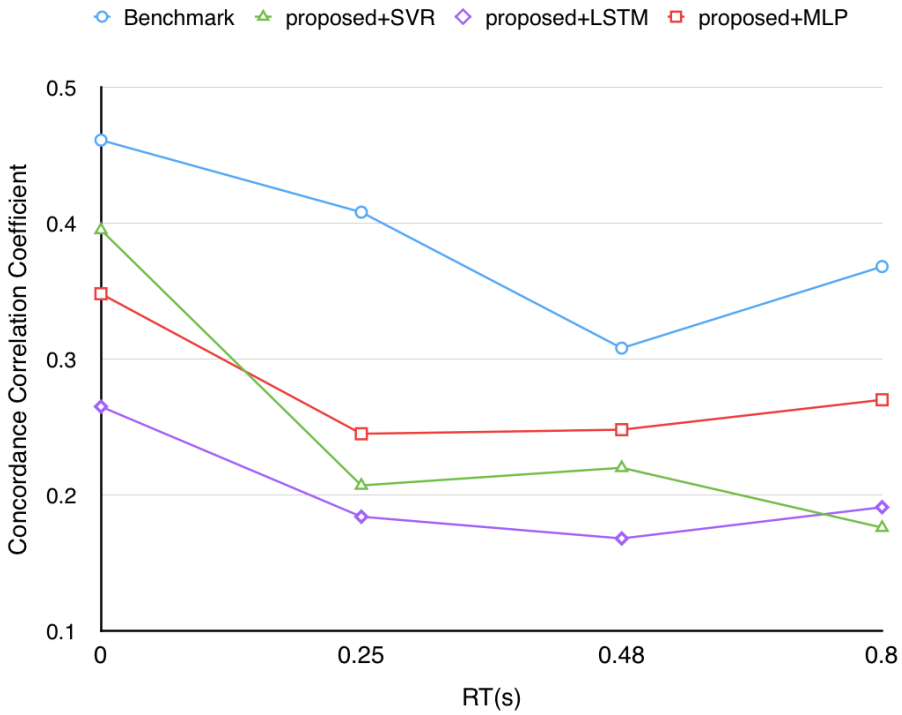


Figure 5.7 – Performance comparison between proposed and benchmark systems as a function of RT for valence

DNNs. Overall, an improvement of up to 13% could be observed for arousal estimation with the proposed-LSTM system. For valence estimation, however, all proposed schemes were outperformed by the benchmark.

5.5.4 Experiment IV: Reverberation-plus-noise conditions

In realistic settings, speech is often corrupted by both reverberation and ambient noise. In this experiment involving both distortion types, it can be seen from Table 5.3 that both the benchmark and proposed systems are severely affected by noise and reverberation. Notwithstanding, the proposed system outperformed the benchmark across the majority of the tested conditions for both arousal and valence dimensions. Overall, an improvement of up to 51% and 50% could be observed with the proposed-LSTM system and the proposed-MLP system over the benchmark for arousal and valence estimation, respectively.

Table 5.3 – Performance comparison of the benchmark and the proposed system under varying reverberation-plus-noise conditions

			0dB		5dB		10dB		15dB		20dB	
RT (s)	Noise type	Features (Model)	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence
0.25	Airport	Benchmark (SVR)	0.218	0.066	0.263	0.035	0.302	0.053	0.321	0.076	0.414	0.151
		Proposed (SVR)	**0.394**	**0.167**	**0.487**	**0.191**	0.588	**0.251**	0.617	**0.274**	0.738	**0.239**
		Proposed (MLP)	0.312	0.166	0.383	0.138	0.629	0.232	0.701	0.270	0.748	**0.257**
		Proposed (LSTM)	0.355	0.121	0.434	0.111	**0.657**	0.172	**0.719**	0.195	**0.761**	0.194
	Babble	Benchmark (SVR)	0.070	0.069	0.165	0.028	0.286	0.022	0.438	0.123	0.487	0.190
		Proposed (SVR)	0.440	0.154	0.444	**0.192**	0.638	**0.211**	0.702	0.234	0.723	0.216
		Proposed (MLP)	0.408	**0.194**	0.451	0.180	0.593	0.195	0.728	**0.241**	0.746	**0.230**
		Proposed (LSTM)	**0.460**	0.170	**0.571**	0.149	**0.691**	0.162	**0.748**	0.187	**0.751**	0.185
	Cafeteria	Benchmark (SVR)	0.383	0.029	0.443	0.024	0.514	0.107	0.526	0.155	0.584	0.215
		Proposed (SVR)	0.397	0.168	0.498	0.148	0.537	0.184	0.588	0.171	0.622	0.190
		Proposed (MLP)	**0.417**	**0.171**	**0.552**	**0.180**	0.580	**0.198**	0.697	**0.212**	0.731	**0.238**
		Proposed (LSTM)	0.391	0.166	0.549	0.201	**0.675**	0.183	**0.726**	0.185	**0.758**	0.184
0.48	Airport	Benchmark (SVR)	0.105	0.123	0.086	0.103	0.151	0.036	0.259	0.051	0.318	0.203
		Proposed (SVR)	**0.498**	0.131	0.474	**0.232**	0.559	0.216	0.696	**0.260**	0.720	0.246
		Proposed (MLP)	0.389	**0.215**	0.389	0.201	0.565	**0.219**	0.708	0.256	0.733	**0.263**
		Proposed (LSTM)	0.437	0.145	**0.478**	0.137	**0.583**	0.149	**0.723**	0.177	**0.754**	0.178
	Babble	Benchmark (SVR)	0.041	0.125	0.120	0.039	0.188	0.089	0.260	0.117	0.388	0.188
		Proposed (SVR)	0.392	0.207	0.399	**0.266**	0.585	0.172	0.665	0.227	0.710	0.218
		Proposed (MLP)	0.363	**0.215**	0.341	0.201	0.543	**0.174**	0.658	**0.244**	0.729	**0.241**
		Proposed (LSTM)	**0.451**	0.169	**0.515**	0.129	**0.641**	0.129	**0.725**	0.156	**0.754**	0.169
	Cafeteria	Benchmark (SVR)	0.212	0.044	0.299	0.016	0.430	0.075	0.427	0.143	0.505	**0.245**
		Proposed (SVR)	0.329	**0.161**	0.433	0.158	0.483	**0.177**	0.540	0.164	0.635	0.190
		Proposed (MLP)	0.283	0.152	0.445	**0.198**	0.489	0.160	0.636	**0.196**	0.703	0.244
		Proposed (LSTM)	**0.357**	0.150	**0.534**	0.196	**0.639**	0.130	**0.674**	0.164	**0.729**	0.166
0.8	Airport	Benchmark (SVR)	0.037	0.016	0.164	0.026	0.178	0.051	0.219	0.009	0.232	0.180
		Proposed (SVR)	**0.359**	0.105	**0.459**	0.172	0.541	0.188	0.608	0.202	0.686	0.190
		Proposed (MLP)	0.318	**0.172**	0.450	**0.198**	0.620	**0.280**	**0.736**	**0.291**	0.730	**0.284**
		Proposed (LSTM)	0.351	0.115	0.457	0.139	**0.633**	0.176	**0.736**	0.200	**0.743**	0.198
	Babble	Benchmark (SVR)	0.066	0.125	0.051	0.038	0.159	0.005	0.249	0.164	0.313	0.178
		Proposed (SVR)	0.277	0.117	0.448	0.144	0.557	0.149	0.656	0.146	0.686	0.159
		Proposed (MLP)	0.289	**0.200**	0.446	**0.180**	0.563	**0.217**	0.728	**0.248**	**0.750**	**0.250**
		Proposed (LSTM)	**0.334**	0.169	**0.514**	0.135	**0.648**	0.166	**0.748**	0.179	**0.750**	0.190
	Cafeteria	Benchmark (SVR)	0.322	0.025	0.395	0.102	0.527	0.143	0.555	0.157	0.581	**0.300**
		Proposed (SVR)	0.311	**0.150**	0.435	0.107	0.504	0.137	0.523	0.114	0.573	0.146
		Proposed (MLP)	0.295	0.112	0.529	**0.193**	0.582	**0.251**	0.672	**0.226**	0.713	0.254
		Proposed (LSTM)	**0.372**	0.140	**0.570**	0.182	**0.642**	0.221	**0.683**	0.190	**0.714**	0.197

5.5.5 Experiment V: Enhanced Speech

For applications in noisy scenarios, it is typical for speech enhancement algorithms to be applied prior to SER [154]. For the experiments herein, the relative convolutive transfer function (RCTF) algorithm was applied to reduce residual reverberation and noise on corrupted emotional speech. Details about the enhancement algorithm are beyond the scope of this work and the interested reader is referred to [155] for more details. As speech enhancement algorithms are fine-tuned to improve perceived quality and intelligibility, it is still unknown what effects enhancement may have on SER performance. Here, we explore the use of the RCTF algorithm prior to benchmark and proposed feature extraction. For brevity, only one noise-only setting is tested (airport noise), one

Table 5.4 – Performance, in terms of CCC, of benchmark features and the proposed system, for predicting valence and arousal after applying RCTF-based speech enhancement to noise-only data.

		0dB		5dB		10dB		15dB		20dB	
Noise type	Features (Model)	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence
Airport	Benchmark (SVR)	0.222	0.048	0.326	0.047	0.455	0.033	0.542	**0.096**	0.594	0.151
	Benchmark (SVR) - Enhanced	**0.423**	**0.130**	**0.539**	**0.084**	**0.651**	**0.053**	**0.667**	0.048	**0.662**	**0.145**
	Proposed (SVR)	0.396	0.175	0.495	0.226	0.592	0.235	0.621	0.235	**0.737**	0.222
	Proposed - Enhanced	**0.400**	**0.204**	**0.504**	**0.277**	**0.598**	**0.299**	**0.627**	**0.301**	0.665	**0.285**

Table 5.5 – Performance, in terms of CCC, of benchmark features and the proposed system, for predicting valence and arousal after applying RCTF-based speech enhancement to noise-plus-reverberation data.

			0dB		5dB		10dB		15dB		20dB	
RT (s)	Noise type	Features (Model)	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence
0.8	Airport	Benchmark (SVR)	0.037	0.016	0.164	0.026	0.178	0.051	0.219	0.009	0.232	**0.180**
		Benchmark (SVR) - Enhanced	**0.334**	**0.158**	**0.553**	**0.161**	**0.648**	**0.107**	**0.669**	**0.100**	**0.663**	0.047
		Proposed (SVR)	0.359	0.105	0.459	0.172	0.541	0.188	0.608	0.202	**0.686**	0.190
		Proposed (SVR) - Enhanced	**0.381**	**0.146**	**0.480**	**0.212**	**0.553**	**0.252**	**0.617**	**0.270**	0.615	**0.259**

noise-plus-reverberation setting ($RT = 0.8$ s and airport noise), and only the support vector regressor is explored. Tables 5.4 and 5.5 show the results obtained after applying the enhancement algorithm on the noise-only and noise-plus-reverberation conditions, respectively. As can be seen, enhancement improves the performance of both the benchmark and proposed systems, particularly at low SNR levels. In high SNR conditions, however, enhancement reduced the performance of the proposed system, particularly for the arousal dimension. This is likely due to the fact that the enhancement algorithm introduced artefacts that were more detrimental than the noise present, thus affecting overall performance.

5.5.6 Experiment VI: Performance on a subset of SEWA dataset

Here, we present the performance of the proposed feature pooling on the dataset described in Section 5.4.1. As the recordings in this database are already considered to be in the wild, noise and reverberation were not added to the speech signal. Results are compared to the baseline features and system used for the AVEC Challenge 2017 [39]. As regressor, we used the same settings provided by the challenge (i.e., a SVR), changing only the input features. The system is trained with the SEWA training set and evaluations are performed on the SEWA development set. As the frame-rate used for this challenge was increased from 40 to 100 ms, the MSR was extracted using the same frame-rate. The best performance was achieved with window length around 5 s. As previously found, this is essential for boosting performance of both *Original scheme 1* and *2*. For instance, *Pooling scheme 1* and *2* went, respectively, from CCC=0.126 to CCC=0.354 and from CCC=0.136 to 0.369

while predicting arousal, whereas for valence, they went from CCC=0.148 to 0.285 and from 0.137 to 0.308. Both pooling schemes outperformed the benchmark, CCC=0.345, for arousal prediction. For valence, the benchmark provided a better result, CCC=0.351.

As mentioned previously, the SEWA database was recorded in the wild, with participants in their homes or offices. Using the blind SNR estimator from the ITU-T P.563 algorithm [46] and the reverberation time estimator in [5], it was found that the noise and reverberation levels of the SEWA database ranged 9-44 dB and 1-1.8 s, respectively. Hence, the results on this second database corroborate those reported in sections IV.B-IV.D for similar levels.

5.6 Conclusion

This chapter has explored spontaneous speech emotion recognition (SER) in the wild, where factors such as noise, reverberation and their combined effects were explored. We show that existing SER systems based on per-frame features (computed from the modulation spectrum), while useful for enacted/posed emotions, perform poorly for spontaneous speech. As such, a feature pooling scheme is proposed that combines information from neighbouring frames. This pooling has significantly contributed to boost SER performance; it has also shown to be extremely important for valence prediction, even in clean conditions. By performing feature pooling, we also showed increased robustness against environmental noise. Compared to a benchmark algorithm from the Audio/Video Emotion Challenge 2016, the proposed feature pooling scheme did better when noise and noise-plus reverberation were present. The gains were more substantial as noise levels went up, thus showing the advantages of the proposed schemes for in the wild SER. Our findings were replicated in a subset of the SEWA dataset, which is considered to be in the wild. The experiments in this database corroborate the main contribution of this chapter, i.e., applying feature pooling is essential to increase performance of the modulation spectrum based features. We found also that for arousal predictions our method outperformed the baseline system used for the Audio/Video Emotion Challenge 2017. In this study, train/test mismatch was explored where clean speech was used for training and degraded speech for testing.

Automatic Speaker Verification from Affective Speech via Estimation of Neutral Speech Characteristics

6.1 Preamble

This chapter is compiled from material extracted from the manuscript submitted to the Speech Communication Journal [J4].

6.2 Introduction

Recent advances in channel compensation techniques [55, 56, 57, 58] and the use of deep learning embeddings, such as the x-vectors [25], have taken voice biometrics to the next level. Notwithstanding, mitigating the effects of affective speech on ASV has remained an open question and robustness towards intra-speaker variability caused by affective speech is still a challenge. Past studies have addressed the detrimental effects of vocal effort [59], language mismatch [60], and speaking style variation [61], but only a handful of studies have specifically proposed methods to mitigate automatic speaker verification (ASV) performance degradation due to expressive speech.

The authors in [62], for instance, proposed to modify prosodic parameters such as duration, pitch and amplitude of affective speech to mitigate the effects of mismatch between enrolment and test

utterances. The system was trained with features extracted from neutral speech and the modified ones extracted from emotional speech. Although the three prosodic parameters used have been widely explored for synthesizing expressive speech, there is no agreement that they can be universally used for all emotions [63]. Moreover, the study seems to neglect the fact that every speaker has unique characteristics in the way emotion is expressed, which makes a speaker-independent ASV solution in emotional environments quite challenging.

In [64], the authors proposed a framework consisting of three classification stages: (1) gender identification, (2) emotion recognition, and (3) speaker verification. Although the authors claimed that the proposed method showed to be superior to speaker verification systems based on gender only or emotion only, the performance of the overall system was shown to be sensitive to errors in these classifiers, thus increasing demand for highly-accurate emotion recognition. In realistic settings, this can be challenging [23]. The authors in [40] used the i-vector framework and proposed to estimate the reliability of ASV systems considering emotional dimensions such as arousal, valence and dominance. For that, the performance of the speaker recognition system is mapped as a function of arousal and valence primitives. A speech emotion recognizer is then trained to predict if emotional content falls into a reliable region. In a follow-up study, the authors then analyzed speaker verification performance in terms of arousal (calm/active), valence (negative/positive) and dominance (weak/strong) [41].

In a more recent study [65], the authors explore transfer learning from the knowledge learned for speaker recognition task to speech emotion recognition (SER). For that they used a pre-trained model to extract x-vectors, which were also used to investigate the impact of emotion on speaker verification. This research brings to light two important aspects of speaker verification in emotional environments. First, it shows that embeddings extracted for speaker recognition carry out relevant emotional cues that might be characterized by how each speaker particularly expresses emotions. In other words, it suggests that speaker recognition may take advantage of speaker specific emotional traits, useful for speech emotion recognition (SER) as well. Second, the work reports high equal error rate (EER) even when the same affective speech (e.g., angry) is used for both enrollment and testing, suggesting that emotion-dependent speaker models may not be enough to mitigate the detrimental effects of affective speech on ASV. As their system is based on a pre-trained x-vector, this issue may be overcome by training the x-vector directly with emotional data instead. However, as shown in [25], embeddings based on x-vectors require a large amount of data labelled with the speaker ID, which is rarely found in emotional datasets.

