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

Affective Human-Machine Interfaces: Towards Multi-lingual, Environment-Robust Emotion Detection from Speech

By

Shruti Rajendra Kshirsagar

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

Evaluation Committee

Internal evaluator and committee president:	Prof. Douglas O'Shaughnessy
External evaluator 1:	Prof. Alessandro Lameiras Koerich
External evaluator 2:	Prof. Kun Qian
Research advisor:	Prof. Tiago H. Falk

Acknowledgements

First and foremost, my deepest gratitude to my supervisor Dr. Tiago H. Falk for the support, opportunities and challenges provided throughout the PhD program, his role as supervisor surpasses all expectations, as such, his mentorship and attitude have been vital for the completion of this thesis, and for my personal development, it is great pleasure to work with him.

I would like also to thank all the members and ex-members of the MuSAE lab for their incredible help, advice, company, and for the fun we have had while discussing ideas, running experiments, rushing for deadlines, debating and so on. In particular, I would like to thank Raymundo Cassani, Abhishek Tiwari, Anderson Avila, Isabela, Joao, Belmir, Oliver, Marilia for their support and camaraderie.

Without my family and friends, I would not be the person that I am today, I would like to largely thank my family, my parents Rajendra Kshirsagar and Shobha Kshirsagar, and my brother Nikhil Kshirsagar; who despite the geographically 12,068 km distance apart but are always in my mind, thanks for giving your love, blessings, support, motivation, advice and care. Lastly, I would like to thanks my Fiance Nikhil Paranjpe for his continuous support and love during my PhD time.

There are not words to describe how grateful I am with every single person who helped to make this work possible. This thesis dedicated to all of you.

Abstract

Recognizing user emotions from speech is something humans do naturally when communicating with each other face-to-face. Machines, however, are still far from perfect when it comes to detecting user emotions, particularly in realistic settings, where ambient noise and room acoustics hamper the signal quality. Moreover, it is common for users from different cultures and speakers of different languages to interact with the same system. Such multi-lingual settings are extremely challenging and result in significant performance drops when a system is e.g., trained on one language and tested on another. As such, cross-cultural emotion recognition (ER) has become an emerging topic in affective computing to improve the generalization power of human-machine interfaces (HMIs). Most of the research conducted to date within the multi-modal emotion recognition domain has relied on controlled environments without any background distractions, such as ambient noise and/or reverberation. However, in real-life situations, signals are often corrupted by environmental factors, which deteriorate ER system performance. Most research has also relied on mono-lingual settings, in which ER models for a particular language are developed and tested for the same language. Once tested in a different language, significant performance drops are observed. This is where this Ph.D. research comes in. In particular, we present new methods and tools to enable not only noise-robustness, but also multi-lingual capabilities for emerging affective HMIs. Here, focus was placed on emotion recognition "in the wild." As will be detailed herein, these achievements came from a combination of new features, multimodal feature fusion, as well as machine learning and domain adaptation schemes. Particularly, in this doctoral thesis, we present the steps towards the development of emotion recognition models for data collected in real-time conditions. To achieve this goal, three main tools have been explored. These include (i) quality aware bag of modulation features, (ii) building domain adaptation schemes, and (iii) using multimodal feature fusion.

First, we propose to combine the bag-of-audio-words methodology with modulation spectrum features for environmental robustness. Second, we take advantage of the inherent quality-awareness properties of modulation spectrum and propose the use of a quality feature as an additional feature to be used by the speech emotion recognizer. The outcome of this exploration showed that the proposed features i) consistently outperforming benchmark systems, ii) providing complementary information to classical features, hence improving performance with feature fusion, and iii) showing robustness against environment and language mismatch. Moreover, we show that when the proposed system is provided with quality information, further improvements are obtained. Overall, the proposed bag of modulation spectrum features are shown to be a promising candidate for "in-the-wild" SER. Secondly, we explore multi-lingual settings. Next, we explore multi-cross lingual settings using German, French, and Hungarian language datasets and data augmentation strategy. We propose to combine the bag-of-word (BOW) approach with DA to improve the cross-language SER system further. Finally, a variant of the CORAL method is proposed, termed N-CORAL. More specifically, both target and source domains are adapted to a third unseen unsupervised domain; in the case of our experiments, Chinese. Experimental results with cross-language SER using CORAL and N-CORAL methods emphasize their effectiveness for both arousal and valence prediction, with the most significant gains occurring for the latter. Lastly, we explore the combination of task-specific speech enhancement and data augmentation as a strategy to improve multimodal emotion recognition in noisy conditions. we showed that multi-modal systems help improve performance for emotion recognition by building noise robustness as well as improving performance over short term windows. It is hoped that the insights presented herein help in the further development of methods for assessment of affective states in real-time conditions.

Keywords: Speech emotion recognition, Cross-language emotion recognition, Multi-modal emotion recognition, domain adaptation, Deep Neural Network

Résumé

Reconnaître les émotions de l'utilisateur à partir de la parole est quelque chose que les humains font naturellement lorsqu'ils communiquent en face-à-face. Les machines, cependant, sont encore loin d'être parfaites lorsqu'il s'agit de détecter les émotions des utilisateurs, en particulier dans des environnements réalistes, où le bruit ambiant et l'acoustique de la pièce entravent la qualité du signal. De plus, il est courant que des utilisateurs de cultures différentes et des locuteurs de langues différentes interagissent avec le même système. De tels paramètres multilingues sont extrêmement difficiles et entraînent des baisses de performances significatives lorsqu'un système est, par exemple, entraîné dans une langue et testé dans une autre. En tant que telle, la reconnaissance des émotions interculturelles (RE) est devenue un sujet émergent dans l'informatique affective pour améliorer le pouvoir de généralisation des interfaces homme-machine (IHM). La plupart des recherches menées à ce jour dans le domaine de la reconnaissance multimodale des émotions se sont appuyées sur des environnements contrôlés sans aucune distraction de fond, comme le bruit ambiant et/ou la réverbération. Cependant, dans des situations réelles, les signaux sont souvent corrompus par des facteurs environnementaux, qui détériorent les performances du système ER. La plupart des recherches se sont également appuyées sur des paramètres monolingues, dans lesquels des modèles ER pour une langue particulière sont développés et testés pour la même langue. Une fois testées dans une langue différente, des baisses de performances importantes sont observées. C'est à cette problématique que ce doctorat souhaite répondre. En particulier, nous présentons de nouvelles méthodes et de nouveaux outils pour permettre non seulement la robustesse au bruit, mais aussi des capacités multilingues pour les IHM affectives émergentes. Ici, l'accent a été mis sur la reconnaissance des émotions "dans la nature". Comme cela sera détaillé ici, ces réalisations sont le résultat d'une combinaison de nouvelles fonctionnalités, de la fusion de fonctionnalités multimodales, ainsi que de schémas d'apprentissage automatique et d'adaptation de domaine. En particulier, dans cette thèse de doctorat, nous présentons les étapes vers le développement de modèles de reconnaissance des émotions pour des données collectées dans des conditions hautement écologiques. Pour atteindre cet objectif, trois outils principaux ont été explorés. Celles-ci incluent (i) un ensemble de caractéristiques de modulation soucieuses de la qualité, (ii) la construction de schémas d'adaptation de domaine et (iii) l'utilisation de la fusion de caractéristiques multimodales.

Tout d'abord, nous proposons de combiner la méthodologie du sac de mots audio avec des caractéristiques de spectre de modulation pour la robustesse environnementale. Deuxièmement, nous tirons parti des propriétés inhérentes de sensibilité à la qualité du spectre de modulation et proposons l'utilisation d'une caractéristique de qualité comme caractéristique supplémentaire à utiliser par le système de reconnaissance des émotions de la parole. Le résultat de cette exploration a montré que les fonctionnalités proposées i) surpassaient constamment les systèmes de référence, ii) fournissaient des informations complémentaires aux fonctionnalités classiques, améliorant ainsi les performances avec la fusion de fonctionnalités, et iii) montrant une robustesse face à l'inadéquation de l'environnement et de la langue. De plus, nous montrons que lorsque le système proposé est fourni avec des informations de qualité, des améliorations supplémentaires sont obtenues. Dans l'ensemble, le sac de caractéristiques du spectre de modulation proposé s'avère être un candidat prometteur pour le SER "à l'état sauvage". Deuxièmement, nous explorons les paramètres multilingues. Ensuite, nous explorons les paramètres multilingues à l'aide d'ensembles de données en allemand, français et hongrois et d'une stratégie d'augmentation des données. Nous proposons de combiner l'approche sac de mots (BOW) avec DA pour améliorer encore le système SER inter-langue. Enfin, une variante de

la méthode CORAL est proposée, appelée N-CORAL. Plus précisément, les domaines cible et source sont adaptés à un troisième domaine caché non supervisé; dans le cas de nos expériences, le chinois. Les résultats expérimentaux avec SER inter-langues utilisant les méthodes CORAL et N-CORAL soulignent leur efficacité pour la prédiction de l'activation physiologique et de la valence, les gains les plus significatifs se produisant pour cette dernière. Enfin, nous explorons la combinaison de l'amélioration de l'audio spécifique à une tâche avec des techniques d'augmentation du jeu de données en tant que stratégie pour améliorer la reconnaissance des émotions multimodales dans des conditions bruyantes. Nous avons montré que les systèmes multimodaux aident à améliorer les performances de reconnaissance des émotions en renforçant la robustesse au bruit ainsi qu'en améliorant les performances sur des fenêtres temporelles courtes. Nous espérons que les découvertes présentées ici contribueront au développement ultérieur de méthodes d'évaluation des états affectifs dans des conditions hautement écologiques.

Mots-clés: Reconnaissance des émotions vocales, reconnaissance des émotions inter-langues, reconnaissance des émotions multimodales, adaptation de domaine, réseau neuronal profond

Contents

List of Figures

List of Tables

List of Abbreviations

AAE	Adversarial Autoencoders
ACC	Accuracy
AER	Automatic emotion recognition
AI	Artificial intelligence
ANS	Autonomic Nervous system
ASR	Automatic speech recognition
AVB	Adversarial Variational Bayes
AVEC	Audio-Video Emotion Challenge
BACC	Balanced Accuracy
BERT	Bidirectional Encoder Representations from Transformers
Bi-LSTM	Bi-directional LSTM
BOAW	Bag-of-audio-words
BOW	Bag-of-Words
CCC	Concordance correlation coefficient
CCC	Matthews Correlation Coefficient
CNN	Convolution neural networks
CNN-LSTM	Convolution neural network-long short term memory
CNS	Central Nervous system
ComParE	Computational Paralinguistics Evaluation Challenge
CORAL	CORrelation ALignment
CV	Cross-validation
DA	Domain adaptation
DANN	Domain adversarial neural network
DBN	Deep belief networks
DFN	Deep feedforward network
DL	Deep learning
DNN	Deep neural networks
ECG	Electrocardiogram
EEG	Electroencephalogram
eGeMAPS	Extended Geneva Minimalistic Acoustic Parameter Set
EmotiW	Emotion in the Wild

ER	Emotion recognition
EVA	l'échelle visuelle analogique
F0	Fundamental frequency
F1	F1-score
FNN	Feedforward neural networks
GD	Gradient descent
GRU	Gated recurrent units
HCI	Human computer interaction
HMI	Human-Machine Interaction
IDF	Inverse document frequency
IHM	l'interaction homme-machine
KCCA	Kernel canonical correlation analysis
LLDs	Low-level descriptors
LSTM	Long short term memory
LSTM-RNNs	Long Short Term Memory-Recurrent Neural Networks
MFCC	Mel-Frequency Cepstral Coefficients
ML	Machine learning
MLM	Masked Language Modeling
MLPs	Multi-layer perceptrons
MSE	Mean squared error
NASA-TLX	NASA Task Load Index
NSP	Next Sentence Prediction
PCA	Principal component analysis
PESQ	Perceptual evaluation of speech quality
PNS	Parasympathetic Nervous system
POLQA	Perceptual objective listening quality assessment
RCTF	Relative convolution transfer function
RECOLA	REmote COLlaborative and Affective interactions
SA	Subspace alignment
SAM	Self Assessment Manikin
SCL	Structural correspondence learning
SEGAN	Speech enhancement generative adversarial network
SER	Speech emotion recognizers
SEWA	Sentiment Analysis
SGD	Stochastic gradient descent
SNS	Sympathetic Nervous system
SOTA	State-of-the-art
SRMR	Speech to reverberation modulation ratio
SSE	Single-channel spectral enhancement

STAI	State-Trait Anxiety Inventory
SVM	Support vector machine
SWAT	Subjective Workload Assessment Technique
TF	Term frequencies
VAE	Variational Autoencoders
VAS	Visual Analog Scale

Synopsis

1: Introduction L'informatique affective apparaît comme un domaine important de l'interaction homme-machine (IHM), car elle fournit à la machine des informations sur l'état affectif de l'utilisateur, rendant ainsi l'interaction plus naturelle et plus humaine [3]. IHM peut être améliorée en ajoutant une intelligence émotionnelle aux systèmes [4]. Cette intelligence a une grande variété d'applications. Dans le domaine de l'éducation, l'évaluation des états affectifs tels que l'ennui ou la perte d'attention peut aider les éducateurs à améliorer les ressources d'apprentissage et à personnaliser la formation pour différents étudiants [5]. En raison de la pandémie de COVID-19, cela est devenu particulièrement pertinent avec l'essor de l'apprentissage en ligne, qui limite l'interaction enseignant-étudiant. La surveillance de la fatigue mentale, quant à elle, peut contribuer à la sécurité des conducteurs en les avertissant lorsqu'ils sont trop fatigués, et des contrôleurs aériens en les avertissant lorsqu'ils ne le sont pas [6]. Parmi les autres domaines dans lesquels le développement de l'intelligence émotionnelle pour les systèmes peut s'avérer pertinent, citons pour exemples, les smarthomes [7], les jeux [8], les systèmes conversationnels intelligents [9] et le neuromarketing [10], pour n'en citer que quelques-uns. En outre, dans les systèmes de réalité virtuelle [11], les systèmes de reconnaissance affectifs

peuvent jouer un rôle crucial dans la mesure de la qualité de l'expérience perçue par l'utilisateur.

La manière dont un utilisateur exprime son état émotionnel peut également varier selon la culture. Ceci est principalement dû au fait que les gens transmettent les émotions sous différentes formes dans différentes cultures. Par conséquent, l'amélioration des performances des systèmes de reconnaissance des émotions (RE) multilingues (interculturels) a suscité un grand intérêt ces derniers temps. En général, la prédiction multilingue des émotions repose sur des outils de normalisation des données, l'adaptation au domaine et l'entraînement multi-tâches pour améliorer la généralisabilité des modèles. Récemment, il a été démontré que les réseaux neuronaux profonds (DNN) donnaient des résultats de pointe pour la RSE monolingue, mais peu de succès ont été obtenus dans l'extraction de caractéristiques qui donnent des résultats cohérents dans différentes conditions sur plusieurs ensembles de données [12].

L'état affectif d'un utilisateur peut être exprimé par de nombreuses modalités différentes, notamment la parole, les gestes/positions et les réponses physiologiques (par exemple, les changements de rythme cardiaque/respiratoire). À ce titre, les systèmes multimodaux ont été présentés comme la solution nécessaire à la mise en place d'IHM affectives fiables, susceptibles de remplacer le traditionnel clavier et la souris. Pour les applications dites "dans la nature", les systèmes multimodaux sont préférés afin de compenser certaines des confusions et d'améliorer la précision globale de l'ER en fournissant au système une certaine redondance et des informations supplémentaires qui ne sont pas disponibles dans les systèmes unimodaux [13, 14]. Cependant, les systèmes multimodaux peuvent être très longs à mettre en œuvre, coûteux à exploiter et potentiellement intrusifs pour les utilisateurs (par exemple, en nécessitant des capteurs corporels) et pour leur vie privée. C'est pourquoi les systèmes de RE basés sur le texte et la parole sont apparus comme deux solutions populaires.

Dans le passé, les états affectifs ont été mesurés à l'aide de questionnaires subjectifs. Ces questionnaires explorent généralement l'état affectif d'un individu sur une base périodique et on y répond

par une évaluation sur une échelle continue ou sur une échelle discrète contenant plusieurs choix prédéfinis. La catégorisation discrète des émotions comprend les six émotions de base : le bonheur, la tristesse, la peur, la colère, le dégoût et la surprise, comme l'ont introduit les auteurs suivants [15]. Différentes échelles multidimensionnelles continues pour la classification des émotions ont été proposées dans les domaines suivants [16, 17, 18]. Le plus accepté de ces modèles est celui de l'éveil et de la valence introduit par [16]. Selon l'approche dimensionnelle, les états affectifs sont systématiquement liés les uns aux autres, comme le montre la figure suivante 1.1. Trois dimensions communes sont l'éveil (ou activation), la valence et la dominance [19]. L'excitation décrit le niveau d'activation (passif ou actif), tandis que la valence décrit le caractère agréable. La dominance, quant à elle, décrit le niveau de contrôle d'une personne pendant l'expérience émotionnelle (faible ou fort). Il a été démontré que la valence et l'éveil couvrent la majorité de la variabilité de l'affect, et ont donc été largement utilisés dans la recherche sur l'informatique affective [20], [21]. Plusieurs questionnaires ont été élaborés pour évaluer l'état affectif, tels que le NASA Task Load Index (NASA-TLX) [22] et le Subjective Workload Assessment Technique (SWAT) [23], L'Inventaire d'état d'anxiété (STAI) [24] pour la mesure de l'anxiété, et l'échelle visuelle analogique (EVA) [25, 26] pour n'en citer que quelques-unes. Diverses méthodes neurophysiologiques basées sur les électroencéphalogrammes (EEG) ont été utilisées pour évaluer l'état affectif. [27, 28]. D'autres signaux corporels résultant des réponses du système nerveux autonome, notamment l'électrocardiogramme (ECG), le signal respiratoire et les signaux cutanés, sont également utilisés. La température, et la réponse galvanique de la peau (GSR) ont également été utilisées pour surveiller différents états affectifs [29].

0.0.1 Défis pour les SER "en liberté"

Les approches existantes de la reconnaissance vocale des émotions ne sont pas parfaites et présentent plusieurs limites, notamment

.La plupart des systèmes RES existants sont basés sur des expériences et des environnements contrôlés en laboratoire. Par conséquent, les performances de ces systèmes se dégradent dans les scénarios du monde réel où les facteurs environnementaux affectent la qualité du signal. Par conséquent, des systèmes tenant compte de la qualité sont toujours nécessaires. En outre, les systèmes existants reposent généralement sur des fonctionnalités développées dans leurs propres conditions. Pour garantir la robustesse au bruit, les systèmes sont entraînés en utilisant une combinaison de données propres et bruitées (un processus appelé entraînement

multi-condition ou, plus récemment, augmentation des données). Bien que cette méthode offre une certaine robustesse, surtout si le type et les niveaux de bruit observés dans les données de test sont similaires à ceux utilisés lors de l'entraînement, les performances restent fortement affectées dans des conditions inconnues (par exemple, des niveaux de réverbération ou des types de bruit variables). Par conséquent, de nouvelles caractéristiques tenant compte de l'environnement ou du bruit sont encore nécessaires pour surmonter cette limitation. Les systèmes RES existants sont généralement monolingues, c'est-à-dire que les émotions ne sont détectées que dans les mêmes langues que celles utilisées pour former les systèmes. Étant donné que des utilisateurs de langues et de cultures différentes interagissent avec la même interface, il est essentiel que les solutions de la prochaine génération soient multilingues. Cependant, la détection multilingue des émotions est une tâche extrêmement difficile, car les précisions rapportées chutent de façon spectaculaire, les émotions pouvant être exprimées différemment selon les langues. C'est pourquoi l'amélioration des performances des systèmes de reconnaissance des émotions multilingues a suscité un intérêt considérable ces derniers temps. L'un des principaux avantages de la modalité vocale est que les progrès récents en matière de conversion automatique de la parole en texte ont permis l'émergence de systèmes multimodaux basés sur la parole et le texte, tout en ne nécessitant la collecte que d'une seule modalité de signal. Cependant, dans les applications du monde réel, où des facteurs environnementaux tels que le bruit additif et convolutif (par exemple, la réverbération d'une pièce) entravent les performances des systèmes multimodaux, la contribution des modalités audio et textuelles dans des conditions de bruit est restée relativement inexplorée.

L'objectif de cette thèse de doctorat est d'aborder ces questions critiques et de développer des solutions pour évaluer les émotions de la parole "dans la nature". La section suivante détaille l'objectif de la thèse, suivie par les contributions et la liste des publications.

0.0.2 Objectifs et contributions de la thèse

L'objectif principal de cette recherche est de créer des outils pour des interfaces homme-machine sensibles aux émotions pour des applications dans des environnements quotidiens. Pour atteindre ce but, trois objectifs ont été abordés :

3. Fournir une conscience environnementale au système SER via des fonctionnalités plus robustes et une conscience de la qualité de la parole au modèle lui-même,
2. Concevoir une stratégie d'adaptation au domaine pour permettre des capacités multilingues pour le système SER développé, et
3. Concevoir un système robuste de reconnaissance multimodale des émotions dans lequel des modules d'amélioration de la parole adaptés aux tâches sont utilisés pour améliorer la précision "dans la nature".

Pour atteindre ces objectifs, trois grandes familles d'innovations ont été proposées, à savoir :

1. De nombreux défis ont émergé au cours des dernières décennies dans le domaine des SER. Bien que les derniers défis aient montré que les réseaux de neurones profonds obtiennent les meilleurs résultats, les caractéristiques d'entrée existantes constituent toujours un goulot d'étranglement et entraînent une dégradation significative des performances dans des scénarios réalistes "dans la nature". Dans la première contribution, nous proposons deux innovations. Tout d'abord, nous proposons de combiner la méthodologie du sac de mots audio avec des caractéristiques du spectre de modulation pour la robustesse environnementale. Deuxièmement, nous tirons parti des propriétés de qualité inhérentes au spectre de modulation et proposons l'utilisation d'une caractéristique de qualité comme caractéristique supplémentaire à utiliser par le reconnaisseur d'émotions vocales.
2. Les systèmes SER existants ont une capacité limitée à traiter les données d'émotion provenant de différentes langues. Une méthode qui a été largement utilisée dans des domaines connexes pour faire face à de telles disparités dans la distribution des données est appelée adaptation au domaine (DA). Alors que les algorithmes DA ont été largement utilisés dans les applications de vision par ordinateur et de traitement du langage naturel, leur utilisation dans la reconnaissance vocale des émotions n'a pas encore été explorée. Dans la deuxième contribution, nous proposons l'utilisation de deux algorithmes DA simples mais efficaces, à savoir l'alignement de corrélation (CORAL) et l'alignement de sous-espace (SA), pour améliorer les performances de la RSC continue dans des contextes multilingues. En utilisant les données d'emotion en allemand, en français et en hongrois, nous montrons les avantages de combiner l'approche du sac de mots proposée avec SA pour améliorer la performance de la SER interlinguistique. Enfin, nous proposons une nouvelle variante de la méthode CORAL, appelée N-CORAL. Ici, les domaines cible et source sont adaptés à une troisième langue

inconnue de manière non supervisée (c'est-à-dire sans avoir besoin d'étiquettes d'émotions supplémentaires) ; dans le cas de nos expériences, la langue chinoise.

3. Les systèmes de reconnaissance multimodale des émotions peuvent s'appuyer sur une combinaison de signaux audio, vidéo, textuels ou physiologiques. Cependant, la collecte de plusieurs modalités de signaux peut être très intrusive, longue et coûteuse. Les progrès récents des systèmes de synthèse vocale et de traitement du langage naturel basés sur l'apprentissage profond ont toutefois permis de développer des systèmes multimodaux fiables basés sur la parole qui ne nécessitent que la collecte de données audio. Cependant, les données audio sont extrêmement sensibles aux perturbations environnementales, telles que le bruit additif, et sont donc confrontées à certains défis lorsqu'elles sont déployées "dans la nature". À cette fin, des algorithmes d'amélioration de la parole ont été développés pour garantir la robustesse au bruit à l'entrée du signal. L'augmentation des données, à son tour, a été déployée pendant la formation pour assurer la robustesse au bruit, mais au niveau du modèle. Dans notre troisième contribution, nous explorons la combinaison de l'amélioration de la parole et de l'augmentation des données pour améliorer la reconnaissance multimodale des émotions dans des conditions bruyantes.

Ces trois innovations ont été décrites dans plusieurs manuscrits, énumérés ci-dessous par ordre chronologique. Le cas échéant, les chapitres de cette thèse dans lesquels ces publications apparaissent sont également indiqués.

0.0.3 Organisation de la thèse

Alors que ce chapitre introductif a présenté les défis de la reconnaissance des états affectifs dans la nature et posé les bases des contributions décrites ici, le reste de cette thèse est structuré comme suit : Le chapitre 2 fournit un aperçu des dernières méthodes de reconnaissance des états affectifs, ainsi qu'une liste des bases de données publiques utilisées dans ce document. Le chapitre 3 présente la première contribution proposant l'utilisation des caractéristiques des sacs de modulation et la prise en compte de la qualité pour les systèmes "dans la nature". Ensuite, le chapitre 4 décrit la deuxième contribution qui traite de l'adaptation au domaine pour la reconnaissance des émotions inter-langues. Dans le chapitre 5, la troisième contribution est décrite. Des systèmes multimodaux sont présentés pour améliorer la robustesse au bruit et les performances de la reconnaissance des

états affectifs en utilisant plusieurs modalités simultanément. Enfin, le chapitre 6 présente les conclusions générales de cette thèse et les domaines de recherche future.

*Chapitre 2 : Background : Systèmes informatiques affectifs basés sur la parole et le texte

Les systèmes d'informatique affective/de reconnaissance de mouvements sont principalement composés de deux modules : le module frontal et le module dorsal. Le module frontal correspond au pipeline de traitement du signal, où le prétraitement, l'amélioration et l'extraction de caractéristiques sont généralement effectués. Le module dorsal, quant à lui, est basé sur le pipeline d'apprentissage automatique, où des facteurs tels que l'adaptation au domaine, l'augmentation des données et la classification/évaluation sont réalisés. Plus récemment, cependant, avec les progrès des réseaux neuronaux profonds de bout en bout, le front-end et le back-end ont fusionné en un seul et le réseau neuronal apprend les représentations des caractéristiques et le mappage du classificateur en une seule étape. Dans cette thèse, l'approche la plus classique est adoptée, comme le montre la Fig.. 2.1, où l'analyse du signal est effectuée en amont et l'apprentissage automatique en aval.

0.0.4 Méthodes d'analyse du signal (Front-end)

Amélioration de la parole Les signaux vocaux peuvent être affectés par des facteurs environnementaux, tels que le bruit additif et convolutif (c'est-à-dire la réverbération de la pièce). Par conséquent, l'amélioration de la parole a été largement utilisée pour contrer ces effets négatifs. Parmi les algorithmes d'amélioration de la parole représentatifs, on trouve l'amélioration spectrale classique à canal unique (SSE). [30], la fonction de transfert de convolution relative (RCTF) [31], le filtrage de Wiener [32] et les méthodes basées sur un modèle statistique [33], ainsi que les modèles plus récents basés sur des DNN, tels que le GAN d'amélioration de la parole (SEGAN)[34] et l'autoencodeur de débruitage récurrent [35].

L'amélioration de la parole peut suivre deux principes : l'amélioration pour les humains ou l'amélioration pour les tâches machine en aval. Lorsqu'il s'agit d'humains, l'amélioration de la qualité du signal et de l'intelligibilité est de la plus haute importance. Il faut donc trouver un compromis entre la suppression du bruit et l'intelligibilité. Cependant, pour les tâches effectuées par des machines, l'intelligibilité n'est pas nécessairement le but ultime. Dans ce cas, l'algorithme effectue un compromis entre la suppression et les performances de la tâche en aval. Dans le cas de la

conversion de la parole en texte, le taux d'erreur sur les mots est généralement utilisé comme critère. Pour la reconnaissance multimodale des émotions vocales, deux approches distinctes peuvent être nécessaires, l'une visant à maintenir l'intelligibilité de la parole pour permettre la préservation des indices émotionnels et l'autre visant à améliorer la conversion parole-texte pour l'analyse textuelle en aval. Dans le premier cas, MetricGAN+ est un réseau neuronal profond à la pointe de la technologie, spécifiquement optimisé pour améliorer la qualité de la parole bruyante. [36]. En particulier, deux réseaux sont utilisés. Le rôle du discriminateur est de minimiser la différence entre un score de qualité prédit (donné par une mesure objective de la qualité de la parole appelée PESQ, [37]) et les scores de qualité réels, garantissant ainsi que le signal amélioré reste de haute qualité et intelligible. Pour ce dernier scénario, l'amélioration de la parole basée sur la cartographie spectrale a été considérée comme l'état de l'art pour les applications de reconnaissance vocale en aval. La méthode d'amélioration basée sur la perte d'imitation [38], par exemple, s'est avéré être la meilleure méthode d'amélioration pour les tâches de reconnaissance vocale en aval.

0.0.4.1 Estimation objective de la qualité de la parole

La mesure de la qualité de la parole de manière automatisée et objective intéresse l'industrie des télécommunications depuis trois décennies. La mesure objective de la qualité de la parole peut être divisée en deux catégories : les méthodes à double extrémité (également appelées intrusives) et les méthodes à extrémité unique (non intrusives). Comme leur nom l'indique, les méthodes à double extrémité nécessitent l'accès à deux signaux, le signal traité bruyant et son homologue original propre. Comme il est souvent impossible d'accéder au signal propre dans la pratique, on a développé des méthodes unilatérales qui ne prennent en entrée que le signal traité bruyant. L'Union internationale des télécommunications (UIT-T) a normalisé les méthodes à deux extrémités et à une extrémité. Les plus récentes sont la recommandation UIT-T P.863, également connue sous le nom d'"évaluation objective de la qualité d'écoute perceptive" (POLQA)" [39], le successeur de P.862 (évaluation perceptive de la qualité de la parole, PESQ [37]), et la Recommandation UIT-T P.563, respectivement. Plus récemment, une mesure unique optimisée pour la parole bruyante et réverbérante a été développée et appelée "rapport de modulation parole/réverbération (norme SRMR)". [40].

0.0.4.2 Extraction de caractéristiques de la parole

Plusieurs caractéristiques ont été utilisées pour le SER, notamment des caractéristiques prosodiques, spectrales et spectrales de modulation. La boîte à outils openSMILE [41] a été largement utilisé par la communauté SER et a servi à extraire des caractéristiques de référence pour la majorité des récents défis SER. La boîte à outils peut être utilisée pour extraire plus de 6000 caractéristiques, mais différents sous-ensembles ont été utilisés pour différentes applications. Par exemple, l'ensemble de caractéristiques pour le défi Interspeech 2011 sur l'état du locuteur comprend un sous-ensemble de 118 descripteurs de bas niveau (LLD). Parmi ceux-ci, les 59 premières caractéristiques comprennent 50 caractéristiques spectrales, cinq caractéristiques liées à la parole et quatre caractéristiques liées à l'énergie, ainsi que 59 caractéristiques supplémentaires correspondant à leurs homologues delta à un pas de temps. [42]. Ces caractéristiques sont extraites sur des trames de 20 ms avec des sauts de 10 ms et ont été utilisées comme points de référence dans plusieurs défis AVEC.

Les coefficients cepstraux de fréquence de Mel (MFCC) ont été largement utilisés dans de nombreuses applications vocales. Il s'agit de coefficients cepstraux calculés après une cartographie de la fréquence mel pour simuler le traitement cochléaire humain. Pour les SER, il est courant d'utiliser des vecteurs de caractéristiques MFCC à 39 dimensions, composés de 13 MFCC, 13 MFCC delta et 13 MFFC double-delta. Ces caractéristiques sont extraites sur des fenêtres de 20 ms, avec une taille de saut de 10 ms, et avec des filtres de 64 mel. Elles ont été utilisées comme références dans les challenges AVEC 2018 et AVEC 2019.

En outre, les caractéristiques prosodiques, notamment la fréquence fondamentale (F0), les mesures d'intensité et les probabilités de vocalisation, ont également été largement utilisées pour les SER. L'ensemble étendu de paramètres acoustiques minimalistes de Genève (eGeMAPS), par exemple, comprend 88 paramètres acoustiques liés à la hauteur, à la sonie, aux segments non vocalisés, à la dynamique temporelle et aux caractéristiques cepstraux. Ces paramètres ont également été utilisés comme points de référence dans plusieurs défis AVEC18, AVEC19 [43], [44].

Plus récemment, une représentation du signal spectral de modulation inspirée de l'auditoire a été proposée pour améliorer la SER, car elle capture la dynamique à long terme du signal de parole [45, 46]. La mesure est la même représentation que celle utilisée dans la mesure de qualité SRMRnorm décrite dans [47], et il a donc été démontré qu'elle offre une certaine robustesse face au bruit ambiant et à la réverbération [46]. Les caractéristiques spectrales de modulation (appelées MSF) ont été

extraites sur des sauts de 256 ms et des sauts de 40 ms, en suivant les étapes de traitement décrites dans [47]. Dans un souci d'exhaustivité, les étapes de traitement du signal impliquées dans le calcul des MSF sont présentées à la Fig. 2.2. La première étape correspond à la normalisation du niveau de la parole à -26 dBov (dB de surcharge) afin de compenser les variations d'énergie indésirables entre les signaux de parole. Ensuite, le signal prétraité est décomposé en 23 signaux de sous-bande à l'aide d'une banque de filtres gammatone à 23 canaux dont les fréquences centrales de filtre vont de 125 Hz à environ la moitié de la fréquence d'échantillonnage [48]. Les bandes passantes des filtres suivent le paradigme de la bande passante rectangulaire équivalente [49]. Ensuite, l'enveloppe de Hilbert est calculée à partir de chacun des 23 signaux de sous-bande en utilisant la transformée de Hilbert. Les enveloppes de Hilbert $HE_j(n), j = 1, \ldots, 23$ sont fenêtrées par une fenêtre de Hamming de 256 ms avec un décalage de 40 ms. Les enveloppes sont ensuite regroupées en 8 bandes suivant les preuves récentes d'une structure de banque de filtres de modulation dans le système auditif humain. On utilise ici un banc de filtres de modulation à 8 canaux dont les fréquences centrales sont également espacées sur l'échelle logarithmique de 4 à 128 Hz. La représentation spectrale finale de la modulation (appelée $E_k(i, j)$) est constituée d'un tenseur d'énergies de modulation $23 \times 8 \times T$, où i indexe le nombre de filtres gammatoniques utilisés ($i = 1, \ldots, 23$), j correspond au nombre de filtres de modulation ($j = 1, \ldots, 8$), et k indexe l'image à analyser ($k = 1, \ldots, T$), où T correspond au nombre total d'images disponibles dans un enregistrement vocal. Le lecteur intéressé est renvoyé à [48] pour des détails complets sur le calcul de cette représentation. À partir de cette représentation, plusieurs caractéristiques peuvent être extraites des directions des fréquences acoustiques et de modulation. Tout d'abord, l'énergie de modulation pour chaque bin de fréquence de modulation acoustique peut être calculée, ce qui donne 23 fois 8 = 184 caractéristiques. Ensuite, six descripteurs spectraux (Θ_1-Θ_6), calculés par trame, sont extraits. Les trois premiers descripteurs mesurent les changements spectraux entre les bandes de modulation, tandis que les trois derniers capturent les changements entre les bandes acoustiques.

Globalement, les trois premiers descripteurs sont calculés sur chacun des huit canaux de modulation, ce qui donne 24 caractéristiques supplémentaires. Les trois derniers, à leur tour, sont calculés pour chacune des cinq bandes groupées, ce qui donne un total de 15 caractéristiques supplémentaires. Un total de 39 descripteurs est ainsi calculé par fichier vocal. Ils sont ensuite ajoutés aux 184 caractéristiques spectrales de modulation pour générer un vecteur final de caractéristiques MSF à 223 dimensions.

0.0.5 Extraction de caractéristiques à partir de texte

0.0.5.1 BERT - Représentations d'encodeurs bidirectionnels à partir de transformateurs

BERT est basé sur un réseau de transformateurs et un mécanisme d'attention [1] qui apprend également les relations contextuelles entre les mots du texte [50]. BERT existe en deux versions : BERTBase et BERTLarge. Le modèle BERTBase utilise 12 couches de blocs de transformateurs avec une dimension cachée de 768 et 12 têtes d'auto-attention ; globalement, il y a environ 110 millions de paramètres entraînables. D'autre part, BERTLarge utilise 24 couches de bloc de transformateurs avec une dimension cachée de 1024 et 16 têtes d'auto-attention, ce qui donne environ 340 millions de paramètres entraînables. Nous utilisons le modèle BERTBase pour l'extraction des caractéristiques du texte et le vecteur d'état caché de BERT est utilisé comme entrée du système de reconnaissance des émotions. Le lecteur intéressé est renvoyé à [1] pour plus de détails sur BERT.

0.0.5.2 TextCNN

TextCNN est un profond modèle d'apprentissage pour les tâches de classification de textes courts et il a été utilisé comme modèle de référence pour la classification de textes [51]. TextCNN transforme un mot en un vecteur à l'aide de mots intégrés, qui sont ensuite introduits dans une couche convolutive, suivie d'une couche de mise en commun maximale et d'une couche de sortie entièrement connectée.

0.0.5.3 Bag-of-Words (or Sac de mots)

La méthode du sac de mots (BOW) est couramment employée dans le traitement du langage naturel [52]. Cette méthode est simple et flexible et peut être utilisée de nombreuses façons pour extraire des caractéristiques des documents. Le BoW représente un texte qui décrit l'occurrence des mots dans un document. Il se compose de deux parties : un vocabulaire de mots connus et une mesure de l'occurrence de ces mots. On l'appelle un "sac" de mots car toute information sur l'ordre ou la structure des mots dans le document est écartée. Le modèle se préoccupe uniquement de savoir si des mots connus apparaissent dans le document, et non dans quel document. Dans

cette méthode, un histogramme des mots est d'abord généré dans un document texte. Ensuite, les fréquences de chaque mot dans un dictionnaire sont calculées, et enfin le vecteur résultant est fusionné en tant que caractéristiques textuelles.

0.0.6 Systèmes de reconnaissance vocale

Afin de générer du texte à partir de la parole, il est nécessaire de disposer d'un système de reconnaissance automatique de la parole à la pointe de la technologie. Dans ce cas, nous utilisons wav2vec 2.0, un système de reconnaissance vocale de bout en bout. Wav2vec 2.0 utilise la forme d'onde brute de la parole comme entrée. Ces données unidimensionnelles passent ensuite par un CNN 1-d multicouche pour générer des vecteurs de représentation de la parole. La quantification vectorielle est ensuite utilisée sur ces représentations latentes pour les mettre en correspondance avec un livre de codes. La moitié des données vocales disponibles est masquée et les données quantifiées restantes sont introduites dans un réseau de transformation.

0.0.7 Pipeline d'apprentissage automatique (Back-End)

Avec l'apprentissage automatique, l'objectif ultime est de développer des modèles qui peuvent bien se généraliser à des conditions inconnues. Si l'on n'y prend garde, un modèle surajusté ou sous-ajusté peut se produire. Dans le cas d'un ajustement excessif, par exemple, le modèle apprend les nuances uniques de l'ensemble de formation, et pas nécessairement les indices émotionnels généraux dans les données. Ainsi, le modèle obtient une grande précision sur l'ensemble d'apprentissage, mais de très mauvais résultats sur l'ensemble de test non vu. Inversement, on parle de sous-adaptation lorsque l'erreur d'apprentissage est très importante, ce qui suggère que le modèle était trop "simple" pour recueillir des indices utiles à partir des données d'apprentissage. Dans la pratique, pour surmonter ces problèmes, un ensemble de validation est couramment utilisé pendant la formation afin de s'assurer que le modèle dispose d'informations "nouvelles" pour évaluer le sur- ou le sous-ajustement du modèle. D'autres outils, tels que les régularisateurs, sont utilisés pour pénaliser les modèles trop complexes, ce qui permet de résoudre le problème du surajustement. L'augmentation des données s'est avérée utile pour l'entraînement des réseaux neuronaux profonds en fournissant au modèle des données d'entraînement dont la distribution est plus diversifiée, ce qui améliore les performances du modèle dans de nouvelles conditions de test. Enfin, l'adaptation au domaine a été

utilisée pour faire correspondre la distribution des données d'entraînement à celle des données de test. De cette façon, le modèle apprend des indices émotionnels qui sont persistants ou " normalisés " dans les ensembles de données. Dans cette thèse, nous explorons l'utilisation de l'augmentation des données et de l'adaptation au domaine comme outils pour améliorer la robustesse des systèmes SER à des conditions de test inédites, y compris le bruit et la langue.

