Neural Conversion of Social Attitudes in Speech Signals

Clément Le Moine Veillon

THÈSE DE DOCTORAT DE SORBONNE UNIVERSITÉ
Spécialité Informatique
ED130 - Ecole doctorale Informatique, Télécommunications et Electronique (Paris)
Sciences et Technologie de la Musique et du Son (UMR 9912) Institut de Recherche et de Coordination Accoustique Musique
Equipe Analyse/Synthèse des Sons.

Présenté par

Clément LE MOINE VEILLON

NEURAL CONVERSION OF SOCIAL ATTITUDES IN SPEECH SIGNALS

Thèse soutenue le 27/02/2022 devant le jury composé de :

M. Thomas HUEBER, Directeur de recherche CNRS, GIPSA-lab	Rapporteur
M. Damien LOLIVE, Professeur, IRISA, Université de Rennes 1	Rapporteur
Mme. Berrak SISMAN, Professeure associée, University of Texas	Examinateur
Mme. Catherine PELACHAUD, Directrice de recherche CNRS, ISIR, Sorbonne Université	Examinateur
M. Carlos BUSSO, Professeur, University of Texas	Examinateur
M. Jaime LORENZO TRUEBA, Chercheur, Amazon	Examinateur

Cette thèse a été dirigée par Axel ROEBEL ET ENCADRÉE PAR NICOLAS OBIN.

When the green woods laugh with the voice of joy,
And the dimpling stream runs laughing by;
When the air does laugh with our merry wit,
And the green hill laughs with the noise of it.

WILLIAM BLAKE, *Laughing Song*

We often refuse to accept an idea merely because the tone of voice in which it has been expressed is unsympathetic to us.

FRIEDRICH NIETZSCHE, *Human, All Too Human*

Abstract

As social animals, humans communicate with each other by transmitting various types of information about the world and about themselves. At the heart of this process, the voice allows the transmission of linguistic messages denoting a strict meaning that can be decoded by the interlocutor. By conveying other information such as attitudes or emotions that connote the strict meaning, the voice enriches and enhances the communication process. In the last few decades, the digital world has become an important part of our lives. In many everyday situations, we are moving away from keyboards, mice and even touch screens to interactions with voice assistants or even virtual agents that enable human-like communication with machines. In the emergence of a hybrid world where physical and virtual reality coexist, it becomes crucial to enable machines to capture, interpret, and replicate the emotions and attitudes conveyed by the human voice.

This research focuses on speech social attitudes, which can be defined - in a context of interaction - as speech dispositions towards others and aims to develop algorithms for their conversion. Fulfilling this objective requires data, i.e. a collection of audio recordings of utterances conveying various vocal attitudes. This research is thus built out of this initial step in gathering raw material - a dataset dedicated to speech social attitudes. Designing such algorithms involves a thorough understanding of what these attitudes are both in terms of production - *how do individuals use their vocal apparatus to produce attitudes?* - and perception - *how do they decode those attitudes in speech?* We therefore conducted two studies, a first uncovering the production strategies of speech attitudes and a second - based on a Best Worst Scaling (BWS) experiment - mainly hinting at biases involved in the perception such vocal attitudes, thus providing a twofold account for how speech attitudes are communicated by French individuals. These findings were the basis for the choice of speech signal representation as well as the architectural and optimisation choices for the design of a speech attitude conversion algorithm. In order to extend the knowledge on the perception of vocal attitudes gathered during this second study to the whole database, we worked on the elaboration of a BWS-Net allowing the detection of mis-communicated attitudes, and thus provided *clean* data for conversion learning. In order to learn how to convert vocal attitudes, we adopted a transformer-based approach in a many-to-many conversion paradigm with mel-spectrogram as speech signal representation. Since early experiments revealed a loss of intelligibility in the converted utterances, we proposed a linguistic conditioning of the conversion algorithm through incorporation of a speech-to-text module. Both objective and subjective measures have shown the resulting algorithm achieves better performance than the baseline transformer both in terms of intelligibility and attitude conveyed.

Keywords: speech attitudes, voice conversion, transformer network, production strategies, best-worst-scaling experiment

Résumé

En tant qu'animaux sociaux, les humains communiquent entre eux en se transmettant divers types d'information sur le monde et sur eux-mêmes. Au cœur de ce processus, la voix permet la transmission de messages linguistiques dénotant un sens strict qui peut être décodé par l'interlocuteur. En transmettant d'autres informations telles que des attitudes ou des émotions qui connotent le sens strict, la voix enrichit et facilite le processus de communication. Au cours des dernières décennies, l'importance des technologies numériques dans nos vies n'a cessé de croître. Dans de nombreuses situations quotidiennes, nous délaissons les claviers, les souris et même les écrans tactiles au profit d'interactions avec des assistants vocaux ou même des agents virtuels qui permettent de communiquer avec les machines comme on le fait avec nos congénères. Avec l'émergence d'un monde hybride où coexistent réalités physique et virtuelle, il devient crucial de permettre aux machines de capter, d'interpréter et de reproduire les émotions et les attitudes véhiculées par la voix humaine.

Cette recherche se concentre sur les attitudes sociales de la parole, qui peuvent être définies dans un contexte d'interaction comme des dispositions vocales envers les autres, et vise à développer des algorithmes pour leur conversion. Pour atteindre cet objectif, des données - c'est-à-dire une collection d'enregistrements audio d'énoncés véhiculant diverses attitudes vocales - sont nécessaires. Cette recherche est donc construite à partir de cette étape initiale de collecte d'une matière première, à savoir un jeu de données dédié aux attitudes sociales de la parole. La conception d'algorithmes de conversion des attitudes vocales implique de comprendre ce qui les définit, à la fois en termes de production - *comment les individus utilisent-ils leur appareil vocal pour produire des attitudes ?* - et de perception - *comment décodent-ils ces attitudes dans la parole?*. Nous avons donc mené deux études, une première mettant en évidence les stratégies de production des attitudes vocales et une seconde - basée sur une expérience de Best Worst Scaling (BWS) - mettant principalement en évidence les biais impliqués dans la perception de ces attitudes vocales, fournissant ainsi une double compréhension de la manière dont les attitudes vocales sont communiquées par les individus français. Ces résultats nous ont permis de motiver notre choix de représentation du signal vocal ainsi que nos choix d'architecture et d'optimisation pour la conception d'algorithmes de conversion des attitudes vocales. Afin d'étendre à l'ensemble de la base de données les connaissances sur la perception des attitudes vocales recueillies lors de cette seconde étude, nous avons travaillé à l'élaboration d'un BWS-Net permettant la détection des attitudes mal communiquées, fournissant ainsi des données *propres* pour l'apprentissage de la conversion. Afin d'apprendre à convertir les attitudes vocales, nous avons adopté une approche basée sur un réseau transformer dans un paradigme de conversion many-to-many utilisant le mel-spectrogramme comme représentation du signal de parole. Les premières expériences ayant révélé une perte d'intelligibilité dans les échantillons convertis, nous avons proposé un conditionnement linguistique de l'algorithme

de conversion en lui incorporant un module de reconnaissance de parole. Des mesures objectives et subjectives ont montré que l'algorithme résultant obtient de meilleures performances que le transformer de référence aussi bien en termes d'intelligibilité et d'attitude véhiculée.

Mots clés: attitudes sociales, conversion de la parole, réseau transformer, stratégies de production, expérience best-worst-scaling

Acknowledgments

First of all, I would like to express my sincere gratitude to my thesis supervisors Prof. Axel ROEBEL and associate Prof. Nicolas OBIN for the continuous support of my Ph.D research, for their scientific expertise and their benevolence.

I would also like to thank Stellantis and, in particular, Luciano OJEDA and Vincent ROUSSARIE for trusting me by funding my research work.

I sincerely thank all the IRCAM staff who have welcomed me over the past four years. I could not have hoped to work on this thesis in a better environment than that in which I evolved during these three years. The curiosity and great competence of the researchers and doctoral students who work at IRCAM in fields as varied as signal processing, psycho-acoustics or cognitive neuroscience make this fabulous place unique. I would particularly like to thank the sound analysis/synthesis team including Rafael FERRO, Frederik BOUS, Antoine LAVAULT, Lenny RENAULT, Guillaume DORAS and Alice COHEN-HADRIA with a special thank to Léane SALAIS and Yann TEYTAUT for their kindness and exceptional support throughout this research. I warmly thank Pablo ARIAS and Victor ROSI whose help went far beyond our research.

I would like to thank my parents, whose unfailing support enabled me to complete this research, as well as my sister for her constant interest in my work, even though it was sometimes difficult to understand. My last and biggest thanks go to Marie who did everything she could to support me through good and bad times and without whom I would never have made it to the end of the road.