Most of the works mentioned above have relied on a single dataset of emotional speech with few speakers to conduct their experiments. As such, it is not clear if these findings persist across datasets, languages, or even unseen speakers. Also, most of these solutions have been based on compensation approaches, thus providing limited performance improvement, as well as negatively affecting the performance of speaker recognition systems for emotionally neutral speech [40]. In this work, we take an alternative approach to avoid these limitations.

First, we analyze the performance of an ASV system across four datasets: the Berlin emotional speech corpus (emodb) [156], the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [157], the Speech Under Simulated and Actual Stress Database (SUSAS) [158], and the MSP-IMPROV corpus [159]. Moreover, we propose a method to blindly mitigate mismatch between neutral and affective speech. For that, a Gaussian mixture model (GMM) is trained with the traditional mel-frequency cepstral coefficients (MFCCs). Each mixture of the GMM is then used to attain an average neutral-speech spectrum. Because the GMM is trained using only neutral speech, it can be seen as the prior probability distribution of acoustic characteristics of neutral speech. Moreover, as our approach is speaker- dependent, the probability distribution carries out speaker traits as well. Given affective speech, we then propose to use the prior probability distribution and the averaged neutral spectrum to minimize the mismatch between neutral and affective speech. It is important to mention that because the proposed solution is agnostic towards the type of emotion, the proposed method is emotion-independent.

The remainder of this chapter is organized as follows. In Section 6.3, we provide background material on the effects of emotions on speech production and speaker verification. In Section 6.4, the proposed method is presented. Section 6.5 then provides details about the experimental setup. Section 6.6 presents the experimental results and discussion. Lastly, Section 6.7 provides the conclusions.

Table 6.1 – Vocal changes due to five basic emotions. Adapted from [6].

	Fear	Anger	Sadness	Happiness	Disgust
Pitch average	Much higher	Much higher	Slightly lower	Higher	Much lower
Pitch range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Normal	Higher	Lower	Higher	Lower
Articulation	Precise	Tense	Slurring	Normal	Normal

6.3 Background Material

6.3.1 Affective Speech

The relationship between the autonomic nervous system (ANS) and human emotions has been a topic of research for many years [160][161]. Although controversial in terms of its generality, some research has shown evidence of emotion-specific ANS activity [162][163]. For example, the thermo-vascular/respiratory study conducted in [164] concludes that changes in temperature can identify two groups of emotions. While positive temperature variations were linked to surprise and anger, negative variations were related to happiness, fear, sadness and disgust. It is well-known that our heart and breathing rates can be influenced by our emotional states [165][163][166], which might have physiological effects on the speech production [167] system.

For example, according to [168] and [161], emotional states can cause dryness of the mouth or larynx, accelerated breathing rates, and muscle tremors. As the larynx regulates the flow of air in and out of the lungs [168], changes in wetness could result in changes of e.g., fundamental frequency (F_0) [168]. Moreover, an increase in respiration rate could result in increased subglottal pressure during speech, thus leading to variations in F_0 modulation during the production of voiced sounds [161]. Short duration segments of speech are another outcome of accelerated breathing rates, thus affecting the normative temporal patterns of speech. Lastly, tenseness and disorganization of the motor response, including tongue, lips and jaw, can influence vocal tract shapes, thus causing variation in the intensity and properties of the output sound.

In fact, alterations of the acoustic characteristics of the speech signal due to varying emotional states have been identified and measured [161]. Table 6.1 summarizes these effects for five basic emotions, according to [6]. As can be seen, the range and the average of F_0 change across different emotions, such as anger and happiness, as does intensity and articulation [161, 169, 170]. Figure 6.1

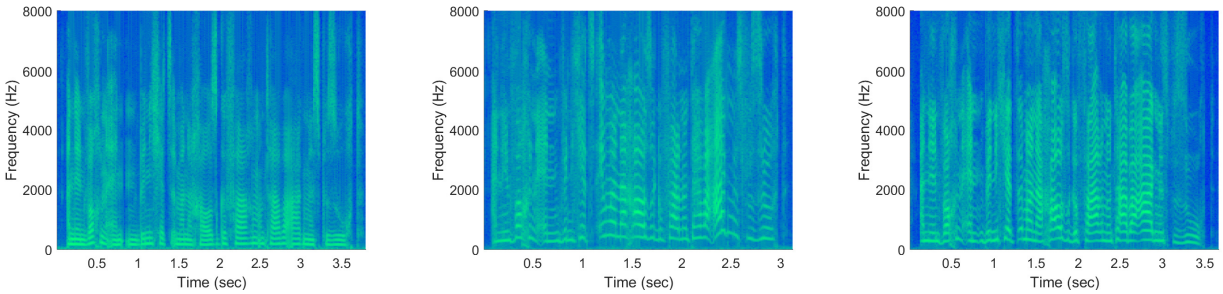


Figure 6.1 – Spectrogram of (a) neutral, (b) happy, and (c) angry speech attained from the Emodb dataset. Spectrograms correspond to the same utterance spoken by the same speaker.

further illustrates these effects. The figure shows the spectrogram of the same utterance, spoken by the same speaker, but under three varying (acted) emotions (from left to right: neutral, happy and angry). As can be seen, the frequency content, F_0, and formant structure vary greatly between the three emotional states. For example, with angry speech, higher average fundamental frequency can be seen when compared to neutral speech (as highlighted in Table 6.1). While this spectral variability may benefit emotion detection from speech, it can be detrimental to speaker recognition systems, as it leads to a mismatch between training and testing utterances.

6.3.2 Speaker verification

Speaker verification systems are concerned with determining whether or not two speech recordings belong to the same speaker. Thus, given an utterance and a claimed identity, the system must detect if the claimer is the genuine speaker or an impostor. A typical speaker verification system is composed of: (1) a front-end, normally dedicated to extraction of relevant features and, more recently, embeddings based on deep neural networks, (2) a back-end where the speaker models are trained, and (3) the module responsible for the scoring and decision process. Figure 6.2 depicts a representative of application of such a system. Typically, a user provides his/her identity, with a speech sample and a transaction request is made. The speaker verification system must then verify if the claimed identity really belongs to the user and authorize (or reject) the request.

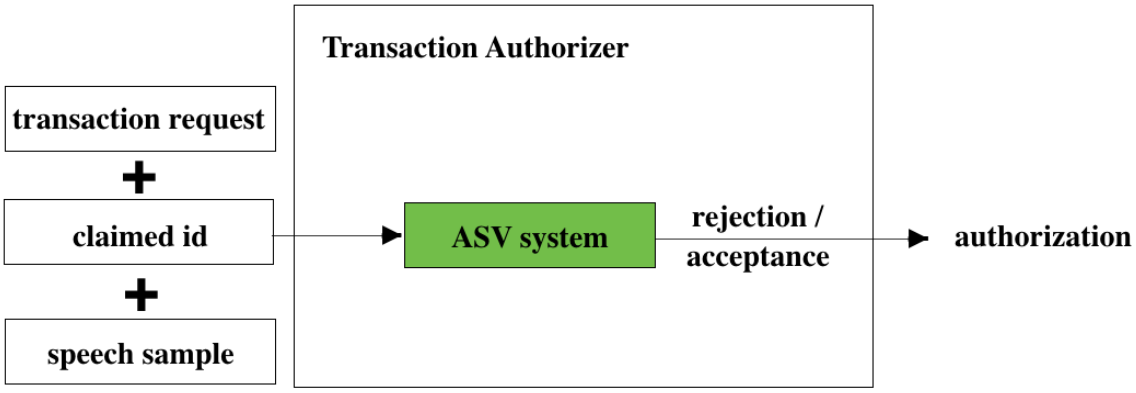


Figure 6.2 – Illustration of speaker verification application.

6.3.3 Affective speech ASV

Figure 6.3 shows the impact of affective speech on a speaker verification system based on i-vectors and the cosine distance as a scoring method [109]. In this preliminary experiment, the Emodb dataset is used (see Section 6.5.2). In this case, speaker models are trained using neutral speech, enrollment is performed with neutral speech, but emotional speech is used for testing. Here, three emotions, namely angry, happy and sad, are used for illustration purposes. As observed, a significant impact on ASV performance is seen with angry and happy emotions resulting in the most severe performance degradation. Similar findings were presented in a recent study [65] where angry speech was shown to result in the worst accuracy across multiple datasets. In that work, enrolling and testing with emotional speech did not show much improvement, thus suggesting that an alternate approach is still needed to minimize the differences between the source domain (i.e., neutral speech) and the target domain (i.e., affective speech) [171]. In the next section, we present the proposed method to mitigate such effects and boost ASV performance.

6.4 Proposed Method

In this section, we detail the proposed method to minimize the mismatch between neutral and affective speech. We start with a description on how to obtain the Gaussian mixture model of neutral speech. Next, we give details on the estimation of the neutral spectrum from affective speech. Lastly, we discuss the use of the proposed model for speaker verification.

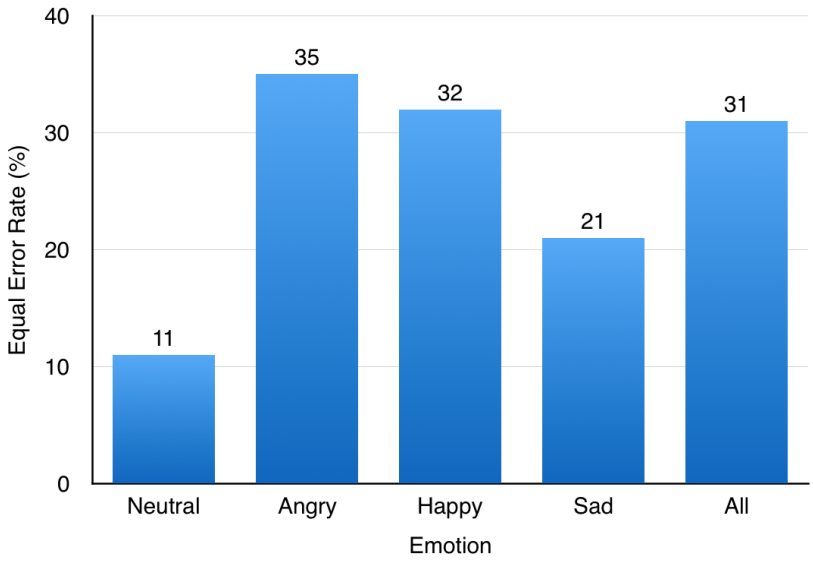


Figure 6.3 – Performance of ASV based on i-vector when training and enrolling with neutral and testing with affective speech.

6.4.1 Gaussian mixture model of neutral speech

Here, we propose to train a GMM based on neutral speech. Such a model is later used to blindly estimate neutral spectral traits from affective speech signals. The block diagram of the adopted procedure to attain the model is depicted by Figure 6.4. The GMM is trained offline (see lower part of Figure 6.4) using only neutral speech, represented here as $s(n)$. After extracting the STFT the attained log-spectrum, $\underline{S}(k,l) = \log(|S(k,l)|)$, is then normalized. This is performed mainly to mitigate issues with signal level differences. Note that the procedure differs from cepstral mean subtraction, which is commonly used to neutralize channel effects. The normalization here solely affects the log-spectral magnitude [172]. The normalized log-magnitude spectrum is attained by subtracting the log-magnitude spectrum, $\underline{S}(k,l)$, by its mean as follows:

$$\overline{S}(k,l) = \underline{S}(k,l) - \frac{1}{K}\sum_{k=1}^{K}\underline{S}(k,l), \tag{6.1}$$

where K is the number of STFT points. In our experiments $K = 512$. Next, 19 mel-frequency cepstral coefficients (MFCC) plus the log energy are attained. Lastly, RASTA filtering is performed in order to mitigate any channel effects [130]. More details about the speech parameterization process is given in Section 6.5.1.

The Gaussian mixture model is then trained with the MFCC-RASTA coefficients, $c_S(l)$. The GMM adopted here contains $M = 1024$ Gaussians with mixture probabilities:

$$\gamma_{l,m}(c_s(l)) = \frac{\pi_m \mathcal{N}(c_s(l)|\mu_m, \sum_m)}{\sum_{j=1}^{M} \pi_j \mathcal{N}(c_s(l)|\mu_j, \sum_j)}, \tag{6.2}$$

where $\lambda = \{\mu_m, \sum_m, \pi_m\}$ are the parameters of a multivariate Gaussian distribution represented by $\mathcal{N}(c_s(l)|\mu_m, \sum_m)$. To attain the average of the short-term log-spectra, $\gamma_{l,m}(c_s(l))$ and $\overline{S}(k,l)$ are combined over several frames of the training data. This leads to M average clean speech log-spectra:

$$\hat{S}_m(k) = \frac{\sum_{l=1}^{L} \gamma_{l,m}(c_s(l))\overline{S}(k,l)}{\sum_{l=1}^{L} \gamma_{l,m}(c_s(l))}, \forall k, m = 1, \ldots M. \tag{6.3}$$

Note that each mixture, m, is associated with a neutral speech spectrum, attained from the weighted average of multiple neutral speech spectra.

6.4.2 Neutral-speech spectrum estimation from affective speech

The upper part of Figure 6.4 provides a description of how to estimate the neutral speech spectrum from the affective signal. For instance, the affective speech signal, $x(n)$, is segmented prior to the extraction of the STFT, $X(k,l)$, and the RASTA-MFCC coefficients, $c_x(l)$. For a given affective speech signal, the neutral log-spectrum is then estimated using the feature vectors (i.e., $c_x(l)$) and the GMM parameters ($\mu_m, \sum_m$ and π_m) computed during the training phase. Then, the likelihood that a feature vector $c_x(l)$ belongs to the m-th mixture can be computed as in (6.2), which leads to a probability, $0 < p_{l,m} < 1$, for each mixture $m = 1,...,M$. This posterior probability can be used to estimate the neutral speech spectra of the l-th frame using the weighted average of the neutral speech spectra, $\hat{S}_m(k)$, associated with each mixture, as described below:

$$\hat{S}(k,l) = \sum_{m=1}^{M} \gamma_{l,m}(c_x(l))\hat{S}_m(k), \forall k. \tag{6.4}$$

As such, $\hat{S}(k,l)$ is an estimation of the neutral speech spectrum from the affective speech, $x(n)$. It is not our intention to claim that this can be used as a voice conversion technique, but we show herein that the proposed method can reduce intra-speaker variability without losing speaker traits useful for the speaker verification task.

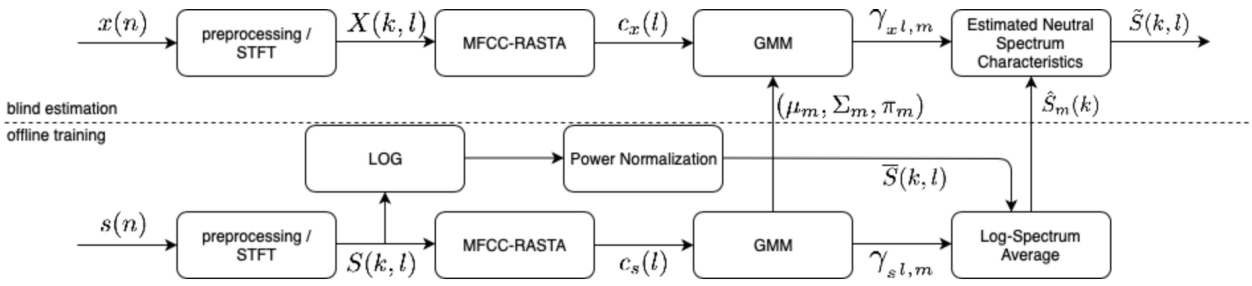


Figure 6.4 – Block diagram illustrating the estimation of neutral spectrum characteristics from affective speech.

6.4.3 Speaker verification using affective speech

Figure 6.5 illustrates the proposed method to mitigate the detrimental effects of affective speech on ASV performance. The first step consists of attaining the neutral speech Gaussian mixture model, as discussed in Section 6.4.2. For that, one GMM is trained per speaker using MFCC-RASTA features extracted from neutral speech. The trained GMM serves as the prior probability distribution of the neutral speech for a given speaker, which can also be seen as the speaker source space. Such a speaker-dependent approach is particularly important to preserve individual traits, thus increasing inter-speaker variability while at the same time decreasing within-speaker variations caused by emotion. As we are focused on a speaker verification task, the speaker model to be used is always known.