0.0.7.1 Augmentation des données

L'augmentation des données, comme son nom l'indique, accroît l'ensemble d'apprentissage en introduisant des échantillons supplémentaires qui ont été corrompus par diverses distorsions. Dans le cas des images, par exemple, il s'agit d'ajouter des versions tournées des images d'entraînement, des versions bruitées, des versions décalées en pixels, pour n'en citer que quelques-unes. Dans le cas de la parole, les distorsions peuvent inclure l'ajout de bruit additif à des rapports signal/bruit (SNR) variables, la convolution avec des réponses impulsionnelles de salle pour simuler la réverbération de la salle, et l'inversion temporelle du signal vocal, pour n'en citer que quelques-unes. L'augmentation des données peut également être utilisée pour ajouter des données synthétiques dans certaines classes, ce qui permet d'atténuer le problème du déséquilibre des données. Globalement, cela permet d'éviter la pénurie de données, d'améliorer la généralisabilité des modèles et de minimiser l'impact des événements rares/extrêmes pendant la prédiction.

0.0.7.2 Domain Adaptation

Les stratégies d'adaptation au domaine visent à atténuer les effets de l'inadéquation formation-test, dans laquelle les conditions disponibles dans les données de formation diffèrent de celles de l'ensemble de test. Dans la littérature sur l'adaptation au domaine, le terme "source" fait référence au domaine dans lequel la formation a lieu. A son tour, le domaine "cible" correspond au domaine d'où proviendront les données de test. Dans cette thèse, nous explorons deux classes d'adaptation de domaine : les méthodes basées sur la corrélation ou l'alignement du sous-espace.

Alignement des sous-espaces Adaptation des domaines DA basée sur l'alignement du sous-espace (SA) vise à trouver un espace caractéristique invariant dans le domaine en apprenant une fonction de correspondance qui aligne le sous-espace source avec le sous-espace cible. [53]. SA aligne linéaire-

ment le domaine source sur le domaine cible dans un sous-espace PCA de dimension réduite. Dans cette méthode, nous créons d'abord des sous-espaces pour les domaines source et cible, puis nous apprenons une correspondance linéaire qui aligne le sous-espace source sur le sous-espace cible.

Alignement des corrélations Adaptation au domaine DA basée sur l'alignement de corrélation (CORAL) fait correspondre les statistiques de premier et de second ordre des données source et cible. Pour ce faire, il calcule d'abord les statistiques du domaine cible, puis soustrait la covariance du domaine cible du domaine source en blanchissant et en recolorant le domaine source.

0.0.7.3 Classification/Régression et figure des mérites

Une fois que l'augmentation des données a été effectuée et que l'adaptation au domaine a été réalisée, le choix de l'algorithme d'apprentissage automatique est fait. Les algorithmes d'apprentissage automatique peuvent prendre la forme de classificateurs si des classes discrètes doivent être prédites (par exemple, les émotions de colère et de tristesse) ou de régresseurs si des variables continues doivent être prédites (par exemple, le niveau d'excitation dans la plage [0,1]). Un classificateur SVM (machine à vecteurs de support) est un modèle d'apprentissage automatique supervisé [54] largement utilisé dans de nombreuses applications de reconnaissance des formes, y compris les tâches SER (par exemple, [46]). Les LSTM, NN ont également été largement utilisés dans la littérature sur la RE.

En revanche, pour les tâches de classification des émotions discrètes, nous nous appuyons sur les scores de précision, de rappel et de F1 comme figures de mérite. Pour la tâche de régression, le CCC a été utilisé comme métrique de permanence.

La section suivante décrit les jeux de données disponibles pour la reconnaissance des émotions.

0.0.8 Données d'émotions

Dans cette thèse, plusieurs jeux de données disponibles publiquement ont été utilisés. Une description des jeux de données utilisés est donnée ci-dessous.

1. **RECOLA** : La première base de données correspond à la base de données REmote COLlaborative and Affective interactions (RECOLA) [55]. Cette base de données a été utilisée lors de

l'édition 2016 du concours AVEC (audio-visual emotion challenge) [56] et est en français. Sur la base d'interactions spontanées et naturalistes recueillies au cours d'une tâche collaborative, six annotateurs ont mesuré l'émotion de manière continue en utilisant une échelle continue dans le temps pour deux primitives de l'émotion, à savoir l'excitation et la valence. Bien que tous les sujets parlaient couramment le français, ils étaient de nationalités différentes (française, italienne et allemande). La base de données offre donc une certaine diversité dans l'expression des émotions. En outre, le nombre total de locuteurs dans l'ensemble de données RECOLA était de 27, dont 16 femmes et 11 hommes. Des statistiques détaillées sur les participants sont disponibles dans [55].

2. **SEWA** : Les deuxième et troisième ensembles de données correspondent aux sous-ensembles en allemand et en hongrois de la base de données Sentiment Analysis in the wild (SEWA). Cette base de données a été utilisée dans les défis AVEC 2017 [57] et AVEC 2019 [44]. Les sujets (par paires d'amis et de parents) ont été enregistrés via une plateforme de chat vidéo dédiée, à l'aide de leurs propres webcams et microphones standard, pendant qu'ils discutaient d'une publicité qu'ils avaient regardée. Les données démographiques détaillées des participants pour les deux ensembles de données sont disponibles dans [58]. Les deux ensembles de données sont divisés en trois parties : 34 fichiers pour la formation, 14 dans l'ensemble de développement et 16 pour le test. La durée des enregistrements dans l'ensemble de données varie de 40 secondes à 3 minutes pour chaque fichier.
3. **SEWA-Chinese** : Le quatrième ensemble de données correspond au sous-ensemble de la langue chinoise du projet SEWA. Le taux d'échantillonnage des enregistrements audio était de 44,1 kHz, et la durée totale des données est de 3:17:52 heures. Il y avait des échantillons audio de 36 participants masculins et 34 participants féminins. Au total, 70 fichiers audio sans étiquette ont été mis à disposition dans le cadre du défi AVEC 2019 [44]. Les données démographiques détaillées sur les participants à cet ensemble de données sont disponibles dans [58].
4. **Emoti-W** : La base de données Emoti-W a été mise à disposition dans le cadre du concours 2017 Emotion Recognition in the Wild Challenge. Emoti-W était en langue anglaise. Dans ce jeu de données, les étiquettes d'émotions sont disponibles pour sept catégories d'émotions : colère, dégoût, peur, bonheur, neutre, tristesse et surprise. Comme il s'agit d'un jeu de données dans la nature, la plupart des enregistrements présentent un certain niveau de bruit

de fond. Les étiquettes de l'ensemble de données du défi EMoti-W ont été créées à partir des sous-titres disponibles dans les films et les séries télévisées.

5. **MELD** : Le jeu de données utilisé pour l'expérimentation dans le chapitre 5 est le Multimodal EmotionLines Dataset (MELD) [59]. Il s'agit d'un jeu de données multimodal de classification des émotions qui a été créé en étendant le jeu de données EmotionLines [60]. MELD contient environ 13 000 énoncés provenant de 1 433 dialogues de la série télévisée 'Friends'. Chaque énoncé est annoté avec des étiquettes d'émotion et de sentiment et englobe des modalités audio, visuelles et textuelles. Le jeu de données MELD contient des conversations, où chaque dialogue comporte des énoncés provenant de plusieurs locuteurs. EmotionLines a été créé en rampant les discussions de chaque épisode, puis en les regroupant en fonction du nombre d'énoncés dans la conversation en quatre groupes d'énoncés.
6. **IEMOCAP** : L'ensemble de données IEMOCAP a été utilisé pour montrer la généralisabilité du modèle proposé dans le chapitre 5. L'ensemble de données IEMOCAP comprend 12 heures de données audiovisuelles provenant de 10 acteurs. Les enregistrements suivent le dialogue entre un homme et une femme dans des sujets scénarisés ou improvisés. Une fois les données audio-vidéo collectées, elles ont été divisées en petits énoncés d'une durée de 3 à 15 secondes, qui ont ensuite été étiquetés par des évaluateurs. Chaque énoncé a été évalué par 3-4 évaluateurs. Le formulaire d'évaluation contenait dix options (neutre, bonheur, tristesse, colère, surprise, peur, dégoût, frustration, excitation, et autres).
7. **AURORA et DEMAND** : Enfin, comme nous souhaitons évaluer la robustesse environnementale de la méthode proposée, nous nous sommes également appuyés sur deux ensembles de données de bruit enregistré afin de corrompre davantage les ensembles de données d'émotion. Pour ce faire, nous avons utilisé les ensembles de données AURORA [61] et DEMAND [62] de sources de bruit enregistrées. En particulier, deux types de bruit ont été utilisés : le babillage d'une personne parlant plusieurs langues et le bruit enregistré à l'intérieur d'un avion commercial. Le bruit a été ajouté à cinq rapports signal/bruit (SNR) différents : 0 dB, 5 dB, 10 dB, 15 dB et 20 dB. Ensuite, afin d'étudier l'effet de la réverbération de la pièce sur le SER, trois réponses impulsionnelles de pièces enregistrées, tirées de [63], ont été utilisées et convoluées avec les fichiers vocaux. Les réponses impulsionnelles correspondent à des pièces dont le temps de réverbération est de $T60 = 0,25,\ 0,48$ et $0,8$ secondes, représentant ainsi une petite, moyenne et grande pièce, respectivement.

Conclusion Ce chapitre a présenté le contexte de la reconnaissance des émotions à partir de signaux vocaux. Tout d'abord, les composants du système de reconnaissance des émotions ont été discutés. Leurs méthodes d'acquisition, leurs propriétés, leur prétraitement et leurs caractéristiques de référence avec leur relation aux états affectifs ont été présentés. Ensuite, une brève description des différents types de systèmes multimodaux ainsi que de leurs avantages et inconvénients a été faite. Ensuite, les différents composants du pipeline d'apprentissage automatique ont été discutés, allant de l'adaptation au domaine, l'augmentation des données et l'évaluation aux mesures de performance utilisées. Enfin, les jeux de données disponibles ont été examinés. Dans l'ensemble, ce chapitre aborde le contexte de la littérature sur la reconnaissance des états affectifs tout en développant les différents outils nécessaires pour faire passer la recherche du laboratoire à la vie réelle.

Chapitre 3 : Sac de caractéristiques du spectre de modulation tenant compte de la qualité pour la reconnaissance robuste des émotions de la parole

Ce chapitre est compilé à partir de documents extraits du manuscrit publié dans le *IEEE Transactions on Affective Computing*. [64].

Dans ce chapitre, nous explorons l'utilité d'un nouvel ensemble de caractéristiques pour les SER robustes à l'environnement, à savoir le sac de caractéristiques spectrales de modulation. Il a été démontré que les caractéristiques spectrales de modulation offrent une certaine robustesse aux facteurs environnementaux (par exemple, le spectre de modulation d'une station radio) [45, 47]). Les sacs de mots audio, en revanche, se sont avérés utiles pour caractériser les états émotionnels à partir des caractéristiques spectrales de la parole, car ils se situent à la frontière entre la caractérisation des informations linguistiques et acoustiques. ([65]). Ici, nous proposons de combiner les deux et montrons que non seulement la précision de la reconnaissance des émotions est améliorée, mais aussi la robustesse aux facteurs environnementaux. De plus, nous montrons que l'information émotionnelle extraite par les caractéristiques proposées est complémentaire à celle obtenue par d'autres mesures conventionnelles. Ainsi, la fusion des caractéristiques permet des améliorations supplémentaires. Enfin, il a été démontré que le spectre de modulation est utile pour l'estimation aveugle de la qualité de la parole. [66]. En tant que tel, nous proposons un système SER tenant compte de la qualité et montrons son avantage pour la reconnaissance des émotions dans des contextes réalistes.

0.0.9 Méthode proposée

La méthode proposée est basée sur la représentation spectrale du signal de modulation qui s'est avérée utile pour la SER : [45, 46]. Nous étudions ici les avantages de la mise en œuvre d'une approche par sac de mots pour les caractéristiques spectrales de modulation et examinons leurs avantages pour le SER "dans la nature". Comme le spectre de modulation a également été utilisé pour la mesure aveugle de la qualité de la parole (aveugle dans le sens où un signal de référence propre n'est pas nécessaire), nous explorons un système tenant compte de la qualité pour améliorer la précision. Nous testons la complémentarité des caractéristiques proposées avec d'autres caractéristiques généralement utilisées via la fusion de caractéristiques..

Experimental Setup

Plusieurs ensembles de données multilingues sur les émotions ont été utilisés. Tout d'abord, les sous-ensembles en langues allemande et hongroise de la base de données Sentiment Analysis (SEWA) ont été utilisés pour les expériences de prédiction d'émotions en continu. Le deuxième ensemble de données sur les émotions utilisé était la base de données anglophone Emoti-W, qui a été mise à disposition dans le cadre du défi 2017 Emotion Recognition in the Wild Challenge. [67]. Nous avons utilisé différents ensembles de caractéristiques comme décrit dans le chapitre et la méthodologie BOW. LSTM et SVM ont été utilisés pour la régression et la classification respectivement. Pour la tâche de prédiction des émotions continues sur la base de données SEWA, la mesure largement utilisée du coefficient de corrélation de concordance (CCC) est utilisée comme figure de mérite. En revanche, pour la tâche de classification des émotions discrètes sur la base de données Emoti-W, nous nous basons sur les scores de précision, de rappel et de F1 comme figures de mérite.

0.0.10 Résultats expérimentaux et discussion

Dans cette première expérience, nous étudions l'effet de différentes tailles de codebook pour les caractéristiques proposées et de référence. Les figures 3.1 et 3.2 représentent les CCC obtenus pour la prédiction de l'excitation et de la valence avec un LSTM pour les mesures proposées et de référence, respectivement. Le tableau 3.1 montre les valeurs CCC obtenues sur notre ensemble de test SEWA allemand avec les caractéristiques proposées et de référence avec et sans agrégation

BoAW. Plusieurs stratégies de fusion de caractéristiques sont également incluses pour tester la complémentarité des ensembles de caractéristiques. Comme on peut le constater, l'agrégation BoAW améliore la précision pour la plupart des caractéristiques testées. Pour le scénario à caractéristique unique, les caractéristiques prosodiques avec BoAW ont montré le CCC le plus élevé pour la valence et l'excitation. Comme prévu, la considération de la qualité n'a pas montré d'amélioration avec les données vocales non traitées lorsque BoAW a été appliqué, mais a amélioré les performances lorsqu'il n'a pas été appliqué, ce qui suggère que l'agrégation BoAW elle-même fournit déjà une certaine robustesse à l'environnement.

Les tableaux 3.3 et 3.4 montre les performances obtenues lorsque les modèles ont été entraînés sur de la parole non traitée et testés sur de la parole bruyante corrompue par des bruits d'aéroport et de babillage, respectivement, à différents niveaux de rapport signal/bruit (c'est-à-dire 0 dB, 5 dB, 10 dB, 15 dB et 20 dB). D'après le tableau, la fusion de différents ensembles de caractéristiques a permis d'obtenir des améliorations supplémentaires et une plus grande robustesse. Les caractéristiques proposées, par exemple, se sont avérées être les plus complémentaires aux caractéristiques prosodiques(100) et ont constamment atteint les plus hauts niveaux de CCC ou de compétitivité par rapport aux autres méthodes fusionnées, en particulier pour la prédiction de l'excitation.

Le tableau 3.7 caractéristiques proposées se sont avérées très complémentaires des caractéristiques prosodiques et, lorsqu'elles sont combinées, des améliorations de performance ont été observées. En fait, pour des niveaux de réverbération plus élevés, la fusion des caractéristiques spectrales de modulation, MFCC et prosodiques a donné le meilleur CCC. Encore une fois, la prise en compte de la qualité a montré des améliorations dans la mesure de la valence et de l'excitation, en particulier dans les systèmes qui ont fusionné les caractéristiques proposées et prosodiques. Les améliorations étaient particulièrement significatives pour la prédiction de la valence. Ces résultats étaient attendus, car le SRMR_{norm} a été conçu à l'origine pour la prédiction de la qualité et de l'intelligibilité de la parole réverbérée [68].

Le tableau 3.6 montre les chiffres de mérite du classifieur proposé sur un problème à 4 classes, où les classes suivantes ont été explorées : heureux, en colère, triste, neutre. Ici, nous présentons les résultats de classification obtenus avec les caractéristiques proposées et de référence sur le jeu de données Emoti-W 2017 (voir Section 3.5.1). Comme on peut le voir, le système proposé avec la conscience de la qualité est capable de surpasser ce système SOTA de 13%.

Le tableau 3.8 montre les résultats lorsque les langues sont appariées, ainsi que lorsqu'elles ne sont pas appariées. Dans ce dernier cas, les résultats avec et sans augmentation des données sont présentés. Comme on peut le constater, un décalage entre les langues de formation et de test peut entraîner une baisse substantielle des performances pour la plupart des combinaisons et fusions de caractéristiques, à l'exception des caractéristiques MFCC BoAW. L'augmentation des données a permis d'améliorer considérablement les résultats, jusqu'à des niveaux comparables à ceux obtenus lorsque les langues étaient appariées, et dans certains cas même supérieurs, notamment pour la dimension de la valence. La prise en compte de la qualité s'est également avérée utile même lorsque l'augmentation des données a été utilisée, en particulier pour la dimension d'excitation et l'ensemble de caractéristiques fusionnées FF6.

Conclusion

Dans ce chapitre, nous explorons la reconnaissance vocale des émotions "dans la nature" où des facteurs environnementaux, tels que le bruit et la réverbération, et des langues différentes sont présents au moment du test, dégradant ainsi les performances du système. Nous montrons l'impact de cette inadéquation formation-test sur les performances de la RSE et proposons un système tenant compte de la qualité, basé sur une nouvelle représentation spectrale de la modulation du sac de mots, qui surpasse plusieurs références. Des expériences sur plusieurs ensembles de données du SER Challenge montrent que les caractéristiques proposées surpassent les performances de plusieurs systèmes de référence, tout en fournissant des informations complémentaires aux caractéristiques conventionnelles. Les ensembles de caractéristiques proposés contiennent intrinsèquement des informations sur la qualité de la parole. Une variante tenant compte de la qualité est également explorée et il est démontré qu'elle améliore encore la prédiction du RSC, à la fois en termes de prédictions de l'excitation/de lavalence et de classification des émotions discrètes. Enfin, nous montrons l'impact de l'augmentation des données sur la robustesse de la non-concordance des langues, mettant ainsi en évidence le potentiel du système proposé pour la SER dans la nature.

Chapitre 4 : Reconnaissance translinguistique des émotions de la parole à l'aide de représentations de type "sac de mots", d'une adaptation au domaine et d'une augmentation des données Dans ce chapitre, nous explorons la SER inter-langues en utilisant des représentations de type sac de mots, l'adaptation au domaine et l'augmentation des données. En particulier, dans ce chapitre de thèse, les contributions suivantes sont apportées :

1. Nous explorons la combinaison de DA et de BOW pour améliorer la SER interlinguistique. Des expériences avec la méthodologie BoW avant ou après l'adaptation au domaine sont réalisées pour évaluer leurs avantages/désavantages. Différentes méthodes d'AD sont explorées pour évaluer leurs effets sur la SER interlinguistique globale. En particulier, la méthode CORrelation ALignment (CORAL) [69], ainsi que les méthodes d'alignement subspatial (SA), sont comparées.
2. Une variante de la méthode CORAL est proposée pour la SER interlinguistique. Cette méthode, appelée N-CORAL, utilise un troisième ensemble de données non étiquetées et non vues pour s'adapter à la fois au domaine et aux données sources, ce qui revient à normaliser les ensembles de données d'entraînement et de test à une distribution commune, comme c'est généralement le cas pour la généralisation à un domaine.
3. Enfin, nous explorons les avantages supplémentaires de l'augmentation des données, en plus de BoW et DA, pour la SER interlinguistique.

0.0.11 Méthode proposée

Cette section décrit la méthode proposée, basée sur la combinaison de la méthodologie des sacs de mots (BOW) et de l'adaptation au domaine pour la SER interlinguistique. La figure 4.1 représente le schéma fonctionnel des deux méthodes étudiées ici, où la méthodologie d'extraction des caractéristiques BOW est étudiée avant ou après l'adaptation au domaine.

0.0.12 Généralisation des domaines avec CORAL

Nous proposons une variante de l'approche décrite dans laquelle un troisième langage est utilisé pour adapter les domaines d'entraînement et de test. La figure 4.2 décrit cette méthode de généralisation de domaine que nous appelons N-CORAL. Dans nos expériences, nous utilisons un jeu de données en langue chinoise comme domaine cible et adaptons les données d'entraînement et de test de trois langues différentes, à savoir l'allemand, le français et le hongrois, à ce domaine commun avant d'entraîner un classificateur SER. Le principal avantage de la méthode proposée est que nous n'avons pas besoin d'accéder aux données de test, comme dans les méthodes précédentes. Les mêmes équations de blanchiment et de recoloration de (6)–(9) sont utilisées, mais maintenant pour une langue commune.

Experimental Setup Pour la prédiction des émotions, nous avons utilisé quatre jeux de données dans quatre langues différentes. Le premier correspond à la base de données RECOLA [55]. Les deuxième et troisième jeux de données correspondent aux sous-ensembles en allemand et en hongrois de la base de données Sentiment Analysis in the wild (SEWA). Cette base de données a été utilisée dans les défis AVEC 2017 [57] et AVEC 2019 [44]. Le quatrième jeu de données correspond au sous-ensemble en langue chinoise du projet SEWA. Dans notre expérience, nous avons spécifiquement utilisé ce jeu de données non supervisé pour la méthode de généralisation de domaine N-CORAL proposée, décrite dans la section 4.4.3.1. Nous échantillonnons à un taux réduit tous les fichiers audio à 16 kHz pour un traitement ultérieur. Ensuite, à des fins d'augmentation des données, nous avons utilisé deux ensembles de données de bruit enregistré pour corrompre davantage les ensembles de données SEWA et RECOLA. Plus précisément, nous avons utilisé les sources de bruit enregistré AURORA [61]. Les architectures utilisées ici ont été motivées par le système de référence AVEC 2019 Challenge décrit dans [44] et comprenaient un Bi-LSTM à deux couches avec des couches cachées de tailles 64 et 32, respectivement. Une activation linéaire a été utilisée dans la couche de sortie, et une fonction de perte basée sur le coefficient de corrélation de concordance (CCC) (voir la section 4.5.4) a été utilisée pour la formation. La mesure de performance utilisée ici est la mesure typique utilisée dans les tâches de SER, c'est-à-dire le *coefficient de corrélation de concordance* (CCC). Enfin, comme mentionné précédemment, nos expériences reposent uniquement sur les partitions d'entraînement et de validation étiquetées des ensembles de données AVEC Challenge. Pour nos expériences, les ensembles de données de formation sont également utilisés pour la formation, mais 20 % de cet ensemble a été mis de côté pour ce que nous appelons " notre ensemble de validation " afin d'effectuer un réglage hyperparamétrique des modèles Bi-LSTM. Pour montrer l'importance des gains obtenus avec la méthode proposée, nous utilisons un test de signification par z-score entre CCC avec un niveau de 95% ($p < 0.05$) ; les comparaisons sont effectuées par rapport au système de référence AVEC 2019.

0.0.13 Résultats expérimentaux et discussion

Le tableau 4.1 présente les résultats de l'étude d'ablation pour plusieurs expériences, dont trois monolingues (formation-test en allemand, hongrois et français), trois multilingues (formation avec l'allemand, le hongrois et le français et test avec chaque langue individuellement), trois multilingues non vus (formation avec deux langues et test avec la troisième non vue), et enfin, les trois conditions

multilingues non vues, mais avec augmentation des données pendant la formation. Pour toutes les expériences, des caractéristiques BOW et un régresseur Bi-LSTM ont été utilisés. Dans le tableau 4.1, l'estimation de la valence est plus difficile que celle de l'excitation, ce qui corrobore les résultats de [70, 71]. De plus, à l'exception du hongrois, la formation multilingue n'a pas amélioré la précision par rapport à la formation monolingue. La présence de la langue de test pendant la formation s'est avérée importante. Enfin, dans le cas où la langue de test n'était pas présente, l'augmentation des données était significative.

Les tableaux 5.3 et 5.4 montrent les résultats inter-langues obtenus dans les différentes conditions explorées ici pour la prédiction de l'éveil et de la valence, respectivement. Les résultats inter-langues obtenus avec différents benchmarks (voir la section 4.5.3) et les systèmes proposés sont rapportés. Comme on peut le constater, dans l'ensemble, la méthode N-CORAL proposée a atteint les valeurs CCC les plus élevées de toutes les méthodes testées, surpassant également plusieurs paramètres multilingues présentés dans le tableau 4.1. L'augmentation des données, en revanche, a amélioré les performances dans la moitié des tâches interlinguistiques, mais en moyenne, elle n'a pas apporté d'avantage significatif pour le paramètre N-CORAL. En outre, on constate que dans les conditions impliquant les tâches interlinguistiques SEWA allemandes et hongroises, les valeurs de CCC les plus élevées ont été atteintes parmi toutes les tâches interlinguistiques testées, en particulier avec la méthode N-CORAL+BOW. Ces résultats suggèrent qu'un tel schéma proposé peut être utile pour la normalisation interlinguistique mais pas nécessairement pour les transcorpus où d'autres facteurs de nuance peuvent être présents. Pour la robustesse inter-corpus et inter-langues, N-CORAL combiné avec BOW et l'augmentation des données a montré les gains les plus significatifs, combinant ainsi les avantages de la méthode N-CORAL+BOW pour la robustesse inter-langues et les avantages de l'augmentation des données pour les facteurs de nuance inter-corpus.

Pour mieux comprendre certains de ces résultats, la Figure 4.4 présente un instantané du spectrogramme de modulation moyen sur plusieurs locuteurs pour trois langues différentes, dans des conditions d'éveil élevé (à gauche) et faible (à droite). Comme on peut le voir, des différences entre les langues peuvent être observées pour les cas d'excitation élevée et faible, ce qui motive le besoin de stratégies inter-langues. En plus des différences entre les langues, des différences peuvent également être observées entre les conditions d'excitation élevée et faible. La figure 4.5 montre les spectrogrammes de modulation pour les conditions de valence élevée (à gauche) et faible (à droite)

dans les trois langues. Comme on peut le voir, les différences sont plus subtiles, ce qui suggère une tâche de classification plus complexe.

Conclusion

Dans ce chapitre de la thèse, nous avons exploré l'utilisation combinée de la méthodologie des sacs de mots, de l'adaptation au domaine et de l'augmentation des données comme stratégies pour contrecarrer les effets négatifs de la reconnaissance des émotions vocales entre les langues (et les corpus). Une nouvelle méthode appelée N-CORAL a également été proposée, dans laquelle toutes les langues sont mises en correspondance avec une distribution commune (dans notre cas, un modèle linguistique chinois). Des expériences avec des langues allemandes, françaises et hongroises montrent les avantages de la méthode N-CORAL proposée, combinée à l'augmentation des données et à BOW pour la SER interlinguistique.. Chapitre 5 : Amélioration de la parole en fonction de la tâche et augmentation des données pour une reconnaissance multimodale des émotions dans des conditions bruyantes. Cependant, l'un des principaux inconvénients des systèmes basés sur la parole (qu'ils soient simples ou multimodaux) est leur sensibilité aux facteurs environnementaux, tels que le bruit additif et convolutif (par exemple, la réverbération d'une pièce). Dans ce chapitre de la thèse, nous avons utilisé un système multimodal. Les méthodes d'amélioration de la parole peuvent avoir deux objectifs très différents. Si elles visent à améliorer l'intelligibilité/la qualité, par exemple, la perception humaine devient le principal facteur déterminant et les améliorations de la qualité/l'intelligibilité sont généralement utilisées comme une figure de mérite (par exemple, [36]). Cependant, si l'amélioration est utilisée pour améliorer les applications de reconnaissance vocale en aval, d'autres mesures de résultats pilotées par la machine, telles que les améliorations du taux d'erreurs de mots, sont plus appropriées. Ainsi, en fonction de la tâche finale, la procédure d'amélioration peut être très différente. Les travaux de [38], par exemple, ont montré que l'amélioration basée sur la perte mimique était optimale pour les tâches de reconnaissance automatique de la parole (ASR) en aval. Ceci étant dit, nous émettons l'hypothèse que pour les systèmes de RAP multimodaux parole-texte, l'utilisation de deux procédures d'amélioration différentes seront utile, avec une procédure axée sur la qualité utilisée pour la branche parole (imitant la façon dont les humains perçoivent les émotions à partir de la parole) et une procédure axée sur la machine pour la branche parole-texte. Nous allons tester cette hypothèse dans cet article. Nous explorerons plus avant les avantages que l'augmentation des données peut offrir, en plus de l'amélioration de la parole, pour les systèmes de reconnaissance vocale multimodale dans la nature.

0.0.14 Méthode proposée

La figure 5.1 représente le schéma fonctionnel du pipeline AER multimodal proposé. Dans le cas qui nous intéresse ici, on suppose que la parole $S(i)$ est corrompue par un bruit de fond additif $N(i)$, ce qui donne un signal vocal bruité $Y(i) = S(i)+N(i)$. Dans le système AER multimodal, la branche supérieure se concentre sur l'extraction des caractéristiques liées aux émotions directement à partir de la composante vocale, tandis que la branche inférieure s'appuie sur un système de reconnaissance automatique de la parole (ASR) de pointe pour générer du texte à partir du signal vocal bruyant. Les caractéristiques sont ensuite extraites des transcriptions de texte. Les caractéristiques de la parole et du texte sont ensuite introduites dans un réseau neuronal profond pour la classification finale des émotions. Comme on sait que la parole bruyante nuit aux performances de l'EAR/ASR, nous incluons également une étape d'amélioration de la parole, une optimisée pour l'amélioration de la qualité de la parole (branche supérieure) et une autre pour l'ASR.

Experimental Setup Le jeu de données utilisé pour l'expérimentation est le Multimodal EmotionLines Dataset (MELD) [59]. Nous avons utilisé le jeu de données IEMOCAP pour montrer la généralisation du modèle proposé. Pour générer du texte à partir de la parole, il est nécessaire de disposer d'un système de reconnaissance automatique de la parole à la pointe de la technologie. Nous utilisons ici wav2vec 2.0, un système de reconnaissance vocale de bout en bout [72]. Nous explorons l'utilisation d'un algorithme d'amélioration optimisé pour la qualité de la branche vocale de la méthode proposée et d'un algorithme ASR optimisé pour la branche de génération de texte. Ensuite, nous nous concentrons sur les trois représentations les plus populaires, à savoir : les caractéristiques prosodiques, eGeMAPS et spectrales de modulation. En particulier, les caractéristiques prosodiques comprennent la fréquence fondamentale (F0), les mesures d'intensité et les probabilités de voisement, car elles ont été largement associées aux émotions [73]. Ensuite, l'ensemble étendu de paramètres acoustiques minimalistes de Genève (eGeMAPS) [74], qui a été largement utilisé dans de nombreux défis récents de reconnaissance des émotions (par exemple, [44, 75, 76]), est également exploré et contient un ensemble de 88 paramètres acoustiques relatifs à la hauteur, à la sonie, aux segments non voisés, à la dynamique temporelle et aux caractéristiques cepstraux. Enfin, les caractéristiques spectrales de modulation sont explorées car elles capturent les périodicités de second ordre dans le signal vocal et il a été démontré qu'elles transmettent des informations émotionnelles [46, 45]. Les caractéristiques spectrales de modulation (appelées MSF) ont été extraites en utilisant une taille de fenêtre de 256 ms et un pas de trame de 40 ms. Le lecteur intéressé est invité à consulter

[48, 45] pour obtenir des détails complets sur le calcul de cette représentation. Le texte a également été utilisé pour déduire le contenu émotionnel de documents écrits et il existe plusieurs méthodes et techniques de pointe. Ici, nous explorons trois méthodes récentes, à savoir BERT (Bidirectional Encoder Representations from Transformers), TextCNN et Bag-of-Words (BoW). Nous nous appuyons sur un réseau neuronal profond entièrement connecté pour la reconnaissance multimodale des émotions. L'exactitude, la précision, le rappel et le score F1 équilibrés sont utilisés comme figures de mérite pour évaluer les performances du classificateur d'émotions proposé.

0.0.15 Résultats expérimentaux et discussion

Le tableau 5.2 montre les performances obtenues pour chaque modalité individuellement. Dans le tableau, la caractéristique appelée "fusion" correspond à la fusion des caractéristiques MSF et eGeMAPS. Nous souhaitons explorer l'ensemble optimal de caractéristiques textuelles et vocales à inclure dans le système final. Nous considérons les modalités de la parole et du texte séparément dans cette étude. Comme on peut le constater, dans des conditions de parole propre et pour une ARE textuelle, les caractéristiques textuelles basées sur BERT ont donné les meilleurs résultats pour toutes les mesures, ce qui corrobore les rapports précédents. [77, 78, 79]. Pour les caractéristiques de la parole, dans des conditions propres, eGeMAPS a montré la meilleure performance globale des trois ensembles de caractéristiques testés, corroborant les résultats de [80]. Cependant, des gains supplémentaires ont été trouvés avec l'ensemble de caractéristiques fusionnées, ce qui suggère la complémentarité des caractéristiques spectrales et spectrales de modulation. Par conséquent, seul l'ensemble de caractéristiques fusionnées est exploré dans la condition de désaccord bruyant.

Cette dernière étude explore les performances du système proposé décrit dans la Figure 5.1, combinant l'amélioration de la parole optimisée pour chaque branche (parole et texte), ainsi que l'augmentation des données pour assurer la robustesse au niveau de l'entraînement du modèle. Les tableaux 5.7 et 5.8 montrent les résultats obtenus dans les lignes étiquetées 'data augmentation only' pour la tâche 1 et la tâche 2, respectivement. Comme on peut le constater, l'augmentation des données seule a déjà amélioré les résultats du REL, corroborant ainsi les conclusions de [81, 82]. Ensuite, nous évaluons les avantages de l'amélioration de la parole. Tout d'abord, nous explorons l'amélioration de la parole sans augmentation des données, ce qui signifie que les modèles ARE sont entraînés uniquement sur la parole propre. Plus précisément, nous pré-traitons les données

de test avec MetricGan+ pour la branche parole et avec l'exhausteur de perte mimique pour la branche texte, comme décrit dans la section 5.3. Les tableaux 5.7 et 5.8 montre les résultats obtenus dans les lignes étiquetées "amélioration seulement". Comme on peut le voir, l'application de l'amélioration de la parole améliore les performances par rapport aux conditions bruyantes, bien que les résultats finaux soient toujours inférieurs à ceux obtenus dans les conditions propres et soient également inférieurs à ceux obtenus avec l'augmentation des données. Les gains sont généralement plus importants à des valeurs SNR faibles, ce qui corrobore les résultats obtenus dans [83]. Comme on peut le constater, les résultats obtenus suggèrent que l'augmentation des données combinée à l'amélioration de la parole peut être une alternative viable pour une reconnaissance automatique robuste des émotions dans la nature sans avoir besoin de modèles de REL très complexes.

Afin de tester le caractère généralisable de la méthode proposée, trois expériences supplémentaires ont été menées le tableau 5.9. L'apprentissage a été effectué sur l'ensemble de données MELD, puis le modèle a été testé sur les données de test IEMOCAP non vues. Comme on peut le constater, le test sur des jeux de données croisés est une tâche extrêmement difficile où la précision des performances peut chuter jusqu'à 50 %. si des stratégies ne sont pas mises en place. Les innovations proposées, en revanche, offrent une certaine robustesse, et des gains de 30 %. et 44 %. ont été trouvés avec le système proposé pour les tâches 1 et 2, respectivement, par rapport à un système sans amélioration de la parole et sans augmentation des données spécifiques à la tâche.

Conclusion Ce chapitre de la thèse a exploré l'utilisation de l'amélioration de la parole spécifique à une tâche, combinée à l'augmentation des données, afin d'assurer la robustesse des systèmes multimodaux de reconnaissance des émotions à des conditions de test inconnues. Les expériences menées sur le jeu de données MELD montrent l'importance de BERT pour l'extraction de caractéristiques textuelles et d'un ensemble spectral de modulation eGEMAPS fusionné pour les caractéristiques audio. L'importance de l'augmentation des données dans la phase de formation et de l'amélioration de la parole spécifique à la tâche dans la phase de test est démontrée, et dans une tâche de classification de l'excitation faible/élevée, la précision dans des conditions bruyantes non vues est proche de celle observée avec la parole propre. Enfin, des expériences sur des bases de données croisées ont montré que les innovations proposées ont permis de réaliser des gains de 40 % par rapport à un système RAE sans amélioration/augmentation. Bien que les résultats suggèrent que l'amélioration spécifique à la tâche, combinée à l'augmentation des données, sont des étapes importantes vers une reconnaissance fiable des émotions "dans la nature", les algorithmes d'amélioration de la parole peuvent encore

être sous-optimaux et supprimer des informations importantes sur les émotions. Par conséquent, les travaux futurs devraient explorer le développement d'algorithmes d'amélioration sensibles aux émotions, capables de trouver un compromis entre la suppression du bruit et la précision de la reconnaissance des émotions.

Chapitre 6 : Conclusions et orientations futures de la recherche

Cette thèse visait à améliorer les performances du système de reconnaissance des affects dans des contextes hautement écologiques. Les principales contributions de cette recherche doctorale sont résumées ici, suivies de possibles domaines de recherche futurs.

0.0.16 Résumé des contributions

Cette thèse contribue à la reconnaissance des émotions vocales "dans la nature" à travers trois innovations principales. Tout d'abord, dans le chapitre 3, nous proposons de combiner la méthodologie audio du sac de mots avec des caractéristiques du spectre de modulation pour la robustesse environnementale. Nous tirons ensuite parti des propriétés inhérentes à la connaissance de la qualité du spectre de modulation et proposons de fournir au système SER une intelligence supplémentaire concernant la qualité du signal vocal d'entrée. Des expériences sont menées avec trois ensembles de données vocales multilingues dégradés par différentes sources et niveaux de bruit, ainsi que par la réverbération de la pièce. Les résultats expérimentaux montrent que les caractéristiques proposées i) surpassent systématiquement les systèmes de référence, ii) fournissent des informations complémentaires aux caractéristiques classiques, améliorant ainsi la performance avec la fusion de caractéristiques, et iii) montrent une robustesse à l'environnement et au changement de langue. De plus, nous montrons que lorsque le système proposé reçoit des informations de qualité, des améliorations supplémentaires sont obtenues, à la fois en termes de prédictions de l'éveil/de la valeur et dans la classification des émotions discrètes. Enfin, nous montrons l'impact de l'augmentation des données sur la robustesse au changement de langue, soulignant ainsi le potentiel du système proposé pour la SER dans la nature.

Ensuite, dans le chapitre 4, nous proposons de combiner la méthodologie du sac de mots (BOW) avec l'adaptation au domaine pour la "normalisation" de la distribution des caractéristiques et l'augmentation des données afin de rendre les algorithmes d'apprentissage automatique plus ro-

bustes dans toutes les conditions de test. Dans ce chapitre, nous nous intéressons particulièrement à la question de l'inadéquation des langues. Nous proposons une nouvelle méthode de généralisation de domaine que nous appelons N-CORAL, dans laquelle les langues de test sont mises en correspondance avec une distribution commune de manière non supervisée ; dans notre cas, avec la langue chinoise. Les expériences menées sur des ensembles de données d'émotions allemandes, françaises et hongroises ont montré que la méthode N-CORAL proposée, en combinaison avec BOW et l'augmentation des données, a atteint la meilleure précision dans la prédiction de l'excitation et de la valence parmi les systèmes testés, soulignant ainsi l'utilité de la méthode proposée pour la reconnaissance vocale des émotions "dans la nature".

Enfin, dans le chapitre 5, nous avons montré que l'amélioration de la parole spécifique à la tâche, combinée à l'augmentation des données, permettait d'obtenir une précision fiable de la reconnaissance multimodale des émotions dans des conditions bruyantes et non visibles. Les résultats obtenus dans des conditions bruyantes étaient proches de ceux obtenus avec une parole propre. Les expériences sur plusieurs bases de données ont montré que les innovations proposées ont permis des gains de 40 % par rapport à un système SER sans amélioration de la parole ni augmentation des données.

0.0.17 Travaux futurs

Dans ce travail, nous avons supposé que les conditions "dans la nature" sont celles où les utilisateurs se trouvent à l'extérieur dans des environnements bruyants ou à l'intérieur dans des pièces réverbérantes. Le bruit additif et la réverbération ont donc été utilisés comme principales sources de distorsions dans les expériences menées ici. Néanmoins, de plus en plus de systèmes de reconnaissance des émotions sont déployés dans les réseaux de télécommunication (par exemple, dans les centres d'appels). À ce titre, les travaux futurs devraient explorer l'impact de différentes altérations du réseau sur la précision de la RSE (par exemple, perte de paquets, perte de communication sans fil).

Au chapitre 4, le système N-CORAL proposé a été évalué en utilisant le chinois comme langue cible. Les travaux futurs devraient explorer l'utilisation d'autres langues, en particulier si la famille de langues cible coïncide avec la famille de langues source. D'autres expériences pourraient égale-

ment explorer l'ajout de plusieurs langues comme cibles (au sein ou entre différentes familles de langues).