Contents

1 INTRODUCTION **9**
- 1.1 Context . . . 10
  - 1.1.1 Foreword . . . 10
  - 1.1.2 IRCAM - *A Unique Place Dedicated to Music, Speech and Sound* . . . 10
- 1.2 General Background . . . 11
  - 1.2.1 Emotions & Attitudes . . . 12
  - 1.2.2 Speech Communication . . . 16
  - 1.2.3 Vocal Expression of Emotions & Attitudes . . . 19
- 1.3 Scope of the Thesis . . . 21
  - 1.3.1 Current Issues & Research Questions . . . 21
  - 1.3.2 A Wide Range of Applications . . . 24
- 1.4 Main Contributions of this Thesis . . . 26
  - 1.4.1 Designing a French Database of Expressive Speech for Social Attitudes . . . 26
  - 1.4.2 Uncovering the Production Strategies and Perception of Vocal Attitudes . . . 26
  - 1.4.3 BWS-Net: Predicting Perceptual BWS Judgements with Neural Networks . . . 27
  - 1.4.4 Sequence-to-Sequence Neural Conversion of Speech Attitudes . . . 28
- 1.5 Outline of the Thesis . . . 29
- 1.6 List of Publications . . . 30

2 STATE-OF-THE-ART **31**
- 2.1 Voice Conversion . . . 32
  - 2.1.1 General Scheme for Voice Conversion . . . 32
  - 2.1.2 Representing Speech Signals . . . 34
  - 2.1.3 Voice Conversion Algorithms . . . 37
  - 2.1.4 Metrics and Methodology for Evaluating Voice Conversion Algorithms . . . 43
  - 2.1.5 Section summary . . . 45
- 2.2 Speech Attitude Recognition . . . 47
  - 2.2.1 General Scheme of Speech Attitude Recognition . . . 47
  - 2.2.2 Speech Emotion Recognition Algorithms . . . 48
  - 2.2.3 Methodology for Evaluating Speech Emotion Recognition Algorithms . . . 49
  - 2.2.4 Section Summary . . . 51

3 ATT-HACK : A DATASET FOR SPEECH SOCIAL ATTITUDES 52
3.1 Genesis of Att-HACK . . . 53
3.1.1 A Need for Expressive Speech Data in French . . . 53
3.1.2 In Defense for a Research on Vocal Communication of Social Attitudes . . . 54
3.1.3 Choice of Speech Attitudes . . . 54
3.2 Design, Methodology and Recording . . . 56
3.2.1 Attitude Communication Context . . . 56
3.2.2 Set of Sentences . . . 57
3.2.3 Recording Sessions . . . 59
3.3 Processing and Preliminary Analyses . . . 60
3.3.1 Data Cleaning and Formatting . . . 60
3.3.2 Att-HACK metadata statistics . . . 61
3.3.3 Investigating pitch patterns underlying the production of vocal attitudes . . . 62
3.4 Discussion . . . 64
3.4.1 Balanced vs Imbalanced Data . . . 64
3.4.2 On Uncovering Data Biases . . . 64
3.5 Chapter Summary . . . 64

4 PRODUCTION STRATEGIES AND PERCEPTION OF SPEECH ATTITUDES 66
4.1 On the need for questioning the term attitude . . . 67
4.2 Anatomical division of the vocal apparatus . . . 68
4.2.1 Vocal fold behaviour . . . 68
4.2.2 Vocal tract actuation . . . 69
4.2.3 Phonetic structure . . . 69
4.3 First Study - Uncovering the production strategies of vocal attitudes . . . 70
4.3.1 Experiment . . . 70
4.3.2 Results . . . 71
4.3.3 Discussion . . . 73
4.4 Second Study - Understanding the perception of vocal attitudes . . . 75
4.4.1 A Perceptual Validation Method Based on Best-Worst-Scaling (BWS) . . . 75
4.4.2 Experiment . . . 77
4.4.3 Preliminary Analysis . . . 79
4.4.4 Understanding the human perception of speech attitudes . . . 84
4.5 Chapter Summary . . . 85

5 NEURAL F0 CONVERSION OF SPEECH ATTITUDES 87
5.1 Related Works on F0 Contours Modelling . . . 88
5.1.1 Pitch Contours (F0) - A First Parametric Approach . . . 88
5.1.2 Multi-level modelling by applying CWT to F0 signals . . . 89
5.1.3 CWT Adaptive Scales . . . 90
5.2 Related Works on Dual-GAN-Based Voice Conversion . . . 91
5.2.1 Generative Adversarial Network . . . 91
5.2.2 Dual implementation of GAN . . . 92
5.3 Contribution . . . 92
5.3.1 Wavelet Kernel Convolutional Encoder . . . 92
5.3.2 Framework overview . . . 93
5.4 Speech Attitude Conversion Experiment . . . 94
5.4.1 Implementation Details . . . 95
5.4.2 A One-to-One Speaker Dependant Conversion Experiment . . . 96

5.4.3 Training Procedure . . . 97
5.5 Results & Discussions . . . 98
5.5.1 Objective Evaluation . . . 98
5.5.2 Subjective Evaluation . . . 100
5.5.3 General discussion . . . 102
5.6 Chapter Summary . . . 104

6 SPEECH ATTITUDE RECOGNITION 108
6.1 Towards Speech Attitude Recognition . . . 110
6.1.1 Mel-spectrogram as Speech Signal Representation . . . 110
6.1.2 Model Architecture . . . 110
6.1.3 Preliminary experiment . . . 112
6.1.4 Results & Discussion . . . 114
6.2 Perceptual Regression Based on BWS Scores . . . 116
6.2.1 About the Possible Uses of Gathered Perceptual Data . . . 117
6.2.2 Proposal for a Perceptual Regressor . . . 117
6.2.3 Experiments with the Perceptual Regressor . . . 119
6.2.4 Results & Discussion . . . 121
6.3 Perceptual Classification based on BWS scores . . . 124
6.3.1 Proposal for a Perceptual Classifier . . . 124
6.3.2 Experiment with the Perceptual Classifier . . . 126
6.3.3 Results & Discussion . . . 127
6.4 Perceptual Metric Learning Based on BWS Raw Judgements . . . 129
6.4.1 Proposal for a Perceptual Arranger . . . 130
6.4.2 Experiments with the Perceptual Arranger . . . 134
6.4.3 Results & Discussion . . . 137
6.5 Chapter Summary . . . 139

7 TRANSFORMER-BASED CONDITIONED VOICE CONVERSION 141
7.1 Related Work - Transformer-Based Voice Conversion . . . 142
7.1.1 Voice Transformer Network's Architecture . . . 142
7.1.2 Voice Transformer Network's Optimization . . . 146
7.1.3 Limitations . . . 147
7.2 Contribution - Speech Attitude Conversion . . . 149
7.2.1 Reformulation in the Scope of Speech Attitude Conversion . . . 149
7.2.2 Linguistic Conditioning of Speech Attitude Conversion . . . 151
7.3 Speech Attitude Conversion Experiments . . . 155
7.3.1 Many-to-Many Experiment for Speech Attitude Conversion . . . 155
7.3.2 Linguistic Conditioning Experiment . . . 156
7.3.3 Selected Configurations & Evaluation Process . . . 157
7.3.4 Objective Evaluation . . . 159
7.4 Perceptual Evaluation of Vocal Attitude Conversion Models . . . 163
7.4.1 Perceptual Experiment . . . 163
7.4.2 Results & Discussion . . . 164
7.5 Chapter Summary . . . 167

8 GENERAL CONCLUSION & FURTHER DIRECTIONS 169
8.1 General Conclusions . . . 169
8.2 Further Directions . . . 171

Chapter 1

INTRODUCTION

Contents

1.1 Context 10
1.1.1 Foreword 10
1.1.2 IRCAM - *A Unique Place Dedicated to Music, Speech and Sound* 10
1.2 General Background 11
1.2.1 Emotions & Attitudes 12
1.2.2 Speech Communication 16
1.2.3 Vocal Expression of Emotions & Attitudes 19
1.3 Scope of the Thesis 21
1.3.1 Current Issues & Research Questions 21
1.3.2 A Wide Range of Applications 24
1.4 Main Contributions of this Thesis 26
1.4.1 Designing a French Database of Expressive Speech for Social Attitudes . . 26
1.4.2 Uncovering the Production Strategies and Perception of Vocal Attitudes . 26
1.4.3 BWS-Net: Predicting Perceptual BWS Judgements with Neural Networks 27
1.4.4 Sequence-to-Sequence Neural Conversion of Speech Attitudes 28
1.5 Outline of the Thesis 29
1.6 List of Publications 30

1.1 Context

This research on the conversion of social attitudes into speech was conducted at the IRCAM-STMS lab, a singular place at the intersection of academic research and creation, making it possible to address the dual demand posed by the need to thoroughly understand speech social attitudes, as well as the technical challenge of voice conversion. This document is the result of inter-disciplinary and collaborative research.

1.1.1 Foreword

As social animals, humans communicate with each other by transmitting various types of information about the world and about themselves, thus creating a common meta-consciousness of their being in the world. At the heart of this process, the voice allows the transmission of linguistic messages denoting a strict meaning that can be decoded by the interlocutor. By conveying other information such as attitudes or emotions that connote the strict meaning, the voice enriches and enhances the communication process. In other words, a subtle and complex meaning can be conveyed by a speaker using only a few words uttered with specific vocal traits that are informative to the recipient. In the last few decades, the digital world has become an important part of our lives. Although it is a world in and of itself, with its own set of rules, functioning and aesthetics, it cannot escape its creator's desire to shape it in his image. The desire to make orality the primary modality of our interactions with machines is at the heart of the anthropomorphic process through which we conceive the digital world. In many everyday situations, we are moving away from keyboards, mice and even touch screens to interactions with voice assistants or even virtual agents that enable human-like communication with machines. In the emergence of a hybrid world where physical and virtual reality coexist, it appears crucial to enable machines to capture, interpret, and replicate the emotions and attitudes conveyed by the human voice.

This research focuses almost exclusively on speech social attitudes, which can be defined - in a context of interaction - as speech dispositions towards others. More specifically, the aim of this research is to develop algorithms that convert these attitudes in speech while preserving other aspects of speech such as what is said and the vocal identity of the speaker. Designing such algorithms involves a thorough understanding of what these attitudes are both in terms of production - *how do individuals use their vocal apparatus to produce attitudes?* - and perception - *how do they decode those attitudes in speech?*. Without addressing these two questions, it cannot be guaranteed that the algorithm is effectively converting the attitudinal aspects conveyed by the speech. Once designed, these algorithms would allow for attitude conversion in recordings of actual human voice but can also be integrated into speech synthesis modules, thus extending the expressive spectrum of synthesized voices. This research has been supported by the French Ph2D/IDF MoVE project on Modelling of Voice Expressivity and application to an expressive conversational agent and funded by the Ile-De-France region and the car manufacturer Stellantis. This research was done with an eye on the existing and potential uses of speech conversion algorithms, although being situated upstream from their actual application in daily life.