Next, the second step refers to the estimation of neutral speech spectra from an affective speech signal. This is achieved using Eq. (6.4). Note that the $\hat{S}_m(k)$, discussed in Section 6.4.1, is the weighted average log-spectrum of neutral speech for a given speaker. For affective speech, $x(n)$, feature vectors, $c_x(l)$, are extracted. The likelihood that the feature vector $c_x(l)$ arises from each mixture M is computed using Eq. (6.2). We then combine this information (i.e., $\hat{S}_m(k)$ and the posterior probability extracted using $c_x(l)$) to estimate the neutral speech spectrum as the weighted average of neutral speech spectra for a given speaker. This can be considered as a transfer learning operation that minimizes the differences between target (affective) and source (neutral) spaces. As shown in Figure 6.5, after estimating the neutral speech spectra for a given affective speech signal, i-vectors are extracted using the procedure described in Section 6.3.2. Note that, because we only care about verifying the speaker ID, a precise mapping between affective and neutral speech is not needed, just a mapping that preserves speaker traits while minimizing the mismatch with training data.

6.5 Experimental Setup

In this section, we present the feature extraction steps, databases, ASV scoring method, baseline system, and figures-of-merit used to gauge the performance of the proposed method.

6.5.1 Feature extraction

Our experiments are based on mel-frequency cepstral coefficients (MFCC). These features are first used to train the GMM, as well as for speaker modeling based on the i-vector framework. Prior to their extraction, speech signals from all datasets are re-sampled to 16 kHz, if necessary. Each speech signal is also segmented into frames of 320 samples (20-ms length for speech signals sampled at 16 kHz), with 50% hop-size (i.e., 160 samples). A Hanning window is also applied on each frame, followed by a pre-emphasis filter of coefficient 0.97. MFCCs are then extracted from each frame. A total of 19 coefficients, plus the log energy are attained, thus leading to a 20-dimensional feature vector for each frame. We emphasize here that first and second order derivatives (delta and double-delta coefficients) were also explored from the MFCC coefficients in order to gauge temporal dynamics information. In our experiments, however, we found no benefit in using first and second derivatives.

For the neutral speech modeling part of the proposed method, which will be discussed next, we also apply the relative spectral (RASTA) filtering technique. This is particularly important to remove irrelevant information that might be "embedded" in the speech signal, such as that characterized by the communication channel [128]. The RASTA filter technique relies on the fact that linguistic content is coded based on the movements of the vocal tract at rates-of-change different from other non-linguistic content. In the past, it has been shown to be efficient in mitigating the negative effects of convolution and additive noise. More details about the RASTA technique can be found in [128].

6.5.2 Emotional Speech Corpora

Emotional speech corpora can be divided into three types: (1) simulated, (2) elicited, and (3) natural. To obtain simulated emotion, professional actors are hired and asked to express neutral sentences in different emotional fashions. In this case, it is more convenient to use discrete emotions

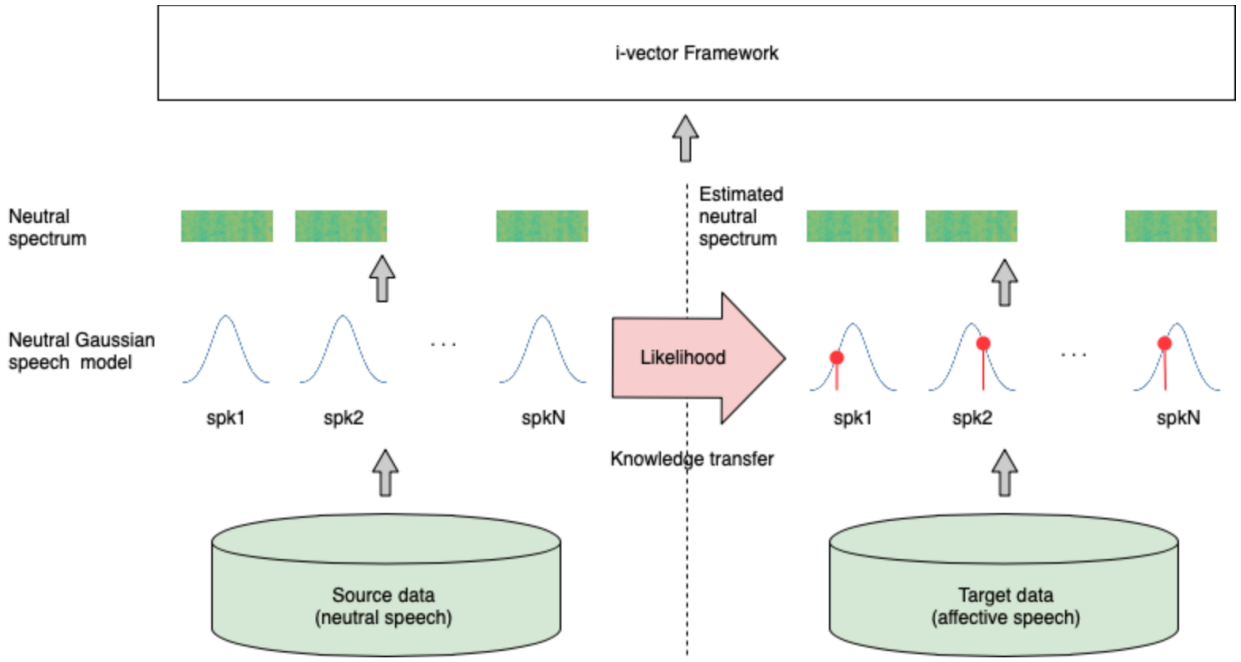


Figure 6.5 – Proposed approach for the speaker verification task in emotional environments.

as labels instead of emotional dimensions such as arousal, valence and dominance. This is because discrete labels are easier to understand and emulate. It also helps when it comes to combining multiple datasets by using the same categorical emotions. Elicited emotional speech, in turn, is attained by creating emotional circumstances. Without their knowledge, speakers are involved in an emotional conversation induced by an anchor. Lastly, natural databases contain speech recorded in spontaneous situations during the speakers' daily lives. These databases tend to be more natural compared to the simulated ones, unless the subjects have knowledge about being recorded. Although there are several emotional speech datasets available publicly, none have been developed with the ASV task in mind, thus are usually very limited in the number of speakers. Hence, in order to conduct the present study, multiple emotional speech datasets had to be utilized, thus not only allowing us to evaluate ASV performance across datasets, but to also increase the number of speakers to levels expected of ASV studies. A brief description of the datasets used in our experiments is given next.

Berlin emotional speech corpus

The Berlin emotional speech corpus [156], hereafter referred to as EMODB, contains simulated emotional speech from 10 professional actors (5 males and 5 females). These actors produced 10

German utterances. The neutral utterances were repeated by each actor in different emotional simulated states. The utterances were common everyday sentences susceptible to be interpreted with different emotions. The recordings took place in an anechoic chamber using high quality recording equipment. The speakers spoke each utterance at 30 cm from the microphone and the speech signals were recorded at 48 kHz rate and later resampled to 16 kHz. Each actor expressed 10 sentences in 7 different emotional states, leading to roughly 700 utterances in total [156]. The corpus comprises the following emotions: **neutral**, **anger**, **happiness**, fear, boredom, **sadness**, and disgust. In bold are the emotions used in our experiments and the main criteria for their selection was the amount of data available for a particular emotion per speaker. During the development of the database, a perception test was also performed to ensure the emotional quality and naturalness of the utterances. A total of 20 subjects participated in this test and were required to listen to emotional utterances and then decide which emotional state best represented the speaker emotional state. Besides being a well known emotional speech corpus, already used in many studies, the EMODB allowed us to validate our method in a language other than English, which was important in order to show applicability and generalization of the proposed solution.

Ryerson Audio-Visual Database of Emotional Speech and Song

The Ryerson audio-visual database of emotional speech and song (also referred to as RAVDESS) is a multimodal database that contains face-and-voice, face-only, and voice-only formats [157]. The speech content was recorded in a professional recording studio at Ryerson University (Toronto, Canada). A Rode NTK vacuum tube condenser microphone was used during the recording sessions. Speech samples were recorded at 48 kHz sampling rate, 16 bit, with files saved in uncompressed wave format. The microphone was placed 20 cm from the participants. A total of 24 gender-balanced professional actors were recruited for the experiment. They cited utterances using a neutral North American English accent. Microphone levels were set by having the actor produce several very angry expressions. A total of 7356 recordings are available comprising the following emotions: calm, happy, sad, angry, fearful, surprise, and disgust [157]. The data was annotated by 247 untrained participants from North America. All emotions are included in our experiments as the number of sentences available per emotion are balanced among speakers.

MSP-IMPROV

The MSP-IMPROV corpus is a simulated multimodal emotional database developed by the Multimodal Signal Processing Laboratory at the University of Texas at Dallas (USA) [159]. The corpus was designed to elicit emotional behaviors with fixed lexical content, but that convey different emotions, which are referred to as target sentences. It is based on spontaneous dyadic improvisations with scenarios designed to elicit realistic emotions. The utterances are cited in English by 12 actors. Their ages ranged between 18 and 21 at the time of the data collection. Audio was recorded at a 48-kHz sampling rate and 32-bit PCM format. The database provides four categorical emotions: neutral, happiness, sadness, and anger. The emotional labeling process was conducted through crowd-sourced perceptual evaluations. The MSP-IMPROV database was collected in an ASHA-certified sound booth, measuring 13 ft x 13 ft. The actors were facing each other, being about two meters apart. Two microphones were used to record the audio with a collar microphone for each actor. More details can be found in [159].

Speech Under Simulated and Actual Stress Database

Although the Speech Under Simulated and Actual Stress (SUSAS) dataset comprises actual and simulated stress conditions, only simulated stress was considered here. The SUSAS database was recorded from 32 speakers (13 female, 19 male) at an 8-kHz sampling rate. Utterances were produced in English. Nine different stress conditions were recorded: neutral, angry, loud, soft, slow, Lombard effect (pink noise presented binaurally at an 85 dB SPL level), fast, and speech produced under two levels of workload: low and high. For more information on the database the interested reader can refer to [158].

6.5.3 ASV system backend and baseline system

The backend of our system is based on probabilistic linear discriminant analysis (PLDA) [173]. It is commonly applied in combination with linear discriminant analysis (LDA). LDA attempts to find directions in a feature space that maximize discriminability between data points. This is achieved by finding a linear combination of features that separates two or more classes. It is closely related to principal component analysis in the sense that it seeks to encounter a linear combination of the

features that best explain the data. It is also used for dimensionality reduction. PLDA provides a probabilistic framework applicable to fixed-length input vectors. It can be seen as a special case of JFA. However, while JFA requires processing the acoustic speech frames as well as their statistics in different mixtures of the UBM, for PLDA the input features are i-vectors that are extracted beforehand. For the scoring process, acoustic features are not required, only two i-vectors from the corresponding utterances are needed [83], thus making PLDA much simpler in implementation.

The baseline consists of a GMM with 32 components and 50 total factors for the total variability matrix. Only neutral speech was used for training the UBM and T matrix as well as for enrolment.

6.5.4 Figures-of-Merit

In conventional ASV, target and nontarget trials exist and the ASV task is to either accept or reject the user (or impostor). From this process, two possible errors are expected [174]:

- false acceptance rate (FAR), or false positive, classifying a non-target trial as a target trial, and
- false rejection rate (FRR), or false negative, classifying a target trial as a non-target trial.

Typically, the equal error rate (EER) has been the measure used to compare the performance of speaker verification systems. The EER refers to the points where both false negative and false positive rates are the same [174]. Additionally, the minimum detection cost functions (DCF) are adopted. These metrics are commonly used for speaker recognition evaluations and are based on the weighted sum of the FAR and FRR metrics, thus representing the cost of making a detection decision. The total cost can be obtained by summing the action-specific costs as below [175]:

$$DCF(\alpha_j) = \sum_{j=1}^{L}\sum_{i=1}^{M} = \pi_i C(\alpha_j|\theta_i) P_{err}(\alpha_j|\theta_i), \tag{6.5}$$

where π_i represents the prior of how often the target and non-target users appears. The cost of assigning the wrong class is given by $C(\alpha_j|\theta_i)$, with α_j representing the decision made by the classification system while θ_i is the actual ground truth. The probability of false rates is given by $P_{err}(\alpha_j|\theta_i)$. The variable L and M are, respectively, the numbers of decision and ground truth [175]. A more familiar terminology used by the ASV research community [176], and adopted in this work,

measures DCF as:

$$DCF(t) = C_{miss}\pi_{tar}P^{asv}_{miss}(t) + C_{fa}(1-\pi_{tar})P^{asv}_{fa}(t), \tag{6.6}$$

where C_{miss} and C_{fa} represent, respectively, the costs of rejecting the target user and accepting a non-target user and t defines the threshold used to accept or reject the target hypothesis for a given decision is score [175].

6.5.5 Test setup and experiments description

In all experiments described herein only neutral speech was used for training the UBM and T matrix. No compensation techniques, such as multi-condition training (i.e., adding emotional speech into the training data) were applied. We adopted a GMM with 32 components and 50 total factors for the total variability matrix. It is important to mention that although the i-vector extractor is typically defined with 2048 Gaussian components and 400 factors [30], these values were not possible given the limited amount of data available per speaker [177]. Moreover, unless pre-trained models are used, as suggested in [65], this limitation in available data makes it unsuitable to explore the use of more recent ASV systems based on deep learning. Such approach will be left for future work.

In order to enforce mismatched conditions, speaker enrolment was performed with neutral speech and affective speech was used only for testing. In the first experiment, the UBM and T matrix were trained with neutral speech from the respective affective datasets. For the EMODB and the RAVDESS datasets, the number of utterances was lower relative to the other datasets. Therefore, segments of roughly 0.5 s from each utterances were extracted to be used during training. For the other two datasets, we considered every single recording to be an utterance. We cared to avoid any overlap between neutral speech used for training and enrollment. Given the limited amount of training data per speaker (expect for the MSP-IMPROV), in our second experiment, we attempt to use the TIMIT dataset to train the UBM and T matrix, freeing more data to be used for enrollment and test. In out last experiment, the proposed solution is evaluated after combining speakers from all datasets. Neutral and angry speech samples from the SUSAS dataset are included in this experiment.

Table 6.2 – Performance of emotional speaker verification in terms of EER (%) and DCF for the baseline and proposed solution for EMODB, RAVDESS and MSP-IMPROV datasets. UBM and T matrix trained with neutral speech from the respective datasets.

	EMODB				RAVDESS				MSP-IMPROV			
	Baseline		Proposed		Baseline		Proposed		Baseline		Proposed	
	EER	*DCF*	*EER*	*DCF*	*EER*	*DCF*	*EER*	*DCF*	*EER*	*DCF*	*EER*	*DCF*
Neutral	10	0.067	2	0.016	16	0.095	13	0.069	5	0.058	1.6	0.008
Angry	31	0.097	0.4	0.001	29	0.1	13	0.076	8.2	0.050	1.1	0.008
Happy	26	0.093	0	0	25	0.1	13	0.059	10	0.074	2.6	0.011
Sad	16	0.082	7.4	0.063	22	0.099	10	0.067	6.8	0.042	2.2	0.006
Calm	14	0.072	2.7	0.029	20	0.099	7.5	0.069	-	-	-	-
Surprise	-	-	-	-	25	0.098	11	0.065	-	-	-	-
Fearful	-	-	-	-	31	0.1	10	0.065	-	-	-	-
Disgust	-	-	-	-	25	0.099	11	0.053	-	-	-	-
Average	19.4	0.102	2.5	0.027	24.1	0.098	11.06	0.065	7.5	0.056	1.8	0.008

6.6 Experimental results and discussion

In the first experiment, the proposed solution is tested on three datasets (EMODB, RAVDESS and MSP-IMPROV). The SUSAS dataset was left out as most of its affective states are related to stress, with the exception of the neutral and angry states. Results reported in Table 6.2 show the EER and DCF values obtained in this experiment for each dataset and across different emotions. The UBM and T matrix were trained using neutral speech from speakers from the respective datasets. Moreover, there was no overlap of data used for training, enrolment, and test. Note that, different from the MSP-IMPROV, the EMODB and the RAVDESS datasets were not designed for the speaker verification task and, therefore, present some drawbacks in terms of utterance length and the amount of data available. This may explain the relatively high EER when testing the baseline system with neutral speech, as mentioned in [65].

For the baseline system, the lowest EER/DCF is always achieved in the matched condition (i.e., enrolling and testing with neutral) regardless of the dataset. We observe a severe degradation of the baseline system when tested on affective speech, specially for happy and angry emotions. Note, for example, EER increasing from 10 %, when tested with neutral speech, to 31%, when tested with angry speech, for the EMODB. A similar trend was encountered for the RAVDESS, which increased EER from 16% to 29%. The proposed method, on the other hand, showed improved performance values, compared to the baseline, almost consistently across all tested emotions. Performance also improved substantially over the baseline. For the EMODB, for example, a reduction from 10% to 2% in terms of EER and from 0.067 to 0.016 in terms of DCF were achieved. For the MSP-IMPROV

Table 6.3 – Performance of emotional speaker verification in terms of EER (%) and DCF for the baseline and proposed solution for EMODB, RAVDESS and MSP-IMPROV datasets. UBM and T matrix are trained with neutral speech from TIMIT.