Enfin, alors que les résultats du chapitre 5 ont montré que l'amélioration de la parole combinée à l'augmentation des données pouvait aider à la reconnaissance des émotions " dans la nature ", une analyse plus approfondie a montré que les algorithmes d'amélioration pouvaient supprimer des indices émotionnels importants du signal vocal. Les travaux futurs devraient explorer le développement d'algorithmes d'amélioration de la parole qui tiennent compte des émotions et qui compensent la suppression du bruit par la précision de la reconnaissance des émotions.

Chapter 1

Introduction

1.1 Affective computing

Affective computing is emerging as a prominent field in Human-Machine Interaction (HMI), as it provides the machine with information about the user's affective state, thus making the interaction more natural and human-like [3]. Emotion recognition is a part of Affective computing field, which helps to determine the mental and psychological state of the users. While positive emotions can help increase the feeling of fulfillment and life satisfaction, negative emotions may hamper quality of life and affect a person's performance in day-to-day activities, potentially leading to depression and anxiety. In fact, the recent COVID-19 pandemic has resulted in a global mental health crisis [84] that will have long-term consequences on society, economy, and healthcare systems. Being able to detect changes in affective states in a timely and reliable manner can allow individuals and organizations to put in place interventions to prevent burnout and depression.

Human-computer interaction (HCI) can be improved by adding emotional intelligence into the systems [4]. In education, assessing affective states such as boredom or loss of attention can help educators improve learning resources and personalize training for different students [5]. Due to the COVID-19 pandemic, this has become especially relevant with the surge in online learning, which limits the teacher-student interaction. Mental fatigue monitoring, in turn, can help driver safety by alerting drivers when they are too tired, and air traffic controllers [6]. Other domains where the development of emotional intelligence for systems can be relevant include smarthomes [7], gaming

[8], intelligent conversational systems [9], and neuromarketing [10], to name a few. Moreover, in real [85] and virtual reality systems [11], affect recognition systems can play a crucial role in measuring a user's perceived quality of experience. Representative examples of affective HMIs already in the market include fatigue level measurement in drivers; anger and stress level measurement of customer voices when dialing in to a call center; and adjustment of teaching strategies and material presentation based on student attention levels. As advances in machine learning, reinforcement learning, and deep learning are pushing us towards more human-like artificial intelligence (AI) systems, the need for robust affective interfaces has become critical. Overall, the affect recognition system can play a key role in individuals' mental health, performance, and overall quality of life. It can also be useful in improving human-computer interaction, allowing technologies to become more intuitive and personalized, ultimately leading to improved learning and other day-to-day experiences.

A user's way of expressing their affective state can also be varied based on the culture. This is mainly due to the fact that people convey emotions in different forms across different cultures. Hence, improving the performance of cross-language (cross-cultural) emotion recognition (ER) systems has gained significant interest recently. As the name suggests, cross-language systems have the ability to deal with different languages by generalizing (or "normalizing") the models. The success of a cross-language ER system relies on powerful features and models that can learn emotional cues across languages. Commonly, cross-lingual emotion prediction has relied on data normalization tools, domain adaptation, and multi-task training to improve the generalization power of the models. Recently, deep neural networks (DNNs) have been shown to result in state-of-the-art results for mono-lingual SER, but little success has been achieved in extracting features that perform consistently over different conditions across multiple datasets [12]. As such, there is ample room for innovations to provide models with the knowledge needed to become language-aware and agnostic.

A user's affective state can be expressed via many different modalities, including speech, gestures/posture, and physiological responses (e.g., changes in heart/breathing rates). As such, multimodal systems have been touted as the needed solution for reliable affective HMIs, thus eventually replacing the traditional keyboard and mouse [86]. For so-called "in the wild" applications, multimodal systems are preferred in order to compensate for certain confounds and to improve overall ER accuracy by providing the system with some redundancy and complementary information not available with unimodal systems [13, 14]. Multimodal systems, however, can be very time-consuming to

implement, costly to run, and potentially intrusive to the users (e.g., requiring on-body sensors) and their privacy. As such, text and speech-based ER systems have emerged as two popular solutions.

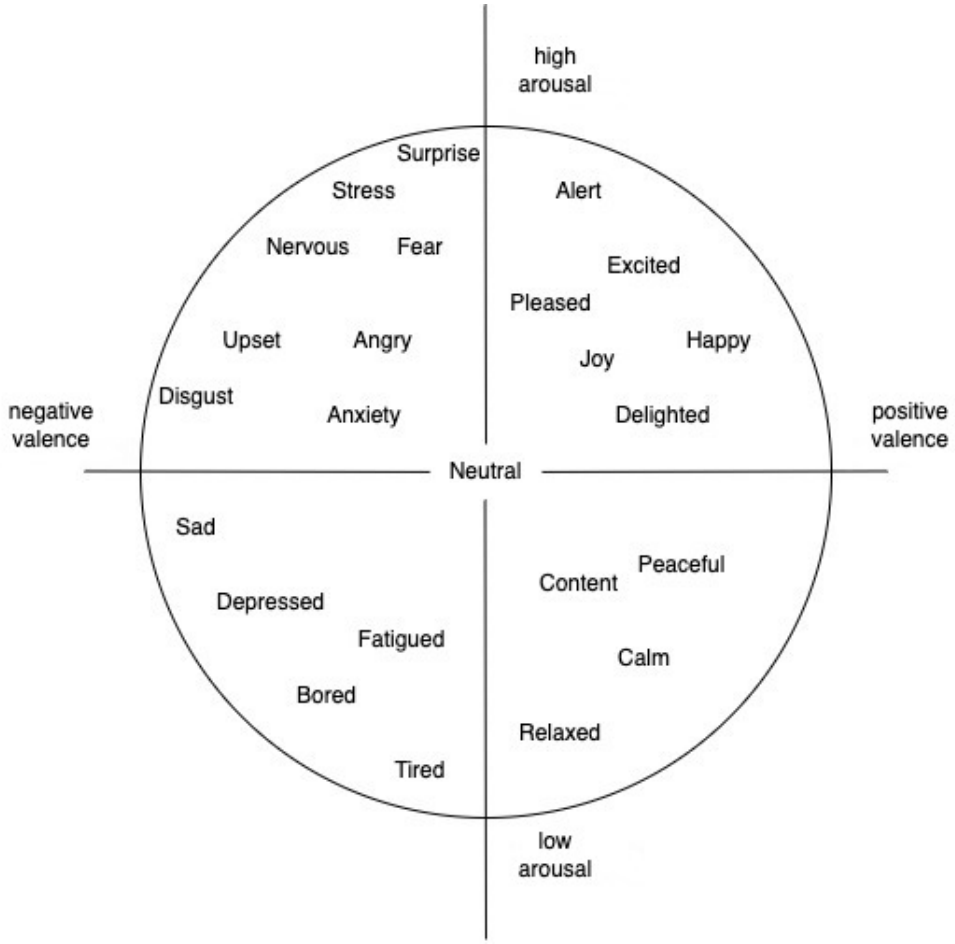

Figure 1.1: 2D emotion model based on the valance and arousal scale for different emotions

Affective states have been measured using subjective questionnaires in the past. These questionnaires usually periodically explore an individual's affective state and are answered by a continuous scale rating or a discrete scale containing several predefined choices. Discrete categorization of emotions includes the six basic emotions: happiness, sadness, fear, anger, disgust, and surprise, as introduced by [15]. Different multi-dimensional continuous scales for emotion classification have been proposed in [16, 17, 18]. The most accepted among these has been the arousal-valence model introduced by [16]. According to the dimensional approach, affective states are systematically related to one another, as can be seen in Figure 1.1. Three common dimensions are arousal (or activation), valence, and dominance [19]. Arousal describes the level of activation (passive or active), whereas valence describes the pleasantness. Dominance, in turn, describes the level of control of a person during the emotional experience (weak or strong). Valence and arousal have been shown to cover

the majority of affect variability, thus have been widely used in affective computing research [20], [21]. Psychological evidence suggests that these two dimensions are inter-correlated [87], [88], [89], [90].

Another reliable dimension for the prediction of emotion which has been widely utilized in affective computing is the liking dimension [91], [92], [93], [94]. Liking refers to how enjoyable the experience is. In the categorical approach, where each affective representation is classified into a single category, complex mental/affective states or blended emotions may be too tricky to handle [95]. Instead, in a dimensional approach, emotion transitions can be easily expressed, and observers can indicate their impression of moderate (less intense) and authentic emotional expressions on several continuous scales. Hence, dimensional modeling of emotions has proven to be helpful in affective content analysis domains [96]. More details on different approaches to model human emotions and their relative advantages and disadvantages can be found in [97],[98]. Typically, the Self-Assessment Manikin (SAM) [99] is a non-verbal pictorial assessment technique that directly measures the pleasure, arousal, and dominance associated with a person's affective state. SAM is a picture-oriented way to assess the different dimensions and can be used independently of language, thus making it possible to use across cultures and countries. The SAMs for valence, arousal, and dominance are shown in Fig. 1.2 [99].

Several questionnaires have been developed for affective state assessment, such as the NASA Task Load Index (NASA-TLX) [22] and the Subjective Workload Assessment Technique (SWAT) [23] questionnaires, The State-Trait Anxiety Inventory (STAI) [24] for anxiety measurement, and the visual analog scale (VAS) [25, 26] to name a few. Subjective questionnaires, despite being straightforward to use for affective assessment, suffer from various limitations, such as:

1. Psychological biases, such as peak-end and recency biases. These relate to the fact that individuals tend to use the most intense region and the final region of an experience with a forgetting "recency" factor that modulates the contribution of the stimulus in the final rating [100, 101].
2. Length of the questionnaire may cause a lack of compliance and careless responses, thus leading to erroneous ratings [102].
3. Questionnaires are usually administered after a task is completed, thus does not capture the real-time experience of the task. While one may increase the sampling rate of the

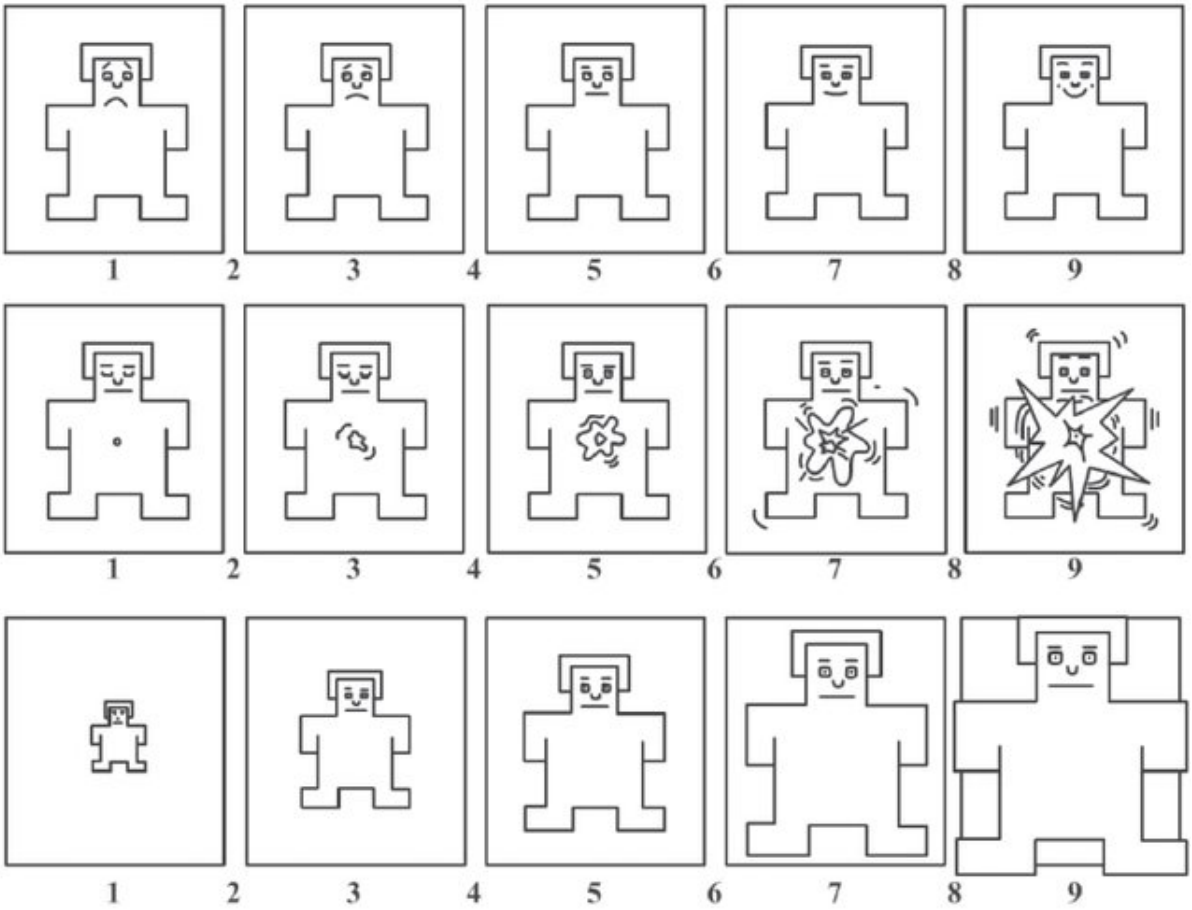

Figure 1.2: SAM for valence (top), arousal (middle) and dominance (bottom) dimensions

questions to provide real-time feedback, this may hinder the performance of the task itself, thus negatively affecting the experiment [6]. Increased frustration due to such hindrances has also been linked to subject dropout from studies.

Various neurophysiological-based methods based on electroencephalograms (EEG) have been employed for affective state evaluations [27, 28]. Other body signals resultant from autonomic nervous system responses, including the electrocardiogram (ECG), respiration signal, skin temperature, and galvanic skin response (GSR) have also been used to monitor different affective states [29]. Typically, physiological signals are also universal across people, thus suffering from comparatively less variability across cultures. However, such systems face three significant challenges, as follows:

1. Data collection of such signals is complex. The subject needs to be present in the lab or needs to wear a device properly to ensure the collected data is usable,
2. Devices to collect data can be costly,

3. Measured signals are highly susceptible to noise, such as movement and confounding factors.

An alternative to affective state monitoring via questionnaires is audio-video processing of data generated by humans. Humans are known to communicate emotions through speech, facial expressions, and body gestures. As a result, a large number of audio-visual databases for emotion recognition have been established by the research community [103, 104, 105]. Typically, audio and video information can be collected with microphones and cameras and a number of different audio-visual affect recognition challenges have emerged recently, including the Audio-Visual Emotion Challenge (AVEC) [43] and the Interspeech COMPARE challenge series [106]. Overall, audio signals provide several advantages over questionnaires and neurophysiological-based methods. For example, compared to subjective questionnaires, these signals can be monitored continuously without interrupting the task. Moreover, compared to physiological-based affective state monitoring, audio-based responses can be easily collected using smartphones and smart devices, which are now widespread around the world in people's homes, cars and offices, thus making data collection more economical and user-friendly. Lastly, relative to video, audio is considered less invasive and less intrusive on user privacy, thus has emerged as a very popular modality for emotion recognition, henceforth referred to as speech emotion recognition (SER).

1.2 Challenges for "in-the-wild" SER

Existing speech emotion recognition approaches are not perfect and have several limitations, including but not limited to:

1. Most existing SER systems have relied on laboratory-controlled settings and experiments. As such, the performance of such systems degrades in real-world scenarios where environmental factors hamper the signal quality. As such, quality-aware systems are still needed. Moreover, existing systems usually rely on features developed in clean conditions. To provide noise robustness, systems are trained using a combination of clean and noisy data (a process termed multi-condition training, or more recently, data augmentation). While this provides some robustness, especially if the noise type and levels seen in the test data are similar to those used during training, performance is still highly affected in unseen conditions (e.g., varying reverberation levels or noise types).

2. Existing SER systems have typically been mono-lingual, where emotions are detected only within the same languages used to train the systems. As users from varying languages and cultures interact with the same interface, it is paramount that next-generation solutions become multi-lingual. Cross-language SER, however, is an extremely challenging task with reported accuracies dropping drastically [70, 107, 108], as emotions can be conveyed differently across languages. Hence, improving the performance of cross-language emotion recognition systems has gained significant interest recently.
3. A major advantage of the speech modality is that recent advances in automated speech-to-text conversion have allowed for multimodal speech-and-text-based systems to emerge while requiring the collection of just one signal modality. However, in the case of real-world applications, where environmental factors, such as additive and convolutional noise (e.g., room reverberation), hamper the performance of multimodal systems, the contribution of the audio and the text modalities under noisy conditions has remained relatively unexplored.

This doctoral thesis research aims to tackle these critical issues, thus developing solutions that enable speech emotion assessment "in the wild." The next section details the objective of the thesis, followed by the contributions and publications list.

1.3 Thesis Objectives and Contributions

The overarching goal of this research is to build tools that enable emotionally-aware human-machine interfaces for applications in everyday settings. To achieve this goal, three objectives have been tackled:

1. Provide environmental awareness to the SER system via more robust features and speech quality awareness to the model itself,
2. Devise a domain adaptation strategy to enable cross-lingual capabilities for the developed SER system, and
3. Devise a robust multimodal emotion recognition system where task-aware speech enhancement modules are utilized to improve accuracy "in the wild."

To reach these objectives, three main families of innovations have been proposed, namely:

1. Numerous Challenges have emerged in the last decades in the field of SER. While the latest Challenges have shown that deep neural networks achieve the best results, existing input features are still a bottleneck and cause severe performance degradation in realistic "in-the-wild" scenarios. In the first contribution, we propose two innovations. First, we propose to combine the bag-of-audio-words methodology with modulation spectrum features for environmental robustness. Second, we take advantage of the inherent quality-awareness properties of the modulation spectrum and propose the use of a quality feature as an additional feature to be used by the speech emotion recognizer. Experiments are conducted with three multi-lingual speech datasets used in recent SER Challenges degraded by different noise sources and levels and room reverberation. Experimental results show the proposed features i) consistently outperforming benchmark systems, ii) providing complementary information to classical features, hence improving performance with feature fusion, and iii) showing robustness against environment and language mismatches. Moreover, we show that further improvements are obtained when the proposed system is provided with quality information. Overall, the proposed bag of modulation spectrum features are shown to be a promising candidate for"in-the-wild" SER.
2. Existing SER systems have limited ability in coping with emotion data from different languages. One method that has been widely used in allied domains to cope with such data distribution mismatches is termed domain adaptation (DA). While DA algorithms have been widely used in computer vision and natural language processing applications, their use within speech emotion recognition has yet to be explored. In the second contribution, we propose the use of two simple yet effective DA algorithms, namely, correlation alignment (CORAL) and sub-space alignment (SA), to improve continuous SER performance in cross-lingual settings. Using emotional data in German, French, and Hungarian languages, we show the advantages of combining the proposed bag-of-word approach with DA to improve cross-language SER performance. Finally, a new variant of the CORAL method is proposed, termed N-CORAL. Here, both target and source domains are adapted to a third unseen language in an unsupervised manner (i.e., without the need for additional emotion labels); in the case of our experiments, the Chinese language. Experimental results show the additional benefits of the proposed N-CORAL method for both arousal and valence prediction.
3. Multimodal emotion recognition systems can rely on a combination of audio, video, text, or physiological signals. Collecting multiple signal modalities, however, can be very intrusive,

time-consuming, and expensive. Recent advances in deep learning-based speech-to-text and natural language processing systems, however, have enabled the development of reliable multimodal systems based on speech *and* text while only requiring the collection of audio data. Audio data, however, is extremely sensitive to environmental disturbances, such as additive noise, and thus faces some challenges when deployed "in the wild". To this end, speech enhancement algorithms have been developed to provide noise robustness at the signal input level. Data augmentation, in turn, has been deployed during training time to also provide noise robustness, but at the model level. In our third contribution, we explore the combination of speech enhancement and data augmentation to improve multimodal emotion recognition in noisy conditions. More specifically, we show the importance of task-aware speech enhancement (i.e., enhancement optimized for speech-to-text versus optimized for quality enhancement) for multimodal SER. Experimental results show the proposed system achieving recognition accuracy in noisy conditions (e.g., 10 dB signal-to-noise ratio babble noise) similar to levels achieved in clean conditions. When compared to a benchmark system without speech enhancement or data augmentation, improvements of 40% could be seen in cross-dataset conditions.

These three innovations have been described in several manuscripts, as listed below in chronological order. Where appropriate, the chapters of this thesis in which these publications appear are also specified.

Articles published in refereed journals

*Related

J1 "Quality-Aware Bag of Modulation Spectrum Features for Robust Speech Emotion Recognition", **Kshirsagar, S.**, Falk, T. H., *IEEE Transactions on Affective Computing*, early acccess, available online July 2022. [64] [Chapter 3]

J2 "Cross-Language Speech Emotion Recognition Using Bag-of-Word Representations, Domain Adaptation, and Data Augmentation", **Kshirsagar S.**, Falk, T. H. (2022), *Sensors, Special Issue on Emotion Recognition Based on Sensors*, accepted with revisions. [109] [Chapter 4]

J3 "Task-Specific Speech Enhancement and Data Augmentation for Improved Multimodal Emotion Recognition Under Noisy Conditions", **Kshirsagar S.**,Pendyala A., Falk, T. (2022), under review *Frontiers in computer Science*, [110] [Chapter 5]

J4 "Modulation Spectral Signal Representation for Quality Measurement and Enhancement of Wearable Device Data: A Technical Note", Tiwari, A., Cassani R., **Kshirsagar S.**, Diana P., Zhu Y. Falk, T. (2022), *Sensors 2022, 22, 4579.* [111]

J5 "COVID-19 Detection via Fusion of Modulation Spectrum and Linear Prediction Speech Features", Yi Zhu, Tiwari A., Monteiro J., **Kshirsagar, S.**, Falk, T. H. (2021), under review *ACM/IEEE Transaction on Audio, Speech and Language Processing.*

*Non-related

J1 "COVID-19 Detection via Fusion of Modulation Spectrum and Linear Prediction Speech Features", Yi Zhu, Tiwari A., Monteiro J., **Kshirsagar, S.**, Falk, T. H. (2021), under review *ACM/IEEE Transactions on Audio, Speech and Language Processing.*

Conference proceedings, challenge reports, and abstracts

- "Speech-based Stress Classification based on Modulation Spectral Features and Convolutional Neural Networks", Avila, A. R., Kshirsagar, S. R., **Tiwari, A.**, Lafond, D., O'Shaughnessy, D., Falk, T. H. (2019, September), *In 2019 27th European Signal Processing Conference (EUSIPCO) (pp. 1-5). IEEE.* [112]
- "Development of the INRS-EMT Scene Classification Systems for the 2020 DCASE Challenge", Gaballah, Amr, Anderson Avila, Joao Monteiro, Parth Tiwari, Shruti Kshirsagar, Falk, T. H. (2019, October). *DCASE challenge 2020)* [113]
- "Modulation Spectral Signal Representation and I-vectors for Anomalous Sound Detection", Gaballah, Amr, Anderson Avila, Joao Monteiro, Parth Tiwari, Shruti Kshirsagar, Falk, T. H. (2019, October). *DCASE challenge 2020)* [114]
- "Multimodal emotion recognition "in the wild"", Shruti Kshirsagar, Falk, T. H. (2019, October), abstract *STARaCOM- Industrial meetup (2019), Montreal, Canada)*

- "Exploring Domain Adaptation for Monolingual and Cross-lingual Speech Emotion Recognition", **Kshirsagar S.**, Falk, T. H. (2019, September), abstract, *ACM Canadian Celebration of women in computing conference (2019), Toronto, Canada*

1.4 Thesis organization

While this introductory chapter has presented the challenges with affective state recognition in the wild and laid out the foundation for the contributions described herein, the remainder of this dissertation is structured as follows: Chapter 2 provides an overview of the state-of-the-art methods in affective state recognition, as well as lists publicly-available databases used herein. Chapter 3 presents the first contribution proposing the use of bag of modulation features and quality-awareness for "in the wild" systems. Next, Chapter 4 describes the second contribution which deals with domain adaptation for cross-language emotion recognition. In Chapter 5, the third contribution is described where multimodal systems are presented to improve noise robustness and performance for affective state recognition by using multiple modalities simultaneously. Lastly, Chapter 6 provides this thesis's general conclusions and future research areas.

It is important to emphasize that this thesis takes the manuscript-style form of presentation, thus the introduction and/or methods sections of some chapters may contain repeated information that the reader may wish to skip as they see fit.

Chapter 2

Background: Speech and Text based Affective Computing Systems

2.1 Introduction

Affective computing/Emotion recognition systems are mainly comprised of two modules: the front-end and the back-end. The front-end corresponds to the signal processing pipeline, where pre-processing, enhancement and feature extraction are commonly performed. The back-end, in turn, relies on the machine learning pipeline, where factors such as domain adaptation, data augmentation, and classification/evaluation are performed. More recently, however, with the advances in end-to-end deep neural networks, the front and back-ends have merged into one and the neural network learns the features representations and classifier mapping in one step. For the purpose of this thesis, the more classical approach is assumed, as depicted by Fig. 2.1, where signal analysis is performed at the front-end and machine learning at the back-end.

In the following sections, we will review the relevant signals analyses used in this thesis, including speech enhancement, recognition, and quality estimation methods, feature extraction (speech and text) and multimodal fusion strategies. Next, we describe the different components of the ML pipeline, including the figures-of-merit used in the experiemnts. Finally, we described the publicly available datasets used in our studies.

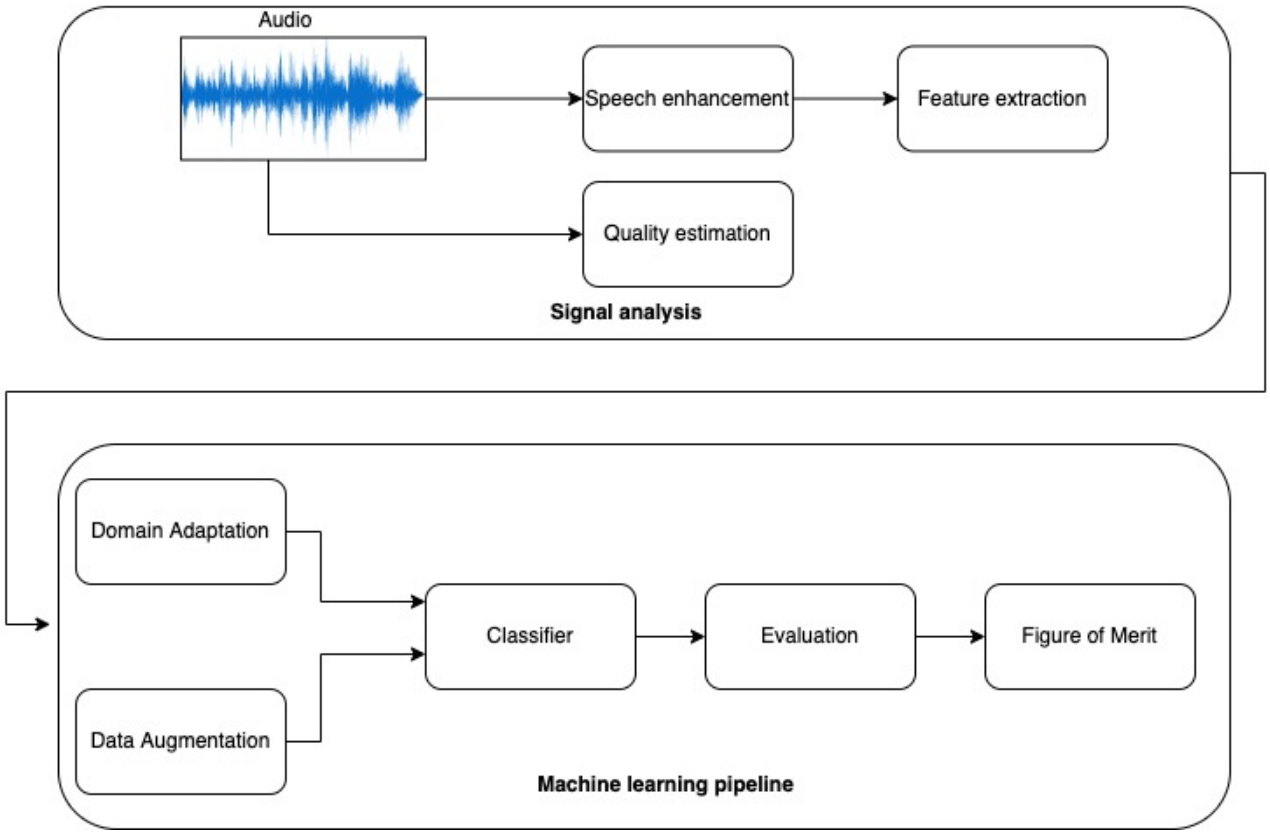

Figure 2.1: Emotion recognition system: (top) front-end and (bottom) back-end modules.

2.2 Signal Analyses Methods (Front-End)

As shown in the top part of Fig. 2.1, the signal analyses components typically include speech enhancement, speech quality estimation, and speech feature extraction. Next, these modules are described in more detail.

2.2.1 Speech Enhancement

Speech signals can be affected by environmental factors, such as additive and convolutional noise (i.e., room reverberation). As such, speech enhancement has been widely used to counter such detrimental effects. Representative speech enhancement algorithms can include the classical single-channel spectral enhancement (SSE) [30], relative convolution transfer function (RCTF) [31], Wiener filtering [32], and statistical model based methods [33], as well as the more recent DNN based models, such as the speech enhancement GAN (SEGAN)[34] and the recurrent denoising autoencoder [35].

DNN-based enhancement is a newer approach that is still gaining popularity. In addition to the more recent ones mentioned above, methods have ranged from multi-layer perceptron [115], to recurrent neural network [116], including long short-term memory (LSTM) models [117, 118], and convolutional neural networks (CNNs) [119].

Speech enhancement can follow two principles: enhancement for humans or enhancement for downstream machine-based tasks. When dealing with humans, enhancing the quality and intelligibility of the signal is of utmost importance, thus a trade-off between noise suppression and intelligibility must be attained. For machine-based tasks, however, intelligibility may not necessarily be the ultimate end goal. In this case, the algorithm trades-off suppression with the downstream task performance. In the case of speech-to-text, word error rates are typically used as the criterion. For multimodal speech emotion recognition, two separate approaches may be needed, one destined to keep speech intelligibility to allow for emotional cues to be preserved and another to improve speech-to-text conversion for the downstream text-based analysis.

The most widely used state-of-the-art enhancement descrition are given as follows:

2.2.1.1 MetricGAN+: A quality-optimized enhancement method

MetricGAN+ is a recent state-of-the-art deep neural network specifically optimized for quality enhancement of noisy speech [36]. In particular, two networks are used. The discriminator's role is to minimize the difference between the predicted quality scores (given by the so-called PESQ, perceptual evaluation of speech quality, rating [37]) and actual PESQ quality scores. PESQ is a standardized International Telecommunications Union full-reference speech quality metric that maps a pair of speech files (a reference and the noisy counterpart) into a final quality rating between 1 (poor) and 5 (excellent). PESQ has been widely used and validated across numerous speech applications.

The generator's role, in turn, is to map a noisy speech signal into its enhanced counterpart. The discriminator and generator models are trained together to enhance the noisy signal in a manner that maximizes the PESQ score of the enhanced signal. MetricGAN+ builds on the original MetricGAN [120] via two improvements for the discriminator and one for the generator. More specifically, for the discriminator training, along with the enhanced and clean speech signals, the noisy speech was

also used to minimize the distance between the discriminator and target objective metrics. The second improvement is that the speech generated from the previous epochs is reused to training the discriminator to avoid the catastrophic forgetting of the discriminator. For the generator, in turn, the learnable sigmoid function was used for mask estimation. The interested reader is referred to [120, 36] for more details on the MetricGAN and MetricGAN+ speech enhancement methods.

2.2.1.2 Mimic loss: an ASR-optimized enhancement method

Spectral mapping-based speech enhancement is an enhancement method specifically optimized for downstream ASR applications [38]. We refer henceforth to this method as 'mimic loss based enhancement' as the model uses mimic loss instead of student-teacher learning, thus the speech enhancer is not jointly trained with a particular acoustic model. According to the developers, the speech enhancer could be used as a pre-processor for any ASR system. The overall system is comprised of two major components: a spectral mapper and a spectral classifier which are trained in three steps.

First, a spectral classifier is trained to predict senone labels from clean speech with a cross-entropy criterion, resulting in a classification loss L_C between predicted and actual senones. The weights of this spectral classifier are then frozen and used in the last step. Second, a spectral mapper is pre-trained to map noisy speech features to clean speech features using a mean squared error (MSE) criterion. This results in a fidelity loss L_F between the denoised features and features from the clean speech counterpart. In [38], log-spectral magnitude components extracted over 25ms windows with a 10-ms shift are used as features, and a deep feed-forward neural network is used for mapping.

Lastly, noisy speech is input to the pre-trained spectral mapper, resulting in a denoised version, which is input to the "frozen" spectral classifier, resulting in a predicted senone. In parallel, the clean speech counterpart is also input to the frozen spectral classifier, resulting in a soft senone label, resulting in a mimic loss L_M between the soft senone label and the predicted senone. The spectral mapper is then retrained using joint loss (L_F and L_M), thus allowing the enhancer to emulate the behavior of the classifier under clean conditions while keeping the projection of the noisy signal closer to that of the clean signal counterpart. The same hyperparameters described in

[38] were used herein. The interested reader is referred to [38] for more details on the mimic loss enhancement method.

2.2.2 Objective Speech Quality Estimation

Measuring speech quality in an automated, objective manner has been of the interest of the telecommunications industry for the last three decades. Objective speech quality measurement can follow two categories: double-ended (or also called intrusive) or single-ended (non-intrusive). As the name suggests, double-ended methods require access to two signals, the noisy processed signal and its original clean counterpart. As access to the clean signal is often not possible in practice, single-ended methods have been developed that take as input only the noisy processed signal. The International Telecommunications Union (ITU-T) has standardized both double- and single-ended methods. The most recent are ITU-T Recommendation P.863, also called "perceptual objective listening quality assessment (POLQA)" [39], the successor of P.862 (perceptual evaluation of speech quality, PESQ [37]), and the ITU-T Recommendation P.563 [121], respectively. More recently, a single-ended measure optimized for noisy and reverberant speech was developed and termed "speech to reverberation modulation ratio (SRMR norm)" [40].

2.2.3 Feature Extraction from Speech

Several features have been used for SER, including prosodic, spectral and modulation spectral features. The openSMILE [41] toolkit has been widely used by the SER community and has been used to extract benchmark features for the majority of the recent SER Challenges. The toolkit can be used to extract 6000+ features, but different subsets have been utilized for different applications. For example, the Interspeech speaker state challenge 2011 feature set comprises a subset of 118 low-level descriptors (LLDs). Of these, the first 59 features include 50-spectral, five-voice, and four energy-related features, along with extra 59 features corresponding to their 1-time step delta counterparts [42]. These features are extracted over 20ms frames with 10ms hops and have been used as benchmarks in several AVEC Challenges.

Mel-frequency cepstral coefficients (MFCCs), in turn, have been widely used across many speech applications. They are cepstral coefficients computed after a mel-scale frequency mapping is per-

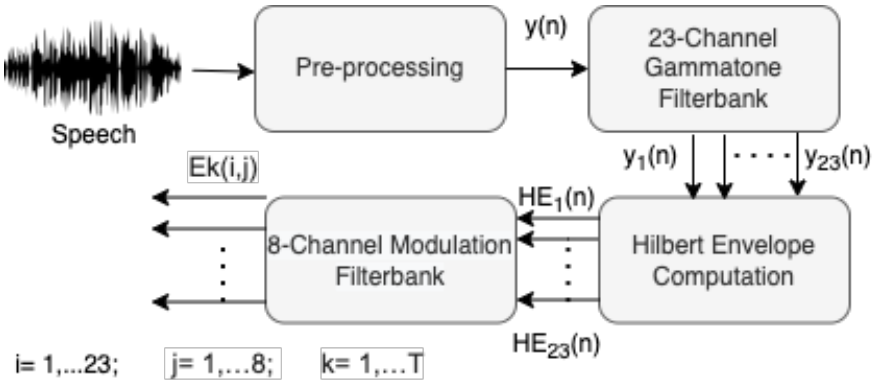

Figure 2.2: Signal processing steps involved in computation of the modulation spectral representation

formed to simulate human cochlear processing [122]. For SER, it is typical for 39-dimensional MFCC feature vectors to be used, comprised of 13 MFCCs, 13 delta MFCCs, and 13 double-delta MFFCs. These features are extracted over 20 ms windows, hop size of 10ms, and with 64 mel filters. They have been used as benchmarks in then AVEC 2018 and AVEC 2019 Challenges.

Moreover, prosodic features including fundamental frequency (F0), intensity measures, and voicing probabilities have also been widely used for SER. The eGeMAPS feature set (extended Geneva Minimalistic Acoustic Parameter Set) for example, comprises 88 acoustic parameters relating to pitch, loudness, unvoiced segments, temporal dynamics, and cepstral features. These have also been used as benchmarks in several AVEC Challenges [57, 43, 44].

More recently, an auditory-inspired modulation spectral signal representation has been proposed for improved SER, as it captures the long-term dynamics of the speech signal [45, 46]. The measure is the same representation used in the SRMRnorm quality measure described in [47], thus has been shown to provide some robustness against ambient noise and reverberation [46]. The modulation spectral features (termed MSFs) were extracted over 256ms windeows and 40ms hops, following the processing steps described in [47].

For the sake of completeness, the signal processing steps involved in the computation of the MSFs can be found in Fig. 2.2. The first step corresponds to normalization of the speech level to -26 dBov (dB overload) in order to compensate for unwanted energy variations across speech signals. Next, the pre-processed signal is decomposed into 23 subband signals using a 23-channel gammatone filterbank with filter center frequencies ranging from 125 Hz to approximately half the sampling frequency [48]. The bandwidths of the filters follow the equivalent rectangular bandwidth

paradigm [49]. Next, the Hilbert envelope is computed from each of the 23 subband signals using the Hilbert transform. Hilbert envelopes $HE_j(n), j = 1, \ldots, 23$ are windowed by a 256 ms Hamming window with a shift of 40 ms. Envelopes are then grouped across 8 bands following recent evidence of a modulation filterbank structure in the human auditory system. Here, an 8-channel modulation filterbank with center frequencies equally spaced in the logarithmic scale from 4 to 128 Hz is used. The final modulation spectral representation (termed $E_k(i, j)$) is comprised of a $23 \times 8 \times T$ tensor of modulation energies, where i indexes the number of gammatone filters used ($i = 1, \ldots, 23$), j corresponds to the number of modulation filters ($j = 1, \ldots, 8$), and k indexes the frame being analyzed ($k = 1, \ldots, T$), where T corresponds to the total number of frames available in a speech recording. The interested reader is referred to [48] for complete details on the computation of this representation.

From this representation, several features can be extracted from both the acoustic and modulation frequencies directions. First, modulation energy for each acoustic-modulation frequency bin can be computed, resulting in $23 \times 8 = 184$ features. Next, six spectral descriptors (Θ_1-Θ_6), computed on a per-frame basis, are extracted. The first three descriptors measure spectral changes across modulation bands, whereas the last three capture changes across the acoustic bands. In particular:

i) Θ_1: specifies the energy distribution of speech along the modulation frequency and is computed as the average modulation energy computed across each of the j modulation bands, i.e.,

$$\Theta_{1,k}(j) = \frac{\sum_{i=1}^{N} E_k(i, j)}{N}, \tag{2.1}$$

ii) Θ_2: specifies the spectral flatness of each modulation band, defined as the ratio of the geometric mean of a spectral energy of modulation channels to the arithmetic mean, i.e.,

$$\Theta_{2,k}(j) = \frac{\sqrt[N]{\Pi_{i=1}^{N} E_k(i, j)}}{\Theta_{1,k}(j)}. \tag{2.2}$$

Values close to unity suggest uniform spectral distributions, whereas values near zero suggest wide spectral amplitude variability.

iii) Θ_3: measures the spectral centroid (or center of mass) of each of the j modulation bands, i.e.,

$$\Theta_{3,k}(j) = \frac{\sum_{i=1}^{N}(i * E_k(i,j))}{\sum_{i=1}^{N}(E_k(i,j))}. \tag{2.3}$$

As mentioned previously, the last three descriptors measure the dynamics of the signal across the gammatone acoustic bands. To reduce the amount of parameters to be computed, instead of measuring the descriptors for each of the 23 gammatone bands, here we group the gammatone filters into five groups: $G1 = 1-4$; $G2 = 5-8$; $G3 = 9-12$; $G4 = 13-18$; and $G5 = 19-23$. Modulation information from channels within the same group are summed together as $\xi_k(l,j) = \sum_{i \in G_l} E_k(i,j)$, $l = 1, \ldots, 5$. Having this said, the next three descriptors are computed as:
iv) Θ_4: corresponds to the spectral centroid computed across modulation channels for each of the five groups, i.e.,

$$\Theta_{4,k}(L) = \frac{\sum_{j=1}^{8}(j.\xi_k(L,j))}{\sum_{j=1}^{8}(\xi_k(L,j))}. \tag{2.4}$$

v) $\Theta_{5,k}(L)$: computes the slope of the modulation energy across each of the five grouped bands via the linear regression coefficient of a first-order polynomial fit across the five groups, and lastly:
vi) $\Theta_{6,k}(L)$: corresponds to the regression error associated with a first-order polynomial fit across the five bands. This captures the rate of change of each acoustic frequency group, thus provides some indication of temporal dynamics. The interested reader is referred to [46] for complete details on the computation of these descriptors.