1.1.2 IRCAM - *A Unique Place Dedicated to Music, Speech and Sound*

The IRCAM - or Institute for Research and Coordination into Acoustics & Music - created in 1977 on the initiative of the French composer Pierre Boulez and the French Ministry of Culture brings together composers, musicians, and researchers for the study and creation around music, speech

and sound. IRCAM is organisationally linked with and located next to the Centre Pompidou in the heart of Paris. Hosted at IRCAM, the STMS lab - or Sciences et Techniques de la Musique et du Son - is a joint research unit which brings together staff from CNRS, Sorbonne University, French Ministry of Culture, and IRCAM focusing on a vast but clearly identified field - music, speech and sound - in a unique context where contemporary creation meets scientific and technological research. The laboratory participates in the renewal of musical expression through the contributions of computer science, acoustics, signal processing, cognitive sciences, neuro-sciences and musicology. Conversely, specific problems posed by contemporary creation give rise to original scientific advances, be they theoretical, methodological, or practical, the scope of which goes far beyond the musical field. Created only a few years ago, IRCAM amplify is the industrial branch of the IRCAM. It creates a bridge between the recognized excellence of research at STMS lab and its potential real-life applications by taking on the process of industrializing and marketing promising lab findings in close collaboration with the researchers.

The human voice, whether spoken or sung, has historically been regarded as a priority research area in interaction with significant concerns about music and artistic creation, building on Pierre Boulez's original idea and the inspiration of emeritus researcher and former head of the Analysis-Synthesis team Xavier Rodet. The original focus of voice research was on the study and synthesis of the singing voice, which led to the creation of the CHANT singing voice synthesizer (Rodet et al., 1984; Mathews and Pierce, 1989) and culminated with the restoration of the castrato Farinelli's singing voice (Depalle et al., 1994). Later, in response to the increasing demand of composers and artists, research on speech gradually evolved (Röbel, 2003; Schwarz, 2003) with the introduction of high-quality speech technologies - ircamAlign, ircamTTS, ircamHTS, and SuperVPTrax - and countless implications for artistic production (Fineberg, 2006; Rohmer, 2007; Gervasoni, 2008; Parra, 2009; Lanza and Pasquet, 2009). The recent advent of neural networks has given a new lease of life to research in voice conversion - which consists in modifying one attribute of the speech while preserving the others - and has led to numerous proposals within the team for speaker identity conversion (Ferro et al., 2020), emotion conversion (Obin et al., 2019) and pitch conversion (Bous and Roebel, 2022).

One of the IRCAM's greatest strengths is that it brings together researchers who study speech from many angles. The perception and sound design team, in particular, featured psycho-acoustics and cognitive neuroscience researchers whose knowledge on speech (Aucouturier et al., 2016; Arias Sarah, 2018; Ponsot et al., 2018b,a; Goupil et al., 2019, 2021a) may enhance the scientific and technical advances that occur in the Analysis-Synthesis team. Within this unique context, the research work therefore tends towards trans-disciplinary collaboration between the different teams.

1.2 General Background

Human vocal expressions have evolved as signals to communicate with others. By continuously modulating the phonatory and articulatory structures of their vocal apparatus, individuals encode a considerable amount of non-verbal information, communicating their emotions such as joy or anger, their attitudes, towards a person such as friendliness or dominance or towards an object such as irony or doubt. In this research work, we propose to investigate social attitudes - i.e. attitudes towards a person - as one component of this non-verbal oral communication. This decision was not made at random and was the outcome of a thorough literature review that sought to examine each of these categories. Before we could situate any of these categories in a study of their oral modality, that is, insofar as they are expressed or communicated vocally, we had to understand

them in a broader sense.

1.2.1 Emotions & Attitudes

Emotion is derived from the Latin *exmovere*, composed of *ex* which means "out of" and *movere* which means "move" or "raise". Thus literally *emotion* can be understood as a change, a movement from one state to another. In this part, we will see that this change can be observed and understood on different scales : physiological, psychological, neural and cognitive. Historically regarded by the majority of scientists and philosophers as an organic instability caused by a disruption of consciousness - an irrelevant epiphenomenon (Skinner, 1953) - emotion has gained great interest among the scientific community from the second half of the 20th century, leading to a considerable amount of research across various disciplines such as biology, psychology, philosophy, anthropology, sociology and, more recently, neuroscience. Without properly attaining a solid consensus, researchers have attempted for centuries to define emotion in a way that is widely accepted (Izard et al., 2006).

Attitude is derived from the Latin *aptus*, which means "fitness" or "adaptedness." Like its by-form *aptitude*, *attitude* also denotes a state of readiness for action. However, the term took on a very distinct meaning as a result of its application in the area of art; it referred to the outward or visible posture (the body position) of a figure in statuary or painting. Like emotions, attitudes are primarily related to a physical lexical field, thus underlying their physical manifestations. It must be noted, however, that attitudes are originally associated with a static posture, whereas emotions are related to dynamics. In modern psychology, the first meaning was preserved in what are frequently referred to as "mental attitudes," and the second meaning in "motor attitudes". The first use of attitude dates back to Spencer and in his First Principles (Spencer, 1860). The concept of attitude, much less important in the history of ideas than that of emotion, became established with the development of social psychology (Allport, 1935) in the early 20th century. However, there is no unambiguous definition of attitude, and the concept has long been criticised for the variety of realities it covers.

Over the course of time, research in many fields has uncovered certain characteristics of both emotions and attitudes, sometimes seeking to distinguish them and sometimes stressing their mutual influence. The following is an attempt to synthesize these findings.

Bodily Evidence of our Emotions & Attitudes

The first material examined in the process of defining emotions was their physiological signature. The sensational theory of emotions, which dates back to Spencer (Spencer, 1881), describes emotions as aggregates or effects whose constituents or causes were sensations. In this line, the Jamesian tradition focuses on the corporeal manifestation of emotions. James stated in 1884 that the human body could be thought of as a *reverberation chamber* in which even the smallest physical changes resonate until they result in the conscious experience of emotion (James, 1884). Our bodies undergo physiological changes in response to a variety of situations, and our experience of these changes is what we refer to as emotion. According to James, emotions would manifest themselves in abrupt physiological changes such as gastrointestinal motility, somatic fluctuations, sweating, pupil dilation, heart rate, cutaneous blood flow (blushing or becoming pale), piloerection, and skin conductance. If we specifically observe the physiology of vocal apparatus, we may see that emotions are experienced as a variety of vocal changes, such as a trembling voice or quick changes in intonation. Emotions would therefore occur as a result of physiological responses to events.

Initially, attitudes were almost exclusively thought of as motor, i.e. as physical manifestations. Lange developed a motor theory according to which the phenomenon of perception was largely a consequence of a certain physical configuration, a motor attitude.(Lange, 1888). Later, Baldwin suggested that motor attitudes were the basis for an understanding of emotional expression (Baldwin, 1895). From this point of view, attitudes were mainly considered as the physical manifestations of emotions. From this perspective, friendliness might be viewed as a certain configuration of facial muscles, for instance involving smiling, along with certain typical vocal traits. To this initial vision, new theories considered attitudes as based on a mental component. The underlying dualism eventually disappeared in favour of a more global view of attitudes.

A whole branch of study known as cognitive evaluation theory was born out of these early theories, which put the observation of physical changes at the center of understanding emotions and attitudes.

Cognitive Evaluation Theory of Emotions & Attitudes

The cognitive evaluation - or appraisal - theory postulates that emotions arise from a personal appraisal of the events experienced. According to these theories, emotion would be a cognition i.e. a complex symbolic processing of information, whether conscious or not. At the heart of those models is the procedure for evaluating the emotional significance (Arnold, 1960; Lazarus, 1968; Scherer, 1999). The standard approach, based on early research by Arnorld and Lazarus (Lazarus, 1968), suggests that individuals employ a limited set of criteria to assess the situations they experience (Frijda, 1999). Depending on how well the individual can handle the so-called circumstance, these criteria that reflect the personal relevance and significance of the encountered situation may lead to either a favorable or negative overall judgment.

Similarly, attitudes have been defined as overall evaluation processes - that is roughly, determining the degree to which an issue, object, or person is liked or disliked (Petty et al., 1981; Eagly and Chaiken, 1993). Attitude may thus also be thought of as a cognition as the evaluation involves processing and summarizing different types of information. There is a close correlation between a person's attitude and the beliefs (Petty et al., 1993) and affects Bodenhausen (1993) they identify with the subject of their attitude. Yet, the affects encompass a broad range of feelings that people can experience such as emotions and moods. Consequently, from this point of view, attitudes are directly linked to emotions. For instance, it appears that the emotions an individual feels towards another person partly determine their attitude towards that person (Mackie et al., 2000). This shows how attitudes and emotions are intertwined, making it challenging to discriminate between the two.

This cognitive theory is based on experiments in which participants are asked to recall past emotional experiences or behaviours, they might also be placed in the conditions to feel such an emotion or to change attitude through stimuli exposure. The inability to distinguish between the antecedents of an emotional experience and its real content - or between those of a specific attitude change and the proper meaning of this attitude - is one shortcoming of these methodologies. Even though it is now acknowledged that a particular stimulus can cause a multitude of subjective emotional experiences (Barrett, 2017), the physiological changes are still regarded as a crucial part of understanding emotions and attitudes.