	EMODB				RAVDESS				MSP-IMPROV			
	Baseline		Proposed		Baseline		Proposed		Baseline		Proposed	
	EER	*DCF*	*EER*	*DCF*	*EER*	*DCF*	*EER*	*DCF*	*EER*	*DCF*	*EER*	*DCF*
Neutral	14	0.095	9.7	0.065	12	0.092	6.2	0.051	10	0.078	1.6	0.033
Angry	40	0.099	5.6	0.033	22	0.1	5	0.047	10	0.086	2.2	0.025
Happy	39	0.1	8.4	0.065	21	0.098	4.9	0.059	16	0.089	6.9	0.051
Sad	24	0.099	13	0.078	20	0.099	5.3	0.055	12	0.076	2.7	0.019
Calm	28	0.089	9.3	0.061	17	0.099	2.7	0.040	-	-	-	-
Surprise	-	-	-	-	21	0.099	4.1	0.065	-	-	-	-
Fearful	-	-	-	-	25	0.099	3.3	0.062	-	-	-	-
Disgust	-	-	-	-	26	0.098	3.3	0.061	-	-	-	-
Average	29	0.096	9.1	0.060	20.5	0.098	4.35	0.055	12	0.082	3.35	0.032

EER was reduced from 5% to 1.6% and DCF from 0.058 to 0.008. For RAVDESS, EER went from 16% to 13% and DCF dropped from 0.095 to 0.069.

Interestingly, for the proposed method, the achieved EER/DCF values were not the lowest in the matched condition (neutral). Experiments with angry samples, considered the most severe emotional state for ASV [65], resulted in lower EER compared to the results with neutral. A possible explanation is that, as our method is speaker-dependent, it preserves the most prominent individual traits, which helps to increase inter-speaker variability. In other words, the detrimental variability caused by some emotions, such as anger, now are conditioned by a strong prior that depends on each speaker.

Table 6.3 reports the results for our second experiment, where the TIMIT dataset was used to train the i-vector extractor. The motivation here is to gain some insight on the impact of using a bigger dataset to train the UBM and T matrix, as well as on the possibility of generalization to unseen speakers. However, since no prior information about the speakers present in each affective dataset was used, EER rose compared to the previous experiment (except for the RAVDESS dataset). Note that the results for the EMODB dataset are, in general, worse compared to the other datasets. It is important to note that, for this experiment, there is also a language mismatch as the EMODB is spoken in German while TIMIT is spoken in English.

Performance also improved substantially over the baseline in matched conditions. For the EMODB, for example, a reduction from 14% to 9.7% in terms of EER and from 0.095 to 0.065 in terms of DCF were achieved. For the MSP-IMPROV, EER was reduced from 10% to 1.6% and DCF from

Table 6.4 – EER (%) and DCF for baseline and proposed solution considering 45 speakers from EMODB, RAVDESS and MSP-IMPROV. Speakers from SUSAS were also considered for neutral and angry speech, totalling 55 speakers.

	45 speakers				*55 speakers*			
	Baseline		Proposed		Baseline		Proposed	
	EER	*DCF*	*EER*	*DCF*	*EER*	*DCF*	*EER*	*DCF*
Neutral	12	0.099	4.3	0.022	22	0.1	10	0.067
Angry	22	0.098	5.2	0.030	27	0.1	16	0.081
Happy	21	0.099	4.3	0.025	-	-	-	-
Sad	20	0.1	5.7	0.031	-	-	-	-
Average	18.75	0.124	4.6	0.027	24.5	0.1	13	0.074

0.078 to 0.033. For RAVDESS, it went from 12% to 6.2% while DCF was reduced from 0.092 to 0.051. The proposed method is also successful in mitigating performance degradation caused by mismatch between enrolment and testing data. Note, for example, EER decreasing from 40% to 5.6% when the proposed method is applied and tested with angry speech for EMODB. Similar trends were encountered for the RAVDESS, with EER deacreasing from 22% to 5%, and from 10% to 2.2% for MSP-IMPROV.

Lastly, to make our evaluation more realistic, the final experiment focuses on an ASV task with a larger number of speakers. As such, the EMODB, RAVESS and MSP-IMPROV datasets were combined, thus totalling 45 speakers; overall, four emotion conditions are considered as they coincided across the datasets. We also considered 10 speakers from the SUSAS dataset under the neutral and angry conditions and tested the proposed method under these two emotional states. Table 6.4 shows the results obtained under these two scenarios. As can be seen, emotion condition mismatch severely degrades performance of the baseline system, whereas little impact is seen for the proposed method. Moreover, under both scenarios tested, the proposed method outperforms the baseline by an average 14% across emotions, with the largest improvement seen for the angry and happy emotion (15%).

In all the experiments described above, we found that angry utterances, usually followed by happy, has the worst impact in the baseline system. As discussed in Section 6.3 and shown in Table 6.1, these emotions are the ones that affect the most the physiology of speech production. Hence, it is expected that they represent the worst type of mismatch.

6.7 Conclusions

In this chapter, we have proposed a new method to compensate for the detrimental effects that emotional speech has on an automatic speaker verification (ASV) system. In particular, we propose a new Gaussian mixture model (GMM) based method to "neutralize" affective speech, thus mitigating the detrimental condition mismatch effects on the speaker verification task. The proposed method, when coupled with a conventional i-vector based ASV, is shown to outperform a baseline by as much as 15%. Experiments are performed across four separate multi-lingual datasets, as well as with a combined larger dataset, and the results obtained consistently show the proposed method outperforming the baseline across up to eight different emotional states. More importantly, the proposed approach does not compromise the ASV performance of neutral speech, an issue commonly observed in other approaches reported in the literature. As more emotional datasets become available, alternate ASV systems based on more recent deep learning architectures may become a viable alternative.

Channel Response Estimation and Residual Neural Network to Detect Physical Attacks

7.1 Preamble

This chapter is compiled from material extracted from the manuscript submitted to the Journal of Computer Speech & Language [J3], also from material that appeared in the Proceedings of Interspeech 2019 [C2].

7.2 Introduction

Automatic speaker verification (ASV) has significantly matured over the last few years [66]. Advances in channel compensation techniques [55][56] and the use of deep learning embeddings, such as x-vectors [25][67], have taken automatic speaker verification to a higher level. The deployment of commercial mobile voice recognition products has already become a reality [68]. To enhance password-based authentication mechanisms, for instance, a number of financial institutions are investing in voice authentication solutions [69]. This is driven mainly by the increased use of mobile devices, as well as by the convenience and non-intrusiveness offered by such technologies. In fact, recent reports predict a continued growth of the mobile biometrics sector due to the increased

consumer demand for safety, especially while using mobile devices for banking transactions and e-commerce [69].

Despite all these advances, malicious spoofing attacks have been recognized as a serious threat to ASV [66]. Characterized by an attempt of a person or a program to illegitimately bypass security by masquerading one's identity, there are growing concerns towards the vulnerability of ASV in the face of spoofing attacks, such as impersonation, replay attacks, speech synthesis, and voice conversion [66]. As such, a handful of initiatives to develop spoof countermeasures have been made lately [70][71]. Many of the efforts in this direction have been focused on developing anti-spoofing techniques to protect ASV systems against speech synthesis (SS) and voice conversion (VC) [70]. In this study, we are particularly interested in countermeasures to replay attacks, which consist of attempts to fool an ASV system by playing back a pre-recorded speech sample. In such circumstances, detecting the replay attack beforehand is crucial to maintain ASV reliability.

Given the emerging interest in the topic, the Automatic Speaker Verification Spoofing and Countermeasures Challenge was created, with the latest versions being in 2017 [71] and 2019 [72]. Henceforth referred to as ASVspoof 2017/2019, these challenges provided common databases, protocols, and metrics to evaluate different countermeasure solutions. Both competitions contained datasets with diverse forms of replay attacks, with the 2019 set providing a larger number of replay attack configurations. Until then, replay attacks, also known as physical attacks, had received little attention from the research community compared to other spoofing modalities (e.g., speech synthesis).

In [73], for example, the authors propose an end-to-end countermeasure solution using the raw waveform. While many countermeasure methods are based on a combination of feature extraction and a back-end classifier, this method dismisses the need for any pre-processing on the speech waveform. Although feature engineering can make more feasible the interpretation of the extracted features, it may neglect rich speech information that might be found in the raw waveform by an end-to-end approach. In [74], the authors proposed a deep residual neural network (ResNet-18) architecture, with a visual attention mechanism on time-frequency representations based on group-delay features, as a countermeasure for replay attacks. Results obtained in terms of equal error rate (EER) were quite low, but only reported on the ASVspoof 2017 dataset. In [75], the authors proposed to complement short-term spectral features with two novel features based on the modulation spectrum. The latter

captures static and dynamic characteristics of the speech signal from the modulation spectrum, which complement short-term spectral features for use in replay detection. The authors in [76], in turn, relied on spectral bitmaps or spectral peaks, which are time-frequency points higher than a pre-defined threshold. The similarity score was attained by computing an element-wise product between the spectral bitmap of the verification sample and stored spectral bitmap templates. More recently, the performance of several features and classifiers was described in [77]. The authors reported results from six magnitude-spectrum and three phase-spectrum-based features on the ASVspoof 2017 replay attack detection challenge, with experiments revealing the superiority of the magnitude-spectrum features over phase-based features for all four classifiers tested. An attentive filtering network combined with a ResNet-based classifier is proposed in [78], thus resulting in better discriminative features both in the time and frequency domains.

Despite the recent advancements in this field, investigating new countermeasure solutions applicable to emerging and more challenging scenarios is still needed. In this work, we propose the use of blind channel spectrum estimation in combination with deep neural network-based classification to detect replay attacks. Considering that in a replay attack the utterance will be acoustically affected by factors such as the room environment, the recording, and the playback devices, it is expected that such effects will generate a unique "signature" in the signal's log-magnitude spectrum. Hence, we propose to detect such spectral signatures by estimating the magnitude response of the channel.

Here, this is achieved by first training a clean speech Gaussian mixture model (GMM). The model is trained using RASTA-filtered mel-frequency cepstral coefficients (RASTA-MFCCs) extracted from several clean speech files, thus allowing us to attain a model of clean spectrum characteristics. The channel response spectrum is then estimated by computing the log-magnitude spectrum average of clean signals and by then subtracting it from the log-magnitude spectrum of the observed signal. As a classifier, we adopted the benchmark GMM to distinguish between true and spoofing utterances. Next, motivated by the recent results obtained with ResNets [79, 80], we also explore the use of such networks.

Experimental results show the proposed method outperforming the benchmarks on both the development and evaluation sets for the ASVspoof 2017 and 2019 datasets. To the best of our knowledge, only few studies have addressed the use of channel estimation as a countermeasure solution. In [81], for example, the authors proposed the use of two low-level descriptors, the constant-

Q cepstral coefficients (CQCC) and the high-frequency cepstral coefficients (HFCC), as input to a convolutional neural network (CNN). The authors claim that the CNN model is estimating the channel conditions although no clear explanation is given regarding how the channel is estimated. The present work provides a more complete investigation of the use of channel estimation, representing an important contribution towards mitigating the problem of spoofing replay attacks. Moreover, compared to our previous work [82], this study (1) presents additional and improved results on an extended dataset; (2) evaluates the impact of the resolution of the channel estimation approach on spoofing detection performance; (3) performs a quality analysis of the two datasets being tested and discusses the impact of signal quality and spoofing detection accuracy; and (4) presents performance improvements with ResNet, while comparing results with a state-of-the-art algorithm.

The remainder of this Chapter is organized as follows. Section 7.3 provides a description of the proposed method and Section 7.4 the proposed deep neural network classifier. In Section 7.5, we present our experiment setup and Section 7.6.1 discusses our experimental results. Section 7.7 concludes the Chapter.

7.3 Blind channel response estimation

In this section, the general ideas behind blind channel response estimation are described, along with the steps to attain the average spectra of clean speech, followed by estimation of the channel response magnitude. Lastly, we give a short description of the MFCC's extraction and some insights on the RASTA filtering procedure.

7.3.1 General Principles

A replay attack is characterized by an attempt to access an ASV system using a pre-recorded speech sample, collected from a bonafide target speaker [66]. A typical scenario is illustrated in Figure 7.1. The utterance recorded from the bonafide speaker is presented to the ASV system. In such circumstance, the recorded utterance is acoustically affected by the different combinations of room environment, microphones, and playback devices. This can be expressed as:

$$x(n) = s(n) * h(n) + v(n), \tag{7.1}$$

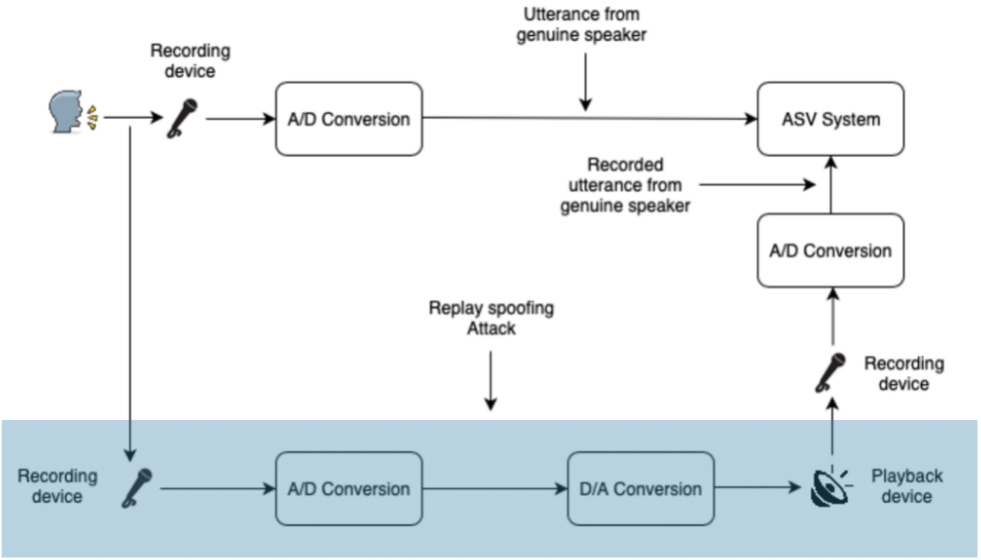


Figure 7.1 – Replay attack scenario.

where n is the time sample and $s(n)$ represents the speech signal from a bonafide access attempt, $h(n)$ denotes the channel impulse response and $v(n)$ refers to the additive noise. Hereafter, we refer to all the possible effects encompassing the spoofed utterance as the channel effect. Thus, given the observed playback waveform, $x(n)$, our goal is to estimate the original acoustic waveform, $s(n)$, as closely as possible. This can be seen as a blind deconvolution problem as neither the clean signal, $s(n)$, nor the channel impulse response, $h(n)$, for a particular recording is known.

By writing Eq. (7.1) in the frequency domain and using the short-time Fourier transform (STFT), we attain:

$$X(k,l) = S(k,l)H(k) + V(k,l), \tag{7.2}$$

with each frequency bin represented by k and the time frame by l. We assume a noiseless environment and hence $V(k,l) \equiv 0$. With prior knowledge of the log-magnitude spectrum average of the clean speech signal, $S(k,l)$, the log-magnitude spectrum of the channel response can be estimated as [130]:

$$\underline{\hat{H}}(k,l) \approx \underline{X}(k,l) - \underline{S}(k,l), \tag{7.3}$$

where $\underline{X}(k,l) = \log(|X(k,l)|)$ and $\underline{S}(k,l) = \log(|S(k,l)|)$. Because $S(k,l)$ is unknown, the performance of the proposed method relies on the clean speech model, which is then used to compute the clean log-magnitude spectrum average, $\underline{\hat{S}}(k,l)$, from the observed signal. Differently from [130],

we kept the frame information while computing the log-magnitude of the channel response. Hence, $\underline{\hat{H}}(k,l)$ represents an estimate of $\underline{H}(k,l)$.

In the remainder of this section we discuss the steps to attain the clean speech model. In the following, we describe how we use the parameters of such a model to estimate the clean log-magnitude spectrum average from the observed signal as well as the log-magnitude spectrum of the unknown channel response $\underline{\hat{H}}(k,l)$.