Overall, the first three descriptors are computed across each of the eight modulation channels, hence resulting in an additional 24 features. The last three, in turn, are computed across each of the five grouped bands, hence totaling an additional 15 features. A total of 39 descriptors are thus computed per speech file. These are then appended to the 184 modulation spectral features to generate a final 223-dimensional MSF feature vector.

2.2.4 Bag-of-Audio-word representation

The bag-of-words (BOW) methodology was initially proposed for natural language processing applications [123]. However, it has gained increased attention recently also for applications involving features extracted at short time scales across numerous modalities. With audio, BOW has been termed bag-of-audio-words (BoAW). The BoAW approach has been utilized for music information

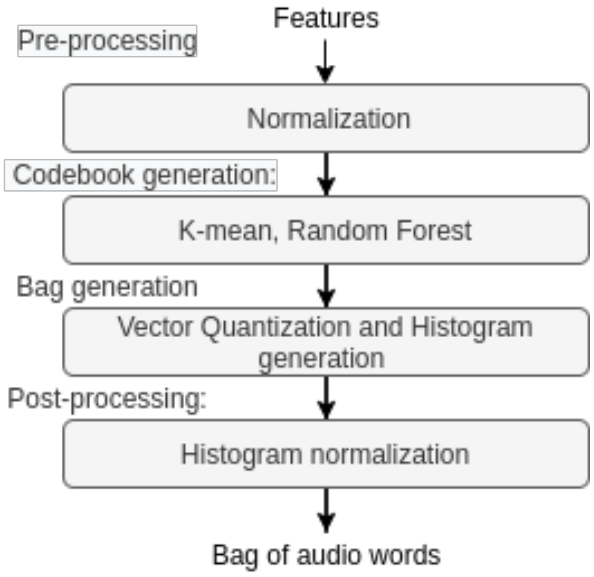

Figure 2.3: Steps for bag of audio word generation.

retrieval [124], and more recently, for SER [125]. BoAW is a methodology that allows for aggregation of short-term features (e.g, MFCC computed every 40 ms) into longer duration properties. Commonly, statistical functionals (e.g., mean, standard deviation, min, max) have been used to aggregate short-term features into utterance-level details. The BoAW approach replaces the functionals and represents the audio features in the form of compact 'audio words' with better resolution than utterance-level aggregation. Classifiers are then trained on histogram features, which represent the frequency of occurrence of the respective *audio words*. Typically, openSMILE-based features have been used within the BoAW approach. Here, we explore the usefulness of replacing such spectral features with modulation spectral ones that have been shown to provide improved robustness against environmental factors.

The bag-of-words methodology has been described in detail in [125] and is summarized here for completeness. The top part of Fig. 2.3 depicts the processing steps involved in the computation of BoAWs. First, relevant feature representations are extracted. As mentioned previously, these have usually been openSMILE-related features, such as the ComParE feature set [126]. Pre-processing is performed where normalization is typically achieved. Next, initial codebooks are generated and vector quantization is performed in order to find the representative audio words to be used for classification. For codebook initialisation, different methods have been explored, including unsupervised methods such as k-means clustering and random sampling, and supervised methods where different codebooks are obtained per class and then aggregated into a final super-codebook. Codebook size

is an important parameter as it dictates the final feature dimensionality and the amount of 'overlap' between audio words. During vector quantization, feature vectors are assigned to the audio word with the smallest Euclidean distance from all codebook entries. A histogram of the frequencies of occurrence of each word in the codebook is then created for each audio segment.

Prior to classification, two post-processing steps can also be applied. Instead of using a hard encoding, soft vector quantization can also be employed (e.g., Gaussian encoding) where the number of occurrences of each word from the codebook (called term frequencies, TF) can be taken into account. This approach is useful in multiple assignments, where a certain number of closest words are also considered along with the closest word. The resulting histograms can then be normalized by applying different strategies that take into account e.g., difference in audio file duration. Commonly-used normalization schemes include logarithmic TF-weighting and inverse document frequency (IDF) weighting [124]. In this work, the openXBOW toolbox was employed [125].

The bottom part of Fig. 2.3 depicts the processing steps of the proposed method, which combines the modulation spectral feature extraction, BoAW computation and classification. For BoAW computation, we start with a codebook size of 500 words to test the usefulness of the proposed method and then fine-tune it for valence and arousal estimation. Random sampling (using the random sampling++ function of openXBOW) was used and the ten closest words in the codebook were used with soft VQ. The number of TFs is compressed by applying logarithmic TF-weighting. These settings showed to lead to improved results in pilot experimentation. Lastly, we explore the use of LSTM and SVM regression/classification algorithms to either estimate valence and arousal values or to predict discrete emotions.

2.2.5 Feature Extraction from Text

Text has also been widely used to infer the emotional content of written material and several state-of-the-art methods and techniques exist. Here, we explore three recent methods, namely BERT (Bidirectional Encoder Representations from Transformers), TextCNN, and Bag-of-Words (BoW). More details about each method are given below:

2.2.5.1 BERT - Bidirectional Encoder Representations from Transformers

BERT is based on a transformer network and attention mechanism [1] that also learns contextual relations between words in the text [50]. BERT comes in two flavours: BERTBase and BERTLarge. The BERTBase model uses 12 layers of transformers block with a hidden dimension of 768 and 12 self-attention heads; overall, there are approximately 110 million trainable parameters. On the other hand, BERTLarge uses 24 layers of transformers block with a hidden size of 1024 and 16 self-attention heads, resulting in approximately 340 million trainable parameters. We employ the BERTBase model for text feature extraction and BERT hidden state vector is used as input to the emotion recognition system. The interested reader is referred to [1] for more details on BERT.

The BERT model has been pre-trained on the Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) task, as shown in Fig. 2.4. The two pre-training objectives allow it to be used on any single sequence and sequence-pair tasks without substantial task-specific architecture modifications. BERT embeddings encode information about parts of speech, roles and syntactic chunks [127], thus naturally learns some syntactic information [128]. BERT layers produce more context-specific representations and the generated embeddings are vectors that encapsulate the meaning of the word; similar words have closer numbers in their vectors. In summary, the embeddings start out in the first layer as having no contextual information. As these embeddings move deeper, they carry more and more contextual information with each layer. In the final layer, information that is specific to BERT's pre-training tasks (the MLM and NSP) are obtained. BERT features are useful to characterize missing words (MLM) or whether the second sentence came after the first (NSP). The BERT embedding can be used as an input to an emotion recognition model.

2.2.5.2 TextCNN

TextCNN is a deep learning model for short text classification tasks and has been used as a baseline model for text classification [51]. TextCNN transforms a word into a vector using word embeddings, which are then fed into a convolutional layer, followed by a max-pooling layer, and a fully connected output layer. In our experiments, TexCNN embeddings were extracted using the model described in [59]. We used three convolutional layers with 64 filters and kernel sizes of 3, 4, and 5 respectively in each layer, followed by max-pooling and finally 150 dense layers to extract the final text features.

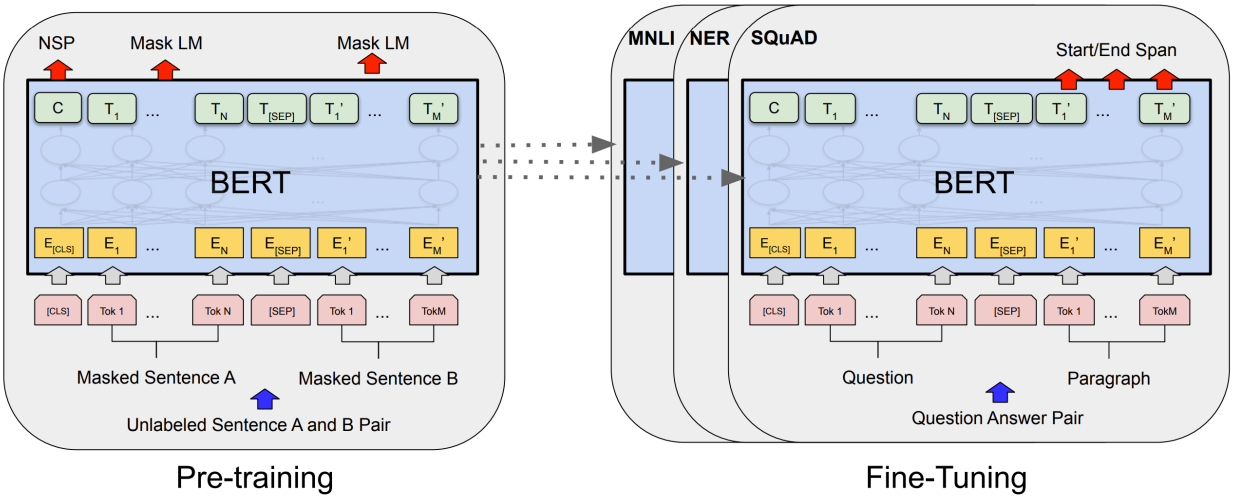

Figure 2.4: BERT model taken from [1]

2.2.5.3 Bag-of-Words:

The bag-of-words (BOW) method is commonly employed in natural language processing [52]. The approach is straightforward and flexible and can be used in many ways to extract features from documents. BoW represents text that describes the occurrence of words within a document. It consists of two parts: a vocabulary of known words and a measure of the presence of these words. It is called a "bag" of words because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not wherein the document. In this method, first, a word histogram is generated within a text document. Next, the frequencies of each word from a dictionary are computed, and finally, the resultant vector is fused as textual features. For our experiment, we used CountVectorizer from the Sklearn library. We obtain a 652 dimension feature vector for each utterance. We used the unigram model for generating the BOW representation.

2.2.6 Speech Recognition Systems

In order to generate text from speech, a state-of-the-art automatic speech recognizer is needed. Here, wav2vec 2.0, an end-to-end speech recognition system, is used [72]. A complete description of the method is beyond the scope of this thesis, hence only an overview is provided; the interested reader can obtain more details from [72]. Wav2vec 2.0 relies on the raw speech waveform as input.

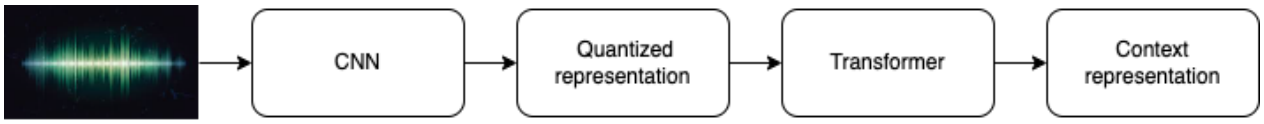

Figure 2.5: Training flow of wav2vec2 ASR.

This 1-dimensional data then passes through a multi-layer 1-d CNN to generate speech representation vectors. Vector quantization is then used on these latent representations to match them to a codebook. Half of the available speech data is masked and the remaining quantized data is fed into a transformer network. By using contrastive loss, the model attempts to predict the masked vectors, thus allowing for pre-training on unlabeled speech data. The model is then fine-tuned on labeled data for the subsequent down-streaming ASR task.

The Wav2Vec2.0 model architecture is shown in Fig. 2.5. It consists of a CNN-based encoder network, a transformer-based context network, and a Vector Quantization module. The interested reader can obtain more information from [72].

2.2.7 Multi-modal Fusion

Multimodal emotion recognition systems have been shown to achieve accuracies higher than those obtained by each modality individually [129]. Kim et al. [130], for example, proposed a new deep learning architecture for audio-video emotion recognition using Deep Belief Networks. In [131], convolutional deep belief network (CDBN) models were proposed that learned salient multi-modal features of expressions of emotions, which significantly improved the performance over state-of-the-art methods. In [132], the author employed different modalities to learn non-linear transformations that shared the subspace such that the representation maximizes the ratio of between and within modality covariance of observation. In [133], Ringevel et al. utilized different architectures of the LSTM network and compared two different fusion approaches. They showed that the prediction of valence requires longer analysis windows than arousal and that decision-level fusion leads to better performance than feature-level fusion.

It is well known that different signals carry complementary information for mental state monitoring [4, 134, 135]. The concatenation of multiple modalities, their associated features, and decisions in order to perform an analysis task is referred to multi-modal fusion [136, 137, 19, 138, 139, 140, 141]. Multi-modal fusion can provide added robustness against sensor failures, as other signal streams

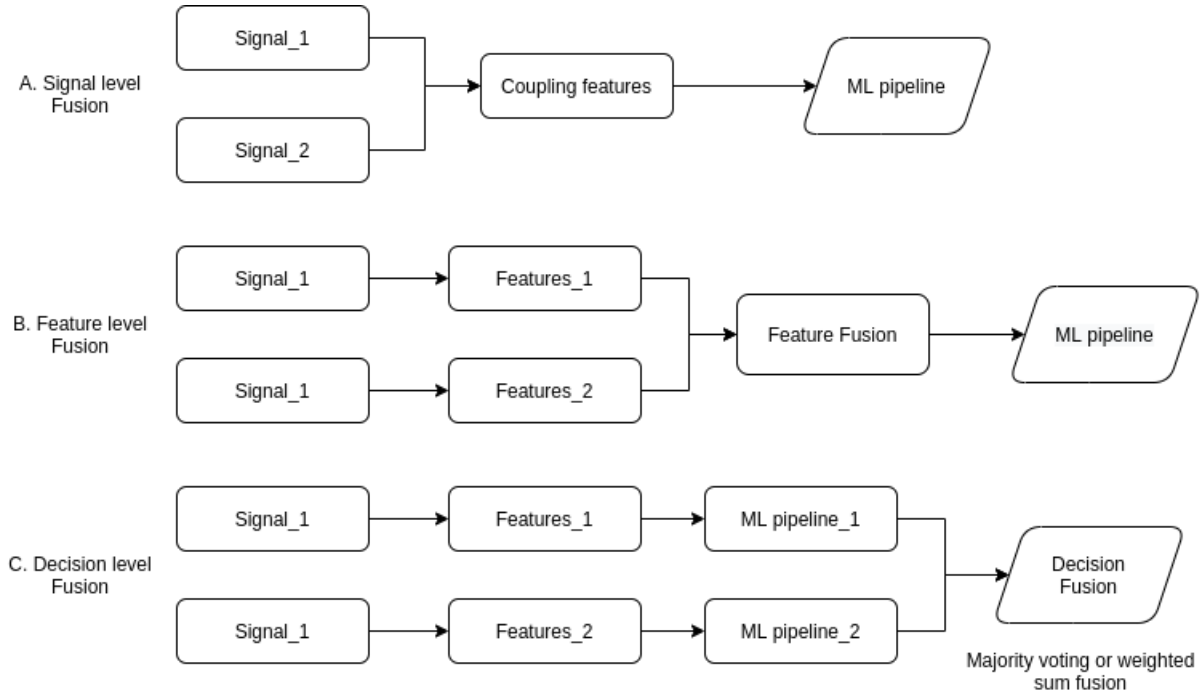

Figure 2.6: Different fusion strategies: (top) signal level, (middle) feature level, and (bottom) decision level.

may compensate for poor signal quality in one modality. Multi-modal information can be combined at different levels, as shown in Fig. 2.6.

Signal level (Fig. 2.6-a) fusion can be used when dealing with multiple signals coming from a very similar modality source. For example, different physiological signals often show coupling behavior as they may be governed by similar underlying mechanisms. One of the most well-known mechanisms is the coupling between respiration and heart rate. This type of fusion has as advantage the fact that no loss of information occurs, as the signal is directly processed. Feature level fusion (Fig. 2.6-b), in turn, relies on features from the different modalities to be separately extracted and later combined before being input to the machine learning pipeline. This is one of the most commonly used fusion approaches [4, 134], but may have some constraints imposed by temporal synchrony between modalities. Lastly, decision level fusion (Fig. 2.6-c) combines the output of the machine learning pipelines optimized for each of the modalities via methods such as majority voting or weighted methods. This method provides robustness against artifacts contaminating a specific modality [142] and may take signal quality into account [143]. With decision fusion, the issue of time synchronization is minimized and different machine learning algorithms may be used for different modalities.

2.3 Machine Learning Pipeline (Back-End)

As shown in Fig. 2.1, the SER system back-end is comprised typically of the machine learning pipeline along with modules needed to ensure robustness of the model to unseen conditions, including data augmentation and domain adaptation, and the model evaluation steps.

With machine learning, the ultimate goal is to develop models that can generalize well to unseen conditions. If care is not taken, model overfitting or underfitting can occur. Overfitting, for example, has the model learn unique nuances of the training set, and not necessarily general emotional cues from the data. As such, the model achieves high accuracy on the training set, but very poor results on the unseen test set. In contrast, underfitting occurs when the training error is very large, suggesting the model was too "simple" to gather any useful cues from the training data. In practice, to overcome these issues, a validation set is commonly used during training to ensure the model has some "unseen" information to gauge the over/underfitting of the model. Other tools, such as regularizers, are used to penalize overly complex models, thus helping with the overfitting problem. Data augmentation has shown to help with the training of deep neural networks by providing the model with training data that has a more diverse distribution, thus improving model performance in unseen test conditions. Lastly, domain adaptation has been used to map the distribution of the training data to match that of the test data. This way, the model learns emotional cues that are persistent or "normalized" across datasets. In this thesis, we explore the use of data augmentation and domain adaptation as tools to improve the robustness of SER systems to unseen test conditions, including noise and language. More details about data augmentation and domain adaptation are given below.

2.3.1 Data Augmentation

Data augmentation, as the name suggests augments the training set by introducing additional samples that have been corrupted by different distortions. With images, for example, this includes adding rotated versions of the training images, noisy versions, shifted pixel versions, to name a few. With speech, distortions can include the addition of additive noise at varying signal-to-noise ratios (SNR), convolution with room impulse responses to simulate room reverberation, and time reversal of the speech signal, to name a few examples. Data augmentation can also be used to add synthetic

data in certain classes, thus helping alleviate the data imbalance problem. Overall, it prevents data scarcity, improves the generalization capacity of the models, and minimizes the impact of rare events/outliers during prediction.

2.3.2 Domain Adaptation

Domain adaptation strategies aim to alleviate the effects of train-test mismatch in which the conditions available in training data differ from those of the test set. In the domain adaptation literature, "source" refers to the domain in which training takes place. In turn, the "target" domain corresponds to the domain in which the test data will come from. Chang et al [144], for example, employed a so-called DCGAN to extract and learn useful feature representations from unlabeled data from a different domain, which eventually lead to better generalization capability. Deng et al, in turn, employed an autoencoder to find a common feature representation across domains [145]. More recently, Mohammed et al [146] employed a domain adversarial neural network (DANN) for emotion recognition. DANN is based on finding a consistent feature representation for the source (train) and target (test) domains. Couple Generative Adversarial Network (GAN), in turn, has been used to train a couple of generative models that learn the joint data distribution across the two domains. Swami et al [147] proposed the generate-to-adapt architecture, a joint adversarial discriminate approach that transfers the information of the target distribution to the learned embedding using a generator discriminator pair. Recently, CycleGAN [148] has also been proposed where two generators and two discriminators are used: generator G converts input from the X to the Y domain, whereas generator F converts inputs from Y to X. Lastly, Choi et al proposed StarGAN [149], a method that can perform image-to-image translations for multiple domains using only a single model. In this thesis, we explore two classes of domain adaptation: correlation- or subspace-alignment based methods.

2.3.2.1 Subspace Alignment Domain Adaptation

Subspace alignment (SA) based DA aims to find a domain invariant feature space by learning a mapping function that aligns the source subspace with the target one [53]. SA linearly aligns the source domain to the target domain in a reduced-dimension PCA subspace. In this method, we first create subspaces for both source and target domains and then learn a linear mapping that

aligns the source subspace with the target subspace. This allows comparisons of the source domain data directly with the target domain data and to build classifiers on source data and apply them to the target domain. With SA-DA source data S, target data T, source labels LS, and subspace dimension D, source data S and target data T are first projected into the D dimension via the following equations:

$$X_S = PCA(S, D), \tag{2.5}$$

$$X_T = PCA(T, D), \tag{2.6}$$

where X_S and X_T represent the projected source and target data, respectively. Moreover,

$$X_A = X_S X_S' X_T, \tag{2.7}$$

where X_S' represents the transpose of the projected source data. Lastly:

$$S_A = SX_A, \tag{2.8}$$

$$T_T = TX_T. \tag{2.9}$$

In our experiments, the resultant feature embedding of source S_A and target domain T_T can serve as input to a deep learning-based SER system.

2.3.2.2 Correlation alignment Domain Adaptation

Correlation alignment (CORAL) based domain adaptation [150] matches the first and second-order statistics of the source and target data. It does this by first calculating the statistics of the target domain and then subtracts the covariance of the target domain from the source domain by whitening and recoloring the source domain. More specifically, the source domain feature matrix $\mathcal{D}_{source}$ is first whitened and given by $\mathcal{D}^w_{source}$, i.e.,:

$$\mathcal{D}^w_{source} = \mathcal{D}_{source} * C^{-\frac{1}{2}}_{source}, \tag{2.10}$$

where * represents matrix multiplication.

The matrix is recolored ($\mathcal{D}_{source}^{adapted}$) by:

$$\mathcal{D}_{source}^{adapted} = \mathcal{D}_{source}^{w} * C_{target}^{\frac{1}{2}}, \tag{2.11}$$

where C_{source} and C_{target} are given by:

$$C_{source} = \Sigma_{source} + I, \tag{2.12}$$

$$C_{target} = \Sigma_{target} + I. \tag{2.13}$$

Here, I corresponds to the identity matrix and Σ_{source} and Σ_{target} to the covariance matrices of the source and target domains, respectively.

2.3.3 Classification/Regression

Once data augmentation is done and domain adaptation is achieved, the choice of machine learning algorithm is made. Machine learning algorithms can take the form of classifiers if discrete classes are to be predicted (e.g., angry versus sad emotions) or regressors if continuous variables are to be predicted (e.g., the arousal level in the [0,1] range). Here we review the classifiers/regressor algorithms used throughout this thesis for different applications.

2.3.3.1 Support vector machines

A support vector machine (SVM) classifier is a supervised machine learning model [54] widely used in many pattern recognition applications, including SER tasks (e.g., [46]). SVM classifiers rely on different kernel functions to nonlinearly map the original features to a high-dimensional space where data can be well classified using a linear classifier. Typical kernels include linear, polynomial, and Gaussian radial basis functions, to name a few.

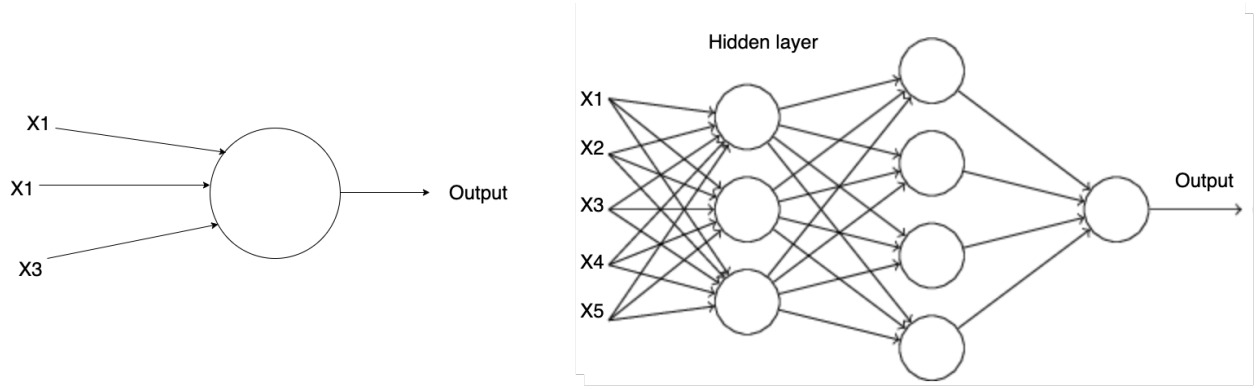

Figure 2.7: A perceptron (left) and FNN (right) taken from [2].

2.3.3.2 Feed-forward neural networks (FNN)

With feedforward neural networks (FNN) information moves in only one direction: forward, from the input nodes, through the hidden nodes (if any), and to the output nodes. Feed-forward neural networks are a category of statistical learning models [151]. In particular, a FNN can be represented by the following equation:

$$h1 = f * (W1 * X + b1), \tag{2.14}$$

where $h1$ is the layer output, $f*$ is a non-linear function, X is the input, and $W1$ and $b1$ are the weight and bias parameters of the layer, respectively. The non-linear mapping function $f*$ can take the form of a sigmoid or a hyperbolic tangent, for example. Its role is to perform sequential non-linear projections of the input on the previous layer, as shown in Figure 2.7. The multiple neuron layers help the model learn the input's abstract information.

The backpropagation algorithm [151] is used to learn the parameters of FNN in a supervised manner. With backpropagation, network parameters are adjusted iteratively to minimize a cost function between its input and desired output. The most commonly used optimization procedure for training FNNs (or DNNs in general) is the gradient descent (GD) or stochastic gradient descent (SGD) methods. More recently, variants of this architecture have been developed to account for temporal dynamics in sequential data. Some representative examples are discussed in the following sections.

2.3.3.3 Convolution neural networks

Convolutional neural networks (CNN) have become very popular [151] and have been employed in 2D image classification, speech and audio recognition, and video processing tasks. CNNs are comprised of two layer types: convolution and pooling. Convolutional layers are used to map from previous to current layer and serve as a weighted sum of the input features from previous convolutional layers and passed through a non-linearity such as ReLU [151]. The pooling layer, in turn, is used for dimensionality reduction by taking the maximum or average of a set of neighboring feature maps via subsampling by merging semantically similar features.

2.3.3.4 Recurrent neural networks

Recurrent neural networks (RNNs) are a family of neural networks designed to process sequential data. They can be seen as a feed-forward neural network with a parameter-sharing scheme, which makes the output corresponding to a specific input dependent on previously seen examples. This aspect makes recurrent neural networks an excellent alternative to model sequential data. However, it is known that under certain conditions vanilla recurrent neural networks present difficulties to learn long-term dependencies due to vanishing or exploding gradients [152]. Therefore, alternatives were proposed to deal with that issue, including the long short-term memory LSTM-RNNs [153],[154]. LSTMs have been widely used in speech applications, including automatic speech recognition [155], speaker verification [156], and SER [43]. The 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:

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

$$i_t = \sigma(\mathbf{b}_i + \mathbf{U}_i \mathbf{h}_{t-1} + \mathbf{W}_i \mathbf{x}_t), \tag{2.16}$$

$$o_t = \sigma(\mathbf{b}_o + \mathbf{U}_o \mathbf{h}_{t-1} + \mathbf{W}_o \mathbf{x}_t), \tag{2.17}$$

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 gates parameters. σ is the sigmoid function.

The hidden state $\mathbf{h}_t$ is given according to:

$$\mathbf{h}_t = o_t \odot \tanh(c_t), \tag{2.18}$$

where c_t is the LSTM cell state defined as:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t. \tag{2.19}$$

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

$$\tilde{c}_t = \sigma(\mathbf{b}_c + \mathbf{U}_c \mathbf{h}_{t-1} + \mathbf{W}_c \mathbf{x}_t). \tag{2.20}$$

The interested reader is referred to [151] for more details about CNNs and RNNs.

2.3.4 Model Evaluation

To ensure that the model generalizes well, model evaluation typically assumes the test set is completely unseen to the model. Larger datasets are typically divided into three disjoint sets: training, validation, and test, ensuring that no data leakage is achieved between partitions. Oftentimes with SER applications, the test set is comprised of conditions not present in the training/validation sets, such as different noise types or levels, different speakers, or different recording equipment. For smaller datasets, in turn, it may be difficult to partition the available data into three disjoint sets. In such cases, cross-validation techniques are commonly applied. With k-fold cross-validation, for example, the training set is split into k smaller sets. Then, for each of the "folds", the model is trained using $k-1$ of the folds and the k-th fold is held out as the test set. This process is repeated until all the k folds have been used for testing, leading to k performance values, which can then be averaged and reported. Another type of cross-validation evaluation is termed leave-one-subject-out, where in this case, data from all but one subjects is used for training and the held-out subject data is used for testing. This is repeated for all subjects available in the dataset.

2.3.5 Figures-of-merit

The performance measure used here is the typical metric used within continuous emotion prediction task, i.e., the *concordance correlation coefficient* (CCC). This figure-of-merit combines Pearson's correlation coefficient ρ with the square difference between the mean of the two compared time series:

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{2.21}$$

where μ and σ stand for the first and second order statistics of times series x and y, which correspond to the emotion predictions and their corresponding subjective ratings, respectively.

In turn, for discrete emotion classification tasks, we rely on the precision, recall, and F1 scores as figures-of-merit. Precision measures the number of correct positive predictions made and is given by:

$$Precision = \frac{TP}{TP + FP}, \tag{2.22}$$

where TP corresponds to true positives and FP to false positives. Recall measures the number of correct positive predictions made out of all positive predictions that could have been made. This is also called true positive rate or sensitivity and given by

$$Recall = \frac{TP}{TP + FN}, \tag{2.23}$$

where FN corresponds to false negatives. The F1 score, in turn, is computed as the harmonic mean between precision and recall and is useful in binary tasks where classes are unbalanced. It is given by:

$$F1 = \frac{2 * Precision + 2 * Recall}{Precision + Recall}. \tag{2.24}$$

Lastly, balanced accuracy is given as the arithmetic mean of recall and specificity (true negative rate) which, in turn, is given by:

$$Specificity = \frac{TN}{TN + FP}, \tag{2.25}$$

where TN corresponds to true negatives. As such balanced accuracy (BA) is given as:

$$BA = \frac{Recall + Specificity}{2}. \tag{2.26}$$

The interested reader is referred to [157] for more details on these classical performance metrics.

2.4 Emotion Datasets

In this thesis, several publicly-available datasets have been used. A description of the datasets used is given below.

1. **RECOLA**: The first database corresponds to the REmote COLlaborative and Affective interactions (RECOLA) database [55]. This database was used during the 2016 audio-visual emotion challenge (AVEC) [56] and is in the French language. Based on spontaneous and naturalistic interactions collected from a collaborative task, six annotators measured emotion continuously using a time-continuous scale for two emotion primitives, namely arousal, and valence. Even though all subjects were fluent French speakers, they came from different nationalities (French, Italian, and German), thus the database provides some diversity in the expression of emotion. Also, the total number of speakers in the RECOLA dataset were 27, out of which 16 were females, and 11 were males. The detailed participant statistics are available in [55]. The subjective labels were originally available with a frame rate of 40 ms. We aggregated five consecutive frames to generate a frame rate of 200ms for analysis via averaging. The RECOLA database is partitioned into three disjoint sets: training, development, and test, each containing 5-minute duration speech files from nine speakers.
2. **SEWA**: The second and third datasets correspond to the German and Hungarian language subsets of the Sentiment Analysis in the wild (SEWA) database. This database was used in the AVEC 2017 [57] and AVEC 2019 [44] challenges. Subjects (in pairs of friends and relatives) were recorded through a dedicated video chat platform, using their own standard web-cameras and microphones while they discussed an advert they had watched. The detailed participant demographics for both datasets are available in [58]. Both datasets are divided into three parts: 34 files for training, 14 in the development set, and 16 for testing. The duration of the recordings in the dataset range from 40 seconds to 3 minutes for each file. In our experiments, we only used the training and development parts, as the labels were not available for the test set. The SEWA dataset has valence and arousal labels available with a frame rate of 100 ms, thus, we aggregated two consecutive frames to remain consistent with the frame durations used with the RECOLA dataset via averaging. Both the RECOLA and

SEWA datasets were recorded with a sampling rate of 44.1 kHz. Further details about the dataset can be found in [58, 158].

It is important to highlight that for our experiments, we relied on only the training and development subsets, as emotion labels are not available for the test set. From this data, six listeners rated the liking, valence, and arousal levels of each recording with a resolution of 100 ms using a joystick and a continuous scale. All annotators were native speakers of the language of the files they were listening to. Each dimension was annotated separately based on the video chat recordings that were shown in random order to each annotator. A final "ground truth" value was found based on the annotations of each of the six listeners. A detailed description of this process can be found in [44].

3. **SEWA-Chinese**: The fourth dataset corresponds to the Chinese language subset of the SEWA project. The audio recordings sample rate was 44.1 kHz, and the total data duration is 3:17:52 hours. There were audio samples from 36 male and 34 female participants. A total of 70 audio files without labels were made available through the AVEC 2019 [44] challenge. Detailed participant demographics for this dataset are available in [58]. In our experiment, we specifically used this unlabeled dataset in Chapter 4 for the proposed N-CORAL domain generalization method described in Section 4.4.3.1. We downsample all audio files to 16 kHz for further processing.
4. **Emoti-W**: The Emoti-W database was made available for the 2017 Emotion Recognition in the Wild Challenge. Emoti-W was in the English-language. In this dataset, emotion labels are available for seven emotion categories: anger, disgust, fear, happiness, neutral, sadness, and surprise. As an "in-the-wild" dataset, most of the recordings present some level of background noise. The labels for the EMoti-W challenge dataset were created from the closed captions available in movies and TV series. In particular, affect related words were searched in the closed captions and used to generate weak emotion labels. For the test set, data from sitcom TV series were used to add to the variety in the environments in which the data was recorded. When partitioning the data between training, validation and test sets, subjects, movies, and TV sources were separated to avoid any data leakage. The data is available in audio and video formats and the audio sampling frequency is 48 kHz; videos are available in MPEG-2 format with 25 frames per second. The training set is comprised of 773 samples, the validation set of 383 examples, and the test set of 653 samples. Complete

details about the Emoti-W dataset can be found in [67]. Again, we use only the labeled training and development subsets in our experiments.

5. **MELD**: The dataset used for experimentation in chapter 5 is the Multimodal EmotionLines Dataset (MELD) [59]. It is a multimodal emotion classification dataset which has been created by extending the EmotionLines dataset [60]. MELD contains approximately 13,000 utterances from 1,433 dialogues from the TV series 'Friends'. Each statement is annotated with emotion and sentiment labels and encompasses audio, visual, and textual modalities. The MELD dataset contains conversations, where each dialogue has utterances from multiple speakers. EmotionLines was created by crawling the discussions from each episode and then grouping them based on the number of statements in conversation into four groups of utterances. Finally, 250 dialogues were sampled randomly from each group, resulting in the final dataset of 1,000 dialogues. The utterances in each dialogue were annotated with the most appropriate emotion category. For this purpose, the six universal emotions (joy, sadness, fear, anger, surprise, and disgust) were considered. This annotation list was extended with two additional emotion labels: neutral and non-neutral.

   Each utterance was annotated by five workers from the Amazon Mechanical Turk platform. A majority voting scheme was applied to select a final emotion label for each utterance. While the MELD dataset has labels for several emotions, here we focus on two specific binary tasks to gauge effects across the valence and arousal dimensions. More specifically, we first focus on two tasks: (1) anger versus sad classification to explore the benefits of the proposed tool for low/high arousal classification and (2) joy versus sad classification for positive-valence-high-arousal and negative-valence-low-arousal characterization. As such, the MELD dataset was split into three disjoint sets: training, test, and development. These were split as follows:

   (a) Training: angry (1109 samples), joy (1743 samples), and sad (682 samples)
   (b) Validation: angry (153 samples), joy (163 samples), and sad (111 samples)
   (c) Testing: angry (345 samples), joy (402 samples), and sad (208 samples)

6. **IEMOCAP**: The IEMOCAP dataset was used to show the generalizability of the proposed model in Chapter 5. The IEMOCAP dataset has 12 hours of audio-visual data from 10 actors where the recordings follow the dialogue between a male and a female actor in both scripted or improvised topics. After the audio-video data was collected, it was divided into small utterances of length between 3 to 15 seconds, which were then labeled by evaluators. Each utterance was evaluated by 3-4 assessors. The evaluation form contained ten options (neutral,

happiness, sadness, anger, surprise, fear, disgust, frustration, excitement, and others). We consider only three in our experiments: anger, sadness, and happy, so as to remain consistent with the previous MELD data experiments and to be able to directly test the models trained on the MELD dataset. To this end, the dataset was split into three disjoint sets: training (70%), development (10%), and test (20%). These were split as follows:

(a) Training: angry (772 samples), happy (416 samples), and sad (758 samples);

(b) Validation: angry (111 samples), happy (60 samples), and sad (110 samples);

(c) Testing: angry (220 samples), happy (119 samples), and sad (216 samples).

7. **AURORA and DEMAND**: Lastly, as we are interested in gauging the environment robustness of the proposed method, we have also relied on two recorded noise datasets in order to further corrupt the emotion datasets. For this purpose, the AURORA [61] and DEMAND datasets [62] of recorded noise sources were used. In particular, two noise types were used: multi-talker babble and noise recorded inside a commercial airplane. Noise was added at five different signal-to-noise ratios (SNRs): 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB. Next, in order to study the effect of room reverberation on SER, three recorded room impulse responses taken from [63] were used and convolved with the speech files. The impulse responses corresponded to rooms with reverberation times of $T60 = 0.25$, 0.48 and 0.8 seconds, thus representing a small, medium and a large-size room, respectively.

2.5 Conclusion

This chapter presented the background of emotion recognition using speech signals. First, the emotion recognition system components were discussed. Their acquisition methods, properties, pre-processing and benchmark features with their relation to affective states were presented. This was followed by a short description of different types of multimodal systems and their advantages and disadvantages. Next, the different components of the machine learning pipeline were discussed, ranging from domain adaptation, data augmentation, and evaluation to performance metrics used. Following this, the available datasets were discussed. Overall, this chapter discusses the background of affective state recognition literature while expanding upon the various tools required for moving research from “in-the-lab” to “in-the-wild”.

Chapter 3

Quality-Aware Bag of Modulation Spectrum Features for Robust Speech Emotion Recognition

3.1 Preamble

This chapter is compiled from material extracted from the manuscript published in the *IEEE Transactions on Affective Computing* [64].

3.2 Introduction

Affective computing has emerged as a prominent field in human-machine interaction (HMI), providing machines with information about the user's affective state, thus making the interaction more natural and human-like [3]. Representative examples of affective HMIs already in the market include driver fatigue level measurement systems in cars; customer anger and stress level measurement systems based on voice inputs to call centers; and adjustment of teaching strategies and material presentation based on student attention levels. As advances in machine learning, reinforcement learning, and deep learning are pushing us towards more human-like artificial intelligence (AI) systems, the need for robust affective interfaces has become critical.

With affective computing, emotional states are commonly classified as discrete or continuous [97]. Discrete emotional states, for example, can be categorized as happy, fearful, surprised, and sad, to name a few. Continuous categorization, in turn, assumes that affective states are systematically related to one another [97] and can be divided into three primitives, namely: valence, arousal, and dominance. The arousal primitive describes the level of activation (passive or active) and relates to the intensity (either positive or negative) of the current affective condition. Valence, in turn, represents the pleasantness level, while dominance, the degree of control exerted by the affective condition. While the majority of the affect variability can be covered by the valence and arousal primitives, their measurement in realistic settings is challenging and discrete emotional modeling has become more popular. More details on the advantages and disadvantages of the two methods can be found in [97].

Recently, with the advances seen with deep machine learning, voice-based control has gained popularity, with voice assistant applications emerging in smart speakers, smartphones, cars, and even wearables. This has opened doors for new speech-based affective computing systems, also known as speech emotion recognizers (SER). Existing SER systems, however, are still very sensitive to environmental factors, such as background noise, room reverberation, and competing speakers [12]. Typically, to provide environmental robustness, systems are trained using a combination of unprocessed and processed/noisy data (a process termed multi-condition training and, more recently, as data augmentation for deep learning approaches). While this provides some robustness, especially if the noise type and levels present in the test data are similar to those used during training, performance is still highly affected with unseen conditions. As such, environment-aware or environment-agnostic systems are still drastically needed.

In this chapter, we explore the usefulness of a new feature set for environment-robust SER, namely bag of modulation spectral features. Modulation spectral features have been shown to provide some robustness against environmental factors (e.g., [45, 47]). Bag of audio words, in turn, have been shown useful in characterizing emotional states from speech spectral features by being at the border of characterizing linguistic and acoustic information [65]. Here, we propose to combine both and show that not only improved emotion recognition accuracy is achieved but also improved robustness to environmental factors. Moreover, we show that the emotional information extracted by the proposed features is complementary to those obtained from other conventional measures. As such, further improvements are achieved with feature fusion. Lastly, the modulation spectrum has

shown to be useful for blind speech quality estimation [66]. As such, we propose a quality-aware SER system and show its advantage for emotion recognition in realistic settings.