However, numerous studies are now appearing to cast doubt on the Darwinian idea that it is possible to infer the emotion felt by an individual from a straightforward analysis of the physical signals he or she conveys (Barrett, 2011). As a result, an important distinction between what is genuinely felt - and materialized by a certain neural activity - and what occurs through various physiological signs, whether physical or audible, has to be made in the comprehension of emotions and attitudes.

Ancient Neural Pathways of Emotions & Attitudes

In the quest for a definition of emotions, affective neuroscience - which examines the neurological underpinnings of affects [1] - has brought new insights uncovering specific brain systems that are, at least partially, devoted to emotion. In 1986, Panksepp suggested the existence of several distinct brain circuits linked to primary emotions (Panksepp, 1987). Panksepp's research on non-human animal emotions led to uncovering seven systems located in the diencephalon [2], each of which is dependent on distinct neural substrates and is essentially similarly present in all mammals: seeking, rage, fear, lust, care, panic and play. Causal stimulation (electrical or chemical) of such systems - in both humans and animals - has been shown to influence emotional feelings and behaviours (Panksepp et al., 2004). Panksepp postulates that these circuits enable the collection of information coming from both inside and outside the body which is then processed and transmitted to sub-circuits responsible for individual's behaviour. Only a thin line seems to exist between what Pranksepp refers to as emotions and specific attitudes like being careful or playful. Meanwhile, Ekman proposed six basic emotions (Ekman, 1999) - sadness, happiness, fear, anger, surprise and disgust - providing evidence for distinctive patterns of autonomic nervous system (ANS) activity for anger, fear and disgust (Ekman et al., 1983) as well as for sadness (Levenson et al., 1991).

The existence of such distinctive patterns - at a neurological level - tends to highlight the role of emotions - and attitudes to a lesser extent - in the evolution of the human species. Indeed, humans have experienced a plethora of issues throughout evolution, including those brought on by birth, death, conflict, and even seduction. In some of these circumstances, deliberating carefully and logically about how to respond is not always the best course of action. However, experiencing an emotion - such as fear or an attitude - such as being seductive - would facilitate quicker contextual adaptation with minimal conscious cognitive effort (Tooby and Cosmides, 2008).

Furthermore, it appears that these neural patterns fall well short of accounting for the wide range of emotions experienced by human individuals throughout time and between cultures (Izard, 2007). A consensus emerged among scientists on a distinction between basic emotions - mentioned above and which refers to affective processes generated by evolutionarily old brain systems (Izard, 2007) - and more subtle emotions referred to as emotion schemas - that involve high order cognitive processes such as complex cognitive processing of the stimulus based on conceptual construction. According to Izard, emotion schemas can be defined as dynamic interactions between cognitive components (Izard, 2008) that differ across individuals and cultures (Izard, 2007). These schemas depend on each person's subjective experience and personality, and usually emerge during early development (Izard, 2007, 2008).

A Constructivist Theory

The theory of constructed emotions constitutes the last significant contribution on emotions. Proposed by Barrett, this theory (Barrett, 2017) is part of a broader understanding of how the human brain operates. The brain interprets information coming from the world, including the rest of the body, by constructing concepts, i.e. collections of embodied, whole-brain representations that predict upcoming sensory environment events and the appropriate course of action to deal with them. The so called interoceptive network - that involves several regions of the human brain - organizes all the sensory data coming from environment, from internal tissues and organs, from hormones in the blood and also from the immune system, labels them, and relates them to concepts. Once formed, a concept allows for a perception or an experience thus explaining the cause of sensory events and directing action, i.e. the concept categorizes the sensory event. In this way, the brain

[1] Affect is a broad concept that encompasses emotions, moods and even attitudes.

[2] The diencephalon (or interbrain) is a division of the forebrain (embryonic prosencephalon).

uses past experience to construct a categorization (Barrett, 2006, 2015) that best fits the situation to guide action. According to Barrett, emotions do not stray from this functioning. We might experience an emotion in many different contexts and circumstances, each of which will be a unique instance of that emotion. To categorize and make sense of the impending sensory inputs, the brain produces a concept of this emotion.

This theory gives a glimpse of other aspects of the role played by emotions - and attitudes - in human functioning. For instance, emotions act as a glue for our memories, imbuing each moment with the emotion we were experiencing at the time. Emotions also have an impact on our thought processes, sometimes in positive ways and sometimes not (Matsumoto et al., 2006). Thus, emotions can no longer be seen as reactions or interpretations to external stimuli. They are constructed by the brain from a condensed summary of the information it receives. Within this perspective, emotions not only influence immediate behavior but also forecast future behavior. Emotions also play an orchestration role by interacting with various systems (perception, attention, inference, learning, etc.), activating some while deactivating others, preventing cognitive chaos (multiple competing systems operating at once), and enabling individuals to respond to environmental cues in a coordinated manner (Levenson, 1999).

From Cognition to Social Cognition

In The Politics (Aristotle. and Lord, 1984), which dates back to the fourth century BC, Aristotle refers to humans as political animals. Indeed, since the dawn of time, humans have lived in groups, forming large or small social and political communities. This social appetite of humans is reflected in their multiple and varied interactions whether they occur in family, professional or political contexts. Individuals evolve in various social spheres where they must communicate with others for a variety of purposes such as to persuade the other or to provide the other information. This communication process - which can occur through several channels or modalities (such as orality) at a time - involves a twofold movement of expression and perception. The foundations of this process will be detailed in the following section. Both emotions and attitudes are the object of communication processes, we express them to others - through modulating our facial muscles, body postures and vocal traits - and we perceive those of others - by decoding cues from face and body movements and speech acoustic signals (Elfenbein and Ambady, 2002).

Social cognition aims to understand the cognitive mechanisms that govern this social order. Warmth and competence have been uncovered as main dimensions underlying social cognition (Fiske et al., 2007). When interacting with someone, we first consider whether they have good intentions - warmth - and then consider whether they have the capacity to act on their judgment - competence. Both attitudes and emotions can be understood within this paradigm. However, it is clear that we are no longer talking about these categories as cognitions - mechanisms of information processing - or the feeling associated with it, we are talking about what transpires of it in the social sphere. As pointed out by Ekman in (Ekman, 1999), emotions obviously do occur without any evident signal, because individuals can at a very large extent, inhibit the appearance of a signal. He also adds that emotional expression can occur without actual emotional feeling. An important distinction then appears, between what is felt by individuals, and what they chose and are able to express which may be totally different. Sometimes, we express emotions that we do not genuinely feel, we portray emotions which potentially relate to totally different neural bases (Anikin and Lima, 2017). The expression of emotions involves a delicate balance between the desire to communicate or not a specific emotional content - that we feel for real or not - and the ability to control one's vocal apparatus, face and body muscles to achieve this.

In the light of social cognition, emotions and attitudes take on a different role for us humans, as their communication influences what the receiver is likely to feel or behave. For instance , they

can be seen as regulators of social interaction (Bradshaw, 1986). Emotions and attitudes inform us about the others and interpersonal relationships and provide incentives for desired social behavior (Keltner et al., 2003).

1.2.2 Speech Communication

Communication is derived from the Latin *communicare* which covers several meanings : "take part", "share", "be in communion with", "transmit" or even "propagate". This polysemy reflects the variety of theories that aim to explain the underlying mechanisms that govern communication. This section aims to provide the keys to understanding the communication of emotions and attitudes. This implies considering, in the first instance, the communication process as a whole.

Human Communication - A Mechanistic View

From a mechanistic point of view, communication can be described as the complex process in which a message is transmitted from an emitter to a receiver. Throughout the development of information theory in the 1940s, Shannon built a formal model of communication, accounting for systems dedicated to information transmission. As depicted in Figure 1.1 : to communicate, an emitter encodes a message into a signal which is transmitted through a channel and decoded by a receiver. During transmission the message is more or less altered depending on the quality of the encoding, decoding and transmission channel.

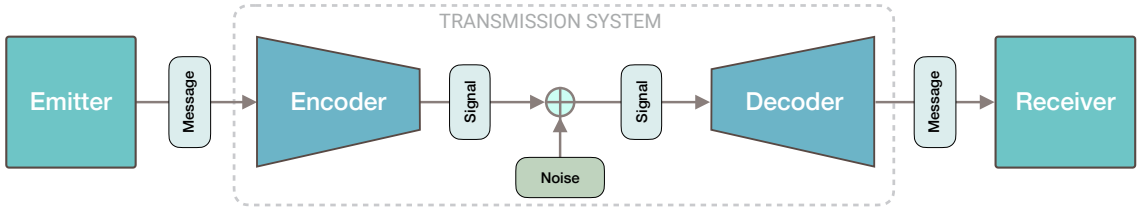

Figure 1.1: Schematic diagram of a communication system, after [Shannon, 1948]

This mechanistic view of the *transmission* process has constituted a starting point for numerous studies trying to describe human communication. In 1954, Wilber Schramm pointed out a major limitation to Shannon's model: unless and until a receiver is able to understand or decode the information the sender wants to communicate, the communication is compromised. Following this observation, he proposed a model - depicted in Figure 1.2 - inspired by Shannon's in which *interpersonal communication* was described as a two way process involving at least two individuals. From that point onwards, the understanding of human communication will include what is referred to as *retroaction* process: a feedback loop that allows the sender to know whether the message has properly been decoded. Human communication is no longer akin to a simple transmission process but must be understood as a collaborative process.