7.3.2 Log-Magnitude Spectrum from a Clean Speech Model

The steps to estimate the channel response log-magnitude spectrum are depicted in Figure 7.2. Such estimation depends on two things: (1) the trained GMM clean speech model with M mixtures and (2) a known average log-magnitude spectrum associated with each mixture. The lower half of Figure 7.2 describes this process. In order to mitigate channel effects, speech parameterization is performed based on the RASTA-filtered Mel-Frequency Cepstral Coefficients (MFCC-RASTA). As we discuss in more detail in subsection 7.3.5, the main reason for adopting these features comes from the fact that they are known for being robust and less impacted by channel effects [172]. Thus, to train the GMM clean speech model, we use the feature vectors $c_s(l)$ extracted from clean speech signals represented by $s(n)$. The M-mixture GMM is denoted by its means, μ_m, by the diagonal covariance matrix, $\sum_m$, and by its weights, π_m. The probability that a particular feature vector, $c_s(l)$, belongs to the m-th mixture is given by:

$$p_{l,m}(c_s(l)) = \frac{\pi_m \mathcal{N}(c_s(l)|\mu_m, \sum_m)}{\sum_{j=1}^{M} \pi_j \mathcal{N}(c_s(l)|\mu_j, \sum_j)}, \tag{7.4}$$

where $\mathcal{N}(c_s(l)|\mu_m, \sum_m)$ represents a multivariate Gaussian distribution. Note that, for a given speech signal, $s(n)$, it leads to a matrix of probabilities $M \times L_S$, $p_{l,m}(c_s(l))$, containing the M mixture probabilities for the L_S available speech frames.

A normalization step is also applied on the log-magnitude spectrum. This is performed mainly to mitigate issues with signal level differences. Note that this procedure differs from the cepstral mean subtraction, which is commonly used to neutralize channel effects. The normalization here solely affects the log-spectral magnitude [172]. The normalized log-magnitude spectrum is attained

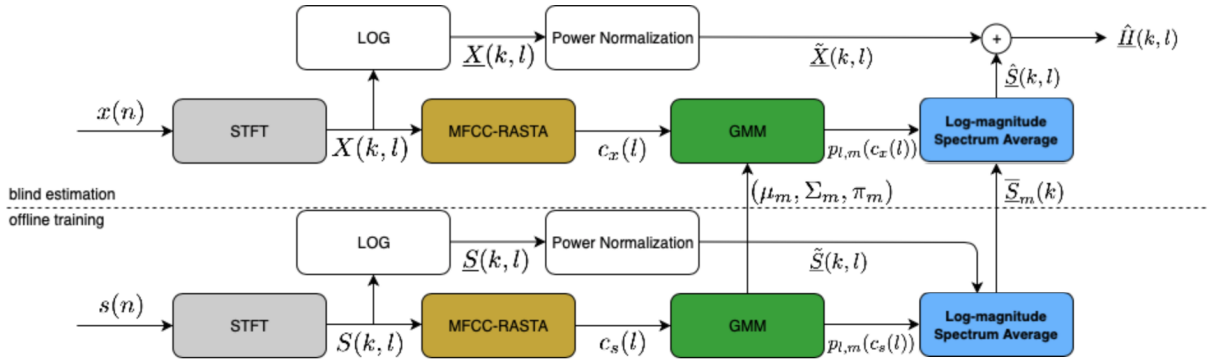


Figure 7.2 – Diagram for estimating the log-magnitude spectrum of the channel response.

by subtracting the log-magnitude spectrum, $\underline{S}(k,l)$, by its mean as follows:

$$\tilde{\underline{S}}(k,l) = \underline{S}(k,l) - \frac{1}{K}\sum_{k=1}^{K}\underline{S}(k,l), \tag{7.5}$$

where the number of STFT points is defined by K. Note that the same normalization process is applied to the observed signal, $x(n)$, in order to obtain its normalized log-magnitude spectrum, $\tilde{\underline{X}}(k,l)$. As our goal is to attain an average of the short-term log-spectra, $p_{l,m}(c_s(l))$ are combined with $\tilde{\underline{S}}(k,l)$. The set of M average clean speech log-magnitude spectra can be obtained as:

$$\overline{\underline{S}}_m(k) = \frac{\sum_{l=1}^{L} p_{l,m}(c_s(l))\tilde{\underline{S}}(k,l)}{\sum_{l=1}^{L} p_{l,m}(c_s(l))}, \forall k, m = 1, \dots M, \tag{7.6}$$

where each mixture, m, is associated with a clean speech spectrum, attained from the weighted average of multiple clean speech spectra assigned to a particular mixture. $\overline{\underline{S}}_m(k)$ has dimension $M \times K$ and its rows represent the average log-magnitude spectrum corresponding to the m-th mixture.

7.3.3 Channel Response Estimation

The upper part of Figure 7.2 provides a description of how to estimate the unknown log-magnitude spectrum of the channel response. For that, the GMM parameters, $\lambda = \{\mu_m, \sum_m, \pi_m\}$, and the log-magnitude spectrum average, $\overline{\underline{S}}_m(k)$, previously estimated, are used. Similarly to Section 7.3.2, the observed speech signal, $x(n)$, potentially altered by channel effects, is segmented into overlapping frames and the same pre-processing steps are taken prior to extracting the STFT, $X(k,l)$, and the respective RASTA-MFCC coefficients. The likelihood that a feature vector, $c_x(l)$, belongs to the

m-th mixture can be computed as in Eq. 7.4, which leads to a probability, $0 < p_{l,m} < 1$, for each mixture $m = 1, ..., M$. Again, for a given observed speech signal, $x(n)$, a matrix of probabilities $M \times L_x$, $p_{l,m}(c_x(l))$, containing the M mixture probabilities for the L_x available speech frames is obtained. It is important to mention that $\sum_m p_{l,m} = 1$. These probabilities are then used to estimate the clean log-magnitude spectrum average, $\hat{\underline{S}}(k,l)$, of the l-th frame using the weighted average of the clean-speech spectra, $\overline{\underline{S}}_m(k)$, as described below:

$$\hat{\underline{S}}(k,l) = \sum_{m=1}^{M} p_{l,m}(c_x(l))\overline{\underline{S}}_m(k), \forall k. \tag{7.7}$$

Considering $\hat{\underline{S}}(k,l) \approx S(k,l)$, the estimated log-magnitude of the channel response, $\hat{\underline{H}}(k,l)$, can be computed according to Eq. 7.3. Here, $p_{l,m}$ acts as a selection matrix and generates a weighted average spectrum based on the class probabilities and the average spectrum templates of clean speech. Note that the accuracy of the proposed method is sensitive to the number of mixtures. For example, if we consider $M = 1$, the likelihood of a frame belonging to that mixture is $p_{l,m} = 1, \forall l$. Therefore, we conclude from Eq. 7.7:

$$\overline{\underline{S}}_m(k) = \frac{\sum_{l=1}^{L} p_{l,m}(c_x(l))\tilde{\underline{S}}(k,l)}{\sum_{l=1}^{L} p_{l,m}(c_x(l))} = \frac{1}{L}\sum_{l=1}^{L} \tilde{\underline{S}}(k,l). \tag{7.8}$$

Therefore, the accuracy of the channel estimation is affected by how accurate the clean speech log-magnitude spectrum average is obtained, which can be improved by increasing the number of mixtures, allowing the detection of more complex distortions.

7.3.4 Log-magnitude spectrum of the channel response for bonafide and spoofed utterances

The time-frequency representation involved after the estimation of the log-magnitude spectrum of the channel response is given in Figure 7.3. The estimated log-magnitude spectrum of the channel response signal for a bonafide utterance is denoted by Figure 7.3-a, whereas Figure 7.3-b and Figure 7.3-c represent, respectively, the log-magnitude spectrum for the observed speech signal, $\tilde{\underline{X}}(k,l)$, and its clean estimated counterpart, $\hat{\underline{S}}(k,l)$. A spoofed utterance is denoted by Figure 7.3-d, Figure 7.3-e and Figure 7.3-f, respectively, representing the log-magnitude spectrum for the channel

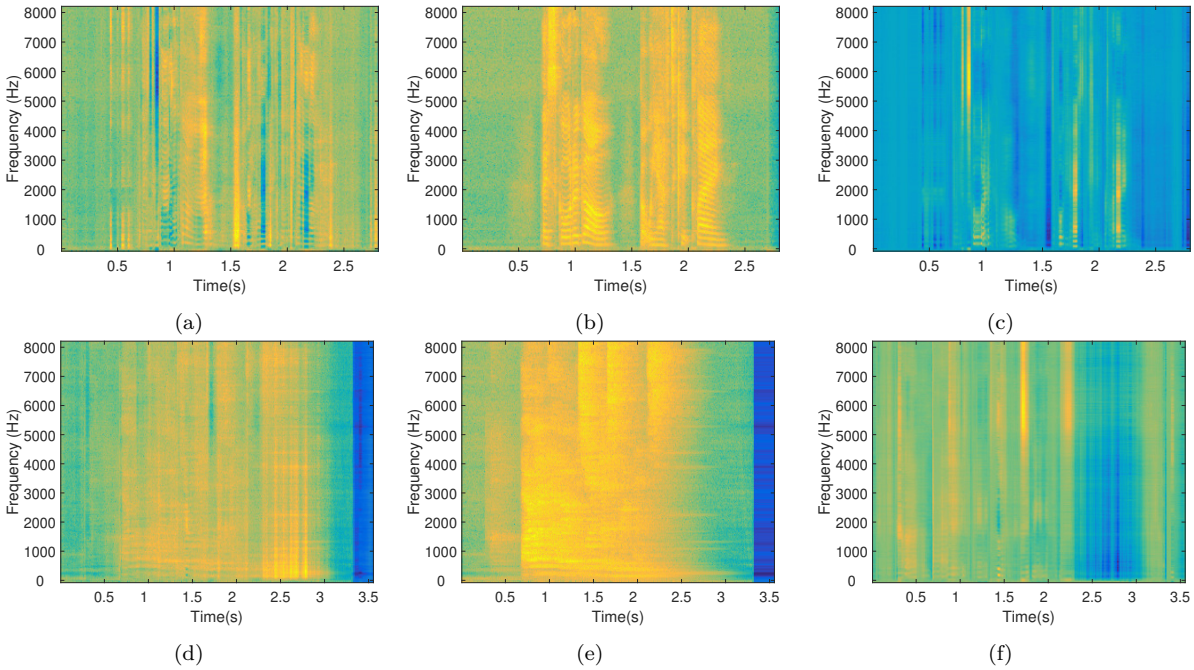


Figure 7.3 – Log-magnitude spectrum of the (a) estimated channel response signal, (b) observed speech signal and the (c) estimated clean speech signal from a bonafide utterance, also the log-magnitude spectrum of the (d) estimated channel response signal, (e) observed signal and (f) estimated clean speech signal from a spoofed utterance.

Table 7.1 – Speech parameterization configuration.

	Configuration
frame length	32 ms (512 samples)
frame step	16 ms (256 samples)
STFT	257 bins
MFCC	13 coef. (with energy)

response signal, the observed speech signal, and the clean speech signal. Compared to the bonafide speech, we can observe that the speech content (i.e., regions with higher energy) for the spoofed utterance is smeared out across frequency and time, which leads to higher average energies for the spoofed speech files. By scrutinizing Figure 7.3-b and Figure 7.3-e, i.e., the bonafide and spoofed utterances, respectively, we can observe that the formants are much less evident for the spectrum representing the spoofed utterance, which is also captured by the log-magnitude spectrum of the estimated channel response in Figure 7.3-a and Figure 7.3-d. These are some of the potential cues to be used by the classifier to distinguish between bonafide and spoofed utterances.

7.3.5 Mel-Frequency Cepstral Coefficients Extraction and RASTA filtering

Mel-frequency cepstral coefficients (MFCCs) are considered the most popular set of features for many tasks involving speech analysis, such as speech and speaker recognition. Prior to their extraction, the speech signal, s(n), is normalized, segmented into frames, and pre-emphasised. In our experiments, the speech signals (sampled at 16 kHz) are first segmented into frames of 32 ms length (i.e., 512 samples), with 16 ms hop-size (i.e., 256 samples). Prior to computing the STFT, $S(k,l)$, the speech signal is pre-emphasized by a filter of coefficient 0.97, which is meant to balance low and high frequency magnitudes. From each frame, the MFCCs are extracted according to:

$$c_n = \sum_{m=1}^{M} [Y_m] cos \frac{\pi n}{M} \left(\ m - \tfrac{1}{2} \ \right), n = 1, 2, 3, ..., N, \tag{7.9}$$

where c_n is the n^{th} mel-cepstral coefficient and Y_m refers to the log-energy of the m^{th} filter. A total of 12 coefficients plus the log energy are attained, leading to a 13-dimensional vector for each frame. As can be seen from (7.9), the MFCC representation is based on a short-term log-power spectrum and a cosine transformation on the nonlinear mel frequency scale. Table 7.1 summarizes the configuration used for extracting the STFT and MFCC.

7.4 Residual neural network

The ResNet architecture adopted in this work is based on the model proposed in [80], and described in Section 2.5. Table 7.2 gives details about the model used. It consists of a 3 $\times$ 3 convolution layer, nine residual blocks, and a Global Average Pooling (GAP) layer followed by a fully connected layer with a sigmoid function. For regularization, we considered L2 with weight decay and spatial dropout [178] with the following rates: 0.001 and 0.3, respectively. The main motivation to use spatial dropout was to mitigate high correlation between feature map activations [179]. Randomly dropping pixels within a feature map has little effect on reducing dependency between feature map activations [180]. An alternative is to drop a larger region as opposed to individual pixels. When spatial dropout is used it prevents the neural network from using nearby pixels to recover information as an entire feature map is dropped with probability p. It is important to mention that although an

Table 7.2 – ResNet architecture adopted in this work.

layer name	output size	layers
conv1	257 x 200	3×3, 16, stride 1
conv2.x	257 x 200	$\begin{bmatrix} 3 \times 3, 16 \\ 3 \times 3, 16 \end{bmatrix} \times 3$
	3×3 max pool, stride 2	
conv3.x	129 x 100	$\begin{bmatrix} 3 \times 3, 32 \\ 3 \times 3, 32 \end{bmatrix} \times 3$
	3×3 max pool, stride 2	
conv4.x	65 x 50	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 3$
	1×1	average pool, 3250-d fc, sigmoid

entire filter might be removed at training time, all filters are present during testing time and with less dependency [74].

7.5 Experimental Setup

In this section, we present the datasets used throughout our experiments, the adopted benchmark features and classifiers, as well as the figure-of-merit.

7.5.1 Database Description

Four datasets are involved in our experiments, all (re)sampled at 16 kHz. The first two correspond to the ASVspoof 2017 and 2019 (Physical Access only) datasets [71, 181]. For the former, we adopted version 2.0, which contains speech files that simulate simple replay spoofing attacks that require no special expertise to be performed, nor special equipment, and thus could be implemented by anyone. It presents 61 distinct replay configurations, which are categorized in three levels, with each level (i.e., low, median and high) representing how difficult it is to detect the attack. These configurations are a combination of acoustic environment (e.g., home, anechoic, office), recording, and playback devices. According to [182], there is a considerable number of unseen replay configurations in the evaluation subset and the conditions used closely represent those observed "in-the-wild." The number

of replay configurations for the training and the development sets is significantly smaller compared to the number of configurations for the evaluation set, which comprises about 57 different replay configurations in total [71].

For the ASVspoof 2019 (Physical Access) dataset, in turn, multiple reverberant acoustic environments were considered while presenting either bonafide or spoofed recordings to the ASV microphone. While this characterizes more complex scenarios as the acoustic configuration is now defined by the combination of three categories of reverberation, three room sizes, and three different speaker-to-microphone distances, the characteristics of the speech signals are more closely related as they were synthetically generated by convolving clean speech samples with synthetic room impulse responses. The dataset presents an additional nine different replay configurations based on three categories of attacker-to-talker recording distances and three categories of loudspeaker quality. Both datasets are disjointly divided into training, development and evaluation sets. The number of bonafide utterances is 5,400 for the training and development sets. The number of spoofed utterances was 24,300 for the development set and twice that amount for the training set. More details about the dataset can be found in [181].

The last two datasets were used solely to train the clean speech model and were not used to train the back-end models; these correspond to the TIMIT database [183] and the noise speech database [115]. The TIMIT database has 630 speakers of eight different American English dialects. The database contains ten utterances recorded from each speaker, totaling approximately 500 utterances within its training set. TIMIT was further complemented with the clean speech files available in [115]. The noise speech database is a clean and noisy parallel speech dataset, developed for the purpose of training speech enhancement algorithms. It contains pairs of clean and noisy speech samples from 28 speakers (14 males and 14 females), all from the same accent region (England), taken from the larger Voice Bank corpus [118].

7.5.2 Benchmark features

Two benchmark features were considered in this work. The primary one is the constant-Q cepstral coefficients (CQCCs). Introduced for spoofing attack detection in [184], these features have been used in the baseline systems in the 2017 and 2019 spoof attack detection challenges [71, 72]. The features refine the time-frequency resolution based on the constant-Q transform (CQT), which will

define the size of the bandwidth frequency, thus making them more perceptually motivated. Note that for the STFT, the quality factor Q_c [185] for the center of the frequency band f_c is defined as:

$$Q_c = \frac{f_c}{\delta_f}, \tag{7.10}$$

with δ_f being the frequency bandwidth. For a fixed width, the quality factor increases together with the center frequency. This is not well aligned with human perception, which is known to have a constant-Q factor between 500 Hz and 20 kHz [185]. Therefore, the CQT was introduced in [186] and later refined in [187].