3.3 Related works

3.3.1 Feature representations

While SER systems have been widely explored over the last decade, it is still not clear which features are optimal, especially with recent advances in end-to-end deep learning models. For example, [159] showed the importance of prosodic features to encode speaker affective states. A number of studies have also explored the spectral changes associated with different emotions. Williams and Stevens [160], for example, showed that arousal levels correlate with overall signal energy, the energy distribution across the frequency spectrum (also corroborated by [159]), and the duration of pauses. The work in [73] showed that the happy emotion is linked to high energy levels at the high-frequency range, while the sad emotion to lower energy levels at the same frequency range.

Over the last decade, the Interspeech Computational Paralinguistics Evaluation (ComParE) Challenge has run annually and a number of new features and feature sets have been found based on the openSMILE toolkit [161]. These include the so-called GeMAPS [74], eGeMAPS [74], and ComParE [126] sets, as well as the IS11 [162] feature subsets, to name a few. Generally, these sets combine a number of so-called low-level descriptors and include voicing, pitch, mel-frequency cepstral coefficients (MFCCs), spectral, and energy features, along with their delta and double-delta derivatives. Moreover, in order to obtain utterance-level descriptors, several statistical functionals are used (e.g., mean, min, max). In the most recent 2021 ComParE Challenge, the ComParE feature set, combined with the bag-of-audio-word representation for these features and their delta representations, were proposed as benchmark measures [163].

In addition to prosodic and spectral feature representations, modulation spectral features have also shown to be useful for emotion recognition and to outperform MFCCs [46]. Recently, temporal pooling of modulation spectral features, both via averaging or via deep neural networks, showed improved robustness to “in-the-wild” conditions [45] for both valence and arousal prediction. Here,

we explore if the bag-of-audio-words principle, applied to modulation spectral features, can further assist with SER performance in noisy settings.

3.3.2 Classification and data augmentation

Beyond the feature sets used, classifiers also play an important role in SER accuracy. Recently, various deep neural network (DNN) architectures have been explored, including recurrent neural networks (RNN), convolution recurrent neural networks (CRNN), convolution neural networks (CNN), convolution neural network-long short term memory (CNN-LSTM), long short term memory (LSTM), and deep belief networks (DBN). Other architectures borrowed directly from the computer vision field, such as ImageNet, Inception, and Resnet, and CNNs in general, have also been explored for feature embedding (e.g., [164, 165, 166]. In [167], for example, an RNN was proposed and shown to better consider long-range contextual effects and the uncertainty around the emotional labels. Mel-spectral features, combined with CNNs, in turn, have been extensively explored for SER, with the work described in [168, 169] showing the usefulness of the self-attention mechanism for extracting emotionally-informative time segments. Recently, handcrafted features vs. deep learning-based features were compared for affect recognition tasks, and no clear winner was found [170], thus suggesting that improved features are still needed.

One limitation of the use of deep learning for SER tasks lies on the limited availability of large datasets. Unlike speech recognition, where thousands of hours of speech data can be collected, emotional speech is harder to collect and label, hence available SER datasets are relatively small. This limits the complexity of the models used, increases the chances of model overfitting, and lowers the generalization capacity of the models to unseen data. To this end, data augmentation has been explored as a viable alternative. Data augmentation increases the amount of training data available by slightly modifying the existing data (e.g., by adding noise, time-reversing the signals) or by creating synthetic data (e.g., from text-to-speech systems). This has shown to be useful for speech recognition in general [155], for keyword spotting [155], and was recently shown to also assist with SER tasks by augmenting existing data via vocal tract length perturbation of the available audio files [171]. Typically, acoustic room simulators [172], manipulations of the spectrogram using image transformations, and/or different filterbank representations [173] are used for data augmentation.

Here, we explore the use of data augmentation as a tool to provide robustness against language mismatch.

3.3.3 SER Challenges

In an effort to advance the SER field, the last decade has seen an emergence of different SER Challenges, including the Audio-Video Emotion Challenge (AVEC) series and the Emotion in the Wild (EmotiW) (e.g., [57, 43, 44, 67]), as well as the aforementioned Interspeech ComParE Challenge series. The winning systems in these challenges can provide insights regarding the top features and classifiers, as well as the gap still present with "in-the-wild" SER.

For example, the winning system in the AVEC 2016 Challenge [174] proposed the use of high-level acoustic, visual, and physiological features from low-level descriptors using a sparse coding method. The speech features used corresponded to prosodic, voice quality, and MFCC features. The winner of the AVEC 2017 Challenge [175], in turn, relied on a multi-task learning approach to predict multiple emotional dimensions using a shared representation. From the speech modality, OpenSMILE based IS10 and Soundnet [176] features were used. The winner of the AVEC 2018 [177] explored efficient deep learning features extracted from different modalities and relied on a LSTM network to capture their long-term temporal information. Several multimodal interaction strategies were explored. For the speech modality, a CNN-based audio embedding feature set (based on VGG [178]) was used. More recently, the winner of the AVEC 2019 proposed an unsupervised adversarial domain adaptation approach to bridge the gap across different cultures for emotion recognition [179]. Again, CNN-based audio embedding features were used.

Other systems that achieved high accuracy during these Challenges relied on convolutional LSTMs [180] and different feature representations, such as bag-of-audio features and a sparse histogram vector for encoding a previously-learned codebook of templates. For example, Jiyoung et al. [168] proposed a convolutional LSTM with spatio-temporal attention, whereas Mohammed et al. [181] presented a bag-of-words representation with an autoencoder for both dictionary creation and matrix assignment. The Emoti-W challenge [67], in turn, provides researchers with a platform to evaluate their methods on 'in the wild' data. For example, the latest winner [182] relied on the 1582 dimensional IS10 audio features. In all cases, benchmark systems are provided to allow Challenge participants to gauge the advantages of their proposed systems. These benchmark systems typi-

cally rely on openSMILE based audio features, bag-of-word representations as front-end, and either support vector machines or LSTM classifiers as back-end. These benchmark systems are detailed in [57, 43, 67].

3.4 Proposed Method

The proposed method is based on the modulation spectral signal representation that has shown useful for SER [45, 46]. Here, we explore the benefits of implementing a bag-of-word approach on top of modulation spectral features and explore their benefits for "in the wild" SER. Since the modulation spectrum has also been used for blind speech quality measurement (blind in the sense that a clean reference signal is not needed), we explore a quality-aware system for improved accuracy. We test the complementarity of proposed features to other typically used features via feature fusion.

3.4.1 Bag-of-words methodology

BoAW is a methodology that allows for aggregation of short-term features (e.g, MFCC computed every 40 ms) into longer duration properties. Commonly, statistical functionals (e.g., mean, standard deviation, min, max) have been used to aggregate short-term features into utterance-level details. As mentioned previously, the BoAW approach replaces the functionals and represents the audio features in the form of compact 'audio words' with better resolution than utterance-level aggregation. Classifiers are then trained on histogram features, which represent the frequency of occurrence of the respective *audio words*. More detail can be found in 2.2.4.

3.5 Experimental Setup

Here, we describe the experimental setup, including databases, classifiers, benchmarks, and figures-of-merit used.

3.5.1 Databases

Several multi-lingual emotion datasets were used. First, the German and Hungarian language subsets of the Sentiment Analysis (SEWA) database were used for the continuous emotion prediction experiments. The second emotion dataset used was the English-language Emoti-W database made available for the 2017 Emotion Recognition in the Wild Challenge [67]. Again, we use only the labeled training and development subsets in our experiments. More detail is available in 2.4

As we are interested in gauging the environment robustness of the proposed method, we have also relied on two recorded noise datasets in order to further corrupt the SEWA and Emoti-W datasets. For this purpose, the AURORA [61] and DEMAND datasets [62] of recorded noise sources were used. The more details is mentioned in 2.4

3.5.2 Classification and Regression Models

Here, we explore two different SER tasks: continuous emotion estimation and discrete emotion classification. As we are interested in testing the effectiveness of the proposed features, we rely on two classical and widely used machine learning algorithms in the SER field. In particular, for continuous emotion recognition we utilize an LSTM deep neural network and for classification, a support vector classifier. While a complete description of these methods is beyond the scope of this chapter, a brief description of the models is given for the sake of completeness. The interested reader is referred to the references below for more details.

3.5.2.1 LSTM-RNN Regression Model

The architecture used here was motivated by the AVEC 2019 Challenge benchmark system described in [44] and is comprised of a two-layer single-directional LSTM with hidden layers of size 64 and 32. We also experimented with a bi-directional LSTM (Bi-LSTM), but in pilot experiments the added complexity did not bring any benefits over an LSTM. As such, all of our experiments rely on an LSTM classifier. Linear activation was used in the output layer and a concordance correlation coefficient (CCC)-based loss function (see Sect. 4.5.4) was used for training. The model was trained for 1000 epochs using the whole sequence. We used our validation set to select the

final hyper-parameters, including the optimizer (RmsProp, Adam, SGD), learning rate (0.01, 0.001, 0.0001) and dropout (0.1-0.5 in 0.1 increments). The final learning rate of 0.001 with RMS prop was found to be useful in the experimentation with a dropout rate of 0.3. Experimentation codes are available on github[1].

3.5.2.2 Support Vector Classification

SVM classifiers rely on different kernel functions to nonlinearly map the original features to a high-dimensional space where data can be well classified using a linear classifier. For the experiments herein, the libsvm function of scikit-learn was used with a radial basis function (RBF) kernel. Default parameter choices were kept for simplicity, hence the results reported herein represent lower bounds on what could be achieved with the SVMs. The parameters used were: C (regularization) = 1.0, and gamma = 'scale'. The interested reader is referred to [54] for more details about the used parameters.

3.5.3 Benchmark systems

Several benchmark features are extracted using the openSMILE toolkit. Benchmark features serve two purposes: (1) gauge the benefits of the proposed method and (2) test the complementarity of the proposed features to existing ones. The first feature set used as a benchmark is the Interspeech speaker state challenge 2011 feature set (IS11), which consists of 118 low-level descriptors. These include 50-spectral, five-voice, and four energy-related features, along with an additional 59 features corresponding to their delta coefficients [162]. These features are extracted over 20ms frames with 10ms hops, as proposed by the AVEC Challenge. The eGeMAPS feature set is also explored as a secondary benchmark set. These are the so-called extended Geneva Minimalistic Acoustic Parameter Set and comprise 88 acoustic parameters relating to pitch, loudness, unvoiced segments, temporal dynamics, and cepstral features and have been used in [57, 43, 44]. Third, the widely-used mel-frequency cepstral coefficients (MFCC) are used. In particular, the 39-dimensional feature vectors comprised of 13 MFCCs, 13 delta MFCCs, and 13 double-delta MFFCs. These features were extracted over 20 ms windows, hop size of 10ms, and 64 filters, and their use is motivated by

[1] `https://github.com/shrutikshirsagar/Bag-of-word-SER`

AVEC 2018 and AVEC 2019 results. Lastly, the fourth benchmark set used is comprised of prosodic features, including fundamental frequency (F0), intensity measures, and voicing probabilities.

Benchmark systems, in turn, comprise combinations of these features and either an LSTM or an SVM classifier. For example, the AVEC 2017 Challenge benchmark system was comprised of GeMAPS features and a support vector machine backend [57]. The AVEC 2018 and AVEC 2019 Challenges, on the other hand, relied on combined MFCC and GeMAPS features and an LSTM backend [43, 44]. For comparison purposes, here we also compute the BoAW representation for these benchmark measures and use an LSTM classifier as an additional benchmark system. In addition to these different Challenge benchmarks, we also explore several other state-of-the-art (SOTA) systems, including modulation spectral features pooling and SVM regression [45], GeMAPS features with a relevance vector machine [183] and an autoregressive exogenous model backend [184], MFCC and GeMAPS features with a BiLSTM classifier [185], a BiLSTM model [186] with SoundNet [175] and bottleck features [187], and lastly, a transformer neural network trained on 34 features, corresponding to 13 MFCCs, 13 chromagram-based features, zero-crossing rates, energy, the entropy of energy, spectral centroid, and speed, spectral entropy, spectral flux, and spectral roll-off [188]. The model relies on bidirectional gated recurrent units with self attention.

3.5.4 Figures-of-Merit and Testing Setup

For the continuous emotion prediction task on the SEWA database, the widely used concordance correlation coefficient (CCC) measure is used as a figure-of-merit. More detail is given in 2.3.5. In turn, for the discrete emotion classification task on the Emoti-W database, we rely on the precision, recall, and F1 scores as figures-of-merit. Lastly, as mentioned previously, our experiments rely only on the labeled training and validation partitions of the various Challenge datasets. For our experiments, the training datasets are also used for training, but 20% of this set has been set aside for what we call 'our validation set' in order to perform hyper-parameter tuning of the LSTMs. As such, the Challenge validation set is used as 'our test set'. With this partitioning, our training, validation, and test sets become disjoint to assure no data leakage. To show the significance of the obtained gains with the proposed method, we use a significance z-score test between CCCs with a 95% level ($p < 0.05$); comparisons are made against the AVEC 2018 eGeMAPS-BoAW-plus-LSTM benchmark.

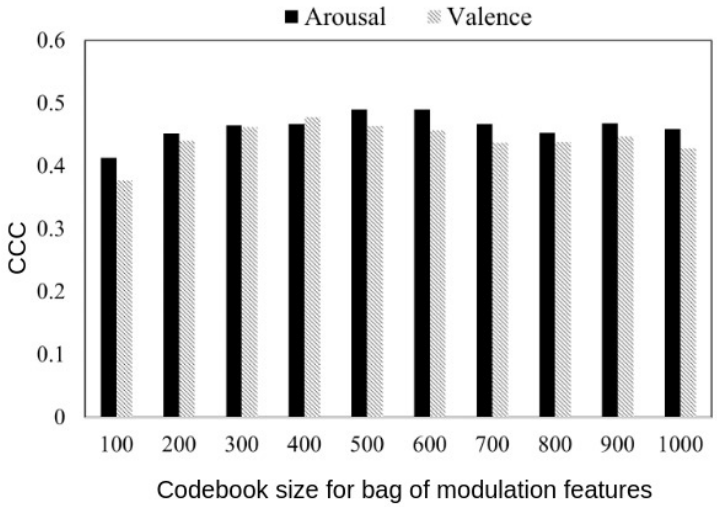

Figure 3.1: Experimental results with different codebook sizes for the proposed modulation features using a LSTM model for arousal and valence prediction

3.6 Experimental Results and Discussion

In this section, we present the experimental results for both unprocessed and noisy speech signals and discuss our findings.

3.6.1 Optimal codebook sizes

In this first experiment, we investigate the effect of different codebook sizes for the proposed and benchmark features. As a first experiment, we use unprocessed speech files for training and testing. For this experiment, we employed the SEWA-German dataset. Figures 3.1 and 3.2 depict the obtained CCC for arousal and valence prediction with an LSTM for the proposed and benchmark measures, respectively. As can be seen, for the proposed feature, an optimal codebook size of 500 is seen for arousal, whereas 400 is better for valence. A codebook size of 500 showed the best combined valence-arousal prediction capability, hence it will be used throughout the remainder of the experiments. Lastly, for the benchmark features, it can be observed that optimal codebook sizes of 500 are observed for MFCCs, 200 for the IS11 feature set, and 100 for the prosodic and the eGeMAPS feature sets, respectively. A codebook size of 500 for prosodic features was shown useful for valence. As such, these codebook sizes are used throughout the remainder of the experiments.

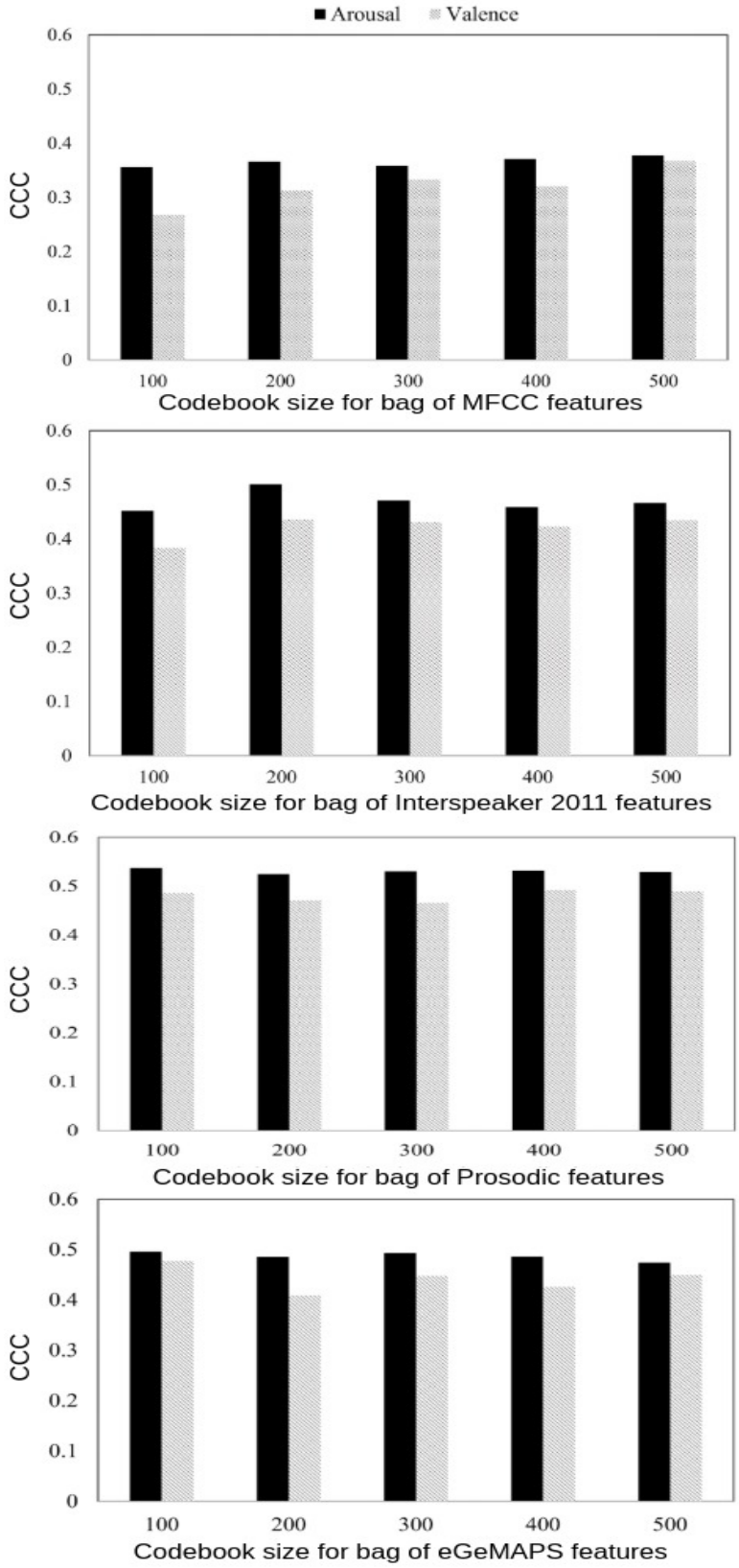

Figure 3.2: Experimental results with different codebook sizes for different benchmark measures using a LSTM model for arousal and valence prediction. From top to bottom: MFCC, IS11, prosodic, and eGeMAPS feature sets.

3.6.2 Unprocessed speech: Continuous emotion prediction

Table 3.1 shows the CCC values obtained on our German SEWA test set with the proposed and benchmark features with and without the BoAW aggregation. Several feature fusion strategies are also included to test the complementarity of the feature sets. These included FF1: MSF and eGeMAPS; FF2: MSF and IS11; FF3: MSF and MFCC; FF4: MFCC and Prosodic; FF5: MSF and Prosodic; FF6: MSF, MFCC and Prosodic. Quality-aware versions are indicated by a '+Q' label in the Table. We explored different machine learning algorithms, including support vector machines, LSTMs and bidirectional LSTMs, but we only include here (for the sake of brevity) the results with the LSTM as it showed improved results across the majority of tested conditions. For fair comparisons, all results have relied on a LSTM regression model. In the Table, values between parentheses indicate the number of audio words used in the BoAW approach for the fused feature sets.

As can be seen, BoAW aggregation improves accuracy for most of the features tested. For the single feature scenario, prosodic features with BoAW showed the highest CCC for both valence and arousal. As expected, quality-awareness did not show any improvement with unprocessed speech data when BoAW was applied, but did improve performance when it was not applied, hence suggesting that BoAW aggregation itself already provides some environment robustness. Feature fusion, in turn, showed to improve CCC significantly and suggests that the proposed features provide complementary SER information to existing benchmark measures. In particular, for unprocessed speech, the highest CCC was achieved when the proposed features were combined with prosodic features. Again, quality-awareness showed little improvement when BoAW was applied, but consistently showed to improve performance when BoAW was not used, thus further corroborating the robustness of the BoAW approach.

Overall, the best system for arousal was comprised of a system with the fusion of the proposed and prosodic features, achieving a $CCC = 0.556$, thus significantly outperforming the benchmark. For valence prediction, the fusion of proposed and prosodic features also showed one of the highest CCC, with a very small improvement seen once MFCCs were further included and quality awareness was used. Overall, for unprocessed speech, fusion of the proposed features and prosodic ones achieved the highest accuracy. Table 3.2 further presents the results obtained with several SOTA systems. As can be seen, five of them are able to significantly outperform the benchmark. Notwith-

Table 3.1: Performance comparison of different feature representations in terms of CCC with and without BoAW computation for unprocessed speech conditions. LSTM regression models were used and highest values are indicated in bold. Significantly better results relative to the benchmark are highlighted by an asterisk.

	With BoAW		Without BoAW	
Features	Arousal	Valence	Arousal	Valence
eGeMAPS	0.496	0.477	0.375	0.398
IS11	0.501	0.436	0.152	0.186
MFCC	0.378	0.368	0.403	0.361
MSF	0.490	0.464	0.315	0.319
MSF + Q	0.490	0.464	0.385	0.330
Prosodic(100)	0.537*	0.486	0.364	0.327
FF1	0.481	0.453	0.404	0.376
FF1 + Q	0.477	0.455	**0.418**	**0.403**
FF2	0.508*	0.479	0.386	0.308
FF2 + Q	0.497	0.441	0.392	0.315
FF3	0.434	0.427	0.354	0.308
FF3 + Q	0.434	0.427	0.362	0.315
FF4	0.521*	0.491*	0.300	0.282
FF5(100)	0.528*	0.512*	0.308	0.276
FF5(500)	**0.556***	0.504*	0.308	0.276
FF5(500) + Q	0.540*	0.514*	0.293	0.290
FF6(500)	0.529*	0.502*	0.275	0.273
FF6(500) + Q	0.532*	**0.518***	0.311	0.305

standing, the proposed method still outperforms all SOTA methods by as much as 3% (e.g., relative to [184] for arousal). As can be seen, the tested SOTA methods achieve accuracy levels inline with those obtained by the tested benchmarks (per Table 3.1) despite requiring substantially higher computational resources and training times (e.g., the transformer system in [188] has almost twice as many weights to relative to the LSTM-based benchmark). Given these results, in subsequent experiments we will rely on the benchmark features coupled with an LSTM backend to allow for more direct and fair comparisons with the proposed method.

3.6.3 Processed speech: Continuous emotion prediction in mismatched conditions

In this section, we investigate the robustness of the proposed features against noisy inputs. As such, unprocessed speech is used to train the various SER models and processed (noisy) speech is used for testing. Such mismatch conditions are known to significantly degrade SER performance.

Table 3.2: Performance comparison of different state-of-the-art methods and AVEC challenge baseline in terms of CCC. Significantly better results relative to the benchmark are highlighted by an asterisk.

SOTA	Arousal	Valence
[57]	0.344	0.351
[43] (MFCC)	0.282	0.306
[43] (eGeMAPS)	0.421	0.398
[45]	0.369	0.308
[183]	0.494	0.507
[186]	0.356	0.396
[175]	0.527*	0.504
[187]	0.533*	0.466
[184]	0.540*	0.502
[185] (MFCC)	0.501	0.452
[185] (eGeMAPS)	0.479	0.375
[188]	0.406	0.368

3.6.3.1 Additive Noise

Tables 3.3 and 3.4 show the performance achieved when models were trained on unprocessed speech and tested on noisy speech corrupted by airport and babble noises, respectively, at various different SNR levels (i.e., 0 dB, 5 dB, 10 dB, 15 dB and 20 dB). The optimal codebook sizes detailed in Section 3.6.1 are used. As prosodic features achieved reliable results with codebook sizes of both 100 and 500 (in the FF5 feature fusion setting) in the clean matched case, both values are also tested in the mismatched condition. The codebook sizes tested are reported within parentheses in the table and all results are based on the German SEWA dataset.

As can be seen, with increasing noise levels, performance degrades for all feature sets, as expected. The drop in performance is particularly notable for the prosodic BoAW features, which can be difficult to be measured under such harsh scenarios [80]. Similarly, the IS11 set is comprised of several voice quality and pitch-related features, thus are also highly affected by noise. MFCC BoAW features, on the other hand, showed some robustness to noise, thus corroborating previous findings [74]. When used alone, the eGeMAPS feature set resulted in the best arousal and valence prediction CCC for the 0 dB noise setting, hence corroborating previous reports of their robustness to noise [80].

The proposed bag of modulation spectral features, in turn, showed to reliably estimate valence for very low SNR values and achieved competitive results for both valence and arousal above 10 dB SNR for both babble and airport noise. Babble noise, which has speech-like modulation spectral

Table 3.3: Performance comparison of different feature representations in terms of CCC with and without BoAW computation for mismatch train-test conditions with airport noise at different SNR levels. LSTM regression models were used for all methods and highest values are indicated in bold. Significantly better results relative to the benchmark are highlighted by an asterisk.

	Airport (0dB)		Airport (5dB)		Airport (10dB)		Airport (15dB)		Airport (20dB)	
Features	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence
eGeMAPS	0.132	0.169	0.136	0.140	0.235	0.147	0.237	0.142	0.267	0.206
eGeMAPS-BoAW	**0.299**	**0.275**	0.305	0.210	**0.448**	**0.415**	0.459	0.429	0.462	0.452
IS11	0.149	0.186	0.266	0.173	0.271	0.227	0.272	0.234	0.292	0.221
IS11-BoAW	0.175	0.203	**0.354***	**0.348***	0.403	0.412	0.436	0.462*	0.453	0.473*
MFCC	0.244	0.227	**0.354***	0.273*	0.357	0.345	0.366	0.380	0.389	0.375
MFCC-BoAW	0.263	0.214	0.337	0.291*	0.351	0.343	0.350	0.348	0.367	0.360
MSF	0.158	0.126	0.163	0.089	0.167	0.156	0.197	0.058	0.212	0.062
MSF-BoAW	0.178	0.260	0.293	0.325*	0.429	0.385	0.457	0.431	0.465	0.445
Prosodic(100)	0.096	0.047	0.190	0.181	0.260	0.183	0.236	0.145	0.256	0.178
Prosodic(100)-BoAW	0.066	0.055	0.283	0.225*	0.400	0.362	**0.519***	**0.469***	**0.526***	**0.490***
FF1	0.193	0.194	0.172	0.195	0.176	0.210	0.303	0.258	0.329	0.281
FF1-BoAW	**0.293**	**0.303**	0.281	0.387*	0.415	0.378	0.417	0.357	0.488*	0.354
FF2	0.215	0.165	0.209	0.173	0.230	0.164	0.287	0.193	0.291	0.197
FF2-BoAW	0.217	0.264	0.347*	0.292*	0.431	0.405	0.428	0.415	0.443	0.452
FF3	0.147	0.087	0.127	0.111	0.154	0.139	0.166	0.175	0.234	0.165
FF3-BoAW	0.279	0.216	0.298	0.255*	0.354	0.371	0.395	0.395	0.394	0.392
FF5	0.139	0.129	0.141	0.138	0.147	0.131	0.186	0.169	0.195	0.157
FF5-BoAW(100)	0.258	0.196	**0.401***	**0.479***	0.476*	0.439*	0.497*	0.464*	0.524*	0.479*
FF5-BoAW(500)	0.177	0.129	0.400*	0.364*	0.427	0.412	0.511*	0.480*	0.525*	0.499*
FF5+Q	0.142	0.122	0.158	0.145	0.165	0.156	0.174	0.169	0.173	0.160
FF5-BoAW(500) + Q	0.226	0.237	0.391*	0.327*	**0.449**	**0.455***	0.502*	0.482*	0.519*	0.489*
FF6	0.174	0.139	0.145	0.193	0.149	0.198	0.181	0.155	0.201	0.219
FF6-BoAW(100)	0.248	0.235	0.368*	0.322*	0.447	**0.455***	0.459	0.463*	0.474*	0.483*
FF6-BoAW(500)	0.267	0.248	0.388*	0.385*	0.447	0.447*	0.512*	**0.503***	0.518*	0.508*
FF6 + Q	0.195	0.190	0.162	0.186	0.159	0.173	0.195	0.177	0.226	0.196
FF6-BoAW(500) + Q	0.192	0.186	0.394*	0.373*	**0.449**	0.422	**0.524***	0.494*	**0.536***	**0.525***

characteristics [47], caused the greatest performance drop with the proposed features. To further investigate the effect of babble noise on the modulation spectrum, Fig. 3.3 depicts the average modulation spectrograms for unprocessed high (top-left) and low (top-right) valence speech files, as well as noisy high (bottom-left) and low (bottom-right) valence speech files. As can be seen, noise substantially affects speech, especially in low valence conditions, thus corroborating the drops in performance. Similar findings were observed for arousal. Overall, BoAW computation over most of the single features showed improved performance across noise types and levels.

Moreover, fusion of different feature sets showed to result in additional improvements and increased robustness. The proposed features, for example, showed to be the most complementary to prosodic(100) features and consistently achieved either the highest CCC or competitive levels compared to other fused methods, especially for arousal prediction. Prosodic features have been linked to arousal [189] and emotions [189], hence complement modulation spectral cues. Moreover, the addition of the $SRMR_{norm}$ quality metric consistently helped improve CCC for both valence

Table 3.4: Performance comparison of different feature representations in terms of CCC with and without BoW computation for mismatch train-test conditions with babble noise at different SNR levels. LSTM regression models were used for all methods and highest values are indicated in bold. Significantly better results relative to the benchmark are highlighted by an asterisk.

	Babble (0dB)		Babble (5dB)		Babble(10dB)		Babble (15dB)		Babble (20dB)	
Features	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence
eGeMAPS	0.129	0.251	0.166	0.280	0.182	0.192	0.204	0.201	0.227	0.280
eGeMAPS-BoAW	**0.276**	**0.278**	0.269	0.245	**0.499**	0.458	0.499	0.427	0.448	0.437
IS11	0.169	0.191	0.175	0.161	0.238	0.198	0.267	0.235	0.251	0.278
IS11-BoAW	0.118	0.102	0.297*	0.305*	0.418	0.415	0.436	0.462*	0.453	0.473*
MFCC	0.107	0.174	0.296*	0.207	0.269	0.249	0.282	0.250	0.311	0.265
MFCC-BoAW	0.214	0.153	0.290*	0.264*	0.277	0.296	0.357	0.351	0.313	0.331
MSF	0.086	0.068	0.119	0.111	0.174	0.133	0.289	0.277	0.330	0.313
MSF-BoAW	0.104	0.162	0.318*	0.318*	0.460	0.426	0.457	0.379	0.456	0.444
Prosodic(100)	0.120	0.139	0.173	0.164	0.234	0.152	0.239	0.162	0.248	0.186
Prosodic(100)-BoAW	0.179	0.144	**0.383***	**0.360***	0.469	**0.476***	**0.525***	**0.495***	**0.527***	**0.489***
FF1	0.161	0.226	0.291*	0.227	0.174	0.176	0.359	0.307	0.327	0.283
FF1-BoAW	0.279	0.218	0.341*	0.258	0.466	0.419	0.469	0.453*	0.481*	0.474*
FF2	0.143	0.136	0.144	0.193	0.224	0.176	0.310	0.207	0.313	0.240
FF2-BoAW	0.181	0.185	0.248	0.271*	0.453	0.420	0.485	0.431	0.490	0.445
FF3	0.139	0.107	0.202	0.207	0.249	0.243	0.198	0.118	0.202	0.195
FF3-BoAW	0.226	0.137	0.238	0.250	0.387	0.363	0.487	0.387	0.497*	0.392
FF5	0.116	0.122	0.140	0.128	0.149	0.175	0.173	0.275	0.266	0.253
FF5-BoAW(100)	**0.312***	**0.249**	0.418*	**0.374***	**0.509***	0.485*	0.494	0.491*	0.501*	0.493*
FF5-BoAW(500)	0.189	0.137	0.350*	0.352*	0.446	0.458	0.500	0.475*	**0.514***	0.497*
FF5 + Q	0.120	0.143	0.128	0.190	0.149	0.258	0.173	0.275	0.281	0.265
FF5-BoAW(500) + Q	0.185	0.137	0.356*	0.321*	0.448	0.449	**0.499**	0.481*	0.501*	0.506*
FF6	0.162	0.193	0.158	0.233	0.16	0.263	0.175	0.198	0.288	0.217
FF6-BoAW(100)	0.235	0.145	**0.382***	0.341*	0.481	0.469	0.469	0.464*	0.478*	0.453*
FF6-BoAW(500)	0.192	0.158	0.375*	0.361*	0.451	0.438	0.497	0.493*	0.484*	0.506*
FF6 + Q	0.194	0.213	0.152	0.241	0.156	0.292	0.233	0.196	0.295	0.231
FF6-BoAW(500) + Q	0.156	0.183	0.371*	**0.376***	0.471	**0.500***	0.459	**0.509***	0.508*	**0.522***

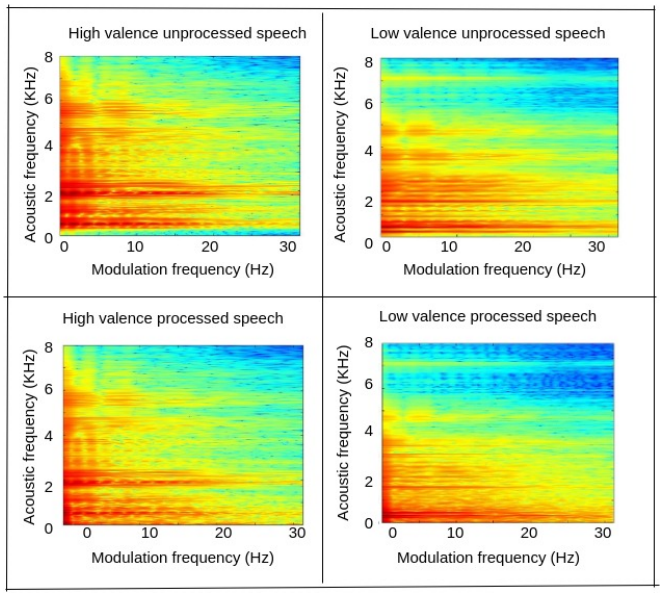

Figure 3.3: Average modulation spectrogram plots for unprocessed (top row) and processed speech (bottom row) for high (left column) and low valence (right column) emotional states.

and arousal when fused with the proposed features and prosodic(500) features, thus highlighting the importance of quality-awareness for SER tasks "in the wild". The benefits of the quality-aware method, however, were not as pronounced with the babble noise conditions. As mentioned above, it is known that the measure can be sensitive to speech-like noise sources [190], hence other quality measures could be explored in the future.

Lastly, for a final comparison we tested the transformer-based SOTA on the airport 5 dB and babble 0 dB conditions and found CCC values of 0.381 for arousal and 0.303 for valence, and 0.263 (arousal) and 0.241 (valence), respectively. As can be seen, the proposed method with feature fusion is able to outperform this very recent SOTA for both valence and arousal in highly noisy mismatched conditions.

3.6.3.2 Reverberation (Convolutive noise)

Table 3.7 lists CCC values achieved when the mismatch is due to room reverberation, i.e., models are trained on unprocessed speech and tested on reverberant speech with varying reverberation times (RT). Again, experiments are conducted on the German SEWA dataset. As can be seen, when exploring sets of single features, the impact of increasing reverberation times was less pronounced relative to the additive noise case for all feature sets. This finding suggests that the BoAW method may provide some inherent robustness to reverberation mismatch. Interestingly, prosodic BoAW feature performance achieved the highest CCC for both valence and arousal. OpenSMILE utilizes a cepstrum based method for pitch tracking which has been shown robust to reverberation [191]; this could explain the high accuracy achieved with the prosodic features alone.

As previously, the proposed features showed to be highly complementary to prosodic features and when combined, performance improvements were observed. In fact, for higher reverberation levels, fusion of modulation spectral, MFCC, and prosodic features showed to achieve the best CCC. Here again, quality-awareness showed improvements in valence and arousal measurement, in particular in systems that fused the proposed and prosodic features. The improvements were particularly important for valence prediction. These results were expected, as SRMR_{norm} was originally tailored for quality and intelligibility prediction of reverberant speech [68]. Overall, the quality-aware methods achieved results at par with those obtained with unprocessed speech, hence emphasizing the robustness of the proposed measures. As previously, to further compare the obtained results

Table 3.5: Performance comparison of different features in terms of CCC with and without BoAW computation for mismatch train-test conditions with reverberation. LSTM regression models were used and highest values are indicated in bold. Significantly better results relative to the benchmark are highlighted by an asterisk.

	RT= 0.25 s		RT = 0.48 s		RT = 0.80 s	
Features	Arousal	Valence	Arousal	Valence	Arousal	Valence
eGeMAPS	0.229	0.186	0.194	0.152	0.239	0.188
eGeMAPS-BoAW	0.396	0.396	0.403	0.394	0.442	0.408
IS11	0.247	0.261	0.137	0.152	0.259	0.202
IS11-BoAW	0.349	0.296	0.298	0.231	0.330	0.262
MFCC	0.302	0.269	0.271	0.236	0.339	0.281
MFCC-BoAW	0.397	0.389	0.431*	0.421*	0.419	0.387
MSF	0.262	0.228	0.228	0.246	0.252	0.267
MSF-BoAW	0.381	0.346	0.333	0.303	0.375	0.373
Prosodic	0.261	0.246	0.230	0.208	0.227	0.225
Prosodic-BoAW	**0.534***	**0.477***	**0.516***	**0.469***	**0.533***	**0.480***
FF1	0.219	0.209	0.248	0.284	0.275	0.295
FF1-BoAW	0.446*	0.406	0.453*	0.396	0.436	0.389
FF2	0.251	0.209	0.228	0.246	0.266	0.236
FF2-BoAW	0.438*	0.367	0.408	0.316	0.415	0.359
FF3	0.204	0.208	0.216	0.247	0.214	0.240
FF3-BoAW	0.399	0.385	0.412	0.385	0.411	0.368
FF5	0.175	0.209	0.196	0.209	0.239	0.217
FF5-BoAW(100)	0.513*	0.501*	0.465*	0.476*	0.489*	0.497*
FF5-BoAW(500)	0.529*	0.489*	**0.524***	0.479*	0.533*	0.464*
FF5+Q	0.189	0.209	0.217	0.234	0.247	0.256
FF5-BoAW(500)+Q	**0.541***	0.508*	0.520*	0.484*	0.528*	0.500*
FF6	0.221	0.235	0.223	0.256	0.247	0.263
FF6-BoAW(100)	0.476*	0.448*	0.472*	0.467*	0.484*	0.440*
FF6-BoAW(500)	0.515*	**0.521***	0.502*	0.527*	0.532*	0.501*
FF6+Q	0.240	0.252	0.243	0.276	0.262	0.277
FF6-BoAW(500)+Q	0.523*	0.503*	0.504*	**0.508***	**0.538***	**0.503***

with a recent SOTA, the transformer based system of [188] is tested in the RT=0.48s condition and a *CCC* of 0.350 and 0.275 was achieved for arousal and valence, respectively. As can be seen, the proposed methods are shown to substantially outperform this recent SOTA method.

3.6.4 Processed speech: discrete emotion classification

Here, we present the classification results achieved with the proposed and benchmark features on the Emoti-W 2017 dataset (see Section 3.5.1). As the recordings of this dataset are already considered to be "in the wild," no further degradations are included. Table 3.6 shows the figures-

Table 3.6: Performance comparison of different features with BoAW computation for the Emoti-W discrete emotion dataset in terms of precision, recall, and F1-score. An SVM was used for all methods and highest values are indicated in bold.

Features	Precision	Recall	F1
eGeMAPS	50.54	43.91	42.12
IS11	43.50	41.53	39.94
MFCC	48.49	**44.26**	41.29
MSF	49.62	43.16	**42.38**
Prosodic(500)	**51.46**	43.17	42.24
FF1	54.80	46.57	45.36
FF2	48.66	44.95	43.20
FF3	49.15	43.31	40.42
FF4(500)	51.84	42.97	41.65
FF4(500) + Q	51.86	43.04	41.66
FF6(500)	57.47	46.59	46.20
FF6(500) + Q	**57.77**	**46.86**	**46.31**

of-merit of the proposed classifier on a 4-class problem, where the following classes were explored: happy, angry, sad, neutral. Overall, 121-Angry, 144-happy, 107-sad, 141-neutral audio files were available for training and 64-angry, 63-happy, 61-sad, and 63-neutral audio files were available for testing. As observed with the German SEWA dataset, prosodic features when used alone showed to achieved the highest performance, followed closely by eGeMAPS and the proposed features. The fused feature set, however, showed the best overall performance with the proposed features fused with MFCC and prosodic features achieving the best precision, recall, and F1-scores. Quality awareness showed only some slight improvements, likely due to the same issues highlighted previously (e.g., presence of speech-like noise), hence alternate quality metrics could be explored in the future. For comparisons with one of the top performing benchmarks in Table 3.1, the system in [183] is tested and shown to achieve F1 = 41%. As can be seen, the proposed system with quality awareness is able to outperform this SOTA system by 13%.