Human Communication - A Cognitive View

The *enaction* theory, at the core of human communication, was proposed by the neurobiologist Francisco Varela in the 1980's. This theory states that knowledge about the world is not predefined data that individuals produce representations of, but rises during sensorimotor interactions between individuals and their environment (Varela, 1996). Studies on the development of language

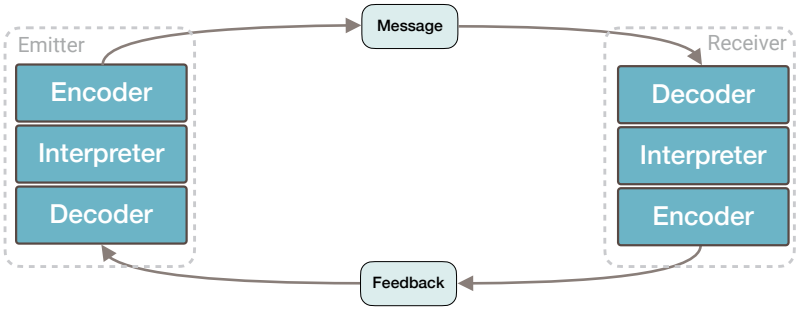

Figure 1.2: Schematic diagram of a communication system, after [Schramm, 1954]

in children, from gestation onwards, show how languages are enacted in humans: they emerge in the mutual and reciprocal rooting of perception and action, through the tactile, sonic and rhythmic experience of the child with its environment (Aden, 2013). By naming things in the world to share an experience with others, we perceive ourselves as perceiving at the same time as we create a shared social space. While communicating, an individual is at the same time a sender and a receiver, a speaker and a listener of his/her own message. Within this scope communication can be understood as a meta-consciousness of our being in the world. Communication can no longer be understood as a simple transmission of information, it is the action of making a common world emerge and thus cannot be thought outside the specific *context* in which it occurs. During an interaction, two individuals communicate conditionally to their mutual past experience as well as the present context of speech and thus construct a meaning - i.e. an interpretation - which is specific to that interaction.

Language - A Human Ability to Communicate

To communicate, humans encode information using sign systems - i.e. sets of symbols that are shared among the individuals involved in the communication process. As soon as they are shared, these systems, potentially of all types, can make inter-individual communication possible. For instance, *language* is the ability of people to express their thoughts and communicate with each other by means of a system of conventional vocal and/or graphic signs. This system - referred to as a *tongue* - is formed by a lexicon (a set of words) and a grammar (a set of operating rules). While language, as ability, is innate and universal in humans, language, as sign system, is learned and differs between groups of individuals, there are more than 6000 of them around the world. Although it has been challenged by the case of sign language, the fundamental particularity of language lies in the ability to associate a sound with a unit of meaning. For example, the vocalization yum-yum means I am hungry in English. In linguistics - the study of language - the concept of sign developed by Ferdinand de Saussure is formed by a signified and signifier, the first designating the mental representation of the concept associated with the sign - *vorstellung* which means the idea - and the latter designing the mental representation of the form and material aspect of the sign - *lautbild* literally the auditory image . The signified of a sign is distinguished from its referent, the object (or set of objects) designated by the sign (de Saussure, 1916). The signified can take two forms: *denotation*, the literal meaning of a term and *connotation*, all the elements of meaning that can be added to this literal meaning. According to Saussure, the relationship between the signifier and the signified is arbitrary, e.g. the concept of water (signified) has several signifiers throughout world (eau in French, agua in Spanish ...).

Multi-Modality of Human Communication

As the most usual context for inter-human communication, face-to-face interaction is also the context in which human language is learned and evolves. Within this context, communication not only conveys information through a single channel but is about exchanging a plethora of multimodal signals. We communicate through a complex orchestration of multiple articulators and modalities: messages are dispatched and received, potentially, via all anatomical and cognitive means available to humans to communicate (Mehrabian, 1972; Andersen, 1999). The communicated signals are mainly produced through gesture, speech, and touch modalities, and symmetrically perceived through vision, hearing, and touch perceptual senses. Articulators, in the classical sense, refer to the vocal organs above the larynx (i.e., the tongue, lips, teeth, and hard palate). In a multimodal perspective, the definition can be extended to other body organs such as the head, the face including the forehead and eyebrows, the upper and lower eyelids, the muscles around the nose, cheeks, and mouth, the hands, arms, and shoulders, the upper torso, and, in principle, the lower torso, legs, and feet, although they are less systematically involved in the communication process. Each conveyed information is processed through a hierarchy of representational levels. In oral language modality, for instance, information is encoded by speakers through semantic, syntactic, morphologic, phonetic levels and symmetrically decoded according to this same inverted hierarchy through complex psycholinguistic processes (binding of multiple, temporally offset signals under tight time constraints posed by a turn-taking system (Holler and Levinson, 2019)). Human communication is therefore made of the co-production of a set of signs conveyed through different modalities (speech, gesture, ...) which are co-intergrated through different human sensory dimensions within a mutual dynamic process.

Vocal Communication

Among all the dimensions of human communication, orality is the one that has been studied the most, mostly through the analysis of its different transcripts. Historically, research on communication initiated in the West and therefore focused on Western languages. Having the particularity of being transcribable (through writing) - i.e. immediate access to the language - these languages were an opportunity to study the oral dimension of communication. With the arrival of sound recording technology the analysis of audio signals was made possible, thus reinforcing the research community's interest in vocality. In contrast, the study of the non-vocal dimensions of communication has proved to be more challenging because they are not strictly speaking governed by a language although gestural modality is mistakenly referred to as body language.

If language, as a system of signs, is a pre-requisite invariant for any communication, conversely, modulation is also essential to it. At the heart of voice communication, modulation allows semantic differentiation (Uriel et al., 1968) : each acoustic modulation - i.e. variation of a set of vocal parameters - has a potential communicative value that results in the modulation of the meaning of the utterance itself. We modulate our voice to produce meaning. Modulation therefore occurs at different levels, from semantics to acoustics, the variations of a given level being bijectively related to those of the others. For instance there is a bijective relationship between the speed of vibration of the vocal cords (anatomic mechanism) and the fundamental frequency variations (speech parameter). In order to understand these modulations, we have to go back to the anatomical functioning of the vocal apparatus. If we divide this functioning anatomically, we can distinguish two sub-parts : 1) the glottal source at the origin of the so-called prosodic modulations associated with vocal parameters such as intonation, rhythm or intensity and 2) the vocal tract that causes timbral and articulatory modulations associated with the spectral envelope of the speech signal.

Anatomy	Signal	Domain	Information
1. **glottal source**	pitch, rythm, intensity	prosody	1.1. **linguistic** *(focus, syntax)* 1.2. **paralinguistic** *(emotion)* 1.3. **extralinguistic** *(speaking style)*
2. **vocal tract**	spectral enveloppe	timbre, articulation	2.1. **linguistic** *(phonemes)* 2.2. **extralinguistic** *(speaker identity)*

Table 1.1: Overview table of the functioning of speech

Studies in general linguistics generally distinguish three main domains in speech communication: namely linguistics, paralinguistics and extralinguistics. As developed in the Table 1.1, each of these domains is related to specific anatomic regions of the human vocal apparatus as well as specific speech signal parameters conveying different types of information.

- The **linguistic domain** governs all information conveying the strict meaning, i.e. the linguistic message. This message is acoustically encoded into a series of temporal acoustic units referred to as phonemes, i.e. sounds associated with textual units. The temporal articulation, within and between these different units, is governed by prosodic patterns (accent, focus, syntax, ...) that are specific to the linguistic content of the encoded message.

- The **para-linguistic domain** governs any information that contributes to the meaning but is not determined by mere linguistic content i.e. any contextual information that specifies the meaning such as emotions or attitudes. This information is acoustically encoded through prosody, notably intonation, rythm and intensity (Puts et al., 2006, 2007; Chen et al., 2004; Li and Wang, 2004; House, 2005; Feinberg et al., 2005; Xu et al., 2013) although it may also involve specific vocal tract behaviour (Feinberg et al., 2005; Puts et al., 2007). Long considered to interfere with vocal communication, certain para-verbal phenomena (hesitations, repetitions, breathing or even laughter) have been rehabilitated by recent studies in their communicative function (Brennan and Schober, 2001; Vettin and Todt, 2004; Campbell and Erickson, 2004). Another example is speaking style as an adaptation to the audience, for example in a professional setting.

- The **extra-linguistic domain** governs any information relating to individual characteristics of the speaker such as identity, individual speaking style and socio-geographical origin. Recent studies have highlighted the prevalence of individual characteristics such as personality traits in speech signals and furtherly shown that by listening to a basic speech sample such as Hello , individuals can predict personality traits of an unknown speaker (McAleer et al., 2014).

1.2.3 Vocal Expression of Emotions & Attitudes

The ability to vocally express emotions or attitudes is crucial for human communication. The meaning of a phrase is connoted by emotion, and thus significantly enhances communication by making it easier and more fluid. Understanding how emotions are expressed through speech is essential if we work to make it possible for machines to capture, interpret, and reproduce them.

Modulation on Top of Linguistic Message?

As suggested by Fonagy in the double speech coding scheme shown in Figure 1.3, emotions communicated vocally can be seen as a modulation of linguistic information, an overlay of encoding that connotes - or suggests an interpretation of - strict meaning. The linguistic content of an utterance determines its overall acoustic form : phonemes are temporally articulated according to specific prosodic patterns. Multiple variations can be applied to this basic form : para-linguistic (emotions, ...) but also extra-linguistic (speaker identity, ...). Thus, the acoustic realisation of the emotion is conditional on that of the linguistic content.