Prior to extraction, the frequency bins of the CQT must be converted from geometric to linear space as an attempt to emulate the human auditory system [184]. This can be seen as a resampling operation as described in [185]. Widely used for music processing, the CQCCs are the application of the CQT on the cepstrum. To extract the CQCCs, an inverse transformation to the discrete Fourier transform (DFT) must be applied. This process is similar to the discrete cosine transformation used while extracting MFCCs [184]. The CQCCs can be then extracted according to:

$$CQCC(p) = \sum_{l=1}^{L} \log |X^{CQ}(l)|^2 \cos \frac{p(l-\frac{1}{2})\pi}{L}. \tag{7.11}$$

The linear frequency cepstral coefficients (LFCC) are the second baseline feature used in our experiments. Similar to MFCCs, LFCCs are also based on the discrete cosine transform (DCT) computation of the log-magnitude of the filter outputs as described in Eq. (7.9). The main difference from mel-scale frequencies, which give more detail to low frequencies, is that center frequencies of the LFCC are linearly spaced, thus each frequency band provides the same detail [188].

7.5.3 Dimensionality Reduction

Throughout our experiments, principal component analysis (PCA) was used only on the proposed features to reduce dimensionality and improve performance, especially for the GMM-based classifier. Dimensionality reduction is an important step in pattern recognition as in a high dimensional space many variables are interrelated. Thus, by applying PCA, we aim at finding a subspace of lower dimensionality where most of the variation and uncorrelated variables are ordered and kept in

the few first dimensions. In fact, PCA projection maximizes the variance of the projected points [189]. In our experiments, the principal components (PC) are learned from the training data and then the projection matrix is applied on the development and evaluation sets. We tested different feature dimensions. The proposed feature set was reduced from its original dimensionality (i.e., 257 dimensional spectrum) to 32, 64 and 90 dimensions. We found that 64-dimensional feature vectors and 90-dimensional ones are best for the ASVspoof 2017 and ASVspoof 2019 datasets, respectively.

7.5.4 Benchmark (Back-end) Classifier

As a benchmark (back-end) classifier, the score given by the difference of log-likelihoods predicted by two Gaussian mixture models, representing bonafide and spoofed models, was used. This is the same decision process adopted by the ASVspoof 2019 challenge. Note that these GMMs differ from the GMM used in the proposed method to model clean speech spectra. A set of 512 Gaussian components for each of the two GMMs is considered. The two mixture models are diagonal and trained using the *Expectation-Maximization* algorithm independently on genuine and attack utterances. The described score for a given utterance represented by a feature vector $\mathbf{y}$ is defined as:

$$\text{Score}(\mathbf{y}) = \log \frac{P(\mathbf{y}|\lambda_g)}{P(\mathbf{y}|\lambda_s)}, \tag{7.12}$$

where λ_g and λ_s are the GMM parameters for genuine and spoof attack, respectively, and $P(\mathbf{y}|\lambda_g)$ and $P(\mathbf{y}|\lambda_s)$ are the likelihoods of $\mathbf{x}$ predicted by each mixture model. To avoid confusion with the GMM used for the proposed channel response estimator, this step will henceforth be referred to as the "GMM back-end classifier." Note that this back-end approach is used for both the proposed method and the benchmarks. In the proposed method, however, the feature vectors $\mathbf{x}$ correspond to the estimated channel responses, whereas for the benchmarks they are either the CQCC or the LFCC benchmark features.

7.5.5 Figures-of-merit

In biometric security, performance is commonly evaluated using the equal error rate (EER). It requires the computation of the false negative rate (FNR), the false positive rate (FPR) and a threshold. When the rates are equal, the common value is referred to as the equal error rate. The

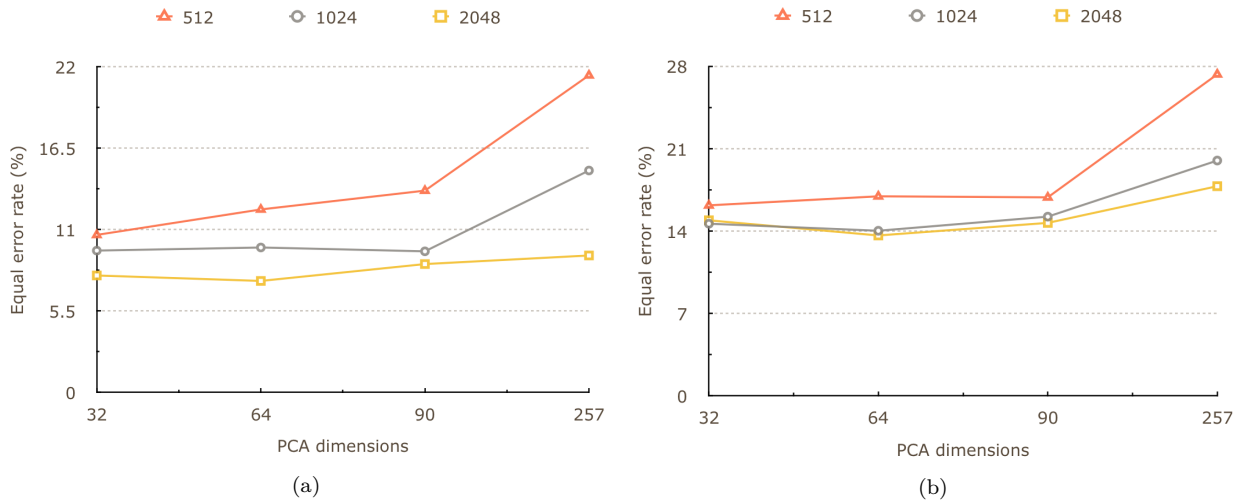


Figure 7.4 – Impact of the number of Gaussian components (i.e., 512, 1024 and 2048) for the GMM of the clean speech model and PCA dimensionality on the performance of the proposed method for the ASVspoof 2017 (a) development and (b) evaluation sets.

value indicates that the proportion of false acceptances is equal to the proportion of false rejections. The EER was the metric used during the ASVspoof 2017 Challenge for performance evaluation and is the adopted evaluation criteria to compare the performance of our proposed system and the benchmarks. The best achieving model is the one that, for a given threshold, provides the lowest EER [190].

A second metric is adopted throughout the experiments. This metric, namely tandem detection cost function (herein referred to as t-DCF), is used to evaluate two combined systems. In the ASVspoof 2019 Challenge, these two systems are the spoof countermeasure and the ASV system (provided by the organisers). Six parameters define the t-DCF: false alarm and miss costs for both systems, and prior probabilities of target and spoof trials. More details about the t-DCF can be found in [175].

7.6 Experimental Results and Discussion

In this section, we describe two experiments along with the respective discussions on the achieved results. The first experiment is performed on the ASVspoof Challenge 2017 dataset, where we compare the performance of the proposed method to the baseline systems. Then, similar experiments are performed on the ASVspoof Challenge 2019 dataset. We also discuss the role of perceptual quality on the performance of our model.

Table 7.3 – Results in terms of EER (%) for replay attack detection on evaluation set of the ASVspoof Challenge 2017. Clean speech model based on 2048 Gaussian components.

Back-end	Features	PCA	Development	Evaluation
GMM	CQCC	-	11.31	27.32
	LFCC	-	7.72	33.04
	MFCC	-	15.83	40.12
	Proposed$_{train}$	64	7.57	13.60
	Proposed$_{train+dev}$	64	-	16.89
ResNet	Proposed$_{train}$	-	-	14.21
ResNet	Proposed$_{train+dev}$	-	-	**11.64**

7.6.1 Experiment I: Performance on ASVspoof 2017

The performance of the proposed method on the ASVspoof 2017 Challenge dataset is presented in Figure 7.4. All the results here are based on the benchmark classifier, i.e., a GMM back-end classifier with 512 components as described in Section 7.5.4. The y-axis gives the EER and the x-axis the number of dimensions after applying PCA. Only a subset of the investigated dimensions are shown for brevity (i.e., 32, 64 and 90, and when no PCA is applied, hence 257). The number of Gaussian components used to train the GMM clean speech model and to estimate the log-magnitude of the channel response included 512, 1024 and 2048. This has direct impact on the accuracy of the channel estimation as discussed in Section 7.3.3. Note also that Figure 7.4-a refers to the results for the development set, while Figure 7.4-b is for the evaluation set.

As can be seen, for the development set the number of Gaussian components has considerable impact on the performance of the proposed method. The impact is not as pronounced for the evaluation set, suggesting that 1024 components may suffice, with some additional gains achieved with 2048 components. Moreover, PCA showed to be an important aspect for the proposed method, especially when lower numbers of Gaussian components were tested. Overall, the PCA dimensionality achieved a "sweet spot" at around 64 components for both the ASVSpoof 2017 evaluation and development sets.

In Table 7.3, we compare our best results with the benchmark solutions. All systems are trained on the benchmark GMM back-end classifier, except for the system listed in the last row, which is based on the ResNet classifier discussed in Section 7.4. For the GMM back-end classifier, results are first reported on the development and evaluation sets separately. As can be seen, in the development set, the proposed method achieved the lowest EER, 7.57%, closely followed by the LFCC, 7.72%.

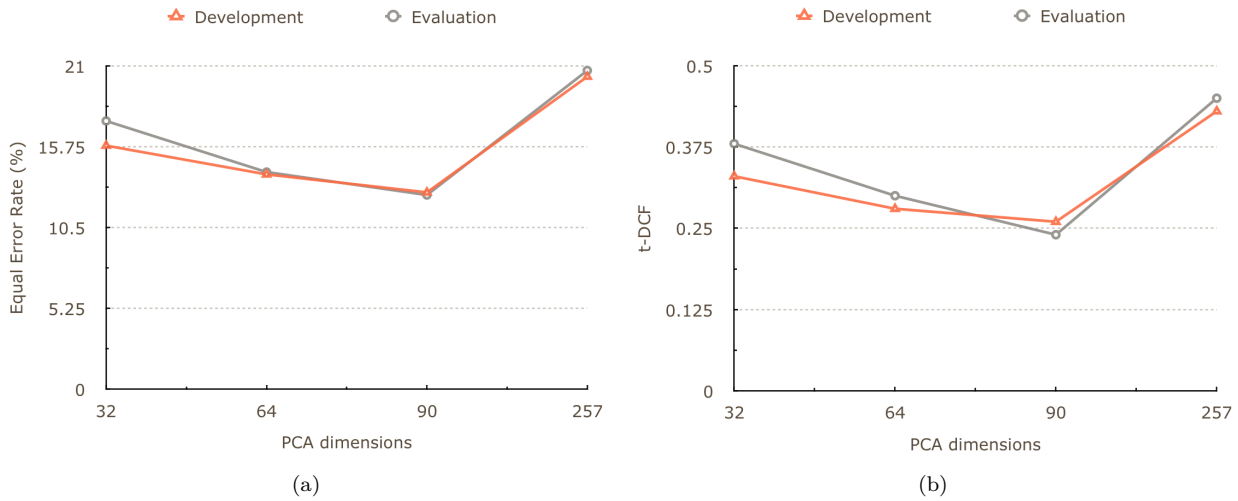


Figure 7.5 – Impact of PCA on the performance of the proposed method for the ASVSpoof 2019 (a) development and (b) evaluation subsets.

In the evaluations set, in turn, the proposed method achieved substantially lower EER, 13.60%, relative to the best baseline system performance, 27.32%, the CQCC features. Lastly, by increasing the training set to include both the training and development sets showed to not be beneficial to the GMM back-end. The proposed ResNet classifier, however, was able to achieve an EER of 11.64%, hence a decrease of 14.4% relative to the results achieved with the proposed method and a GMM back-end classifier when trained with only the training set. In the case of the ResNet, when trained with the extended dataset, 20% of samples are randomly separated to be used as validation set. Then, early stopping is adopted as a criteria to stop training. Hence, training is stopped after no improvement is found in the validation set after a specific number of epochs.

Moreover, it can be seen from the table that the proposed method is the least affected by the differences between the evaluation and development sets. As mentioned previously, the training set is comprised of three replay attack configurations, whereas the development set 10 and the evaluation set 110. The proposed method, for example, is able to achieve an EER of 13.6% on the evaluation set, which is 16.4% lower than what the MFCC-based benchmark achieved on the *development* set, thus showing the advantages of the proposed method to unseen conditions.

7.6.2 Experiment II: Performance on ASVSpoof 2019

In the second experiment, we evaluate the performance of our system on the ASVSpoof Challenge 2019 dataset. In Figure 7.5, the EER and the t-DCF values are reported for different PCA dimensions,

as previously, for the (a) development and (b) evaluation sets. Motivated by the results from the previous experiment, a GMM clean speech model with 2048 mixtures was used. As with the previous experiment, PCA dimensionality reduction was shown to be important, with EER dropping as high as 40% by using 90 components. Unlike the previous experiment, here 90 principal components was shown to be the "sweet spot" for the ASVSpoof 2019 dataset. Moreover, the performance of the proposed system was also shown to be similar for both the evaluation and development sets.

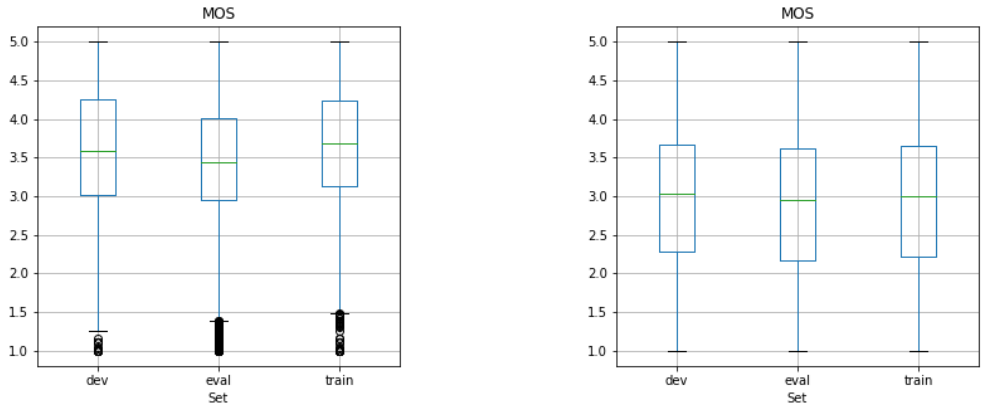
Table 7.4 presents the performance of our proposed method and of the benchmarks on the evaluation set. Here, we explore several different configurations. For the ResNet classifier, all models are trained using both training and development sets. We present two solutions. The first is based on the diagram presented in Figure 7.2, where power normalization is used at training and channel estimation phases. For the second solution, i.e., Proposed2, we removed the power normalization step from the channel estimation phase. Our intuition is that power normalization also removes some of the channel effects from the log-magnitude spectrum and therefore by removing this step, improved channel response estimation can be achieved. Our hypothesis was confirmed for the two back-end being tested. Results went from 0.2481 and 12.62% to 0.2748 and 10.32%, in terms of t-DCF and EER, for the GMM backend. As can be seen, further improvement was attained with the ResNet system combined with the features without power normalization and trained with a larger training set. Overall, a t-DCF of 0.1086 and an EER of 4.26% was achieved, thus a reduction of 55% and 58%, respectively, when compared to the proposed system but with a GMM back-end classifier. Note that no PCA operation was performed for the ResNet. Although we have tested its performance with the same optimal PCA components used to train the GMM, i.e., 90 components, we found no benefits in applying PCA.

7.6.3 Impact of database quality on spoofing detection

As seen from Tables 7.3 and 7.4, the benchmark methods achieved very different results across the two datasets, with EER values on the evaluation set of ASVSpoof 2017 over double the EER attained on the evaluation set of the ASVSpoof 2019 dataset (e.g., 13.24% vs 40.12% for the MFCC benchmark). This impact was almost nonexistent with the proposed method, where an EER of 13.6% was achieved on the 2017 dataset and an EER of 12.62% was achieved on the 2019 set. Notwithstanding, when comparing the ResNet results across the datasets when training and

Table 7.4 – Performance comparison for replay attack detection on the evaluation set of the ASVSpoof Challenge 2019. Clean speech model is based on 2048 Gaussian components.