3.6.5 Data augmentation to reduce language train/test mismatch

With deep learning, data augmentation has shown to be extremely beneficial to improve accuracy and provide robustness against train/test mismatch. Here, we explore the impact that data augmentation can have on the generalization capability of the LSTM regressor; for the sake of brevity, focus is placed on cross-lingual SER. To this end, we augment the unprocessed speech training dataset with the two different noise types used (i.e., babble and airport) at five different

Table 3.7: Performance comparison of different features with BoAW computation for mismatch train-test conditions with reverberation. LSTM regression models were used. Ar.=arousal; Val.=valence.

	RT = 0.25s		RT = 0.48s		RT = 0.80s	
Features	Ar.	Val.	Ar.	Val.	Ar.	Val.
eGeMAPS	0.396	0.396	0.403	0.394	0.442	0.408
IS11	0.349	0.296	0.298	0.231	0.330	0.262
MFCC	0.397	0.389	0.431	0.421	0.419	0.387
MSF	0.381	0.346	0.333	0.303	0.375	0.373
Prosodic(100)	0.534	0.477	0.516	0.469	0.533	0.480
FF1	0.446	0.406	0.453	0.396	0.436	0.389
FF2	0.438	0.367	0.408	0.316	0.415	0.359
FF3	0.399	0.385	0.412	0.385	0.411	0.368
FF5(100)	0.513	0.501	0.465	0.476	0.489	0.497
FF5(500)	0.529	0.489	0.524	0.479	0.533	0.464
FF5(500) + Q	0.541	0.508	0.520	0.484	0.528	0.500
FF6(100)	0.476	0.448	0.472	0.467	0.484	0.440
FF6(500)	0.515	0.521	0.502	0.527	0.532	0.501
FF6(500) + Q	0.523	0.503	0.504	0.508	0.538	0.503

Table 3.8: Performance comparison of different features in terms of CCC with and without data augmentation under matched and mismatched language conditions. LSTM regression models were used and highest values are indicated in bold. A=arousal; V=valence. Significantly better results relative to the benchmark are highlighted by an asterisk.

Train-	Hungarian		German		German + aug.		German		Hungarian		Hungarian + aug.	
Test	Hungarian		Hungarian		Hungarian		German		German		German	
Features	A	V	A	V	A	V	A	V	A	V	A	V
eGeMAPS-BoAW	0.273	0.153	0.121	0.174	0.276	0.095	0.496	0.477	0.355	0.317	0.419	**0.429**
IS11-BoAW	0.260	**0.158**	0.137*	0.110	0.306*	0.180*	0.501	0.436	0.199	0.179	0.383	0.323
MFCC-BoAW	0.261	0.130	**0.257***	**0.207***	**0.366***	**0.197***	0.378	0.368	0.259	0.134	0.403	0.378
MSF-BoAW	0.275	0.079	0.113	0.009	0.216	0.089	0.490	0.464	0.317	0.215	0.407	0.340
Prosodic-BoAW	**0.304***	0.139	0.061	0.099	0.273	0.125*	**0.537***	**0.486**	**0.357**	**0.374**	**0.424**	0.324*
FF1-BoAW	0.288*	**0.170***	0.128	0.166	0.337*	0.162*	0.481	0.453	**0.430***	0.315	0.447*	0.320
FF2-BoAW	0.273	0.149	0.118	0.116	0.309*	0.201*	0.477	0.455	0.236	0.259	0.301	0.289
FF3-BoAW	0.299*	0.162	0.184*	0.177	0.342*	**0.220***	0.508*	0.479	0.236	0.187	0.367	0.229
FF5-BoAW	0.302*	0.145	0.106	0.180	0.322*	0.186*	**0.556***	0.504*	0.381*	0.354*	0.436*	0.367
FF5-BoAW + Q	0.312*	0.146	0.165*	0.166	0.326*	0.189*	0.540*	0.514*	0.395*	**0.367***	**0.489***	**0.480***
FF6-BoAW	**0.334***	0.066	0.178*	0.146	0.260*	0.162*	0.529*	0.502 *	0.144	0.305	0.451*	0.431
FF6-BoAW + Q	0.327*	0.095	**0.191***	**0.290***	**0.364***	0.162*	0.532*	**0.518***	0.320	0.309	**0.489***	0.433

SNR levels (SNR=0-20 dB, 5dB increments), with three reverberation levels (RT= 0.25, 0.48, and 0.80), and 12 noise-plus-reverberation conditions (2 noises × 3 SNR levels (0, 10, and 20 dB) × 2 RT values (0.25, 0.8)). Overall, a total of 1248 training audio files were used for training, hence a 26-fold increase relative to the case reported in Section V.B. To show the effectiveness of augmentation on cross-language SER, Table 3.8 shows the obtained results when languages are matched, as well as when they are mismatched. In this latter scenario, results with and without data augmentation are reported. In all cases, the test set is unprocessed.

As can be seen, a mismatch between train/test languages can cause a substantial decrease in performance for most feature and feature fusion combinations, with the exception of the MFCC BoAW features. Data augmentation substantially improved the results, to levels comparable to those achieved when the languages were matched, and in some cases, even higher, especially for the valence dimension. Quality awareness was also shown to be useful even when data augmentation was used, in particular for the arousal dimension and the FF6 fused feature set. Overall, these findings suggest that the proposed features, combined with quality awareness and data augmentation during training could also be useful resources for "in-the-wild" SER tasks with language mismatch.

3.7 Conclusion

In this chapter, we explore "in-the-wild" speech emotion recognition where environmental factors, such as noise and reverberation, and different languages are present at test time, thus degrading system performance. We show the impact that this train-test mismatch has on SER performance and propose a quality-aware system based on a new modulation spectral bag-of-words feature representation that outperforms several benchmarks. Experiments on several SER Challenge datasets show the proposed features outperforming several benchmark systems, as well as providing complementary information to conventional features. The proposed feature sets inherently carry speech quality information, thus a quality-aware variant is also explored and shown to further improve SER prediction, both in terms of arousal/valence level predictions and in discrete emotion classification. Finally, we showed the impact that data augmentation had on language mismatch robustness, thus highlighting the potential of the proposed system for "in-the-wild" SER.

Chapter 4

Cross-Language Speech Emotion Recognition Using Bag-of-Word Representations, Domain Adaptation, and Data Augmentation

4.1 Preamble

The content in this chapter is extracted from the manuscript published in the MDPI Sensors journal [109].

4.2 Introduction

Speech emotion recognition (SER) is an emerging field in affective computing that, as the name suggests, has as goal the detection or characterization of speaker emotional states based on the analysis of the speech signal alone. SER can have applications across a wide range of domains, from call centers, to smart cars, healthcare, and education, to name a few. Emotion-aware human-machine interfaces are starting to emerge in the market via start-ups, such as audEERING,

Nemesysco, Nexidia, and Emospeech. "In the wild" SER, however, is still very challenging as there are a number of parameters that can vary between training and testing conditions, including but not limited to: types of emotions collected, labeling schemes, sampling rates, environmental conditions, microphone settings, speakers, as well as spoken languages and cultural background, just to name a few. These cross-corpus changes are known to severely hamper SER performance [70, 107, 192, 108].

While many studies have explored the issue of cross-corpus SER (e.g., [193, 194, 70, 107, 195]), in this chapter, we focus on the mismatch due to different languages. Commonly, cross-lingual emotion prediction has relied on three methods: feature normalization [194, 196, 192, 197, 198], domain adaptation (DA) [199, 200, 201], or transfer learning [107, 108, 202, 203]. We pay particular focus on domain adaptation methods, which have seen great success in computer vision tasks (e.g., [204]) but are still under-explored in SER tasks. Domain adaptation improves the generalization of the SER system by minimizing the distribution shift between the source (training) and target (testing) data, including shifts due to varying languages. DA separates the two domains via measures of maximum mean discrepancy, correlation distances, or even by creating a shared representation of the source and target data [205].

Alternately, bag-of-word (BoW) [43] and data augmentation methodologies [44, 64] have also been explored as ways to remove cross-corpus biases. BoW has been used for text-based sentiment analysis, as well as multimodal emotion recognition systems [206, 207], and was recently explored for cross-lingual SER tasks [43], [64], [64]. Data augmentation, in turn, has shown to provide some robustness against cross-corpus mismatches, including cross-languages [44]. While DA, BoW, and data augmentation have been explored individually for cross-language SER tasks in the past, their combinations have yet to be explored. This chapter aims to fill this gap, and experiments with SER tasks in German, Hungarian, Chinese, and French are performed. In particular, in this chapter, the following contributions are made:

1. We explore the combination of DA and BOW for improved cross-language SER. Experiments with the BoW methodology before or after domain adaptation are performed to assess their advantages/disadvantages. Different DA methods are explored to gauge their effects on overall cross-language SER. In particular, the CORrelation ALignment (CORAL) [69] method, as well as the subspace alignment (SA) methods, are compared.

2. A variant of the CORAL method is proposed for cross-language SER. The method, termed N-CORAL, makes use of a third unseen unlabelled dataset/language to adapt both domain and source data, thus, in essence, normalizing both training and test datasets to a common distribution, as typically done with domain generalization.
3. Lastly, we explore the added benefits of data augmentation, on top of BoW and DA, for cross-language SER.

The remainder of this chapter is organized as follows: Section 4.3 describes related literature in cross-language SER, multi-lingual, and data augmentation for SER, DA, and domain generalization. Section 4.4 describes the proposed method along with materials and methods used, section 4.5 presents experimental setup. Section 4.6 presents the experimental results and discusses them, and, lastly, Section 4.7 draws the conclusions.

4.3 Related work

In this section, we describe related work dealing with cross-language, multi-lingual training and data augmentation, domain adaptation, as well as domain generalization aspects of speech emotion recognition.

4.3.1 Cross-language SER

The primary goal of an SER system is to detect emotions within a speech signal. As available datasets are typically recorded in one language, the majority of existing systems have reported mono-lingual results. Under cross-corpus conditions, however, especially with systems trained on one language and tested on another, system performance can decay drastically [70, 107, 192, 108]. The work in [194], for example, proposed the use of feature and speaker normalization to remove language effects from the SER system and experiments across six languages showed the importance of multi-lingual training paradigms. The work in [71], on the other hand, showed the importance of feature selection for cross-language SER. The experiments in [208] showed that gender- and language-specific models could be used to improve cross-language SER accuracy between Mandarin and other western languages. They reported higher performance in cross-language families compared to within-language families, suggesting some universal cues of emotional expression regardless of

language. Overall, cross-language SER accuracy has shown to be higher for the arousal dimension relative to valence [70, 71].

In the work described in [209], subspace alignment-based domain adaptation schemes were used to map language-specific SER models to unseen languages. Chiou and Chen, in turn, explored data normalization using histograms for improved cross-language SER [197]. Furthermore, the authors in [210] used deep belief networks to learn generalized features across different languages and showed improved cross-corpus SER performance. Ning et al., in turn, showed that universal feature representations could be achieved with bidirectional long short-term memory (Bi-LSTM) neural networks with shared hidden layers trained on English and Mandarin speech [211]. Several other works (e.g., [212], [213], [214]) have explored the use of end-to-end deep neural networks trained on multi-lingual data for improved cross-language SER. Multi-lingual SER can also be seen as a form of data augmentation; thus the next section focuses on multi-lingual training and data augmentation for cross-lingual SER.

4.3.2 Multilingual training and data augmentation for SER

Hesam et al. showed the benefits of using language identification coupled with multi-language SER models [198] for cross-language SER. In fact, multi-language training (i.e., training models with data from more than one language) has shown to attain reliable SER predictions for unseen languages [215, 216, 217, 218]. Schuller et al. showed the effect of selecting only the most prototypical examples when training cross-dataset SER systems [219] and later showed the importance of fusing the outputs of different deep learning systems [220]. More recently, convolutional neural networks (CNN) with attention have been proposed for cross-language SER [107, 195], where multi-language training showed to improve cross-language SER performance. In [107, 221], data augmentation also showed to improve cross-language SER accuracy.

4.3.3 Domain adaptation for SER

Domain adaptation aims to improve the generalization capacity of models by adapting the domain shift of source or target data, thus minimizing the differences in the feature space between both domains. Zhang et al. tackled cross-language SER by separately normalizing the features of

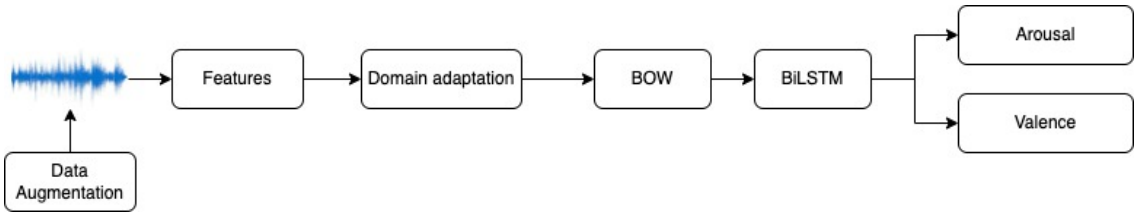

Figure 4.1: Block diagram of the two explored cross-language SER systems combining BOW and domain adaptation.

each speech corpus [217]. Hassan et al., in turn, employed kernel mean matching to increase the weight of the train data to match that of the test data distribution [199]. Zong et al. used least square regression to remove the projected mean and covariance differences between the source data and unlabeled target samples while learning the regression coefficient matrix [222], thus proposing a domain-adaptive least-squares regression model for cross-corpus SER. Song et al. proposed a novel DA method based on dimensionality reduction to create a similar feature space for both source and target domains [223]. Abdelwahab et al., in turn, explored a model-based DA method in which supervised adaptation of a support vector machine (SVM) classifier was performed via access to small amounts of target domain data [224].

Furthermore, the work in [200] proposed the kernel canonical correlation analysis (KCCA) on principal component subspaces for DA. They first projected the source and target to the feature space using PCA applied on the combined source and target domains. Then, they used KCCA to maximize the correlation between both. Song et al., in turn, proposed a nonnegative matrix factorization-based DA for cross-language SER [201]. More specifically, the authors proposed an algorithm that aimed to represent a matrix formed by data from both source and target domains as two nonnegative matrices whose product was an approximation of the original matrix. In order to ensure that the differences in the feature distributions of the two corpora were minimized, they regularized this factorization by the maximum mean discrepancy. Moreover, Abdelwahab and Busso [225] proposed a semi-supervised approach by creating ensemble classifiers. In this method, each classifier focuses on a different feature space, thus, learning the discriminant features for the target domain.

4.3.4 Domain generalization for SER

Domain generalization differs from domain adaptation, in which train and test domains are mapped to a common space where the feature representation is more robust to the variations between the domains. In [226], a sparse autoencoder method was used for feature transfer learning for SER. A common emotion-specific mapping rule is first learned from a small set of labeled data in a target domain. Then, this rule is applied to emotion-specific data in a different domain. Deng et al. [227], in turn, used auto-encoders to find a common feature representation between the source and target domains by minimizing the reconstruction error on both domains. Later, Mao et al. [228] proposed to learn a shared feature representation by sharing the class priors across domains. The work in [229] proposed Universum autoencoders, where the Universum loss is added to the reconstruction loss of an auto-encoder to reduce the reconstruction and classification error on both source and target domains. Deng et al. also presented a denoising autoencoder-based approach for cross-language SER [230, 231]. In fact, several variations of autoencoders have been investigated for cross-language SER, including variational autoencoders (VAE) [232], adversarial autoencoders (AAE) [233], and adversarial variational Bayes (AVB) [233].

4.4 Proposed Method

This section describes the proposed method based on the combination of bag-of-word (BOW) signal methodology and domain adaptation for cross-language SER. Figure 4.1 depicts the block diagram of the two methods explored herein, where the BOW feature extraction methodology is explored before or after domain adaptation. More detail about each individual block is described next.

4.4.1 Speech feature extraction

SER systems rely on different speech feature representations. In previous AVEC Challenges, hand-crafted features have been compared against feature representations obtained directly from end-to-end deep neural networks. It has been observed that hand-crafted features still outperform deep spectrum based features, for example [43]. This is likely due to the fact that existing emotion-

labeled datasets are fairly small compared to other domains, such as speech recognition, in which large amounts of data are available to allow for accurate feature representations to be obtained directly from the model. As the datasets used herein are fairly small, we employ two popular feature representations, namely: eGeMAPS and modulation spectral features (MSF).More detail can be found in section 2.2.3. We fuse these two feature sets into a final feature vector of dimension 246, of which 23 correspond to eGeMAPS and 223 to MSF features.

4.4.2 Bag-of-words methodology

In audio processing, BOW has been utilized with the term bag-of-audio-words, where LLDs are extracted from the audio signal and then codebook quantized [234]. Generally, statistical functionals such as mean, standard deviation, minimum, and maximum have been widely employed to represent frame-level features into utterance-level features. BOW is an alternative representation that aggregates frame-level features into utterance-level ones using different clustering methods. The more information can be found in 2.2.4. Here, we explore the usefulness of applying the BOW methodology after domain adaptation for cross-language SER. More details about the BOW procedure can be found in [125, 64]. The code for BOW generation can be found at `https://github.com/shrutikshirsagar/cross-language-SER`. In particular, we employed the Z-score standardization and random sampling for codebook generation. The random sampling-based codebook generation is much faster than the k-mean clustering-based algorithm.

4.4.3 Domain adaptation/generalization

Here, two domain adaptation methods are explored, and one domain generalization method is proposed. More details are provided in 2.3.2

4.4.3.1 Domain generalization with CORAL

We propose a variant of the described approach where a third language is used to adapt both the train and test domains. Figure 4.2 depicts this domain generalization method which we term N-CORAL. In our experiments, we utilize a Chinese language dataset as the target domain and adapt the train and test data of three different languages, namely German, French and Hungarian,

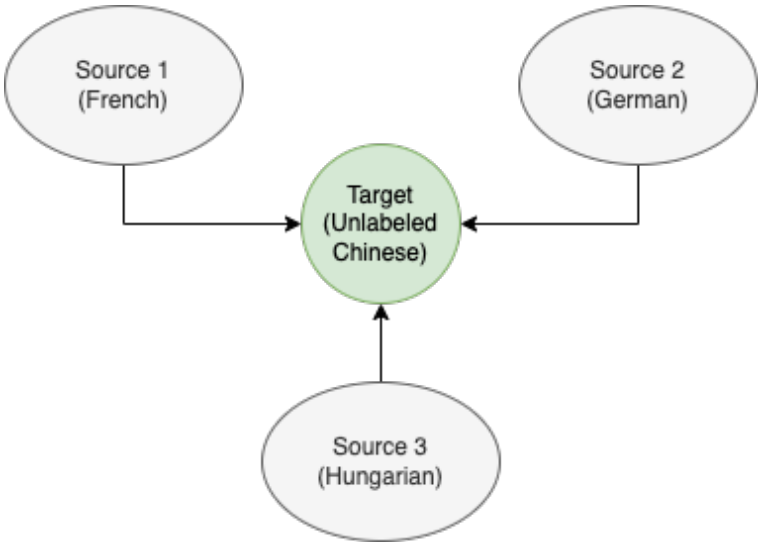

Figure 4.2: Proposed N-CORAL based domain generalization strategy for cross-language SER.

to this common domain before training an SER classifier. The main advantage of the proposed method is that we do not need access to the test data, as in previous methods. The same whitening and recolouring equations from (6)–(9) from 2.3.2 are used, but now to a common language.

4.5 Experimental setup

In this section, we describe the databases used, proposed regression model architectures, benchmark systems, and figure-of-merit used to gauge system performance.

4.5.1 Databases

For emotion prediction, we employed four datasets in four different languages. The first corresponds to the REmote COLlaborative and Affective interactions (RECOLA) database [55]. The second and third datasets correspond to the German and Hungarian language subsets of the Sentiment Analysis in the Wild (SEWA) database. The fourth dataset corresponds to the Chinese language subset of the SEWA project. Next, we used the recorded noise dataset AURORA [61] for data augmentation purposes to further corrupt the SEWA and RECOLA datasets. More detail about the above databases and data augmentation setup is given in section 2.4.

4.5.2 Regression Model

A bidirectional LSTM, or BiLSTM, is a sequence processing model that consists of two LSTMs: one taking the input in a forward direction and the other in a backward direction. BiLSTMs effectively increase the amount of information available to the network, improving the content available to the algorithm. BiLSTMs and LSTMs have been widely used in speech applications (e.g., [211, 155, 156, 43].

We employed the benchmark architecture from the AVEC 2018 [43] and AVEC 2019 Challenges described in [44]. We used two-layer BiLSTM with hidden layers of sizes 64 and 32, respectively. A whole sequence was used for training and experimented with 1000 epochs. A concordance correlation coefficient (CCC)-based loss function (see Section 4.5.4) was used for training inspired by the AVEC 2018 and 2019 benchmark systems. We used TensorFlow with KERAS as a backend. The implemented model for experimental validation can be found at `https://github.com/shrutikshirsagar/cross-language-SER`. Finally, for the hyper-parameter search, we used the validation set. We experimented with three widely used optimizers, including RmsProp, Adam, and SGD, with three different learning rates (0.01, 0.001, 0.0001) and dropout (0.1-0.5 in 0.1 increments).

4.5.3 Benchmark systems

Several benchmarks are used to gauge the benefits achieved with the proposed SER system. In particular, SA-DA alone [53] and CORAL-DA alone are used as benchmarks [150], as well as BOW alone, and no processing. We also used the AVEC 2019 challenge baseline [43] as an additional benchmark, as it relied on a bag-of-words methodology but over eGeMAPS features alone, together with an LSTM regressor. The AVEC 2019 challenge benchmark was also employed. In the AVEC 2019 baseline, the LSTM model was trained on a German + Hungarian language dataset. They also used bag-of-word features on the eGEMAPS feature set to aggregate over longer window sizes. Furthermore, in order to demonstrate the usefulness of the proposed approach, we compared it with several other methods, including: transfer learning approach based on principal component analysis (PCA), as in [235]; canonical correlation analysis based method (KCCA), as in [200]; and structural correspondence learning (SCL), as in [236]. Lastly, we use only data augmentation as a benchmark system.

4.5.4 Figure-of-Merit, Testing Set-up, and Experimental Aims

The performance measure used here is the typical metric used within SER tasks, i.e., the concordance correlation coefficient (CCC). More detail is given in section2.3.5

For the experimental setup, we have employed only the labeled training and validation partitions of the AVEC Challenge datasets. More specifically, we used AVEC challenge training data as our training data and further divided this training data with 80-20% for hyper-parameter tuning of Bi-LSTM models. Also, our test set is the Challenge validation set. In the end, once we found appropriate parameters, including optimizer, learning rate, and drop-out, for the model's final training, we joined our training -validation set (which was earlier divided as 80-20%). We also showed the significance of the obtained results using a z-score test between CCCs In particular; we used a 95% level ($p < 0.05$) against the AVEC 2019 benchmark system.

We start the experiments with an ablation study aimed at measuring the upper-bound achieved per language using mono-lingual models where the same language is used for training and testing. Next, we examine the impact of multi-lingual training, where multiple languages are combined during training. Next, we explore the impact of including the bag-of-words methodology, domain adaptation schemes, and data augmentation, both individually and combined. Lastly, we experiment with the proposed N-CORAL method.

4.6 Experimental Results and Discussion

In this section, we present the experimental results and then discuss our findings in light of existing literature.

4.6.1 Ablation Study

As an ablation study, we explored mono-lingual and multi-lingual training experiments to obtain "upper bounds" on what could be achieved for the tested languages and datasets without any domain adaptation strategies in place. Mono-lingual refers to experiments where the language of the test samples are the same as those used during training. Multi-lingual, in turn, combines multiple

Table 4.1: Ablation study results for mono-lingual, multi-lingual with matched test language, and multi-lingual with unseen test language experiments, without and with ('+Aug') data augmentation.

Train	Test	Arousal	Valence
German	German	0.450	0.363
AVEC 2019	German	0.434	0.455
Multi-matched	German	0.399	0.318
Multi-unseen	German	0.067	0.150
Multi-unseen + Aug	German	0.179	0.187
Hungarian	Hungarian	0.123	0.145
AVEC 2019	Hungarian	0.291	0.135
Multi-matched	Hungarian	0.263	0.154
Multi-unseen	Hungarian	0.147	0.037
Multi-unseen + Aug	Hungarian	0.241	0.240
French	French	0.772	0.418
AVEC 2019	French	0.323	0.144
Multi-matched	French	0.538	0.186
Multi-unseen	French	0.046	0.045
Multi-unseen +Aug	French	0.157	0.164

languages during training and tests them individually with either matched or unseen languages. Table 4.1 presents the ablation study results for several experiments, including three mono-lingual (train-test in German, Hungarian and French), three multi-lingual (train with German-Hungarian-French and test with each language individually), three unseen multi-lingual (train with two languages and test with the third unseen), and lastly, the three unseen multi-lingual conditions, but with data augmentation during training. For all experiments, BOW features and a BiLSTM regressor were used.

For these experiments, 34 audio files for training and 14 audio files for testing were used from the SEWA-German and SEWA-Hungarian datasets, whereas nine audio files were used for training, and nine audio files were used for testing from the RECOLA-French dataset. As can be seen from Table 4.1, valence estimation is more challenging compared to arousal, corroborating findings in [70, 71]. Moreover, with the exception of the Hungarian language, multi-language training did not help improve accuracy over the mono-lingual settings. Having the test language present during training showed to be important. Lastly, in the case where the test language was unseen, data augmentation showed to be important.

Table 4.2: Performance comparison of arousal estimation with different explored schemes in terms of CCC. Bi-LSTM regression was used for all methods. Highest values are indicated in bold and significantly better results relative to benchmark are highlighted by an asterisk.

Systems	Settings	Arousal						
		G-H	G-F	F-H	F-G	H-F	H-G	Avg
Benchmark	AVEC 2019	0.160	0.143	0.134	0.312	0.021	0.698	0.244
	No processing	0.118	0.128	0.144	0.237	0.045	0.711*	0.230
	BOW only	0.179*	0.131	0.155*	0.320	0.115*	0.749*	0.274*
	PCA	0.130	0.146	0.097	0.125	0.028	0.717*	0.207
	KCCA	0.180*	0.228*	0.123	0.082	0.180*	0.674	0.244
	SCL	0.124	0.165*	0.141	0.198	0.037	0.766*	0.238
	SA	0.140	0.151	0.122	0.148	**0.195***	0.762*	0.253
	CORAL	0.125	**0.236***	0.124	0.161	0.119*	0.729*	0.249
	Data augmentation	0.201*	0.129	0.220*	0.119	0.150*	0.447	0.211
Proposed	SA + BOW	0.193*	0.154	0.188*	0.248	0.109*	0.765*	0.276*
	CORAL + BOW	0.268*	0.167*	0.138	0.433*	0.138*	0.739*	0.313*
	N-CORAL	0.207*	0.127	0.167*	0.278	0.148*	0.733*	0.276*
	N-CORAL + BOW	**0.282***	0.126	0.175*	0.464*	0.124*	**0.787***	**0.326***
	N-CORAL + BOW + Aug	0.193*	0.189*	**0.241***	**0.369**	0.156*	0.480	0.271*

4.6.2 Proposed System

Tables 5.3 and 5.4 show the cross-language results obtained under the different conditions explored herein for arousal and valence prediction, respectively. Cross-language results achieved with different benchmarks (see Section 4.5.3) and the proposed systems are reported. In the Table, column labeled 'G-H' means that German and Hungarian languages were used for training and testing, respectively. As previously, for training, we used the thirty-four (train) audio files from SEWA-G and SEWA-H datasets and nine French language audio files from the RECOLA dataset. For testing, we used 14 audio files from the SEWA-G and SEWA-H test sets and nine from RECOLA. Significant improvements relative to the AVEC 2019 benchmark are reported with an asterisk. All results rely on the BiLSTM model.

As can be seen, for arousal, on average all of the proposed methods significantly outperformed the AVEC 2019 benchmark system. This benchmark is based on BOW applied to eGeMAPS features combined with an LSTM regressor. Interestingly, the other benchmark based on applying only BOW on the combined eGeMAPS and MSF features, together with a BiLSTM, also showed to significantly improve arousal prediction accuracy over the AVEC 2019 benchmark, thus highlighting the importance of the MSF features for the task at hand. Moreover, comparing the two proposed domain adaptation methods (SA+BOW and CORAL+BOW), on average, the CORAL-based method achieved the highest CCC. The left and right-side plots in Figure 4.3 depict the histograms of one

Table 4.3: Performance comparison of arousal estimation with different explored schemes in terms of CCC. Bi-LSTM regression was used for all methods. Highest values are indicated in bold and significantly better results relative to benchmark are highlighted by an asterisk.

Systems	Settings	Valence						
		G-H	G-F	F-H	F-G	H-F	H-G	Avg
Benchmark	AVEC 2019	0.046	0.112	0.200	0.073	0.090	0.671	0.198
	No processing	0.014	0.104	0.204	0.130*	0.109*	0.719*	0.213*
	BOW only	0.074*	0.133*	0.260*	0.153*	0.141*	0.745*	0.251*
	PCA	0.031	0.160*	0.137	0.104*	0.048	0.712*	0.198
	KCCA	0.069*	0.129	0.165	0.069	0.057	0.641	0.188
	SCL	0.024	0.157*	0.169	0.092*	0.071	0.771*	0.214*
	SA	0.033	0.126	0.128	0.117*	0.117*	0.782*	0.217*
	CORAL	0.065*	0.168*	0.315*	0.113*	0.09	0.726*	0.246*
	Data augmentation	0.107*	0.141*	0.214	0.075	0.154*	0.32	0.165
Proposed	SA + BOW	0.094*	0.139*	0.114	0.130*	0.158*	0.778*	0.235*
	CORAL + BOW	0.128*	**0.200***	0.371*	0.202*	0.123*	0.681*	0.284*
	N-CORAL	0.062*	0.125	0.143	0.131*	0.078	0.752*	0.215*
	N-CORAL + BOW	**0.141***	0.169*	0.217*	**0.310***	0.051	**0.799***	**0.281***
	N-CORAL + BOW + Aug	0.129*	0.131*	**0.352***	0.247*	**0.169***	0.473	0.267*

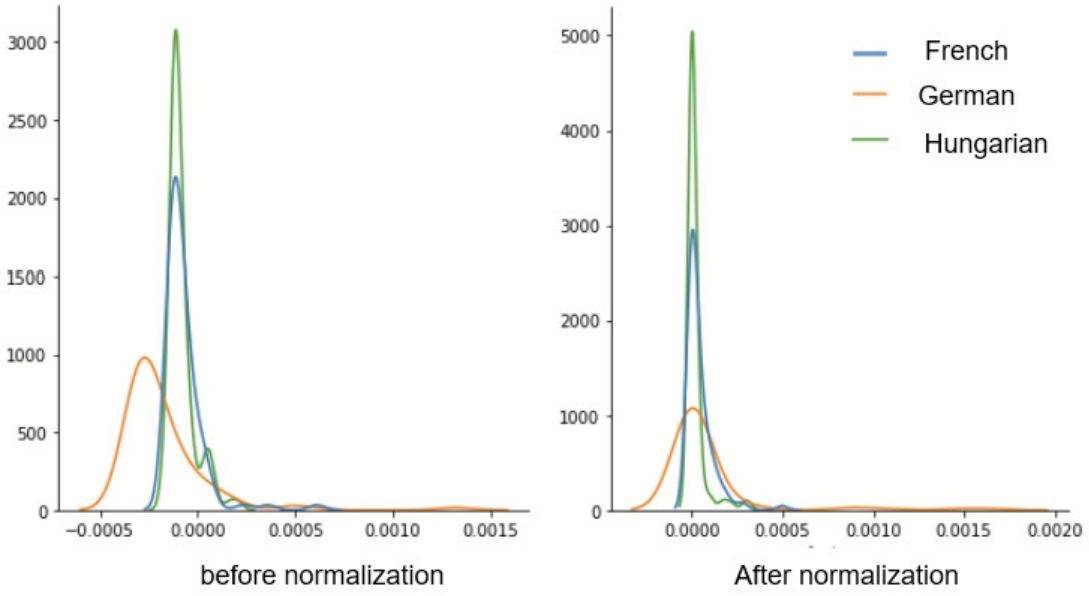

Figure 4.3: Illustration of the effects of CORAL on the distribution of one MSF feature for French, German, and Hungarian languages. Plots on the left are before normalization and on the right are after normalization.

MSF feature for French, German, and Hungarian languages before and after CORAL normalization, respectively. As can be seen, CORAL reduces the shift between the distributions across language, resulting in improved cross-language accuracy.

Overall, the proposed N-CORAL method achieved the highest CCC values of all tested methods, also outperforming several multi-lingual settings shown in Table 4.1. This was followed closely by

the proposed CORAL+BOW setting. Data augmentation, in turn, helped improve performance for half of the cross-language tasks, but on average, it did not provide any significant advantage for the N-CORAL setting. Moreover, it can be seen that in the conditions involving the SEWA German and SEWA Hungarian cross-language tasks achieved the highest CCC values across all tested cross-language tasks, especially with the N-CORAL+BOW method. These findings suggest that such a proposed scheme can be useful for cross-language normalization but not necessarily for cross-corpus where other nuance factors may be present. For cross-corpus and cross-language robustness, N-CORAL combined with BOW and data augmentation showed the most significant gains, thus combining the benefits of the N-CORAL+BOW method for cross-language robustness and the benefits of data augmentation for cross-corpus nuance factors.

Moreover, comparing the results from the tables, it can be seen that valence prediction is a more challenging task compared to arousal prediction, thus corroborating previous findings [215, 70]. Notwithstanding, the proposed methods showed to reduce this gap across many of the cross-language tasks. Overall, all of the proposed methods achieved CCC values significantly better than most benchmarks. Similar to the arousal prediction case, the proposed N-CORAL and CORAL+BOW settings achieved the highest average results. For valence prediction, data augmentation only helped for two of the six tested cases.

To better understand some of these findings, Figure 4.4 depicts a snapshot of the average modulation spectrogram across multiple speakers for three different languages for both high (left) and low (right) arousal conditions. As can be seen, differences across languages can be seen for both high and low arousal cases, thus motivating the need for cross-language strategies. Apart from the language differences, differences can also be seen between the high and low arousal conditions. Figure 4.5, on the other hand, shows modulation spectrograms for high (left) and low (right) valence conditions across the three languages. As can be seen, the differences are more subtle, thus suggesting a more complex classification task.

Furthermore, it can be observed that that while utilizing only data augmentation to compensate for cross-language issues can provide some improvements relative to doing nothing, the gains are typically substantially lower than applying other domain adaptation strategies. This can be somewhat expected as the augmentation strategies comprised adding noisy versions of the same language.

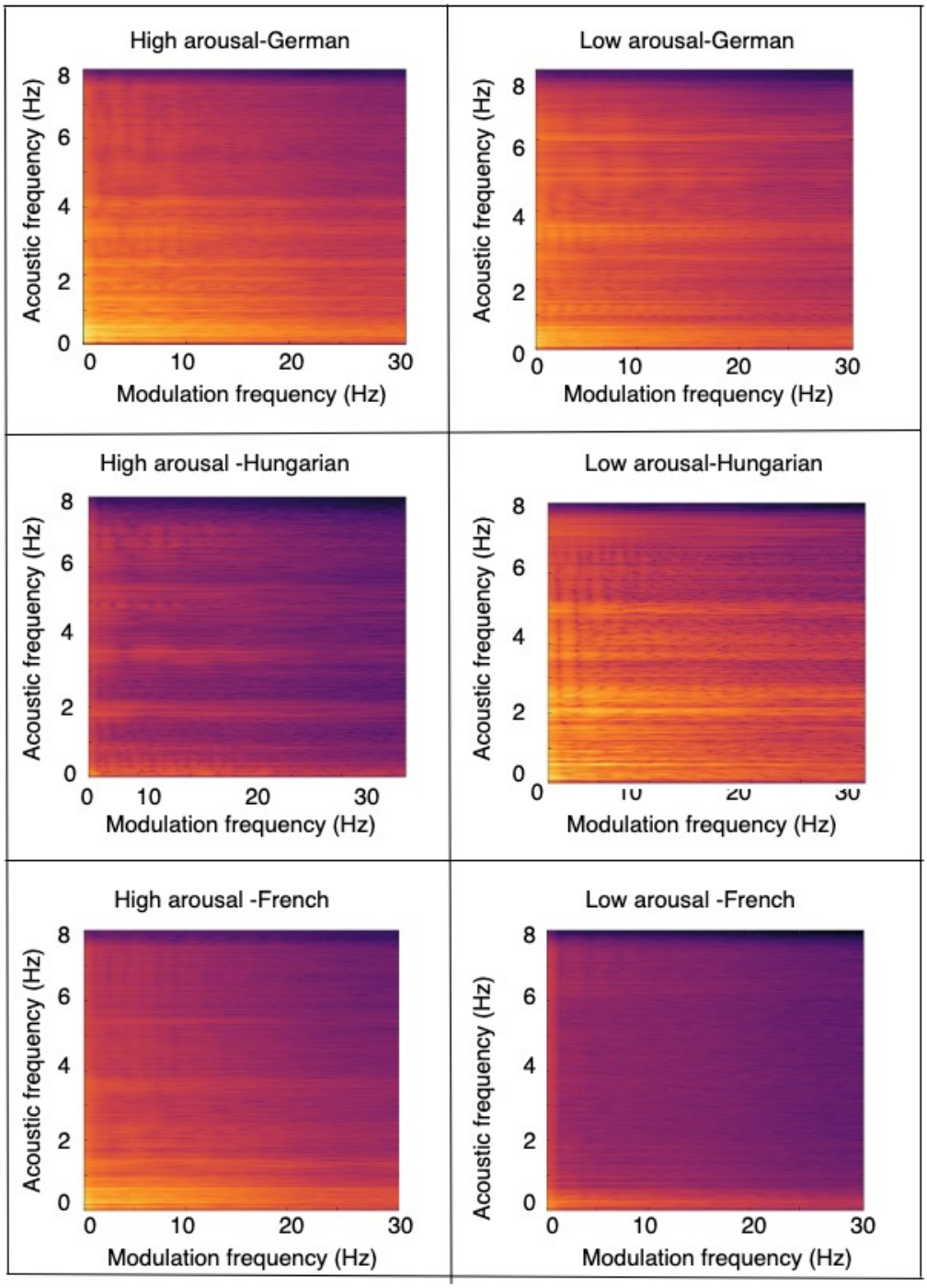

Figure 4.4: Average modulation spectrogram for German (top), Hungarian (middle) and French (bottom) language for high (left) and low (right) arousal conditions.