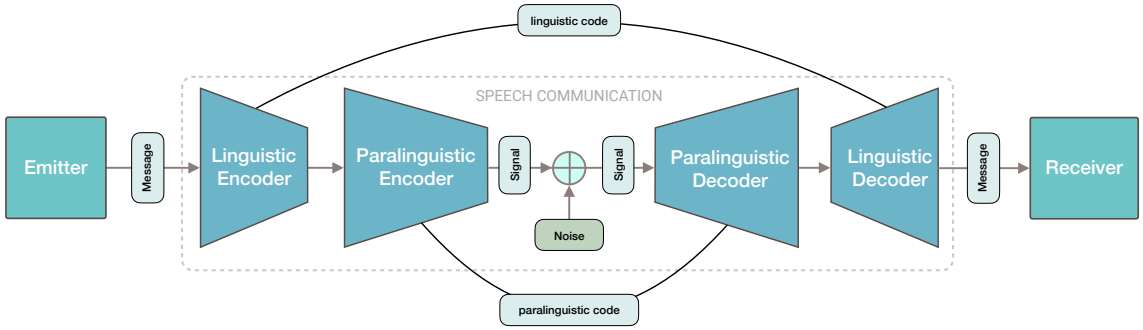

Figure 1.3: Double speech coding scheme, after [Fonagy, 1983]

An Oral Language of Emotions & Attitudes?

Unlike linguistic information, emotional information is not strictly speaking reducible by means of a language and this is so for many reasons. If we make the analogy with linguistics, we would have a signified, the concept associated with an emotion, and a signifier, its acoustic realisation. Can we say that all concepts of emotion, including the most elaborate ones (schemas), have an associated acoustic form? The answer seems to be no, insofar as individuals have neither a lexicon nor a grammar that allows the communication of the full range of their emotions. There is not a proper meaning-sound equivalence, as we know it in most languages. Moreover, vocal communication of emotions has been shown to be highly individual, which is also consistent with an absence of emotional language - which would imply that it is shared among individuals.

However, it is a fact that individuals are able to capture emotional information in speech signals. Therefore, there must exist a number of signs - more or less elaborate acoustic signifiers - shared within larger or smaller cultural communities and which allow individuals to decode the emotions communicated. Individuals are said to share strategies (neural, cognitive, anatomical) that they use to transcribe their emotions into acoustic signifiers.

Acoustic Signature of Emotions & Attitudes

Emotions are known to be conveyed mostly by affective prosody according to three main speech parameters: intonation (through pitch contours), rhythm (speech rate, syllable durations) and intensity (airflow energy). Among these parameters, intonation has particularly been identified for its role in speech emotional (Banse and Scherer, 1996; Bachorowski and Owren, 1995; Bänziger and Scherer, 2005) and attitudinal (Ponsot et al., 2018b) expression. In fact, algorithmically manipulating an individual's voice in real-time by incrementing their mean pitch and pitch variation

(with dynamic inflections) can change the emotional state of the speaker (Aucouturier et al., 2016). Speech rate is also linked to mood, to the extent that there are significant negative correlations between depression tests and speech-rate (Cannizzaro et al., 2004), with slower speech-rate associated with higher levels of depression. Finally, timbre, studied for instance by the presence of rough cues, the production of which by the vocal apparatus is usually due to the saturation of the vocal folds, is an acoustic cue often used as an expression of arousal in humans, but also across species, extending to mammals (Fitch et al., 2002).

Universality of Speech Emotions & Attitudes

Several studies have revealed evidence of universality for basic communicative signals. For instance, individuals from a culturally isolated Namibian village in southwest Africa and westerners can both recognize the archetypal vocalizations conveying Ekman's fundamental emotions (anger, disgust, fear, joy, sadness, and surprise) (Sauter et al., 2010). This universality seems to extend beyond the domain of emotions, in (Bryant et al., 2018) listeners from 21 societies across six world regions were able to differentiate whether laughter produced by English speakers was fake or real.

On the other hand, a lot of research works tend to relativize or at least restrict the universalist thesis to a few fundamental signals, like laughter. According to Barrett, emotions are not universal, but vary from culture to culture (Barrett, 2006), they are not triggered by any event but built out of sensations at large. Emotions emerge as a result of a complex interaction between the physical characteristics of the body, a flexible brain that adapts to any environment it develops in, and the culture and upbringing that create that environment. Vocally, this theory translates into a diversity of strategies used by individuals to produce emotions. These strategies may depend on individual criteria (Bachorowski, 1999) such as lived experience, tastes and, in general, habitus (Bourdieu, 1972). In particular, an emotion that has never been experienced cannot be conceptualised or perceived, let alone produced and communicated, by an individual. Emotion is not a given in human beings, individuals do not all have the same emotional granularity. For a given emotion some will have (and have learned) a very rich concept, including for instance different degrees of intensity, while others will have a poor concept allowing only few nuances. This granularity is reflected in the way emotions are communicated vocally, through the level of complexity of the strategies - notably anatomical - employed to produce emotions.

1.3 Scope of the Thesis

In this section we first present the main research questions which this work attempts to answer in the light of current issues in Voice Conversion (VC). In a second part, we give a non-exhaustive list of current and potential applications of voice conversion.

1.3.1 Current Issues & Research Questions

Two main questions underpin this research: *What is a vocal attitude?* and *How to convert the attitude conveyed in a speech signal?*

Answering these questions requires data, i.e. a collection of audio recordings of utterances conveying various vocal attitudes. Since there was no database specifically dedicated to vocal attitudes available on the first day of this thesis, the gathering of such data was a requirement for any attempt to provide a response to any of these concerns. Thus, this thesis is built out of this initial step in gathering raw material - a dataset dedicated to speech social attitudes.

What Is a Vocal Attitude?

First of all, it should be stated that we will essentially restrict the scope of this question to the field of speech signal acoustic analysis. In this context, we will not undertake the ambitious task of questioning the ontology of vocal attitudes - i.e. what they are in general. Instead, we will attempt to glean what can be inferred about its substance from the available data, following a phenomenological approach. Furthermore, we will leave aside the psychological, cognitive and neurological aspects that underlie the existence of these vocal attitudes. It is important to note that the quest for such a definition is primarily intended to guide our choice of a speech signal representation that is suited to the attitude conversion task, i.e. a representation that feeds attitude conversion algorithms.

In the first place, we may answer this question by identifying the speech parameters that encode attitudinal information in the speech signal and how they encode it, we would thus identify acoustic correlates of vocal attitudes. Presently, there are two questions that underlie the determination of these correlates. The first is to ask how individuals produce these attitudes, i.e. how do individuals use their vocal apparatus to produce acoustic signifiers of the attitudes they want to communicate to others? Within this scope, attitudes are no longer just defined through their acoustic correlates but through the cognitive and anatomical mechanisms that underlie their production and that reveal people's social intentions. The second question aims to understand how individuals perceive the attitudes communicated to them, i.e. to determine the vocal parameters and their variations that allow individuals to decode the attitude conveyed in an utterance. By understanding the mechanisms at work at each end of the communication chain, we expect to provide a definition of vocal attitudes as communicated.

Establishing such a definition makes it possible, on the one hand, to provide knowledge about vocal attitudes. In particular, for each attitude studied we expect a stereotypical profile - i.e. an acoustic signature that ensures the attitude is efficiently communicated on average between individuals - to emerge. This would also enable us to take a reflective look at the data collected, and in particular to assess the overall content of the database with respect to the attitudinal profiles uncovered. An attempt may also be made to identify samples in the database that do not meet the established definition of vocal attitudes. In addition, such a definition ought to be the basis upon which we model vocal attitudes for conversion purposes, i.e., how attitudinal data is represented in order to be efficiently processed by a conversion model. In particular, such a representation should at least reflect the variations in the parameters underlying the production and perception of vocal attitudes. Finally, we may consider assessing the effectiveness of our voice attitude conversion models by comparing the yielded converted utterances with the established attitude definition.

How to Convert the Attitude Conveyed in a Speech Signal?

To begin, we must define what is meant here by *convert*. Similar to a *transformation*, a *conversion* focuses on a single or several characteristics of an object that are meant to be changed. However, a *conversion* differs from a *transformation* in that it maintains consistency between the characteristics of the altered object and unaltered ones so that the object's global nature is preserved. In this case, converting the attitude conveyed by a speech signal means transforming that attitude while preserving all other aspects of that signal and the intrinsic consistency of the signal. This last point is the most critical, it ensures that the converted signal sounds as the same speaker pronouncing the same linguistic message but with a different attitude in a natural way. The preserved aspects thus include the linguistic content, i.e. what is being said, but also the identity of the speaker, his or

her accent and individual speaking style as well as anything else that is part of speech signal but which differs from the attitude. Given how many different and varied attributes a single utterance can express, this constraint is very strong and makes speech conversion challenging.

In order to perform attitude conversion, we adopt the voice conversion neural paradigm which consists in learning a conversion function - in this instance a neural model - that maps source and target representations of speech signals. In most cases, the learning of such a function will be performed on so-called parallel data - i.e. each source utterance matches a target one conveying the same linguistic message and pronounced by the same speaker but conveying another attitude. Voice Conversion is thus formalized as a three-step process: 1) Representations are extracted from speech signals, 2) a mapping is learnt between source and target representations, 3) the converted signal is synthesized from the converted representation. In this specific framework, a twofold question arises: what type of representations and algorithm will we use to learn this mapping?