Backend	Features	t-DCF	EER (%)
GMM	CQCC	0.2454	11.04
	LFCC	0.3017	13.54
	MFCC	0.2729	13.24
	STFT	0.3126	14.28
	$\text{STFT}_{train+dev}$	0.2551	10.83
	Proposed_{train}	0.2481	12.62
	Proposed2_{train}	0.2748	10.32
	$\text{Proposed2}_{train+dev}$	0.2275	10.33
ResNet	$\text{CQCC}_{train+dev}$	0.1256	5.39
	$\text{LFCC}_{train+dev}$	0.1291	5.73
	$\text{MFCC}_{train+dev}$	0.1360	5.70
	$\text{STFT}_{train+dev}$	0.1282	5.10
	$\text{Proposed2}_{train+dev}$	**0.1086**	**4.26**



(a) ASVSpoof 2017 2.0

(b) ASVSpoof 2019

Figure 7.6 – Boxplots providing the overall estimated MOS distribution for (a) ASVSpoof 2017 v2.0 and (b) the ASVSpoof 2019 datasets.

development sets were used for training, the achieved results were almost three times lower on the 2019 dataset than on the 2017 one. We hypothesize that beyond the number of training samples, the quality of the datasets may have an impact on this mismatch. This should be expected, as the former contained "in-the-wild" conditions, whereas the latter contained mostly synthetically-generated reverberant speech. To further investigate this, Figure 7.6 provides the ITU-T Rec. P.563 MOS scores for the 2017 (subplot a) and 2019 (subplot b) datasets, for the training, development, and evaluation sets.

As can be seen, the ASVSpoof 2017 dataset has an average MOS almost 0.5 higher than the ASVSpoof 2019 dataset; has several "outlier" samples of very low quality, especially for the training set; has a lower spread of scores (almost 0.5 MOS); and has an evaluation set with a different quality profile than the training and evaluation sets. Hence, while the higher quality signals may have resulted in the lower EER values reported for the benchmarks in the 2017 development dataset (relative to the evaluation set of ASVSpoof 2019), the mismatch in the quality of the train/development sets to that of the evaluation set is likely the culprit of the large difference seen in the evaluation sets across both datasets. This impact is reduced with the proposed method. In turn, the similarity among all the three subsets in the ASVSpoof 2019 database can explain the similar plots observed in Figure 7.5. The higher quality spread of the 2019 dataset may also explain the need for a higher number of principal components (90 vs. 64).

Moreover, it seems that training neural network models with data across a larger, but more even spread of signal quality (e.g., roughly 1 MOS spread for 2017, from 3-4; and 1.5 for 2019, from 2.1-3.6) can lead to improved spoofing detection. While the ASVSpoof 2019 dataset had samples with an average lower quality, few outlier points existed, thus providing three data subsets with an even quality spread. The mismatch between the train/development and evaluation conditions on the 2017 dataset, as well as the numerous outlier points, likely contributed to the higher EER achieved with ResNet. The differences across the two datasets and the sensitivity seen to the mismatch in quality profiles makes it difficult for cross-database experiments.

7.6.4 Impact of external data and comparisons with Challenge participants

To show that the proposed blind channel response estimation works without prior knowledge of the challenge dataset, which could confer performance advantage to our method, we decided to use external data to train it. However, as no external data was allowed during the 2017 and 2019 challenges, it raises the question as to whether training our clean speech models on external data would provide a large advantage. To verify this claim, we experimented with building clean speech models using only the data provided for the ASVSpoof 2019 challenge. Thus, the clean speech samples from the TIMIT database and from the noise speech database, presented in Section 7.5.1 and used solely to train the proposed method, were replaced by clean speech samples, from genuine speakers, presented in the ASVSpoof 2019 challenge dataset. Once the clean speech model was

trained, the estimation of the log-magnitude spectrum of the channel response was computed as explained in Section 7.3.3. We tested both back-ends: GMM and ResNet. It was found that for the GMM back-end approach, an EER of 15.8% was achieved, thus suggesting some advantage of using *de facto* clean speech data to train the models. Note that performance reported in Figure 7.5-a, EER of 20.71, is much lower. Therefore, as we expected, using prior knowledge of the challenge dataset helps to improve performance of the proposed method. Notwithstanding, no changes in performance were seen for the ResNet configurations.

Relative to the 2019 Challenge participants, it can be seen that the results reported here are in line with the results achieved by the systems based on single feature and single classifier. While the obtained results are lower than those achieved with the ensemble classifier based method proposed by the Challenge winner (t-DCF=0.0096 and EER=0.39%) [**?**], the simplicity of the proposed method based on a single feature type and a single classifier could provide some advantages for "in-the-wild" scenarios, such as robustness against mismatch conditions, as seen in Table 7.3. Moreover, we believe that further improvements should be achieved with the combination of different classifiers and features. Nevertheless, the exploration of fusion strategies is left for future work.

7.7 Conclusions

In this Chapter, we proposed the use of blind channel response estimation as a new approach for replay attack detection. Our assumption is that the nuances of the acoustic ambience, microphones and playback devices present in the spectrum contain enough information to distinguish between a bonafide and a spoofed attack. We explored a baseline back-end based on Gaussian mixture models, as well as a deep residual neural network classifier. Experiments on the ASVSpoof 2017 and the ASVSpoof 2019 Challenge datasets show the proposed methods outperforming several challenge baseline systems as well as providing improved robustness against train/evaluation set mismatch. A discussion on the effects of signal quality on spoofing detection performance is also reported, thus providing some preliminary insights on how to best train models for the task at hand.

Conclusions and Future Work

This thesis aimed at increasing human experience while using speech-based technologies. Its main contributions, limitations and future work are described next.

8.1 Contribution and Results

This thesis contributes with new approaches for instrumentally assessing the quality of a speech signal and also to increase the reliability of speech-based technologies. In the first part of the thesis, a new full-reference instrumental quality based on the i-vector framework is proposed. Results showed the proposed method providing performance in line with benchmark algorithms, bypassing the need for time alignment between reference and processed signals. A no-reference instrumental quality measure is also proposed in Chapter 4. Experimental results showed the proposed method outperforming several non-intrusive benchmark algorithms, and achieving accuracy aligned with intrusive algorithms, without the need for a clean reference signal. In the second part of the thesis, focus was placed on improving the performance and reliability of speech applications in-the-wild. First, an environment-robust speech emotion recognition system is proposed. A feature pooling scheme based on modulation spectral features and that combines information from neighbouring frames is proposed in Chapter 5. This pooling approach significantly contributed to boost SER performance in the presence of noise end reverberation. In Chapter 6, an ASV system meant to be robust towards affective speech is proposed. Results show the proposed method outperforming the traditional i-vector baseline by as much as 15%. Finally, a countermeasure for physical access attacks is proposed. Our method is based on a front-end blind channel response estimation combined with a

deep residual neural network classifier. Experiments on the ASVspoof 2017 and the ASVspoof 2019 Challenge datasets show the proposed methods outperforming several challenge baseline systems.

8.2 Limitations and Future Work

The results obtained in the first part of this work were mainly based on distortions caused by background noise and reverberation and on a handful number of datasets. Although, highly correlated with subjective scores, knowing the performance of the proposed models on a larger number of distortions is desirable, as well as how it would perform on additional speech quality datasets. Also, the impact of network impairments was not covered thus the behaviour of the proposed models under these conditions is still unknown. This makes it difficult to fully compare the performance of our method with standardized methods, such as PESQ and POLQA, which were tested on a larger range of distortions. Another important analysis that was not covered in this thesis was the impact of the bandwidth on the proposed i-vector framework. Assessment is based on wide-band signals and the behavior of our model for different bandwidths is still unknown. Instrumental measures such as PESQ and POLQA can operate in different modes, each optimized for the respective speech signal bandwidth.

In summary, regarding the first part of this thesis, we intend to:

1. Evaluate and optimize the proposed speech quality estimator based on i-vector on a large number of distortions as well as on additional speech quality datasets
2. Test and optimize the proposed i-vector framework on different bandwidth modes

In the second part of this thesis, the problem of reliability is addressed. Although speech was the main focus of this thesis, for the SER task discussed in Chapter 5, multi-modal models are often able to confer more reliability to emotion recognition systems. Most of the emotion recognition challenges, for example, include not just audio but also image and sometimes text data. Thus, improved results can be obtained combining the proposed model with other modalities. In Chapter 6, the problem of emotional ASV using the traditional i-vector framework is addressed. More recent approaches based on end-to-end ASV and on embeddings based on more recent deep neural network architectures were not explored. Considering the new advances in deep learning, such approaches are expected

to lead to improved results. As more emotional datasets become available, alternate ASV systems based on more recent deep learning architectures may become a viable alternative. For the task of physical access attacks, the best performance in the literature is based on ensembled models; these were not explored in our experiments. Certainly, further improvements can be achieved by fusing such systems with the solutions proposed in Chapter 7.

Therefore, regarding the second part of this thesis, we plan to:

1. Explore multi-modal approaches to boost SER performance and reliability
2. Investigate recent end-to-end ASV approaches
3. Explore ensembled methods as a countermeasure to physical access attacks

Lastly, this work left unexplored some recent deep neural network architectures, such as variational autoencoders (VAE) and adversarial neural networks. Future work should explore their use for quality measurement and reliability improvement tasks. Lastly, future work should explore the applicability of so-called x-vectors, as they have recently shown to outperform i-vectors for speech recognition tasks.

Bibliography

[1] S. Möller. *Assessment and prediction of speech quality in telecommunications*. Springer Science & Business Media, 2012.

[2] L.F. Gallardo. *Human and automatic speaker recognition over telecommunication channels*. Springer, 2015.

[3] A. Raake. *Speech quality of VoIP*. Wiley Online Library, 2006.

[4] S. Möller and et al. Speech quality estimation: Models and trends. *IEEE Signal Processing Magazine*, 28(6):18–28, 2011.

[5] T.H. Falk and W.-Y. Chan. Temporal dynamics for blind measurement of room acoustical parameters. *IEEE Transactions on Instrumentation and Measurement*, 59(4):978–989, 2010.

[6] K. Izdebski. *Emotions in the human voice, volume 3: culture and perception*, volume 3. Plural Publishing, 2008.

[7] C.S. Peirce. *Collected papers of charles sanders peirce*, volume 2. Harvard University Press, 1960.

[8] J.G. Beerends and et al. Subjective and objective assessment of full bandwidth speech quality. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:440–449, 2019.

[9] S. Pascual, A. Bonafonte, and J. Serra. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017.

[10] S. Chakrabarty and E.A.P. Habets. Time–frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(4):787–799, 2019.

[11] M.T. Vega and et al. A review of predictive quality of experience management in video streaming services. *IEEE Transactions on Broadcasting*, 64(2):432–445, 2018.

[12] C. Sloan and et al. Objective assessment of perceptual audio quality using visqolaudio. *IEEE Transactions on Broadcasting*, 63(4):693–705, 2017.

[13] M. Chinen and et al. Visqol v3: An open source production ready objective speech and audio metric. *arXiv preprint arXiv:2004.09584*, 2020.

[14] ITU-T. Recommendation G.113: Transmission impairments due to speech processing, February 2007.

[15] P. Gastaldo, R. Zunino, and J. Redi. Supporting visual quality assessment with machine learning. *EURASIP Journal on Image and Video Processing*, 2013(1):54, 2013.

[16] B. Cauchi and et al. Perceptual and instrumental evaluation of the perceived level of reverberation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 629–633. IEEE, 2016.

[17] ITU-T. Recommendation P.800: Methods for subjectiuve determination of transmission quality, February 1998.

[18] A.R. Avila and et al. Performance comparison of intrusive and non-intrusive instrumental quality measures for enhanced speech. In *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 1–5. IEEE, 2016.

[19] D. Wu and et al. Millimeter-wave multimedia communications: challenges, methodology, and applications. *IEEE Communications Magazine*, 53(1):232–238, 2015.

[20] C. Bello and et al. From speech quality measures to speaker recognition performance. In *Iberoamerican Congress on Pattern Recognition*, pages 199–206. Springer, 2014.

[21] D. Garcia-Romero and et al. Using quality measures for multilevel speaker recognition. *Computer Speech & Language*, 20(2-3):192–209, 2006.

[22] Anderson R Avila, Md Jahangir Alam, Douglas D O'Shaughnessy, and Tiago H Falk. Investigating speech enhancement and perceptual quality for speech emotion recognition. In *INTERSPEECH*, pages 3663–3667, 2018.

[23] A.R. Avila and et al. Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild. *IEEE Transactions on Affective Computing*, 2018.

[24] V. Roto and et al. User experience white paper: Bringing clarity to the concept of user experience. In *Dagstuhl Seminar on Demarcating User Experience*, page 12, 2011.

[25] D. Snyder and et al. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE, 2018.

[26] J. Villalba and et al. Analysis of speech quality measures for the task of estimating the reliability of speaker verification decisions. *Speech Communication*, 78:42–61, 2016.

[27] H. Gamper and et al. Intrusive and non-intrusive perceptual speech quality assessment using a convolutional neural network. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 85–89. IEEE, 2019.

[28] B. Cauchi and et al. Non-intrusive speech quality prediction using modulation energies and lstm-network. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 27(7):1151–1163, 2019.

[29] A.R. Avila and et al. Intrusive quality measurement of noisy and enhanced speech based on i-vector similarity. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–5. IEEE, 2019.

[30] N. Dehak and et al. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2010.

[31] D. Garcia-Romero, X. Zhou, and C.Y. Espy-Wilson. Multicondition training of gaussian plda models in i-vector space for noise and reverberation robust speaker recognition. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4257–4260. IEEE, 2012.

[32] N. Dehak and et al. Cosine similarity scoring without score normalization techniques. In *Odyssey*, page 15, 2010.

[33] C. Jin and R. Kubichek. Vector quantization techniques for output-based objective speech quality. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 491–494. IEEE, 1996.

[34] B. Schuller and et al. The interspeech 2009 emotion challenge. In *Interspeech*, pages 312–315, 2009.

[35] B. Schuller, S. Steidl, and A. Batliner. The interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[36] A. Dhall and et al. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the ACM International Conference on Multimodal Interaction*, pages 423–426, 2015.

[37] A. Dhall. Emotiw 2019: Automatic emotion, engagement and cohesion prediction tasks. In *2019 International Conference on Multimodal Interaction*, pages 546–550, 2019.

[38] M. Valstar and et al. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 3–10. ACM, 2016.

[39] F. et al. Ringeval. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 3–9. ACM, 2017.

[40] S. Parthasarathy and C. Busso. Predicting speaker recognition reliability by considering emotional content. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 434–439. IEEE, 2017.

[41] S. Parthasarathy and et al. A study of speaker verification performance with expressive speech. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5540–5544. IEEE, 2017.

[42] S. Möller and A. Raake. *Quality of experience: advanced concepts, applications and methods*. Springer, 2014.

[43] ITUTP Recommendation. Itu-t recommendation p.10/g.100, "vocabulary for performance and quality of service. amendment 2: New definitions for inclusion in recommendation itu-t p.10/g.100. *Vocabulary for performance and quality of service*, 2008.

[44] K. Brunnström and et al. Qualinet white paper on definitions of quality of experience. Technical report, 2013.

[45] S. Baraković and L. Skorin-Kapov. Survey and challenges of qoe management issues in wireless networks. *Journal of Computer Networks and Communications*, 2013, 2013.

[46] ITU-T. Single-ended method for objective speech quality assessment in narrow-band telphony applications, 2004.

[47] T.H. Falk and et al. Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools. *IEEE signal processing magazine*, 32(2):114–124, 2015.

[48] A.R. Avila and et al. Non-intrusive speech quality assessment using neural networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 631–635. IEEE, 2019.

[49] M. H. Soni and H. A. Patil. Novel deep autoencoder features for non-intrusive speech quality assessment. In *Signal Processing Conference (EUSIPCO), 2016 24th European*, pages 2315–2319. IEEE, 2016.

[50] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[51] G. Hinton and et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[52] A Krizhevsky, I. Sutskever, and G Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[53] N. Bostrom and E. Yudkowsky. The ethics of artificial intelligence. *The Cambridge handbook of artificial intelligence*, pages 316–334, 2014.

[54] P. Rosalind. Affective computing: challenges. *International Journal of Human-Computer Studies*, 59(1):55–64, 2003.

[55] M.H. Rahman and et al. Improving plda speaker verification performance using domain mismatch compensation techniques. *Computer Speech & Language*, 47:240–258, 2018.

[56] A. Misra and J.H.L Hansen. Modelling and compensation for language mismatch in speaker verification. *Speech Communication*, 96:58–66, 2018.

[57] W.B. Kheder and et al. Fast i-vector denoising using map estimation and a noise distributions database for robust speaker recognition. *Computer Speech & Language*, 45:104–122, 2017.

[58] A. Avila and et al. Investigating the use of modulation spectral features within an i-vector framework for far-field automatic speaker verification. In *Telecommunications Symposium (ITS), 2014 International*, pages 1–5. IEEE, 2014.

[59] M. Sarria-Paja and T.H Falk. Fusion of bottleneck, spectral and modulation spectral features for improved speaker verification of neutral and whispered speech. *Speech Communication*, 102:78–86, 2018.

[60] J. Novoa and et al. Robustness over time-varying channels in dnn-hmm asr based human-robot interaction. In *INTERSPEECH*, pages 839–843, 2017.