Lastly, while the proposed N-CORAL method showed to achieve the best performance across several valence and arousal prediction tests, further improvements may be achieved if other target languages are used. Here, the Chinese unlabeled data from the SEWA dataset was used as it was available together with other emotion labeled subsets. As Chinese belongs to a different family of

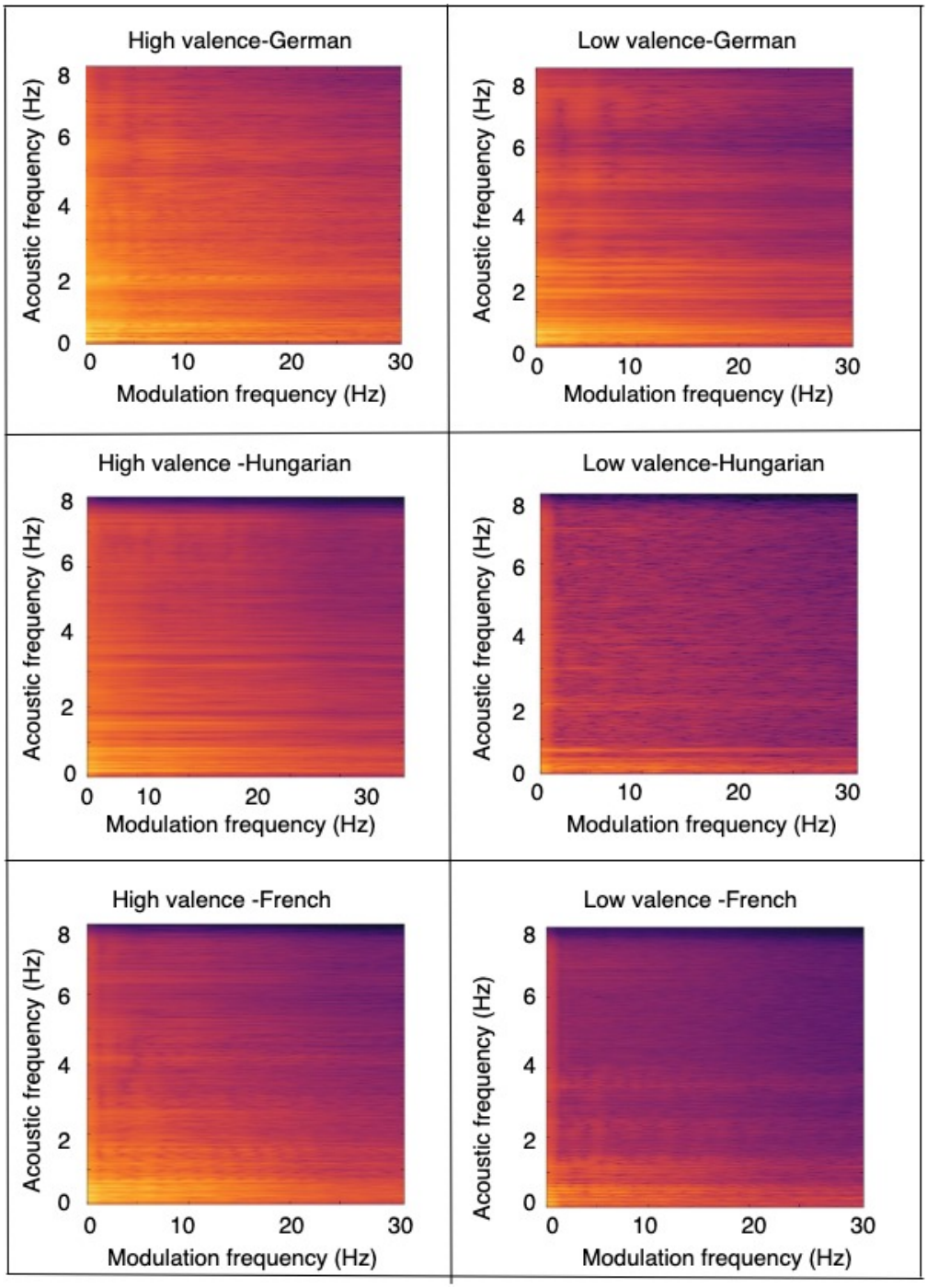

Figure 4.5: Average modulation spectrogram for German (top), Hungarian (middle) and French (bottom) language for high (left) and low (right) valence conditions.

languages then the other source languages experimented with here, future work may explore the use of different target languages belonging to the same family as the source languages to see if further gains can be achieved.

4.7 Conclusions

In this chapter, we explored the combined use of bag-of-words methodology, domain adaptation, and data augmentation as strategies to counter the detrimental effects of cross-language (and cross-corpus) speech emotion recognition. A new method termed N-CORAL was also proposed in which all languages are mapped to a common distribution (in our case, a Chinese language model). Experiments with German, French, and Hungarian languages show the benefits of the proposed N-CORAL method, combined with data augmentation and BOW for cross-language SER.

Chapter 5

Task-Specific Speech Enhancement and Data Augmentation for Improved Multimodal Emotion Recognition Under Noisy Conditions

5.1 Preamble

This chapter is compiled from material extracted from the manuscripts under review at *Frontiers in computer Science* [110].

5.2 Introduction

Affective human-machine interfaces are burgeoning as they provide more natural interactions between the human and the machine [3]. Emotion recognition (ER) systems have seen applications across numerous domains, from marketing, smart cities and vehicles, to call centres and patient monitoring, to name a few. In fact, the COVID-19 pandemic has resulted in a global mental health crisis that will have long-term consequences to society, economy, and healthcare systems [84]. Being

able to detect changes in affective states in a timely and reliable manner can allow individuals and organizations to put in place interventions to prevent, for example, burnout and depression [237].

ER systems can rely on a wide range of modalities, including speech, text, gestures/posture, and physiological responses (e.g., via changes in heart/breathing rates). For so-called "in the wild" applications, multimodal systems are preferred in order to compensate for certain confounds and to improve overall ER accuracy by providing the system with some redundancy and complementary information not available with unimodal systems [13, 14]. Multimodal systems, however, can be very time consuming to implement, costly to run, and potentially intrusive to the users (e.g., requiring on-body sensors with physiological data collection) and their privacy [136]. Notwithstanding, with audio inputs, one may be able to devise a multimodal speech-and-text system with the use of an advanced speech-to-text system, thus relying on a single input modality. As such, text and speech have emerged as two popular ER modalities.

Recent advances in deep learning architectures, such as transformers [238], have redefined the performance envelope of existing ER systems. In fact, most state-of-the-art systems today rely on deep neural network architectures in some way. For example, for text-based systems, self attention and dynamic max pooling has been proposed by [77]. The widely-used Bidirectional Encoder Representations from Transformers (BERT) model [1], in turn, has been used to detect cyber abuse in English and Hindi texts [239]. The work by [240, 241], in turn, relies on recurrent neural networks (RNN) to better consider long-range contextual effects and to better model the uncertainty around emotional labels. For speech-based ER systems, in turn, mel-spectral features combined with a convolutional neural networks (CNNs) have been extensively explored, specially with self-attention mechanisms to extract emotionally-informative time segments (e.g., [242]). Long-short term memory networks (LSTM) have also been extremely popular (e.g., [243, 244, 245]) and end-to-end solutions have also been explored [246].

As mentioned previously, one major advantage of the audio modality is that recent advances in automated speech-to-text conversion have allowed for multimodal speech-and-text-based systems to emerge while requiring the collection of just one signal modality [247]. Text and speech have been shown to be very useful modalities for multimodal ER systems [188]. In this regard, attention-based bidirectional LSTM models [248], bi-directional RNNs [249], transformer-based models [250], multi-level multi-head fusion attention mechanisms [251], graph-based CNNs [252], gated-recurrent units

[253], early and late fusion strategies [254], and cross-modal attention [255] have been explored as strategies to optimally combine information from the two modalities.

One major disadvantage of speech-based systems (either uni- or multi-modal), however, is their sensitivity to environmental factors, such as additive and convolutional noise (e.g., room reverberation). These factors can be detrimental to ER systems [256, 188]. Commonly, speech enhancement algorithms are applied at the input level stage to minimize environmental factors for "in-the-wild" speech applications. Enhancement methods can range from more classical methods, such as spectral subtraction and Wiener filtering [30, 31], to more recent deep neural network (DNN) based ones (e.g., [34, 35, 116, 119]). The use of speech enhancement for ER "in-the-wild" has shown some benefits (e.g., [45]).

Speech enhancement methods can have two very different purposes. If aimed at improving intelligibility/ quality, for example, human perception becomes the main driving factor and quality/intelligibility improvements are typically used as a figure of merit (e.g., [36]). However, if enhancement is used to improve downstream speech recognition applications then other machine-driven outcome measures, such as word error rate improvements, are more appropriate. As such, depending on the final task, the enhancement procedure can be very different. The work by [38], for example, showed that mimic loss-based enhancement was optimal for automatic speech recognition (ASR) downstream tasks. Having this said, it is hypothesized that for multimodal speech-and-text ER systems the use of two different enhancement procedures will be useful, with a quality-driven one used for the speech branch (mimicking how humans perceive emotions from speech) and a machine-driven one for the speech-to-text branch. We will test this hypothesis herein.

Lastly, with deep learning based approaches showing the latest state-of-the-art results, data augmentation has emerged as a useful technique to make systems more robust to "in-the-wild" distortions at the model training stage (e.g., [155]). With data augmentation, the training set is increased multi-fold by applying certain transformations to the available training signals, including time-reversal, time-frequency masking, pitch alterations, background noise addition and reverberation corruption, to name a few. For ER specifically, the work by [257] showed that vocal track length perturbations served as a useful data augmentation strategy. In this chapter, we further explore the advantages that data augmentation can provide, in addition to speech enhancement, for multimodal "in-the-wild" ER.

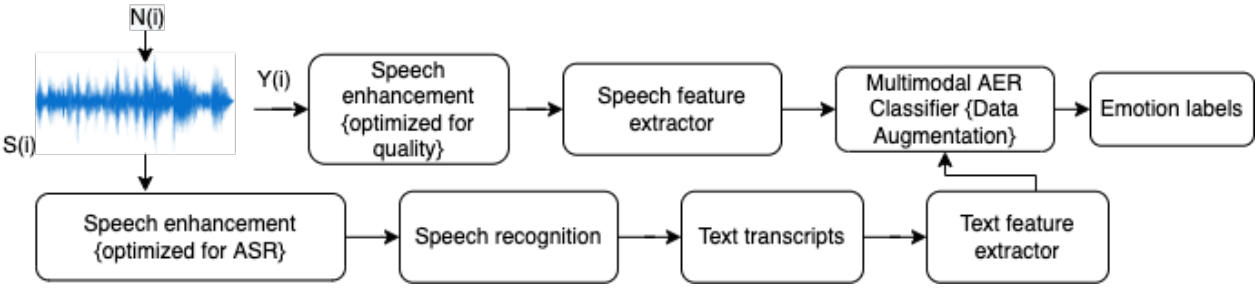

Figure 5.1: Experimental pipeline for ER using audio and text features

The remainder of this chapter is organized as follows. Section 5.3 describes the proposed system. Section 5.4 describes the experimental setup. Experimental results and a discussion are presented in Section 5.5 and conclusions in Section 5.6.

5.3 Proposed method

Figure 5.1 depicts the block diagram of the proposed multimodal ER pipeline. In the case of interest here, speech $S(i)$ is assumed to be corrupted by additive background noise $N(i)$, resulting in noisy speech signal $Y(i) = S(i) + N(i)$. With the multimodal ER system, the top branch focuses on extracting emotion-relevant features directly from the speech component, whereas the bottom branch relies on a state-of-the-art automatic speech recognizer (ASR) to generate text from the noisy speech signal. Features are then extracted from the text transcripts. We concatenated Speech and text features, then these concatenated features are input to a deep neural network for final emotion classification. As noisy speech is known to corrupt ER/ASR performance, here we also include a speech enhancement step, one optimized for speech quality improvement (top branch) and another for ASR. Each sub-block is described in detail in the subsections to follow:

5.3.1 Speech enhancement

Enhancement and noise suppression has been widely used across many different speech-based applications. In human-to-human communications, the goal of enhancement is to improve the quality of the noisy signal, not only to increase intelligibility, but also to improve paralinguistic characterization that humans do so well, such as emotion recognition. In human-to-machine interaction (e.g., ASR), however, improving quality may not be the ultimate goal, and instead, improvement

in downstream system accuracy could be regarded as a better optimization criterion. Here, we explore the use of a quality-optimized enhancement algorithm for the speech branch of the proposed method and an ASR-optimized algorithm for the text generation branch. The two algorithms used are described next:

5.3.1.1 MetricGAN+: A quality-optimized enhancement method

MetricGAN+ is a recent state-of-the-art deep neural network specifically optimized for quality enhancement of noisy speech and shown to outperform several other enhancement benchmarks [120, 36]. In particular, two networks are used. The discriminator's role is to minimize the difference between the predicted quality scores (given by the so-called PESQ, perceptual evaluation of speech quality, rating [37]) and actual PESQ quality scores. PESQ is a standardized International Telecommunications Union full-reference speech quality metric that maps a pair of speech files (a reference and the noisy counterpart) into a final quality rating between 1 (poor) and 5 (excellent). PESQ has been widely used and validated across numerous speech applications.

The generator's role, in turn, is to map a noisy speech signal into its enhanced counterpart. The discriminator and generator models are trained together to enhance the noisy signal in a manner that maximizes the PESQ score of the enhanced signal. MetricGAN+ builds on the original MetricGAN [120] via two improvements for the discriminator and one for the generator. More specifically, for the discriminator training, along with the enhanced and clean speech signals, the noisy speech was also used to minimize the distance between the discriminator and target objective metrics. The second improvement is that the speech generated from the previous epochs is reused to train the discriminator to avoid the catastrophic forgetting of the discriminator. For the generator, in turn, the learnable sigmoid function was used for mask estimation. The interested reader is referred to [120, 36] for more details on the MetricGAN and MetricGAN+ speech enhancement methods.

5.3.1.2 Mimic loss: an ASR-optimized enhancement method

Spectral mapping-based speech enhancement is an enhancement method specifically optimized for downstream ASR applications [38]. We refer henceforth to this method as 'mimic loss based enhancement' as the model uses mimic loss instead of student-teacher learning, thus the speech

enhancer is not jointly trained with a particular acoustic model. We use this enhancement model as it has been shown to be a useful pre-processing method for many ASR systems, thus offers some flexibility on the choice of ASR model to use [38]. The overall system is comprised of two major components: a spectral mapper and a spectral classifier which are trained in three steps.

First, a spectral classifier is trained to predict senone labels from clean speech with a cross-entropy criterion, resulting in a classification loss L_C between predicted and actual senones. The weights of this spectral classifier are then frozen and used in the last step. Second, a spectral mapper is pre-trained to map noisy speech features to clean speech features using a mean squared error (MSE) criterion. This results in a fidelity loss L_F between the denoised features and features from the clean speech counterpart. [38] relied on log-spectral magnitude components extracted over 25ms windows with a 10-ms shift as features and a deep feed-forward neural network for mapping.

Lastly, noisy speech is input to the pre-trained spectral mapper, resulting in a denoised version, which is input to the "frozen" spectral classifier, resulting in a predicted senone. In parallel, the clean speech counterpart is also input to the frozen spectral classifier, resulting in a soft senone label and a mimic loss L_M between the soft senone label and the predicted senone. The spectral mapper is then retrained using joint loss (L_F and L_M), thus allowing the enhancer to emulate the behavior of the classifier under clean conditions while keeping the projection of noisy signal closer to that of the clean signal counterpart. The same hyperparameters described by [38] were used herein. The interested reader is referred to [38] for more details on the mimic loss enhancement method.

5.3.2 Automatic speech recognition

In order to generate text from speech, a state-of-the-art automatic speech recognizer is needed. Here, wav2vec 2.0, an end-to-end speech recognition system is used [72].More detail can be found in section 2.2.6.

5.3.3 Speech feature extractor

Several ER systems have been proposed recently, and they have relied on different speech feature representations. Here, we focus on the three most popular representations, namely: prosodic,

eGeMAPS, and modulation spectral features. The interested reader is referred to section 2.2.3 for complete details on the computation of this representation.

5.3.4 Text feature representations

Text has also been used to infer the emotional content of written material and several state-of-the-art methods and techniques exist. Here, we explore three recent methods, namely BERT (Bidirectional Encoder Representations from Transformers), TextCNN, and Bag-of-Words (BoW). More detail can be found in section

5.3.5 Multimodal ER classifier

Here, we rely on a fully connected deep neural network for multimodal emotional recognition. Three dense layers (of dimensions 256, 128, 32) were used, plus a final classification layer. A dropout rate of 0.6 was used, batch normalization was performed after every layer, and class weights of [1, 1.8] were assigned during training. Grid search was performed on the validation set to obtain the optimal hyperparameters. Rmsprop, Adam, and SGD optimizers were explored, and learning rates of 0.01, 0.001, and 0.0001 were tested to find the optimal combination. Once the best parameters were found with the validation set, we reported the best performance on our test data. Experimentation codes are available on github[1]. The network is trained with and without data augmentation in order to explore its effect on "in-the-wild" ER performance.

5.4 Experimental Setup

In this section, we present the setup used in our experiments.

5.4.1 Datasets used

The dataset used for experimentation is the Multimodal EmotionLines Dataset (MELD) [59]. To test the robustness of the proposed methods to "in-the-wild" conditions, the MELD dataset is

[1] `https://github.com/shrutikshirsagar/Speech-enhancement-Audio-Text-ER`

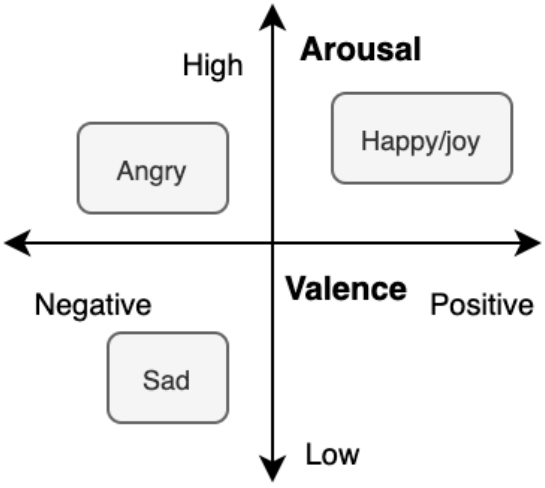

Figure 5.2: Valence-arousal emotional space with the three discrete emotions considered here.

corrupted by multi-talker babble noise and noise recorded inside a commercial airplane at three different SNR levels: 0, 10, and 20 dB. The AURORA [61] and DEMAND noise datasets [62] are used for this purpose. Next, we utilized the IEMOCAP dataset to show the generalizability of the proposed model. More detail on this databases can be obtained from section 2.4.

5.4.2 Benchmark systems

To gauge the benefits of the proposed method, two benchmark systems are used, namely BcLSTM and DialogueRNN. BcLSTM is bi-directional RNN proposed by [249]. It is comprised of a two-step hierarchical training process. First, it extracts embeddings from each modality. For text, GloVe embeddings [258] were used as input to a CNN-LSTM model to extract contextual representations for each utterance. For audio, Openmsile based features, which extracts 6373 dimensional features constituting several LLDs and various statistical functionals of varied vocal and prosodic features [41] were input to an LSTM model to obtain audio representations for each utterance. Next, contextual representations from the audio and text modalities are fed to the BcLSTM model for emotion classification.

DialogueRNN, in turn, employs three stages of gated recurrent units (GRU) to model emotional context in conversations [253]. The spoken utterances are fed into two GRUs: global and party GRU, to update the context and speaker state, respectively. In each turn, the party GRU updates its state

Table 5.1: Benchmark system performance for the two ER tasks based on the MELD dataset

	Task 1				Task 2			
Model	F1-score	Precision	Recall	BA	F1-score	Precision	Recall	BA
bcLSTM	0.70	0.73	0.67	0.72	0.82	0.81	0.83	0.83
DialogueRNN	0.72	0.72	0.72	0.72	0.84	0.84	0.84	0.85
Proposed system	0.74	0.75	0.75	0.73	0.87	0.87	0.87	0.87

based on i) the utterance spoken, ii) the speaker's previous state, and iii) the conversational context summarized by the global GRU through an attention mechanism. Finally, the updated speaker state is fed into the emotion GRU, which models the emotional information for classification. The attention mechanism is used on top of the emotion GRU to leverage contextual utterances by different speakers at various distances. Lastly, our proposed system comprises a feedforward DNN model and a 768- dimensional BERT(base) text feature vector fused (at the feature level) with a 311-dimensional vector comprised of eGEMAPs and MSF features.

5.4.3 Figures-of-Merit

Balanced accuracy, precision, recall, and F1-score are used as figures of merit to assess the performance of the proposed emotion classifier. The interested reader is referred to section 2.3.5 for more details on these classical performance metrics.

5.5 Experimental Results and Discussion

In this section, we present and discuss the obtained experimental results.

5.5.1 Ablation study 1

In this first ablation experiment, we wish to explore the optimal set of text and speech features to include in the final system. We consider speech and text modalities separately in this study. We start with clean speech to find the best feature per modality and, subsequently, test the robustness of such set under unseen noisy conditions. In this study, babble and airport noises are considered. In both cases, the emotion classifier is trained on clean speech only. Table 5.2 shows the perfor-

mance obtained for each modality individually for task 1. In the table, the feature termed 'fusion' corresponds to the fusion of MSF and eGeMAPS features.

As can be seen, for clean speech conditions and text-only ER, BERT-based text features resulted in the best performance across all metrics, hence corroborating previous reports [77, 78, 79]. As such, only BERT features are explored in the unseen noisy conditions. Babble noise is shown to degrade overall performance more severely than airport noise. Overall, BERT based features under 0 dB noise conditions are shown to achieve accuracy inline with that achieved by textCNN features under clean conditions, thus further suggesting improved robustness of the BERT text features. Given this finding, the final proposed system shown in Fig. 5.1 will rely on BERT based text features.

As for speech features, under clean conditions eGeMAPS showed the highest overall performance of the three tested feature sets, thus corroborating findings by [80]. Further gains could be seen with the fused feature set, however, thus suggesting the complementarity of spectral and modulation spectral features. As such, only the fused feature set is explored in the noisy mismatch condition. Moreover, similar to the text features, at low SNR levels, babble noise degraded performance more drastically compared to airport noise. Overall, the achieved performance with text-based features only was higher than what was achieved with audio features alone, thus corroborating the results reported by [188].

5.5.2 Ablation study 2

This second ablation study is an oracle experiment in which one modality in the multimodal system is kept clean and the other is corrupted by noise at varying levels and types. This study will allow us to gauge which modality is most sensitive to environmental factors and would benefit the most from speech enhancement. In all cases, the emotion classifier is trained on clean speech only. Tables 5.3 and 5.4 show the performance obtained for Task 1 and Task 2, respectively.

As can be seen, the fusion of speech and text features in the clean condition (first row in the tables) showed improvements relative to each modality alone (i.e., Table 5.2) by as much as 2% for text and 7% for audio in terms of F1 score for Task 1. Furthermore, using noisy speech to generate "noisy" text resulted in more severe performance degradations for both Tasks, thus suggesting that

Table 5.2: Ablation study 1: Performance comparison of different features for each individual modality. Feature termed 'fusion' corresponds to the fusion of eGeMAPS and MSFs.

Noise type	Feature	F1	Precision	Recall	BA
Text					
Clean	BERT	**0.72**	**0.72**	**0.72**	**0.76**
Clean	TextCNN	0.56	0.56	0.55	0.54
Clean	BoW	0.62	0.62	0.62	0.59
Airport (0dB)	BERT	0.54	0.55	0.54	0.52
Airport (10dB)	BERT	0.60	0.60	0.60	0.57
Airport (20dB)	BERT	0.62	0.62	0.61	0.59
Babble (0dB)	BERT	0.58	0.58	0.58	0.56
Babble (1dB)	BERT	0.61	0.62	0.60	0.58
Babble (20dB)	BERT	0.61	0.63	0.61	0.58
Audio					
Clean	Prosodic	0.62	0.65	0.62	0.61
Clean	eGEMAPS	0.69	0.70	0.69	0.67
Clean	MSF	0.66	0.69	0.66	0.67
Clean	Fusion	**0.69**	**0.70**	**0.69**	**0.71**
Airport_0	Fusion	0.51	0.62	0.57	0.51
Airport_10	Fusion	0.53	0.59	0.54	0.50
Airport_20	Fusion	0.55	0.60	0.56	0.52
Babble_0	Fusion	0.51	0.63	0.63	0.52
Babble_10	Fusion	0.52	0.61	0.55	0.51
Babble_20	Fusion	0.52	0.61	0.56	0.51

more powerful machine-tuned enhancement algorithms may be useful for "in-the-wild" applications to assure the highest possible quality for text generation. Overall, on average, over the two types of noise, a drop of 32%, 24%, and 21% in F1 score was observed at 0, 10, 20 dB SNR levels relative to clean conditions, respectively, for Task 1. On the other hand, corrupting only the speech content had a less pronounced effect. Overall, on average, over the two types of noise, a drop of 16%, 13%, and 9% in F1 score was observed at 0, 10, 20 dB SNR levels over clean conditions for Task 1, respectively.

For Task 2, similar findings were observed. Overall, on average, over the two types of noise, a drop of 65%, 33%, and 28% in F1 score has been observed at 0, 10, 20 dB SNR levels relative to clean conditions, respectively, when only text was corrupted. The drops in accuracy when the audio was corrupted were of 41%, 27%, and 25%, respectively. These findings corroborate those by [259, 188] who showed that text modality achieved higher performance than audio in clean conditions. The drops in accuracy, however, under noisy conditions motivate the need for strategies to improve accuracy in the wild, as in the proposed system.

Table 5.3: Ablation study 2 (Task 1): Performance comparison of multimodal oracle system for low/high arousal classification

Audio	Text	F1-score	Precision	Recall	BA
Clean	Clean	0.74	0.75	0.75	0.73
Clean	Airport (0 dB)	0.57	0.61	0.56	0.58
Clean	Airport (10 dB)	0.61	0.60	0.62	0.58
Clean	Airport (20 dB)	0.62	0.62	0.64	0.59
Clean	Babble (0 dB)	0.58	0.62	0.58	0.59
Clean	Babble (10 dB)	0.61	0.61	0.63	0.58
Clean	Babble (20 dB)	0.61	0.61	0.63	0.58
Airport (0 dB)	Clean	0.65	0.65	0.66	0.62
Airport (10 dB)	Clean	0.65	0.65	0.66	0.63
Airport (20 dB)	Clean	0.68	0.67	0.68	0.65
Babble (0 dB)	Clean	0.62	0.63	0.64	0.60
Babble (10 dB)	Clean	0.65	0.65	0.66	0.62
Babble (20 dB)	Clean	0.68	0.68	0.69	0.65

Table 5.4: Ablation study 2 (Task 2): Performance comparison of multimodal oracle system for joy vs sad classification.

Audio	Text	F1	Prec	Recall	BA
Clean	Clean	0.87	0.87	0.87	0.87
Clean	Airport (0 dB)	0.55	0.60	0.63	0.53
Clean	Airport (10 dB)	0.67	0.70	0.68	0.62
Clean	Airport (20 dB)	0.68	0.68	0.69	0.63
Clean	Babble (0 dB)	0.50	0.55	0.61	0.51
Clean	Babble (10 dB)	0.63	0.65	0.67	0.58
Clean	Babble (20 dB)	0.67	0.67	0.68	0.62
Airport (0 dB)	Clean	0.60	0.71	0.61	0.66
Airport (10 dB)	Clean	0.68	0.69	0.68	0.68
Airport (20 dB)	Clean	0.70	0.70	0.69	0.68
Babble (0 dB)	Clean	0.63	0.71	0.63	0.67
Babble (10 dB)	Clean	0.68	0.69	0.67	0.67
Babble (20 dB)	Clean	0.69	0.69	0.69	0.67

5.5.3 Ablation study 3

This third ablation study is an oracle experiment in which we wanted to test the hypothesis if we need two separate enhancement for improving ASR accuracy. As mentioned earlier, we used quality- (MetricGAN+) and ASR-optimized (mimic loss) enhancement algorithms for the speech and text branches shown in the proposed model in Fig. 5.1. This study will allow us to gauge which combination of speech enhancement is better suited for this task. In all cases, the emotion classifier is trained on clean speech only. Tables 5.5 and 5.6 show the performance obtained for Task 1 and Task 2, respectively.

Table 5.5: Ablation study 3 (Task 1): Performance comparison of enhancement system for angry vs sad classification.

Noise	Enhancement-1	Enhancement-2	F1	Precision	Recall	BA
Airport (0 dB)	MetricGAN+	MetricGAN+	0.60	0.61	0.60	0.60
	MetricGAN+	Mimic-loss	**0.65**	**0.66**	**0.64**	**0.64**
	Mimic-loss	MetricGAN+	0.61	0.61	0.60	0.59
	Mimic-loss	Mimic-loss	0.61	0.62	0.61	0.60
Babble (0 dB)	MetricGAN+	MetricGAN+	0.59	0.60	0.58	0.59
	MetricGAN+	Mimic-loss	**0.62**	**0.62**	**0.64**	**0.59**
	Mimic-loss	MetricGAN+	0.60	0.60	0.61	0.60
	Mimic-loss	Mimic-loss	0.61	0.62	0.60	0.61

Table 5.6: Ablation study 3 (Task 2): Performance comparison of enhancement system for joy vs sad classification.

Noise	Enhancement-1	Enhancement-2	F1	Precision	Recall	BA
Airport (0 dB)	MetricGAN+	MetricGAN+	0.53	0.52	0.55	0.50
	MetricGAN+	Mimic-loss	**0.56**	**0.56**	**0.57**	**0.51**
	Mimic-loss	MetricGAN+	0.54	0.54	0.53	0.52
	Mimic-loss	Mimic-loss	0.55	0.55	0.56	0.52
Babble (0 dB)	MetricGAN+	MetricGAN+	0.56	0.55	0.57	0.51
	MetricGAN+	Mimic-loss	**0.57**	**0.56**	**0.61**	**0.51**
	Mimic-loss	MetricGAN+	0.56	**0.56**	0.56	0.51
	Mimic-loss	Mimic-loss	0.56	0.55	0.58	0.52

As can be seen, for both Task 1 and Task 2, the best combination comprised the use of a quality-optimized enhancement algorithm for the top speech branch and an ASR-optimized (mimic loss) method for the bottom text branch. This combination resulted in the best accuracy for very extreme conditions (i.e., 0 dB SNR levels) and emphasizes the need for task-specific enhancement algorithms for ER.

5.5.4 Overall System Performance

This last study explores the performance of the proposed system described in Figure 5.1, combining speech enhancement optimized for each branch (speech and text), as well as data augmentation to provide robustness at the model training level. Data augmentation methods are useful to solve imbalanced data problems. It also helps the model to learn the complex distribution of the data and helps prevent overfitting. The work by [260] showed that adding noisy versions of the clean speech data to the training set improved speech recognition accuracy in mismatched noisy conditions. Therefore, in this work, we utilized the same strategy. Tables 5.7 and 5.8 show the obtained

results in rows labelled 'Data augmentation only' for Task 1 and Task 2, respectively. As can be seen, data augmentation alone already improved ER results, thus corroborating findings by [81], [82], and [64], [109].

Next, we gauge the benefits of using speech enhancement alone. As before, ER models are trained solely on clean speech. During run time, we pre-process the test data with the MetricGAN+ algorithm for the speech branch and the mimic loss enhancer for the text branch, as described in Section 5.3. Tables 5.7 and 5.8 show the obtained results in rows labelled 'Enhancement only'. As can be seen, applying speech enhancement improves overall performance relative to the noisy conditions, but the final results are still below what was achieved in clean conditions, as well as what was achieved with data augmentation. The gains observed were typically more substantial at low SNR values, thus corroborating results by [83].

In an attempt to better understand the reason behind the poor ER performance with speech enhancement alone, Fig. 5.3 depicts an average modulation spectrogram, from top to bottom, for clean, noisy (airport at 0 dB SNR), MetricGAN+, and mimic-loss enhanced speech for angry (left) and sad (right) emotions, respectively. Modulation spectrograms are a frequency-frequency representation where the y-axis depicts acoustic frequency and the x-axis modulation frequency. From the clean plot, we can see the typical speech modulation spectral representation with most modulation energy lying below 16 Hz [47] and a slowing of the amplitude modulations with the sad emotion [46]. Noise, in turn, is shown to affect the modulation spectrogram by smearing the energy across higher acoustic and modulation frequencies, as suggested by [261]. The enhancement algorithms, however, are not capable of completely removing these environmental artifacts and seem to be introducing other types of distortions that can make the ER task more challenging. Combined, these factors result in the reduced gains reported in the Tables. This was in fact confirmed by listening to the outputs of the MetricGAN+ enhancement algorithm.

Finally, we test the combined effects of speech enhancement and data augmentation, as in the proposed system, to gauge the benefits of noise robustness applied at both the input and model levels, respectively. For Task 1, gains (relative to using each strategy individually) were seen only for the airport noise condition at higher SNR conditions (10 and 20 dB). In fact, with data augmentation alone, accuracy inline with what was achieved with clean speech was obtained. For Task 2, in turn, the proposed model showed improvements over the other methods for almost all tested conditions

Table 5.7: Performance comparison of the proposed method in different noisy test conditions for Task 1

Signal	F1	Precision	Recall	BA
Clean	0.74	0.75	0.75	0.73
Noisy - Airport (0dB)	0.57	0.58	0.57	0.55
Data augmentation only	**0.67**	**0.69**	**0.67**	**0.68**
Enhancement only	0.65	0.66	0.64	0.64
Proposed	0.65	0.65	0.66	0.63
Noisy - Airport (10dB)	0.59	0.64	0.62	0.57
Data augmentation only	0.69	**0.71**	0.69	**0.70**
Enhancement only	0.68	0.67	0.68	0.65
Proposed	**0.71**	**0.71**	**0.71**	0.69
Noisy - Airport (20dB)	0.60	0.65	0.65	0.58
Data augmentation only	0.69	0.70	0.69	0.68
Enhancement only	0.67	0.67	0.67	0.65
Proposed	**0.71**	**0.71**	**0.71**	**0.69**
Noisy - Babble (0dB)	0.59	0.61	0.59	0.57
Data augmentation only	**0.66**	**0.68**	**0.66**	**0.66**
Enhancement only	0.62	0.62	0.64	0.59
Proposed	0.64	0.63	0.64	0.61
Noisy - Babble (10dB)	0.60	0.63	0.61	0.58
Data augmentation only	**0.72**	**0.72**	**0.72**	**0.71**
Enhancement only	0.68	0.68	0.69	0.66
Proposed	0.70	0.70	0.71	0.68
Noisy - Babble (20dB)	0.61	0.63	0.61	0.58
Data augmentation only	**0.74**	**0.74**	**0.74**	**0.72**
Enhancement only	0.70	0.69	0.70	0.67
Proposed	0.70	0.70	0.70	0.69

in terms of F1 score, thus showing the importance of the proposed method to classify between opposing emotions in extremely noisy scenaerios; in the case here, joy versus sad. Notwithstanding, for Task 2 a gap of 23% remained between the best achieved performance and the clean speech accuracy. For comparison purposes, the state-of-the-art DialogueRNN system achieved an F1 score of 0.59 and 0.55 for Task 1 and Task 2, respectively, when corrupted with airport noise at 0 dB. The proposed system, in turn, was able to outperform this benchmark by 10 and 12%, respectively. Overall, the obtained results suggest that data augmentation combined with speech enhancement can be a viable alternative for robust "in-the-wild" automatic multimodal emotion recognition while requiring access to only one signal modality: audio.

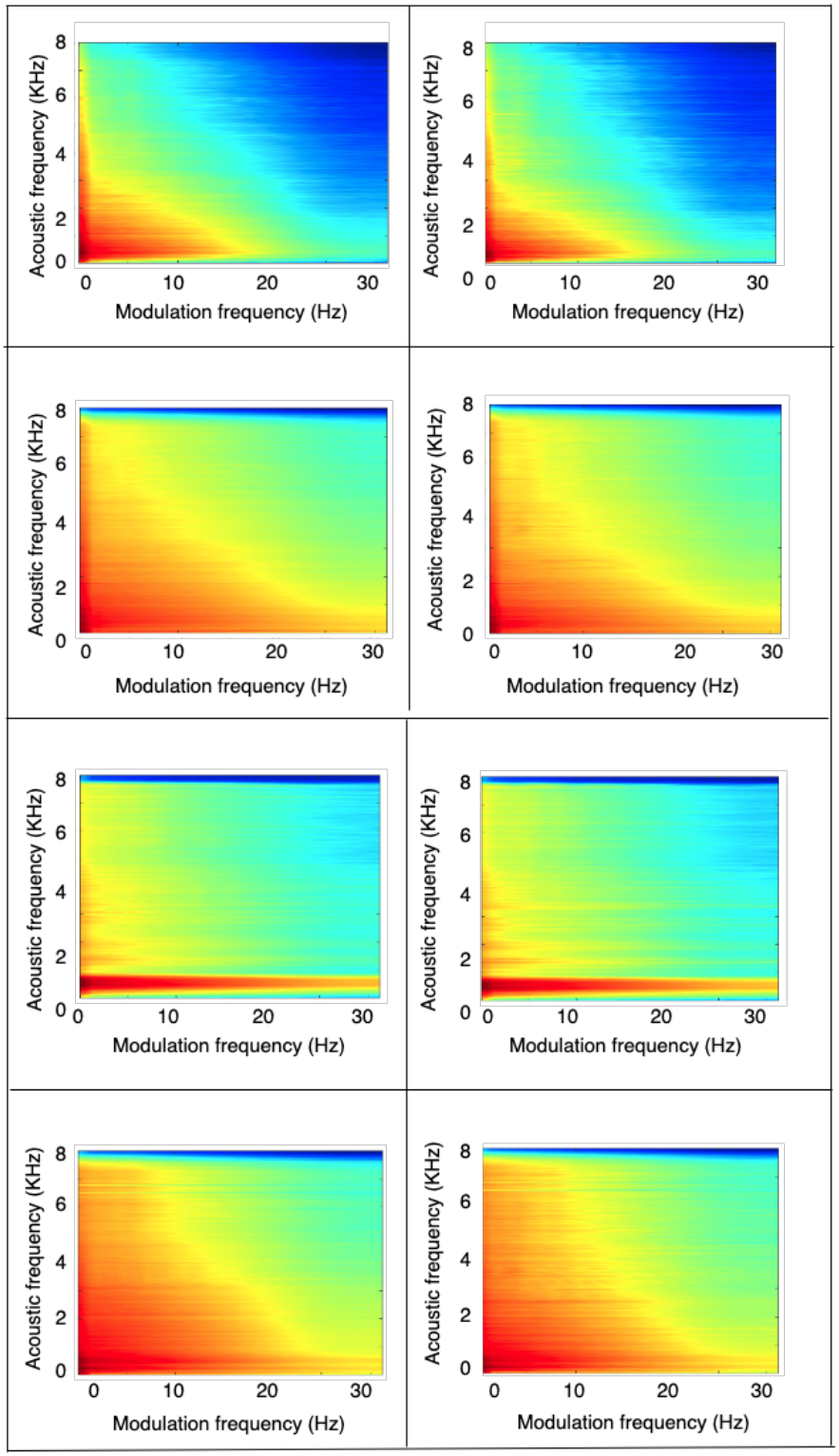

Figure 5.3: Modulation spectrogram for different conditions, from top to bottom: clean, (airport) noisy at 0 dB , MetriGAN+, and mimic-loss enhanced speech. Left plots correspond to angry and right plots to sad emotion.

Table 5.8: Performance comparison of the proposed method in different noisy test conditions for Task 2

Signal	F1	Precision	Recall	BA
Clean	0.87	0.87	0.87	0.87
Noisy - Airport (0 dB)	0.50	0.48	0.53	0.50
Data augmentation only	0.61	**0.64**	0.60	**0.61**
Enhancement only	0.56	0.56	0.57	0.51
Proposed	**0.62**	0.63	**0.62**	0.59
airport (10 dB)	0.55	0.58	0.53	0.51
Data augmentation only	**0.66**	**0.69**	0.65	**0.66**
Enhancement only	0.61	0.63	**0.66**	0.55
Proposed	0.65	0.65	0.64	0.62
Noisy - Airport (20 dB)	0.60	0.61	0.59	0.55
Data augmentation only	**0.67**	**0.69**	0.66	**0.66**
Enhancement only	0.62	0.63	0.65	0.56
Proposed	**0.67**	0.68	**0.67**	0.65
Noisy - Babble (0 dB)	0.54	0.54	0.54	0.51
Data augmentation only	0.58	0.61	0.57	**0.59**
Enhancement only	0.57	0.56	**0.61**	0.51
Proposed	**0.61**	**0.62**	0.60	0.58
Noisy - Babble (10 dB)	0.58	0.59	0.58	0.54
Data augmentation only	0.63	0.65	0.62	0.62
Enhancement only	0.61	0.63	**0.66**	0.55
Proposed	**0.66**	**0.67**	**0.66**	**0.64**
Noisy - Babble (20 dB)	0.61	0.63	0.61	0.56
Data augmentation only	**0.67**	**0.69**	0.67	**0.67**
Enhancement only	0.66	0.67	**0.69**	0.60
Proposed	**0.67**	0.68	0.67	0.64

5.5.5 Generalizability of proposed method

To test the generalizability of the proposed method, three additional experiments have been conducted. First, we retrain the proposed ER model using the IEMOCAP training dataset partition and test it on the IEMOCAP test set to obtain an upper bound on what can be achieved on this particular dataset. Next, to gauge the advantages brought by the proposed system, we retrain the ER system shown in Fig. 5.1 but without the enhancement and data augmentation steps. Training was done on the MELD dataset and the model was then tested on the unseen IEMOCAP test data. This gives us an idea of how challenging the cross-corpus task is when the proposed innovations are not present and should give us a lower bound on what could be achieved cross-corpus. Finally, we tested the full proposed method trained on the MELD dataset and tested on the unseen IEMOCAP test data. Experimental results are reported in Table 5.9. As can be seen, cross-corpus testing is

Table 5.9: Cross-corpus performance for Tasks 1 and 2.

Task	Experiment	F1	Precision	Recall	BA
Task 1	1	0.94	0.94	0.94	0.94
	2	0.49	0.57	0.53	0.55
	3	0.64	0.79	0.67	0.69
Task 2	1	0.85	0.86	0.85	0.85
	2	0.50	0.68	0.52	0.60
	3	0.72	0.72	0.72	0.70

an extremely challenging task where performance accuracy can drop to chance levels if strategies are not put in place. The proposed innovations, on the other hand, provides some robustness, and gains of 30% and 44% could be seen with the proposed system for Tasks 1 and 2, respectively, over a system without task-specific speech enhancement and data augmentation. The gaps to the upper bound obtained with Experiment 1 suggest that there is still room for improvement and emotion-aware enhancement and/or alternate data augmentation strategies may still be needed.