The answer provided to the first research question we asked - i.e. *what is a vocal attitude?* - would have shed light on the parameters and their variations involved in the communication of attitudes. This acoustic signature should then guide our choice of speech signal representation in the context of attitude conversion. A first option is to adopt a parametric approach - we may, for instance, represent speech signals by using certain speech parameters such as fundamental frequency or spectral envelope and thus convert attitudes by learning to change those parameters. The main issue with this approach is that it is then incumbent on the conversion algorithm to learn how to maintain consistency between these parameters within the converted representation. To do so it must learn an implicit coupling between those, which is not straight forward and often implies telling the model about the relative importance of different parameters in the conversion learning optimization. In the case of an incomplete multi-parametric representation - i.e. one that does not allow for signal reconstruction on its own - or even a single-parametric representation, signal reconstruction involves the use of other parameters from the source sentence that have not been modified and are therefore inconsistent with the converted representation. The synthesis of the converted speech signal will therefore necessarily be incoherent and exhibit sound artefacts. Despite these apparent shortcomings, parametric representations have been and still are widely used for the speech conversion task (Sisman et al., 2017; Sisman and Li, 2018; Luo et al., 2019; Kameoka et al., 2021, 2020). A second option is to adopt a non-parametric representation of speech signal, i.e. that makes no assumptions about the parameters underlying speech signal. Among those, *complete* representations designate the ones from which original speech can recovered accurately such as the linear spectrogram. Another one is the mel-spectrogram which presents the advantage of representing all aspects of the speech signal that are perceptually relevant. The mel-spectrogram representation is *homogeneous* - i.e. each of its components are conceptually similar and therefore of equal importance for conversion learning and there is no conceptual difference between the successive temporal frames that compose it, unlike the F0 which has two definitions depending on whether the temporal segment is voiced or unvoiced. These two qualities make it an ideal representation for the task of speech conversion (Zhang et al., 2020; Qian et al., 2019, 2020; Bous et al., 2022; Bous and Roebel, 2022).

The choice of speech signal representation is decisive as to what can be expected to be changed within the signal, and a fortiori as to how the conversion model works. If, for instance, the established definition of speech attitudes entails considering how various temporal segments - such as phonemes - are compressed or expanded so that a specific attitude is conveyed, then converting such attitudes will require learning to map representations of the speech signal of various duration. Technically speaking, this implies that the conversion model's input and output would not be the

same size. This illustration shows how the design of a neural architecture intended for conversion is influenced by the choice of representation of the speech signal. Answering this second research question is therefore essentially a matter of interpreting the definition of vocal attitudes in technical terms - that is, in terms of neural architecture. This involves determining the set of neural layers to be employed, their type, their location within the architecture and their intrinsic characteristics according to the role they are assigned in the speech attitude conversion task. Inherent machine learning issues related to the amount of data, computational resources or training time are very important to be taken into account when developing a conversion algorithm.

1.3.2 A Wide Range of Applications

While the original motivation of voice conversion could be simply novelty and curiosity, the technological advances from statistical modeling to deep learning made a major impact on many real-life applications. Here is an attempt to give a non-exhaustive list of those - existing and potential - applications. It should be noted that the ethical criterion is set aside here, the idea is to say what is technically possible. Moreover, it seems obvious that some of these applications are potentially undesirable, do not answer any need and are object of potential totalitarian drifts and anthropological mutations (Guerouaou et al., 2021).

Understanding Human Interactions

Firstly, voice conversion algorithms constitute a powerful tool for artificially recreating so-called ecological - daily-life - situations in order to understand the social and cognitive mechanisms that govern our social interactions. One can imagine that studies, for now exclusively based on signal processing to transform speech attributes such as smiling (Ponsot et al., 2018a; Arias2020 et al., 2020) or dominance and trustworthiness (Ponsot et al., 2018b), can benefit from deep learning and the impetus it gives to voice conversion. While many technical issues - e.g. real-time - remain, the ability of deep learning-based voice conversion algorithms to faithfully transform high-level attributes of speech such as emotion or gender is a boon to identifying their mental representations and the role they play in inter-human interactions.

Enhancing Inter-Human Interactions

A wide range of applications for voice conversion systems is dedicated to inter-human distant interaction. Voice conversion algorithms are meant to enhance - if not *improve* - communication in such interactions. For a call center employee who spends hours speaking with unpleasant or irritable clients, a voice conversion system can be used to make the interaction more comfortable for both parties. In this use case, it is usually the call center customer's voice that is converted so as to be perceived less aggressive or distant by the caller. Conversely, the employee's voice could be converted so that it better catches the customers' attention. In this specific call center context, many use cases can be imagined leading to several technical achievements. Additionally, a flurry of applications are to be developed as video conferencing becomes more and more prevalent in our lives, partially as a result of Covid19 and the effects it has had on our communication habits at work and in private life. Amongst them, one could imagine auditory equivalents of Instagram and Tik Tok visual filters, for instance vocal emotional filters could be used to enhance our physically distant expression (thus deprived of many communication aspects) or even hide our anxiety or current mood.

Enhancing Human-Machine Interactions

These days, we are more likely to interact with machines as they grow more prevalent in our lives. This new kind of interaction has given birth to a spate of research works which attempt to understand its specificity. Over the last years, the voice has become the main medium for interaction with the machines in our environment. We are getting used to talking to what is referred to as vocal assistants, embedded in phones, cars or TVs, embodied into a robot or an avatar. The number of use cases is considerable and keeps on growing year after year. In number of those cases, the conversion algorithm would be integrated into a speech synthesis engine and would govern certain aspects of speech such as emotion, speaker identity, accent or even gender. As those aspects are very important to us, humans, in order to fully understand the meaning of an utterance. Within this scope, voice conversion systems would be employed to humanize the inner voice of machines thus pursuing a global objective of communication improvement between humans and machines. For such application, voice conversion algorithms can be used in many different ways such as personalized speech synthesis (Kain and Macon, 1998; Zhang et al., 2019). We might also quote audio books for which voice conversion can be used to adapt synthesized voice to linguistic content and characters (Sini et al., 2022). Nonetheless, these improved conversion algorithms have also led to concerns about privacy and authentication. It thus becomes highly desired to be able to prevent one's voice from being improperly utilized with such voice conversion technologies. Several proposals were made in this regard for speaker de-identification (Srivastava et al., 2020) or even speaker disguise (Huang et al., 2021).

Widening the Scope of Audio-Visual Creations

Voice synthesis and conversion have a wide range of uses in the audio-visual field of audio visual such as film and musical production, which is particularly relevant at IRCAM. One could imagine employing a voice system, for example, to alter an actor's performance after an editing stage of the film making process. This can be done for many reasons such as artistic purposes or to enhance poorly played sequences. Moreover, the reproduction of existing voices - through voice mimicry (Wu and Li, 2014) - for documentaries, movies or TV broadcast may be improved by providing speech synthesizers with voice modules aiming to control specific speech attribute synthesis like speaker identity or emotions.

Providing New Therapeutic Approaches

Finally, it has been demonstrated that vocal self-perception plays a role in the emergence of emotions (Goupil et al., 2019), this effect is often referred to as vocal feedback. Consequently, manipulating people's voice may influence their internal states, at least when people do not detect this manipulation (Goupil et al., 2021a). This discovery may lead to a range of applications, particularly in a therapeutic context. For instance, recent unpublished research shows that the recovery of patients with post-traumatic stress disorder (PTSD) can be enhanced by the use of vocal feedback processes. Based on these initial case studies, a field of research could be developed that would lead to many virtuous applications of speech conversion systems. A variety of therapeutic uses of voice conversion such as communication aids for the speech impaired (Veaux et al., 2013) have already been proposed and many others are yet to be imagined.

1.4 Main Contributions of this Thesis

In this section, I list the main contributions made during these three years of research. It should be noted at this point that this research has been nourished by the permanent desire to take into account the variety of aspects that make up the richness of the research object that constitutes vocal attitudes. Some of these aspects led me to move away from my own disciplinary field in order to better understand the object as a whole.

1.4.1 Designing a French Database of Expressive Speech for Social Attitudes

We propose a new expressive speech database in French dedicated to vocal attitudes: Att-HACK. We recorded twenty actors - twelve females and eight males — playing a hundred different sentences in four different attitudes — friendly, distant, dominant, and seductive. The originality of this new database lies in the great variability it offers in terms of production strategies for a given sentence, speaker and attitude. In addition, it features linguistic parallelism, i.e. for a given speaker, the same linguistic content is played in the four attitudes represented in the database, which makes it easily usable in the context of learning conversion models.

1.4.2 Uncovering the Production Strategies and Perception of Vocal Attitudes

This contribution was carried out in close cooperation with Pablo ARIAS, post-doctoral researcher at Glasgow University, Léane SALAIS, PhD student in the Analysis Synthesis team at IRCAM and Victor ROSI, former PhD student in the Perception and Sound Design team at IRCAM. Additionally, it provided an opportunity to hone teamwork and communication skills, which are crucial for the success of group research. I would also like to add that the need to be understood by researchers working in different fields raises the bar for clarity and aids in a broader understanding of the issue at hand.

Designing an Anatomically Based Method for Speech Production Strategies Assessment

This contribution is built in two stages. The first step consists in the development of a speech signal analysis method that aims at understanding vocal production as a global mechanism that goes from social intent to speech parameters modulation. We mean to exploit the bijective relationship between physiological behavior and vocal parameter variation, such as the speed of vocal cord vibration and F0 fluctuation, to identify such production strategies. Our proposed method consists in the extraction of speech parameters from which certain temporal segments are isolated so as for such parameters' variations to be interpreted in the light of vocal apparatus functioning. In particular, the temporal segmentation is carried out using a neural phoneme-to-audio aligner (Teytaut and Roebel, 2021) proposed by Yann TEYTAUT (also a PhD student in the team). This method is fully replicable on any speech data, thus providing a useful speech analysis tool.

Uncovering the Production Strategies of Vocal Attitudes

The second step entails applying this methodology to the Att-HACK data in order to pinpoint the strategies that underpin the production of vocal attitudes. We thus used speech descriptors and group statistics to uncover quantitative prototypes reflecting the speakers' vocal apparatus control. We showed that French speakers share common production strategies to communicate vocal attitudes such as friendliness, dominance, seductiveness or distance. Notably, to our knowledge,

this is the first study to reveal diverging speech attitude production strategies at the articulatory level. These findings led us to reconsider the question of how to represent speech signals in order to learn speech attitude conversion models.