[61] A. Avila and et al. The effect of speech rate on automatic speaker verification: a comparative analysis of gmm-ubm and i-vector based methods. In *12th Audio Engineering Conference (AES-Brazil)*, 2014.

[62] Z. Wu, D. Li, and Y. Yang. Rules based feature modification for affective speaker recognition. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE, 2006.

[63] M. Schröder. Emotional speech synthesis: A review. In *Seventh European Conference on Speech Communication and Technology*, 2001.

[64] I. Shahin. Speaker verification in emotional talking environments based on three-stage framework. In *2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA)*, pages 1–5. IEEE, 2017.

[65] R. Pappagari and et al. x-vectors meet emotions: A study on dependencies between emotion and speaker recognition. *arXiv preprint arXiv:2002.05039*, 2020.

[66] Zhizheng Wu and et al. Spoofing and countermeasures for speaker verification: A survey. *speech communication*, 66:130–153, 2015.

[67] J.S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

[68] Aware.com. Biometrics software. https://www.aware.com/wp-content/uploads/2017/09/BM_product_guide_0917_email-3.pdf, 2017. [Online; accessed 23-August-2019].

[69] Biometricupdate.com. Mobile biometric applications. `https://www.biometricupdate.com/wp-content/uploads/2017/03/special-report-mobile-biometric-applications.pdf`, 2088. [Online; accessed 20-March-2018].

[70] Z. Wu and et al. Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[71] T. Kinnunen and et al. The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. 2017.

[72] http://www.asvspoof.org/. Asvspoof 2019: Automatic speaker verification spoofing and countermeasures challenge evaluation plan. `http://www.asvspoof.org/`, 2019. [Online; accessed 20-March-2018].

[73] H. Dinkel and et al. End-to-end spoofing detection with raw waveform cldnns. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4860–4864. IEEE, 2017.

[74] F. Tom, M. Jain, and P. Dey. End-to-end audio replay attack detection using deep convolutional networks with attention. In *Interspeech*, pages 681–685, 2018.

[75] G. Suthokumar and et al. Modulation dynamic features for the detection of replay attacks. In *Interspeech*, pages 691–695, 2018.

[76] J. Gałka, M. Grzywacz, and R. Samborski. Playback attack detection for text-dependent speaker verification over telephone channels. *Speech Communication*, 67:143–153, 2015.

[77] C. Hanilçi. Features and classifiers for replay spoofing attack detection. In *10th International Conference on Electrical and Electronics Engineering (ELECO), 2017*, pages 1187–1191. IEEE, 2017.

[78] C. Lai and et al. Attentive filtering networks for audio replay attack detection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6316–6320. IEEE, 2019.

[79] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[80] H. Kaiming and et al. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[81] P. Nagarsheth and et al. Replay attack detection using dnn for channel discrimination. In *Interspeech*, pages 97–101, 2017.

[82] A.R Avila and et al. Blind channel response estimation for replay attack detection. *Proc. Interspeech*, pages 2893–2897, 2019.

[83] J. Hansen and T. Hasan. Speaker recognition by machines and humans: A tutorial review. *IEEE Signal processing magazine*, 32(6):74–99, 2015.

[84] Sebastian Möller and Alexander Raake. Telephone speech quality prediction: towards network planning and monitoring models for modern network scenarios. *Speech Communication*, 38(1-2):47–75, 2002.

[85] ITU-T. Recommendation P.862: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, February 2001.

[86] A. Rix, J. Beerends, M. Hollier, and A. Hekstra. Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs. pages 749–752, 2001.

[87] ITU-T. Recommendation P.863: Perceptual objective listening quality assessment: An advanced objective perceptual method for end-to-end listening speech quality evaluation of fixed, mobile, and IP-based networks and speech codecs covering narrowband, wideband, and super-wideband signals, January 2011.

[88] J. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl. Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part ii: Perceptual model. *Audio Eng. Soc.*, 61(6), 2013.

[89] Jianfen Ma, Yi Hu, and Philipos C. Loizou. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. 125(5):3387–3405, May 2009.

[90] JH Janssen. A method for the calculation of the speech intelligibility under conditions of reverberation and noise. *Acta Acustica united with Acustica*, 7(5):305–310, 1957.

[91] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. pages 4214–4217, 2010.

[92] L. Malfait, J. Berger, and M. Kastner. P.563 - the ITU-T standard for single-ended speech quality assessment. 14(6):1924–1934, 2006.

[93] ITU-T. Subjective performance assessment of telephone-band and wideband digital codecs, 1996.

[94] T.H. Falk, C. Zheng, and W.Y. Chan. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1766–1774, 2010.

[95] J. F. Santos, M. Senoussaoui, and T. H. Falk. An improved non-intrusive intelligibility metric for noisy and reverberant speech. pages 55–59, September 2014.

[96] J.L. Hansen and L.M. Levent M. Robust feature-estimation and objective quality assessment for noisy speech recognition using the credit card corpus. *IEEE Transactions on speech and audio processing*, 3(3):169–184, 1995.

[97] S. Wu, T.H. Falk, and W.-Y. Chan. Automatic speech emotion recognition using modulation spectral features. *Speech communication*, 53(5):768–785, 2011.

[98] T. Falk and W.Y. Chan. Modulation spectral features for robust far-field speaker identification. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(1):90–100, 2009.

[99] M. Slaney and et al. An efficient implementation of the patterson-holdsworth auditory filter bank. *Apple Computer, Perception Group, Tech. Rep*, 35:8, 1993.

[100] S.D. Ewert and T. Dau. Characterizing frequency selectivity for envelope fluctuations. *Journal of the Acoustical Society of America*, 108(3):1181–1196, 2000.

[101] 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*, 15(4):1435–1447, 2007.

[102] P. Kenny, G. Boulianne, and P. Dumouchel. Eigenvoice modeling with sparse training data. *IEEE transactions on speech and audio processing*, 13(3):345–354, 2005.

[103] D. Garcia-Romero and C. Y. Espy-Wilson. Analysis of i-vector length normalization in speaker recognition systems. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[104] A. Jain, J. Mao, and M. Mohiuddin. Artificial neural networks: A tutorial. *Computer*, 29(3):31–44, 1996.

[105] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[106] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. *ICML (3)*, 28:1310–1318, 2013.

[107] A. Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[108] K. He and et al. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016.

[109] S. Shum and et al. Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification. In *Odyssey*, page 16, 2010.

[110] National Institute of Standards and Technology (NIST). *Cosine Distance*, accessed July 23, 2020. `https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/cosdist.htm`.

[111] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov), 2008.

[112] T. Halmrast. Sound coloration from (very) early reflections. *Journal of the Acoustical Society of America*, 109(5):2303, 2001.

[113] Joyce W.B. Sabine's reverberation time and ergodic auditoriums. *The Journal of the Acoustical Society of America*, 58(3):643–655, 1975.

[114] S. Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[115] C. Valentini-Botinhao and et al. Noisy speech database for training speech enhancement algorithms and tts models. *University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR)*, 2017.

[116] J. Santos and T.H. Falk. Towards the development of a non-intrusive objective quality measure for dnn-enhanced speech. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2019.

[117] Y. Hu and P. C. Loizou. Subjective comparison and evaluation of speech enhancement algorithms. *Speech Communication*, 49(7-8):588–601, 2007.

[118] C. Veaux, J. Yamagishi, and S. King. The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 1–4. IEEE, 2013.

[119] A. Varga and H.J.M 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*, 12(3):247–251, 1993.

[120] E.A. Lehmann and A.M. Johansson. Diffuse reverberation model for efficient image-source simulation of room impulse responses. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1429–1439, 2009.

[121] H. Hirsch and D. Pearce. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[122] J.F. Santos and T.H. Falk. Speech dereverberation with context-aware recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(7):1236–1246, 2018.

[123] D.S. Williamson and D. Wang. Time-frequency masking in the complex domain for speech dereverberation and denoising. *IEEE/ACM transactions on audio, speech, and language processing*, 25(7):1492–1501, 2017.

[124] B. Wu and et al. A reverberation-time-aware approach to speech dereverberation based on deep neural networks. *IEEE/ACM transactions on audio, speech, and language processing*, 25(1):102–111, 2016.

[125] Y. Hu and P.C. Loizou. Speech enhancement based on wavelet thresholding the multitaper spectrum. *IEEE transactions on Speech and Audio processing*, 12(1):59–67, 2004.

[126] D.E. Tsoukalas, J.N. Mourjopoulos, and G. Kokkinakis. Speech enhancement based on audible noise suppression. *IEEE Transactions on Speech and Audio Processing*, 5(6):497–514, 1997.

[127] ITU-T. Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs, November 2007.

[128] H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE transactions on speech and audio processing*, 2(4):578–589, 1994.

[129] T.H. Falk and W.Y. Chan. Single-ended speech quality measurement using machine learning methods. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1935–1947, 2006.

[130] N.D Gaubitch, M. Brookes, and A.A Naylor. Blind channel magnitude response estimation in speech using spectrum classification. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2162–2171, 2013.

[131] B Series. Recommendation itu-r bs. 1534-3 method for the subjective assessment of intermediate quality level of audio systems. *International Telecommunication Union Radio Communication Assembly*, 2014.

[132] M. Schoeffler and et al. webmushra—a comprehensive framework for web-based listening tests. *Journal of Open Research Software*, 6(1), 2018.

[133] S.W. Fu and et al. Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm. *arXiv preprint arXiv:1808.05344*, 2018.

[134] A.W. Rix. Comparison between subjective listening quality and p. 862 pesq score. *Proc. Measurement of Speech and Audio Quality in Networks (MESAQIN'03), Prague, Czech Republic*, 2003.

[135] M.V. Shcherbakov and et al. A survey of forecast error measures. *World Applied Sciences Journal*, 24(24):171–176, 2013.

[136] ITU-T. Mapping function for transforming p.862 raw result scores to mos-lq, 2003.

[137] P. Kenny and et al. A study of interspeaker variability in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5):980–988, 2008.

[138] ITU-T Recommendation P.835. Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm. *International Telecommunication Union, Geneva*, 2003.

[139] B. Cauchi, I. Kodrasi, R. Rehr, S. Gerlach, A. Jukić, T. Gerkmann, S. Doclo, and S. Goetze. Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech. 2015:1–12, July 2015.

[140] J. Thiemann and et al. Speech enhancement for multimicrophone binaural hearing aids aiming to preserve the spatial auditory scene. *EURASIP Journal on Advances in Signal Processing*, 2016(1):12, 2016.

[141] H. Gunes. Automatic, dimensional and continuous emotion recognition. 2010.

[142] J. Kim. *Bimodal emotion recognition using speech and physiological changes*. Citeseer, 2007.

[143] F.J. Fraga and et al. Towards an eeg-based biomarker for alzheimer's disease: improving amplitude modulation analysis features. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1207–1211. IEEE, 2013.

[144] F. Ringeval and et al. Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pages 1–8, 2013.

[145] J. Thiemann, N. Ito, and E. Vincent. The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings. In *Proceedings of Meetings on Acoustics ICA2013*, volume 19, page 035081. ASA, 2013.

[146] M. Jeub, M. Schafer, and P. Vary. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *16th International Conference on Digital Signal Processing*, pages 1–5. IEEE, 2009.

[147] N. Srivastava and et al. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[148] G. Hinton and et al. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[149] F. Eyben and et al. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. ACM, 2013.

[150] V. Vapnik and et al. Support vector method for function approximation, regression estimation, and signal processing. *Advances in Neural Information Processing Systems*, pages 281–287, 1997.

[151] I. Lawrence and K. Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.

[152] Z. Zhang and et al. Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with lstm neural networks. In *Proceedings of the d17th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3593–3597, 2016.

[153] F. Weninger and et al. Recognition of nonprototypical emotions in reverberated and noisy speech by nonnegative matrix factorization. *EURASIP Journal on Advances in Signal Processing*, 2011(1):838790, 2011.

[154] J. Pohjalainen and et al. Spectral and cepstral audio noise reduction techniques in speech emotion recognition. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 670–674. ACM, 2016.

[155] S. Braun and et al. Late reverberation psd estimation for single-channel dereverberation using relative convolutive transfer functions. In *Proceedings of the IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 1–5, 2016.

[156] F. Burkhardt and et al. A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520, 2005.

[157] S.R Livingstone and F.A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

[158] J. Hansen and et al. Getting started with susas: a speech under simulated and actual stress database. In *Eurospeech*, volume 97, pages 1743–46, 1997.

[159] C. Busso and et al. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80, 2017.

[160] W.B. Cannon. *Bodily changes in pain, hunger, fear, and rage: An account of recent researches into the function of emotional excitement*. D. Appleton, 1916.

[161] C.E. Williams and K.N. Stevens. Emotions and speech: Some acoustical correlates. *The Journal of the Acoustical Society of America*, 52(4B):1238–1250, 1972.

[162] R.W. Levenson and et al. Emotion, physiology, and expression in old age. *Psychology and aging*, 6(1):28, 1991.

[163] P. Ekman, R.. Levenson, and W.V. Friesen. Autonomic nervous system activity distinguishes among emotions. *science*, 221(4616):1208–1210, 1983.

[164] C. Collet and et al. Autonomic nervous system response patterns specificity to basic emotions. *Journal of the autonomic nervous system*, 62(1-2):45–57, 1997.

[165] R. Li-Chern R.L Pan and J.K Li. A noninvasive parametric evaluation of stress effects on global cardiovascular function. *Cardiovascular Engineering*, 7(2):74–80, 2007.

[166] S.D. Kreibig. Autonomic nervous system activity in emotion: A review. *Biological psychology*, 84(3):394–421, 2010.

[167] P. Rajasekaran, G. Doddington, and J. Picone. Recognition of speech under stress and in noise. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.*, volume 11, pages 733–736. IEEE, 1986.

[168] A. Kappas, U. Hess, and K.R. Scherer. Voice and emotion. *Fundamentals of nonverbal behavior*, 200, 1991.

[169] F. Fairbanks and W. Pronovost. An experimental study of the pitch characteristics of the voice during the expression of emotion. *Communications Monographs*, 6(1):87–104, 1939.

[170] G. Fairbanks and L.W. Hoaglin. An experimental study of the durational characteristics of the voice during the expression of emotion. *Communications Monographs*, 8(1):85–90, 1941.

[171] B. Paaßen and et al. Expectation maximization transfer learning and its application for bionic hand prostheses. *Neurocomputing*, 298:122–133, 2018.

[172] N.D. Gaubitch and et al. Single-microphone blind channel identification in speech using spectrum classification. In *2011 19th European Signal Processing Conference*, pages 1748–1751. IEEE, 2011.

[173] S.J.D. Prince and J.H. Elder. Probabilistic linear discriminant analysis for inferences about identity. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[174] D.A.V. Leeuwen and N. Brümmer. An introduction to application-independent evaluation of speaker recognition systems. In *Speaker classification I*, pages 330–353. Springer, 2007.

[175] T. Kinnunen and et al. t-dcf: a detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification. *arXiv preprint arXiv:1804.09618*, 2018.

[176] Brümmer N and J.D. Preez. Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2-3):230–275, 2006.

[177] W. Rao and M.W Mak. Alleviating the small sample-size problem in i-vector based speaker verification. In *2012 8th International Symposium on Chinese Spoken Language Processing*, pages 335–339. IEEE, 2012.

[178] J. Tompson and et al. Efficient object localization using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 648–656, 2015.

[179] S. Park and N. Kwak. Analysis on the dropout effect in convolutional neural networks. In *Asian conference on computer vision*, pages 189–204. Springer, 2016.

[180] A. Labach, H. Salehinejad, and S. Valaee. Survey of dropout methods for deep neural networks. *arXiv preprint arXiv:1904.13310*, 2019.

[181] M. Todisco and et al. Asvspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*, 2019.

[182] H. Delgado and et al. Asvspoof 2017 version 2.0: meta-data analysis and baseline enhancements. 2018.

[183] J.S. Garofolo. Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993*, 1993.

[184] M. Todisco, H. Delgado, and N. Evans. A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients. In *Speaker Odyssey Workshop, Bilbao, Spain*, volume 25, pages 249–252, 2016.

[185] M. Todisco, H. Delgado, and N. Evans. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language*, 45:516–535, 2017.

[186] J. Youngberg and S. Boll. Constant-Q signal analysis and synthesis. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'78.*, volume 3, pages 375–378. IEEE, 1978.

[187] K. L. Kashima and B. Mont-Reynaud. The bounded-Q approach to time-varying spectral analysis. *Dept. of Music, Stanford Univ., Tech. Rep. STAN-M-28*, 1985.

[188] D.A. Reynolds. Experimental evaluation of features for robust speaker identification. *IEEE Transactions on Speech and Audio Processing*, 2(4):639–643, 1994.

[189] I. Jolliffe. *Principal component analysis*. Springer, 2011.

[190] I. Chingovska and et al. Evaluation methodologies for biometric presentation attack detection. In *Handbook of Biometric Anti-Spoofing*, pages 457–480. Springer, 2019.