5.6 Conclusions

This chapter has explored the use of task-specific speech enhancement combined with data augmentation to provide robustness to unseen test conditions for multimodal emotion recognition systems. Experiments conducted on the MELD dataset show the importance of BERT for text feature extraction and a fused eGEMAPS-modulation spectral set for audio features. The importance of data augmentation at the training stage and of task-specific speech enhancement at the testing stage are shown on two binary speech emotion classification tasks. Lastly, cross-corpus experiments showed the proposed innovations resulting in 40% gains relative to an ER system without enhancement/augmentation. While the obtained results suggest that task-specific enhancement, combined with data augmentation are important steps towards reliable "in the wild" emotion recognition, speech enhancement algorithms may still be suboptimal and may be removing important emotion information. As such, future work should explore the development of emotion-aware enhancement algorithms that can trade-off noise suppression and emotion recognition accuracy.

Chapter 6

Conclusions and Future Research Directions

This thesis aimed at improving the performance of affect recognition system in real-time settings. The main contributions of this doctoral research are summarized here, followed by areas of possible future research directions.

6.1 Summary of Contributions

This thesis contributes to "in the wild" speech emotion recognition via three main innovations. First, in Chapter 3, we propose to combine the bag-of-audio-words methodology with modulation spectrum features for environmental robustness. We then take advantage of the inherent quality-awareness properties of modulation spectrum for the SER task. Experiments are conducted with three multi-lingual speech datasets degraded by different noise sources and levels, as well as room reverberation. Experimental results show the proposed features i) consistently outperforming benchmark systems, ii) providing complementary information to classical features, hence improving performance with feature fusion, and iii) showing robustness against environment and language mismatch. Moreover, we show that when the proposed system is provided with quality information, further improvements are obtained, both in terms of arousal/valence level predictions and in dis-

crete emotion classification. Finally, we showed the impact that data augmentation had on language mismatch robustness, thus highlighting the potential of the proposed system for "in-the-wild" SER.

Next, in Chapter 4, we propose to combine the bag-of-words (BOW) methodology with domain adaptation for feature distribution "normalization," and data augmentation to make machine learning algorithms more robust across testing conditions. In this chapter, we are particularly interested in dealing with the issue of cross-language mismatch. We propose a new domain generalization method which we term N-CORAL, in which test languages are mapped to a common distribution in an unsupervised manner; in our case, to the Chinese language. Experiments with German, French, and Hungarian language emotion datasets showed that the proposed N-CORAL method, combined with BOW and data augmentation, achieved the best arousal and valence prediction accuracy of the tested systems, thus highlighting the usefulness of the proposed method for "in the wild" speech emotion recognition.

Lastly, in Chapter 5, we showed that task-specific speech enhancement combined with data augmentation resulted in reliable multimodal emotion recognition accuracy in noisy unseen conditions. Results in noisy conditions approximated those seen with clean speech. Cross-dataset experiments showed the proposed innovations resulting in 40% gains relative to a SER system without speech enhancement or data augmentation.

6.2 Future Work

In this work, we have assumed "in the wild" conditions to be those where users are outdoors in noisy settings or indoors in reverberant rooms, thus additive noise and reverberation have been utilized as the main sources of distortions in the experiments conducted herein. Notwithstanding, more and more emotion recognition systems are being deployed in telecommunication networks (e.g., in call centers). As such, future work should explore the impact of different network impairments on SER accuracy (e.g., packet losses, wireless communication losses). Attention or saliency map could be use for further understanding of important noise-robust frequency region in the modulation spectrum in order to build the noise-robust SER system.

In Chapter 4, the proposed N-CORAL system was evaluated using Chinese as the target language. Future work should explore if other languages are used, especially if target language family

coincides with the source language family. Other experiments could also explore the addition of multiple languages as target (within or across different language families).

Lastly, while the obtained results in Chapter 5 showed that speech enhancement combined with data augmentation could help with "in the wild" emotion recognition, an in-depth analysis showed that the enhancement algorithms could be removing important emotional cues from the speech signal. Future work should explore the development of emotion-aware speech enhancement algorithms where a trade-off between noise suppression and emotion recognition accuracy can be made.

Bibliography

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[2] M. A. Nielsen, *Neural networks and deep learning*. Determination press San Francisco, CA, USA, 2015, vol. 25.

[3] Z. Zeng, J. Tu, M. Liu, T. S. Huang, B. Pianfetti, D. Roth, and S. Levinson, "Audio-visual affect recognition," *IEEE Transactions on multimedia*, vol. 9, no. 2, pp. 424–428, 2007.

[4] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[5] J. Xu and B. Zhong, "Review on portable EEG technology in educational research," *Computers in Human Behavior*, vol. 81, pp. 340–349, 2018.

[6] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," *Neuroscience & Biobehavioral Reviews*, vol. 44, pp. 58–75, 2014.

[7] N. Thakur and C. Y. Han, "Framework for an intelligent affect aware smart home environment for elderly people," *Int. J. Recent Trends Hum. Comput. Interact.(IJHCI)*, vol. 9, no. 1, pp. 23–43, 2019.

[8] B. Kerous, F. Skola, and F. Liarokapis, "EEG-based BCI and video games: a progress report," *Virtual Reality*, vol. 22, no. 2, pp. 119–135, 2018.

[9] K.-J. Oh, D. Lee, B. Ko, and H.-J. Choi, "A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation," in *2017 18th IEEE international conference on mobile data management (MDM)*. IEEE, 2017, pp. 371–375.

[10] M. Yadava, P. Kumar, R. Saini, P. P. Roy, and D. P. Dogra, "Analysis of EEG signals and its application to neuromarketing," *Multimedia Tools and Applications*, vol. 76, no. 18, pp. 19 087–19 111, 2017.

[11] D. Egan, S. Brennan, J. Barrett, Y. Qiao, C. Timmerer, and N. Murray, "An evaluation of heart rate and electrodermal activity as an objective qoe evaluation method for immersive virtual reality environments," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2016, pp. 1–6.

[12] F. Eyben, G. L. Salomão, J. Sundberg, K. R. Scherer, and B. W. Schuller, "Emotion in the singing voice—a deeperlook at acoustic features in the light ofautomatic classification," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–9, 2015.

[13] A. B. Naumann, I. Wechsung, and J. Hurtienne, "Multimodal interaction: Intuitive, robust, and preferred?" in *IFIP Conference on Human-Computer Interaction*. Springer, 2009, pp. 93–96.

[14] M. Parent, A. Tiwari, I. Albuquerque, J.-F. Gagnon, D. Lafond, S. Tremblay, and T. H. Falk, "A multimodal approach to improve the robustness of physiological stress prediction during physical activity," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 2019, pp. 4131–4136.

[15] P. Ekman, W. V. Friesen, M. O'sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti *et al.*, "Universals and cultural differences in the judgments of facial expressions of emotion." *Journal of personality and social psychology*, vol. 53, no. 4, p. 712, 1987.

[16] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[17] C. M. Whissell, "The dictionary of affect in language," in *The measurement of emotions*. Elsevier, 1989, pp. 113–131.

[18] H. Schlosberg, "Three dimensions of emotion." *Psychological review*, vol. 61, no. 2, p. 81, 1954.

[19] A. Jaimes and N. Sebe, "Multimodal human–computer interaction: A survey," *Computer vision and image understanding*, vol. 108, no. 1, pp. 116–134, 2007.

[20] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in personality*, vol. 11, no. 3, pp. 273–294, 1977.

[21] J. A. Russell, M. Lewicka, and T. Niit, "A cross-cultural study of a circumplex model of affect." *Journal of personality and social psychology*, vol. 57, no. 5, p. 848, 1989.

[22] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in psychology*. Elsevier, 1988, vol. 52, pp. 139–183.

[23] A. Luximon and R. S. Goonetilleke, "Simplified subjective workload assessment technique," *Ergonomics*, vol. 44, no. 3, pp. 229–243, 2001.

[24] T. M. Marteau and H. Bekker, "The development of a six-item short-form of the state scale of the spielberger state—trait anxiety inventory (stai)," *British journal of clinical psychology*, vol. 31, no. 3, pp. 301–306, 1992.

[25] F.-X. Lesage, S. Berjot, and F. Deschamps, "Clinical stress assessment using a visual analogue scale," *Occupational medicine*, vol. 62, no. 8, pp. 600–605, 2012.

[26] E. Facco, G. Zanette, L. Favero, C. Bacci, S. Sivolella, F. Cavallin, and G. Manani, "Toward the validation of visual analogue scale for anxiety," *Anesthesia progress*, vol. 58, no. 1, pp. 8–13, 2011.

[27] C. Niemic, "Studies of emotion: A theoretical and empirical review of psychophysiological studies of emotion." *Journal of Undergraduate Research*.

[28] S. M. Alarcao and M. J. Fonseca, "Emotions recognition using EEG signals: A survey," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 374–393, 2017.

[29] R. Castaldo, P. Melillo, U. Bracale, M. Caserta, M. Triassi, and L. Pecchia, "Acute mental stress assessment via short term HRV analysis in healthy adults: A systematic review with meta-analysis," *Biomedical Signal Processing and Control*, vol. 18, pp. 370–377, 2015.

[30] 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," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, p. 61, 2015.

[31] S. Braun, B. Schwartz, S. Gannot, and E. A. Habets, "Late reverberation psd estimation for single-channel dereverberation using relative convolutive transfer functions," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2016, pp. 1–5.

[32] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[33] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.

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

[35] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, 2013, pp. 436–440.

[36] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "Metricgan+: An improved version of metricgan for speech enhancement," *arXiv preprint arXiv:2104.03538*, 2021.

[37] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[38] D. Bagchi, P. Plantinga, A. Stiff, and E. Fosler-Lussier, "Spectral feature mapping with mimic loss for robust speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5609–5613.

[39] J. G. 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 i—temporal alignment," *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 366–384, 2013.

[40] J. F. Santos, M. Senoussaoui, and T. H. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *2014 14th International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2014, pp. 55–59.

[41] 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*. ACM, 2013, pp. 835–838.

[42] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The interspeech 2011 speaker state challenge," 2011.

[43] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud *et al.*, "Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*. ACM, 2018, pp. 3–13.

[44] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, and E. Amiriparian, S.and Messner, "AVEC 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proc. of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019.

[45] A. Avila, Z. Akhtar, J. Santos, D. O'Shaughnessy, and T. Falk, "Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 177–188, 2021.

[46] S. Wu, T. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, no. 5, pp. 768–785, 2011.

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

[48] T. Falk and W.-Y. Chan, "Temporal dynamics for blind measurement of room acoustical parameters," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 4, pp. 978–989, 2010.

[49] B. Glasberg and B. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, vol. 47, 1990.

[50] I. Tenney, P. Xia, B. Chen, A. Wang, A. Poliak, R. T. McCoy, N. Kim, B. Van Durme, S. R. Bowman, D. Das *et al.*, "What do you learn from context? probing for sentence structure in contextualized word representations," *arXiv preprint arXiv:1905.06316*, 2019.

[51] Y. Zhang, Q. Wang, Y. Li, and X. Wu, "Sentiment classification based on piecewise pooling convolutional neural network," *Comput. Mater. Continua*, vol. 56, no. 2, pp. 285–297, 2018.

[52] W. P. Alston, "Philosophy of language," *Philosophy of Language Foundation of philosophy series Prentice-Hall foundations of philosophy series*, 1964.

[53] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Subspace alignment for domain adaptation," *arXiv preprint arXiv:1409.5241*, 2014.

[54] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[55] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[56] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*. ACM, 2016, pp. 3–10.

[57] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 3–9.

[58] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, B. Schuller, K. Star *et al.*, "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild," *arXiv preprint arXiv:1901.02839*, 2019.

[59] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[60] S.-Y. Chen, C.-C. Hsu, C.-C. Kuo, L.-W. Ku *et al.*, "Emotionlines: An emotion corpus of multi-party conversations," *arXiv preprint arXiv:1802.08379*, 2018.

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

[62] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proc. of Meetings on Acoustics ICA2013*, vol. 19, no. 1. ASA, 2013.

[63] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *International Conference on Digital Signal Processing*. IEEE, 2009.

[64] S. R. Kshirsagar and T. H. Falk, "Quality-aware bag of modulation spectrum features for robust speech emotion recognition," *IEEE Transactions on Affective Computing*, pp. 1–14, 2022.

[65] M. Schmitt, F. Ringeval, and B. Schuller, "At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech." in *Interspeech*, 2016, pp. 495–499.

[66] T. Falk, C. Zheng, and W. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[67] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, "From individual to group-level emotion recognition: Emotiw 5.0," in *Proc. of the 19th ACM international conference on multimodal interaction*, 2017.

[68] J. Santos, M. Senoussaoui, and T. Falk, "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *Acoustic Signal Enhancement (IWAENC), 2014 14th International Workshop on*, Sep. 2014, pp. 55–59.

[69] B. Sun, J. Feng, and K. Saenko, "Correlation alignment for unsupervised domain adaptation," in *Domain Adaptation in Computer Vision Applications*. Springer, 2017, pp. 153–171.

[70] S. M. Feraru, D. Schuller *et al.*, "Cross-language acoustic emotion recognition: An overview and some tendencies," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 125–131.

[71] F. Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl, "Cross-corpus classification of realistic emotions–some pilot experiments," in *Proc. 7th Intern. Conf. on Language Resources and Evaluation (LREC 2010), Valletta, Malta*, 2010.

[72] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[73] R. Banse and K. Scherer, "Acoustic profiles in vocal emotion expression." *Journal of personality and social psychology*, vol. 70, no. 3, 1996.

[74] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[75] W. Xue, C. Cucchiarini, R. van Hout, and H. Strik, "Acoustic correlates of speech intelligibility. the usability of the egemaps feature set for atypical speech," 2019.

[76] 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*. ACM, 2016, pp. 3–10.

[77] K. Yang, D. Lee, T. Whang, S. Lee, and H. Lim, "Emotionx-ku: BERT-max based contextual emotion classifier," *arXiv preprint arXiv:1906.11565*, 2019.

[78] L. Stappen, A. Baird, L. Christ, L. Schumann, B. Sertolli, E.-M. Messner, E. Cambria, G. Zhao, and B. W. Schuller, "The muse 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress," *arXiv preprint arXiv:2104.07123*, 2021.

[79] Y. Yang and X. Cui, "BERT-enhanced text graph neural network for classification," *Entropy*, vol. 23, no. 11, p. 1536, 2021.

[80] F. Eyben, F. Weninger, and B. Schuller, "Affect recognition in real-life acoustic conditions-a new perspective on feature selection," in *Proceedings 14th INTERSPEECH, Lyon, France*, 2013.

[81] V. A. Trinh, H. S. Kavaki, and M. I. Mandel, "Importantaug: a data augmentation agent for speech," *arXiv preprint arXiv:2112.07156*, 2021.

[82] M. Neumann and N. T. Vu, "Investigations on audiovisual emotion recognition in noisy conditions," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 358–364.

[83] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, and B. W. Schuller, "Towards robust speech emotion recognition using deep residual networks for speech enhancement." in *Interspeech*, 2019, pp. 1691–1695.

[84] J. Xiong, O. Lipsitz, F. Nasri, L. M. Lui, H. Gill, L. Phan, D. Chen-Li, M. Iacobucci, R. Ho, A. Majeed *et al.*, "Impact of covid-19 pandemic on mental health in the general population: A systematic review," *Journal of affective disorders*, 2020.

[85] K. Laghari, R. Gupta, S. Arndt, J. Antons, R. Schleicher, S. Moller, and T. Falk, "Neurophysiological experimental facility for quality of experience (qoe) assessment," in *Proceedings of International Conference on Quality of Experience Centric Management (QCMan)*, 2013, pp. 1300–1305.

[86] M. Gurban and J.-P. Thiran, "Basic concepts of multimodal analysis," *Multimodal Signal Processing: Theory and Applications for Human-Computer Interaction*, p. 145, 2009.

[87] A. M. Oliveira, M. P. Teixeira, I. B. Fonseca, and M. Oliveira, "Joint model-parameter validation of self-estimates of valence and arousal: Probing a differential-weighting model of affective intensity." *Proceedings of Fechner Day*, vol. 22, no. 1, pp. 245–250, 2006.

[88] N. Alvarado, "Arousal and valence in the direct scaling of emotional response to film clips," *Motivation and Emotion*, vol. 21, no. 4, pp. 323–348, 1997.

[89] R. D. Lane and L. Nadel, *Cognitive neuroscience of emotion*. Oxford University Press, 1999.

[90] P. A. Lewis, H. Critchley, P. Rotshtein, and R. J. Dolan, "Neural correlates of processing valence and arousal in affective words," *Cerebral cortex*, vol. 17, no. 3, pp. 742–748, 2006.

[91] D. A. Ritossa and N. S. Rickard, "The relative utility of 'pleasantness' and 'liking'dimensions in predicting the emotions expressed by music," *Psychology of Music*, vol. 32, no. 1, pp. 5–22, 2004.

[92] R. L. Cardy and G. H. Dobbins, "Affect and appraisal accuracy: Liking as an integral dimension in evaluating performance." *Journal of Applied Psychology*, vol. 71, no. 4, p. 672, 1986.

[93] A. C. North and D. J. Hargreaves, "Liking, arousal potential, and the emotions expressed by music," *Scandinavian journal of psychology*, vol. 38, no. 1, pp. 45–53, 1997.

[94] R. Dietz and A. Lang, "Affective agents: Effects of agent affect on arousal, attention, liking and learning," in *Proceedings of the Third International Cognitive Technology Conference, San Francisco*, 1999, pp. 1–35.

[95] C. Yu, P. M. Aoki, and A. Woodruff, "Detecting user engagement in everyday conversations," *arXiv preprint cs/0410027*, 2004.

[96] Y.-H. Yang, Y.-F. Su, Y.-C. Lin, and H. H. Chen, "Music emotion recognition: The role of individuality," in *Proceedings of the international workshop on Human-centered multimedia*. ACM, 2007, pp. 13–22.

[97] K. R. Scherer, "Psychological models of emotion," *The neuropsychology of emotion*, vol. 137, no. 3, pp. 137–162, 2000.

[98] D. Grandjean, D. Sander, and K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness and cognition*, vol. 17, no. 2, pp. 484–495, 2008.

[99] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[100] S. S. Sabet, C. Griwodz, and S. Möller, "Influence of primacy, recency and peak effects on the game experience questionnaire," in *Proceedings of the 11th ACM Workshop on Immersive Mixed and Virtual Environment Systems*, 2019, pp. 22–27.

[101] U. W. Müller, C. L. Witteman, J. Spijker, and G. W. Alpers, "All's bad that ends bad: there is a peak-end memory bias in anxiety," *Frontiers in psychology*, vol. 10, p. 1272, 2019.

[102] G. Eisele, H. Vachon, G. Lafit, P. Kuppens, M. Houben, I. Myin-Germeys, and W. Viechtbauer, "The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population," *Assessment*, p. 1073191120957102, 2020.

[103] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE international conference on multimedia and Expo*. IEEE, 2005, pp. 5–pp.

[104] H. Gunes and M. Piccardi, "A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior," in *18th International conference on pattern recognition (ICPR'06)*, vol. 1. IEEE, 2006, pp. 1148–1153.

[105] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool, "A 3-d audio-visual corpus of affective communication," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 591–598, 2010.

[106] B. W. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. B. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny *et al.*, "The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats." in *Interspeech*, 2018, pp. 122–126.

[107] M. Neumann and N. T. Vu, "Cross-lingual and multilingual speech emotion recognition on english and french," *arXiv preprint arXiv:1803.00357*, 2018.

[108] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Cross corpus speech emotion classification-an effective transfer learning technique," *arXiv preprint arXiv:1801.06353*, 2018.

[109] S. Kshirsagar and T. H. Falk, "Cross-language speech emotion recognition using bag-of-word representations, domain adaptation, and data augmentation," *Sensors*, vol. 22, no. 17, p. 6445, 2022.

[110] P. A. Kshirsagar, Shruti and T. H. Falk, "Task-Specific Speech Enhancement and Data Augmentation for Improved Multimodal Emotion Recognition Under Noisy Conditions," *Frontiers in computer Science*, 2022.

[111] A. Tiwari, R. Cassani, S. Kshirsagar, D. P. Tobon, Y. Zhu, and T. H. Falk, "Modulation spectral signal representation for quality measurement and enhancement of wearable device data: A technical note," *Sensors*, vol. 22, no. 12, p. 4579, 2022.

[112] A. R. Avila, S. R. Kshirsagar, A. Tiwari, D. Lafond, D. O'Shaughnessy, and T. H. Falk, "Speech-based stress classification based on modulation spectral features and convolutional neural networks," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[113] A. Gaballah, A. Avila, J. Monteiro, P. Tiwari, S. Kshirsagar, and T. H. Falk, "Development of the inrs-emt scene classification systems for the 2020 edition of the dcase challenge (tasks 1a and 1b)."

[114] P. Tiwari, Y. Jain, A. Avila, J. Monteiro, S. Kshirsagar, A. Gaballah, and T. H. Falk, "Modulation spectral signal representation and i-vectors for anomalous sound detection."

[115] S. Tamura and A. Waibel, "Noise reduction using connectionist models," in *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1988, pp. 553–556.

[116] S. Parveen and P. Green, "Speech enhancement with missing data techniques using recurrent neural networks," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 2004, pp. I–733.

[117] A. Maas, Q. V. Le, T. M. O'neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust asr," 2012.

[118] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust asr," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

[119] H. Zhao, S. Zarar, I. Tashev, and C.-H. Lee, "Convolutional-recurrent neural networks for speech enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2401–2405.

[120] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2031–2041.

[121] L. Malfait, J. Berger, and M. Kastner, "P. 563—the itu-t standard for single-ended speech quality assessment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1924–1934, 2006.

[122] B. Logan *et al.*, "Mel frequency cepstral coefficients for music modeling." in *ISMIR*, vol. 270, 2000, pp. 1–11.

[123] F. Weninger, P. Staudt, and B. Schuller, "Words that fascinate the listener: Predicting affective ratings of on-line lectures," *International Journal of Distance Education Technologies (IJDET)*, vol. 11, no. 2, pp. 110–123, 2013.

[124] M. Riley, E. Heinen, and J. Ghosh, "A text retrieval approach to content-based audio retrieval," in *Int. Symp. on Music Information Retrieval (ISMIR)*, 2008, pp. 295–300.

[125] M. Schmitt and B. Schuller, "Openxbow: introducing the passau open-source crossmodal bag-of-words toolkit," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 3370–3374, 2017.

[126] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proc ACM international conference on Multimedia*, 2013, pp. 835–838.

[127] N. F. Liu, M. Gardner, Y. Belinkov, M. E. Peters, and N. A. Smith, "Linguistic knowledge and transferability of contextual representations," *arXiv preprint arXiv:1903.08855*, 2019.

[128] Z. Wu, Y. Chen, B. Kao, and Q. Liu, "Perturbed masking: Parameter-free probing for analyzing and interpreting BERT," *arXiv preprint arXiv:2004.14786*, 2020.

[129] J. D. Ortega, M. Senoussaoui, E. Granger, M. Pedersoli, P. Cardinal, and A. L. Koerich, "Multimodal fusion with deep neural networks for audio-video emotion recognition," *arXiv preprint arXiv:1907.03196*, 2019.

[130] Y. Kim, H. Lee, and E. M. Provost, "Deep learning for robust feature generation in audiovisual emotion recognition," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 3687–3691.

[131] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.

[132] K. Somandepalli, N. Kumar, R. Travadi, and S. Narayanan, "Multimodal representation learning using deep multiset canonical correlation," *arXiv preprint arXiv:1904.01775*, 2019.

[133] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters*, vol. 66, pp. 22–30, 2015.

[134] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 42–55, 2011.

[135] J.-P. Thiran, F. Marques, and H. Bourlard, *Multimodal Signal Processing: Theory and applications for human-computer interaction*. Academic Press, 2009.

[136] N. Sebe, I. Cohen, T. S. Huang *et al.*, "Multimodal emotion recognition," *Handbook of Pattern Recognition and Computer Vision*, vol. 4, pp. 387–419, 2005.

[137] S. K. et al, "Emonets: Multimodal deep learning approaches for emotion recognition in video," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[138] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern recognition*, vol. 38, no. 12, pp. 2270–2285, 2005.

[139] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, 2011.

[140] S. Oviatt, "Advances in robust multimodal interface design," *IEEE computer graphics and applications*, vol. 23, no. 5, pp. 62–68, 2003.

[141] B. Dumas, D. Lalanne, and S. Oviatt, "Multimodal interfaces: A survey of principles, models and frameworks," in *Human machine interaction*. Springer, 2009, pp. 3–26.

[142] W.-L. Zheng, B.-N. Dong, and B.-L. Lu, "Multimodal emotion recognition using EEG and eye tracking data," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2014, pp. 5040–5043.

[143] R. Gupta, M. Khomami Abadi, J. A. Cárdenes Cabré, F. Morreale, T. H. Falk, and N. Sebe, "A quality adaptive multimodal affect recognition system for user-centric multimedia indexing," in *Proceedings of the 2016 ACM on international conference on multimedia retrieval*, 2016, pp. 317–320.

[144] J. Chang and S. Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2746–2750.

[145] J. Deng, R. Xia, Z. Zhang, Y. Liu, and B. Schuller, "Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4818–4822.

[146] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, 2018.

[147] S. Sankaranarayanan, Y. Balaji, C. D. Castillo, and R. Chellappa, "Generate to adapt: Aligning domains using generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8503–8512.

[148] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[149] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8789–8797.

[150] B. Sun, J. Feng, and K. Saenko, "Correlation alignment for unsupervised domain adaptation," in *Domain Adaptation in Computer Vision Applications*. Springer, 2017, pp. 153–171.

[151] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

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

[153] J. Hochreiter and J. Schmidhuber, "Long short-term memory network," *Neural computation*, 1997.

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

[155] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, and A. Coates, "Deep speech: Scaling up end-to-end speech recognition," *arXiv:1412.5567*, 2014.

[156] Q. Wang, C. Downey, L. Wan, P. Mansfield, and I. Moreno, "Speaker diarization with LSTM," in *ICASSP*, 2018.

[157] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.

[158] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, B. Schuller, K. Star *et al.*, "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild," *arXiv preprint arXiv:1901.02839*, 2019.

[159] R. Frick, "Communicating emotion: The role of prosodic features," *Psychological Bulletin*, vol. 97, no. 3, pp. 412–429, 1985.

[160] C. Williams and K. Stevens, "Vocal correlates of emotional states," *Speech evaluation in psychiatry*, pp. 221–240, 1981.

[161] E. F., M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[162] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The interspeech 2011 speaker state challenge," *Proc. INTERSPEECH 2011, Florence, Italy*, 2011.

[163] B. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, and L. Stappen, "The interspeech 2021 computational paralinguistics challenge: Covid-19 cough, covid-19 speech, escalation & primates," *arXiv:2102.13468*, 2021.

[164] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3852–3856.

[165] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 3030–3043, 2017.

[166] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech emotion recognition using self-supervised features," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6922–6926.

[167] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[168] W.-C. Lin and C. Busso, "An efficient temporal modeling approach for speech emotion recognition by mapping varied duration sentences into fixed number of chunks," *Proc. Interspeech 2020*, pp. 2322–2326, 2020.

[169] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *arXiv preprint arXiv:2203.07378*, 2022.

[170] J. Wagner, D. Schiller, and E. André, "Deep learning in paralinguistic recognition tasks: Are hand-crafted features still relevant?" *Proc. Interspeech 2018*, 2018.

[171] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, "CNN+LSTM architecture for speech emotion recognition with data augmentation," *arXiv:1802.05630*, 2018.

[172] C. Kim, A. Misra, K. Chin, T. Hughes, A. Narayanan, T. Sainath, and M. Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home," *Proc. interspeech 2017 (2017), pp. 379-383 year=2017*.

[173] A. Shilandari, H. Marvi, H. Khosravi, and W. Wang, "Speech emotion recognition using data augmentation method by cycle-generative adversarial networks," *Signal, image and video processing*, pp. 1–8, 2022.

[174] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, and T. S. Huang, "Multimodal audio, video and physiological sensor learning for continuous emotion prediction," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 97–104.

[175] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 19–26.

[176] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems*, 2016, pp. 892–900.

[177] J. Zhao, R. Li, S. Chen, and Q. Jin, "Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions," in *Proc. of the 2018 on Audio/Visual Emotion Challenge and Workshop*.

[178] S. Hershey, S. Chaudhuri, D. Ellis, J. Gemmeke, A. Jansen, R. Moore, M. Plakal, D. Platt, R. Saurous, and B. Seybold, "CNN architectures for large-scale audio classification," in *ICASSP*, 2017.

[179] J. Zhao, R. Li, J. Liang, S. Chen, and Q. Jin, "Adversarial domain adaption for multi-cultural dimensional emotion recognition in dyadic interactions," in *Proc. of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019.

[180] J. Huang, Y. Li, J. Tao, Z. Lian, and J. Yi, "End-to-end continuous emotion recognition from video using 3d convlstm networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6837–6841.

[181] M. Senoussaoui, P. Cardinal, and A. L. Koerich, "Bag-of-audio-words based on autoencoder codebook for continuous emotion prediction," *arXiv preprint arXiv:1907.04928*, 2019.

[182] B. Knyazev, R. Shvetsov, N. Efremova, and A. Kuharenko, "Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video," *arXiv:1711.04598*, 2017.

[183] T. Dang, B. Stasak, Z. Huang, S. Jayawardena, M. Atcheson, M. Hayat, P. Le, V. Sethu, R. Goecke, and J. Epps, "Investigating word affect features and fusion of probabilistic predictions incorporating uncertainty in avec 2017," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, 2017, pp. 27–35.

[184] A. Ouyang, T. Dang, V. Sethu, and E. Ambikairajah, "Speech based emotion prediction: Can a linear model work?" in *INTERSPEECH*, 2019, pp. 2813–2817.

[185] M. Macary, M. Tahon, Y. Estève, and A. Rousseau, "On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 373–380.

[186] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 890–897.

[187] J. Huang, Y. Li, J. Tao, Z. Lian, Z. Wen, M. Yang, and J. Yi, "Continuous multimodal emotion prediction based on long short term memory recurrent neural network," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 11–18.

[188] R. A. Patamia, W. Jin, K. N. Acheampong, K. Sarpong, and E. K. Tenagyei, "Transformer based multimodal speech emotion recognition with improved neural networks," in *2021 IEEE 2nd International Conference on Pattern Recognition and Machine Learning (PRML)*. IEEE, 2021, pp. 195–203.

[189] R. Hämäläinen, B. De Wever, T. Waaramaa, A.-M. Laukkanen, and J. Lämsä, "It's not only what you say, but how you say it: Investigating the potential of prosodic analysis as a method to study teacher's talk," *Frontline Learning Research*, vol. 6, no. 3, 2018.

[190] M. Senoussaoui, J. Santos, and T. Falk, "Speech temporal dynamics fusion approaches for noise-robust reverberation time estimation," in *ICASSP*. IEEE, 2017.

[191] K. Środecki, "Evaluation of the reverberation decay quality in rooms using the autocorrelation function and the cepstrum analysis," *Acta Acustica United with Acustica*, vol. 80, no. 3, pp. 216–225, 1994.

[192] V. Hozjan and Z. Kačič, "Context-independent multilingual emotion recognition from speech signals," *International journal of speech technology*, vol. 6, no. 3, pp. 311–320, 2003.

[193] I. Lefter, L. J. Rothkrantz, P. Wiggers, and D. A. Van Leeuwen, "Emotion recognition from speech by combining databases and fusion of classifiers," in *International Conference on Text, Speech and Dialogue*. Springer, 2010, pp. 353–360.

[194] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[195] M. Neumann and N. T. Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *arXiv preprint arXiv:1706.00612*, 2017.

[196] H. Sagha, J. Deng, M. Gavryukova, J. Han, and B. Schuller, "Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5800–5804.

[197] B.-C. Chiou and C.-P. Chen, "Speech emotion recognition with cross-lingual databases," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[198] H. Sagha, P. Matejka, M. Gavryukova, F. Povolnỳ, E. Marchi, and B. W. Schuller, "Enhancing multilingual recognition of emotion in speech by language identification." in *Interspeech*, 2016, pp. 2949–2953.

[199] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: compensating for covariate shift," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1458–1468, 2013.

[200] H. Sagha, J. Deng, M. Gavryukova, J. Han, and B. Schuller, "Cross lingual speech emotion recognition using canonical correlation analysis on principal component subspace," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 5800–5804.

[201] P. Song, W. Zheng, S. Ou, X. Zhang, Y. Jin, J. Liu, and Y. Yu, "Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization," *Speech Communication*, vol. 83, pp. 34–41, 2016.

[202] D. Wang and T. F. Zheng, "Transfer learning for speech and language processing," *arXiv preprint arXiv:1511.06066*, 2015.

[203] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, p. 9, 2016.

[204] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.

[205] I. Redko, E. Morvant, A. Habrard, M. Sebban, and Y. Bennani, *Advances in domain adaptation theory*. Elsevier, 2019.

[206] N. Cummins, S. Amiriparian, S. Ottl, M. Gerczuk, M. Schmitt, and B. Schuller, "Multimodal bag-of-words for cross domains sentiment analysis," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4954–4958.

[207] M. Seo and M. Kim, "Fusing visual attention cnn and bag of visual words for cross-corpus speech emotion recognition," *Sensors*, vol. 20, no. 19, p. 5559, 2020.

[208] Z. Xiao, D. Wu, X. Zhang, and Z. Tao, "Speech emotion recognition cross language families: Mandarin vs. western languages," in *2016 International Conference on Progress in Informatics and Computing (PIC)*. IEEE, 2016, pp. 253–257.

[209] E. M. Albornoz and D. H. Milone, "Emotion recognition in never-seen languages using a novel ensemble method with emotion profiles," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 43–53, 2015.

[210] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," *arXiv preprint arXiv:1801.06353*, 2018.

[211] Y. Ning, Z. Wu, R. Li, J. Jia, M. Xu, H. Meng, and L. Cai, "Learning cross-lingual knowledge with multilingual BLSTM for emphasis detection with limited training data," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5615–5619.

[212] Y. Zhang, Y. Liu, F. Weninger, and B. Schuller, "Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4990–4994.

[213] J. Parry, D. Palaz, G. Clarke, P. Lecomte, R. Mead, M. Berger, and G. Hofer, "Analysis of deep learning architectures for cross-corpus speech emotion recognition." in *INTERSPEECH*, 2019, pp. 1656–1660.

[214] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, "Towards speech emotion recognition" in the wild" using aggregated corpora and deep multi-task learning," *arXiv preprint arXiv:1708.03920*, 2017.

[215] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Using multiple databases for training in emotion recognition: To unite or to vote?" in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[216] X. Li and M. Akagi, "Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model," *Speech Communication*, vol. 110, pp. 1–12, 2019.

[217] Z. Zhang, F. Weninger, M. Wöllmer, and B. Schuller, "Unsupervised learning in cross-corpus acoustic emotion recognition," in *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2011, pp. 523–528.

[218] M. Shami and W. Verhelst, "Automatic classification of expressiveness in speech: a multi-corpus study," in *Speaker classification II*. Springer, 2007, pp. 43–56.

[219] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Selecting training data for cross-corpus speech emotion recognition: Prototypicality vs. generalization," in *Proc. Afeka-AVIOS Speech Processing Conference, Tel Aviv, Israel*, 2011.

[220] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Using multiple databases for training in emotion recognition: To unite or to vote?" in *Twelfth Annual Conference of the International Speech Communication Association*. Citeseer, 2011.

[221] S. Latif, A. Qayyum, M. Usman, and J. Qadir, "Cross lingual speech emotion recognition: Urdu vs. western languages," in *2018 International Conference on Frontiers of Information Technology (FIT)*. IEEE, 2018, pp. 88–93.

[222] Y. Zong, W. Zheng, T. Zhang, and X. Huang, "Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression," *IEEE signal processing letters*, vol. 23, no. 5, pp. 585–589, 2016.

[223] P. Song, Y. Jin, L. Zhao, and M. Xin, "Speech emotion recognition using transfer learning," *IEICE TRANSACTIONS on Information and Systems*, vol. 97, no. 9, pp. 2530–2532, 2014.

[224] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5058–5062.

[225] M. Abdelwahab and C. Busso, "Ensemble feature selection for domain adaptation in speech emotion recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5000–5004.

[226] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2013, pp. 511–516.

[227] J. Deng, R. Xia, Z. Zhang, Y. Liu, and B. Schuller, "Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 4818–4822.

[228] Q. Mao, W. Xue, Q. Rao, F. Zhang, and Y. Zhan, "Domain adaptation for speech emotion recognition by sharing priors between related source and target classes," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 2608–2612.

[229] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 500–504, 2017.

[230] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1068–1072, 2014.

[231] J. Deng, Z. Zhang, and B. Schuller, "Linked source and target domain subspace feature transfer learning–exemplified by speech emotion recognition," in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 761–766.

[232] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion: A preliminary study," *arXiv preprint arXiv:1712.08708*, 2017.

[233] S. E. Eskimez, Z. Duan, and W. Heinzelman, "Unsupervised learning approach to feature analysis for automatic speech emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5099–5103.

[234] S. Pancoast and M. Akbacak, "Softening quantization in bag-of-audio-words," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1370–1374.

[235] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[236] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proceedings of the 2006 conference on empirical methods in natural language processing*, 2006, pp. 120–128.

[237] K. Patrick and J. F. Lavery, "Burnout in nursing," *Australian Journal of Advanced Nursing*, vol. 24, no. 3, p. 43, 2007.

[238] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[239] A. Malte and P. Ratadiya, "Multilingual cyber abuse detection using advanced transformer architecture," in *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*. IEEE, 2019, pp. 784–789.

[240] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Interspeech 2015*, 2015.

[241] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, and H. Prendinger, "Deep learning for affective computing: Text-based emotion recognition in decision support," *Decision Support Systems*, vol. 115, pp. 24–35, 2018.

[242] S. Chen, M. Zhang, X. Yang, Z. Zhao, T. Zou, and X. Sun, "The impact of attention mechanisms on speech emotion recognition," *Sensors*, vol. 21, no. 22, p. 7530, 2021.

[243] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical signal processing and control*, vol. 47, pp. 312–323, 2019.

[244] S. Tripathi, S. Tripathi, and H. Beigi, "Multi-modal emotion recognition on iemocap dataset using deep learning," *arXiv preprint arXiv:1804.05788*, 2018.

[245] F. Haytham, M. Lech, and C. Lawrence, "Evaluating deep learning architectures for speech emotion recognition," *Neural Networks*, 2017.

[246] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[247] Z.-J. Chuang and C.-H. Wu, "Multi-modal emotion recognition from speech and text," in *International Journal of Computational Linguistics & Chinese Language Processing, Volume 9, Number 2, August 2004: Special Issue on New Trends of Speech and Language Processing*, 2004, pp. 45–62.

[248] C. Li, Z. Bao, L. Li, and Z. Zhao, "Exploring temporal representations by leveraging attention-based bidirectional lstm-rnns for multi-modal emotion recognition," *Information Processing & Management*, vol. 57, no. 3, p. 102185, 2020.

[249] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2017, pp. 873–883.

[250] S. Siriwardhana, T. Kaluarachchi, M. Billinghurst, and S. Nanayakkara, "Multimodal emotion recognition with transformer-based self supervised feature fusion," *IEEE Access*, vol. 8, pp. 176 274–176 285, 2020.

[251] N.-H. Ho, H.-J. Yang, S.-H. Kim, and G. Lee, "Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network," *IEEE Access*, vol. 8, pp. 61 672–61 686, 2020.

[252] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, "Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations." in *IJCAI*, 2019, pp. 5415–5421.

[253] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100 943–100 953, 2019.

[254] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 4749–4753.

[255] S. Sangwan, D. S. Chauhan, M. Akhtar, A. Ekbal, P. Bhattacharyya *et al.*, "Multi-task gated contextual cross-modal attention framework for sentiment and emotion analysis," in *International Conference on Neural Information Processing*. Springer, 2019, pp. 662–669.

[256] M. Maithri, U. Raghavendra, A. Gudigar, J. Samanth, P. D. Barua, M. Murugappan, Y. Chakole, and U. R. Acharya, "Automated emotion recognition: Current trends and future perspectives," *Computer Methods and Programs in Biomedicine*, p. 106646, 2022.

[257] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, "CNN+ LSTM architecture for speech emotion recognition with data augmentation," *arXiv preprint arXiv:1802.05630*, 2018.

[258] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[259] L. Kessous, G. Castellano, and G. Caridakis, "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis," *Journal on Multimodal User Interfaces*, vol. 3, no. 1, pp. 33–48, 2010.

[260] H. Hu, T. Tan, and Y. Qian, "Generative adversarial networks based data augmentation for noise robust speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5044–5048.

[261] T. 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*, vol. 18, no. 7, pp. 1766–1774, 2010.