Understanding the Perception of Vocal Attitudes

We conducted a large-scale BWS experiment on a substantial part of Att-HACK in which 100 participants - each evaluating only one attitude - were asked about their perception of the considered attitude. At the end of the experiment, each evaluated speech sample was ranked on a BWS scale ranging from *the less* to *the most perceived as*. First, this study enables to perceptually validate a large part of Att-HACK, thus providing our conversion models with *clean* data. Second, it provides valuable material that can be used to train perceptual regression models, those models being furtherly used to extend perceputal assessment to unevaluated data. Third, it can be employed for perceptual conditioning of attitude conversion algorithms. Fourth, this study revealed that certain speech attributes such as linguistic content or gender do influence speech attitude perception. Given those findings, we attempted to understand how individuals decode attitudes without reaching a satisfying answer. It is likely that they process speech parameters - or aspects - temporally. We look forward to assessing this assumption in future works.

We also plan to complement with studies on felt attitudes - investigating mental representations - to achieve a full-stack understanding of vocal social attitudes.

1.4.3 BWS-Net: Predicting Perceptual BWS Judgements with Neural Networks

The process by which individuals decode attitude in speech signal is highly complex, as our analysis of the perceptual data gathered throughout the Att-HACK BWS experiment has demonstrated. Although understanding them is very challenging, deep learning based algorithms can be employed to reproduce this process implicitly. The objective of this chapter is to design such algorithms that can "artificially" mimic this process.

Vocal Attitude Neural Recognition Based on A Priori Categories

Before considering the collected perceptual data as training material, we wanted to start with a more classical task of recognizing a priori attitude categories - i.e. labels corresponding to the attitudes targeted by the actors during the Att-HACK recording - from mel-spectrogram representations of speech signals. We thus took inspiration from both speech emotion recognition proposals (Chen et al., 2018) and (Li et al., 2019) to design a baseline architecture dedicated to speech attitude recognition. This contribution features an ablation study that highlights the role of attention mechanism in this specific task of attitude recognition, as well as its impact on the algorithm's performances. The obtained architecture served as a starting point for the following three contributions.

The three following contributions are several attempts to artificially mimic the process by which people derive the attitude elicited by an utterance from the speech signal. We no longer consider a priori attitude categories but Att-HACK's attitudes as perceived by individuals.

Design of a Perceptual Regression Model Dedicated to Vocal Attitudes

The first attempt is to design a regression model that predicts the perceptual scores obtained during the BWS experiment carried out on Att-HACK from mel-spectrograms. The model's performance has been shown to be significantly influenced by speaker identification and linguistic content. The Gaussian distribution of BWS scores drives the model to focus on average scores and causes poor predictions for low and high scores. Despite attempts to overcome this learning bias, the model's performance remains relatively modest for the prediction of extreme scores, which makes it unusable for Att-HACK validation purpose, especially for the identification of sentences whose attitude has been miscommunicated. This observed limitation led us to raise questions about the meaning of low and high perceptual scores.

Design of a Perceptual Domain Classifier Dedicated to Vocal Attitudes

This contribution continues where the previous one left off. Here, a domain, or range of scores, rather than a single score, is what is being sought after. As shown in the case of regression, we found great influence of speaker identity as well as linguistic content on the model's performance. In contrast to the regression model, this classification model enables the identification of poor outcomes - i.e. utterances in which the attitude is poorly communicated (low BWS scores) - as well as top ranked outcomes - i.e. utterances that particularily well communicated. In light of the model's performances, we discuss the existence of several perceptual domains, highlighting variations in the nature and degree of the attitude conveyed.

Design of a Perceptual Arranger Dedicated to Vocal Attitudes

Since BWS scores can be considered as projections of actual judgements made by the individuals, we can make the assumption that the first contain less information than the latter. In other access proper mental representations of attitudes and the underlying perceptual space formed by the individuals judgements, we attempted to directly use the raw judgements as training material. We designed a cost function that enables to interpret these judgments as relationships between the distances separating specific points - i.e. speech samples - of a latent space. Through optimizing a model with respect to this custom loss, we have been able to learn a latent space whose structure reflects the judgements made the participants of our BWS experiment. Although this constitutes a significant theoretical contribution - a method than can applied to any BWS data - its application to the perception of attitudes only yields average results. In particular, the model's performance on data not seen during training is modest, we present some possible explanations for this limitation.

1.4.4 Sequence-to-Sequence Neural Conversion of Speech Attitudes

Neural Conversion of Vocal Attitudes Based on Speech Fundamental Frequency

We propose an end-to-end architecture to learn efficiently conversions between attitudes from F0 contours only represented by Continuous Wavelet Transforms (CWT). The designed neural architecture brings together the F0 decomposition and the dual-GAN into a single network, so that the CWT decomposition is optimized in the sense of the dual-GAN objective, and combining separation and reconstruction losses of the resulting decomposition. An application to the voice conversion of social attitudes shows that the proposed approach significantly improves the quality of the conversion by comparison with the CWT-AS (Luo et al., 2019) approach.

We first reevaluated how to represent speech signal for conversion purpose in the light of our careful examination of the strategies used by speakers in Att-HACK to produce speech attitudes. If the whole signal were to be altered, each component of the speech signal should be reflected in the representation. We therefore chose to use the mel-spectrogram as an intermediate representation of the speech signal. A newly proposed neural vocoder (Roebel and Bous, 2022) enables an accurate reconstruction of the signal from this representation.

Adaptation of a State-of-the-Art Voice Transformer Network (VTN) to the Specific Case of Speech Attitude Conversion

We adapted the transformer based architecture proposed in (Kameoka et al., 2021) in the context of speaker identity conversion to our attitude conversion issue. We place ourselves in the so-called many-to-many conversion paradigm, i.e. we learn to convert from any attitude to any other attitude with a single training. For this and the next contribution, we learnt conversions on a single speaker, the many-to-many aspect being enough challenging. Despite the already high quality of the conversions yielded, we identified a substantial issue with linguistic content loss in several of those conversions.

Linguistic Conditioning of the Speech Attitude Conversion Algorithm

Intelligibility appears to be the first criterion that a conversion must satisfy. In other words anyone must be able to decode the linguistic content conveyed in a converted utterance. We have therefore sought to solve this shortcoming encountered in conversions. In order to achieve this we have worked on incorporating a speech recognition module, i.e. a speech-to-text, into our conversion algorithm. Objective measurements have shown the effectiveness of this solution in preserving the linguistic content of the conversions yielded. We also conducted a perceptual experiment that confirmed the trends observed through objective assessment. At the end of this thesis, we dispose of an efficient algorithm for the conversion of vocal attitudes.

1.5 Outline of the Thesis

The document is organized into six interrelated Chapters whose succession is conceptual, i.e the order of Chapters serves to put forward a global vision of the problem posed. This progression is therefore not chronological, as many of the works presented in distinct chapters have been carried out simultaneously.

A state-of-the-art on voice conversion and speech emotion recognition is provided in Chapter 2. A formalisation of both tasks of speech conversion and speech attribute - e.g. attitudes or emotion - recognition is proposed as well as a review of the related literature highlighting the neural network based approaches on which this research is founded. The design of the French dataset Att-HACK dedicated to speech attitudes as well as a first metadata and data analysis is presented in Chapter 3. This dataset can be seen as the foundation on which this while research is built. Chapter 4 - as a next logical step - outlines two studies that provides a first twofold account for how speech attitudes are communicated by individuals. The first one aims to uncover the production strategies of speech attitudes while the other mainly hints at biases involved in the perception of such attitudes. Those two studies provide invaluable information about the Att-HACK data. In Chapter 5, a first attempt of speech attitude conversion based only on the F0 is proposed, it involves an end-to-end neural architecture that brings together the F0 decomposition and the dual-GAN into

a single network. The transition between chapters 5 and the last two chapters - 6 and 7 - of the document marks a paradigm shift both with regards to the modelling of speech signals and to the method employed to learn attitude conversions. Then comes Chapter 6 which proposes to design algorithms to predict attitudinal traits with using the perceptual data gathered in the second study of Chapter 4. The last Chapter 7 provides a new attempt of speech attitude conversion based on Voice Transformer Network - fed with mel-spectrograms - that is linguistically conditioned to improve intelligibility of the conversions.

A general conclusion finally gives a clear overview of the whole document, points limitations of this research and provides insights for future works.

1.6 List of Publications

The research conducted during this thesis led to the publication of four papers in major international conferences.

- Le Moine, C. and Obin, N. (2020). Att-HACK : An Expressive Speech Database with Social Attitudes. In *Speech Prosody*, Tokyo, Japan.
- Le Moine, C., Obin, N., and Roebel, A. (2021). Speaker Attentive Speech Emotion Recognition. In *Proc. Interspeech 2021*, pages 2866–2870. Brno, Czech Republic.
- Le Moine, C., Obin, N., and Roebel, A. (2021b). Towards end-to-end F0 voice conversion based on Dual-GAN with convolutional wavelet kernels. In *EUSIPCO*. Dublin (virtual), Ireland.
- Salais, L., Arias, P., Le Moine, C., Rosi, V., Teytaut, Y., Obin, N., and Roebel, A. (2022). Production strategies of vocal attitudes. In *Proc. Interspeech 2022*, pages 4985–4989. Icheon, Korea.

Nevertheless, many of the results presented in this paper have not yet been published and will be submitted for peer review in the coming months. In particular, we are working on a journal paper that deals with the use of perceptual data for enhancement and control of speech attitudes neural conversion.

Chapter 2

STATE-OF-THE-ART

Contents

2.1 Voice Conversion 32
2.1.1 General Scheme for Voice Conversion 32
2.1.2 Representing Speech Signals 34
2.1.3 Voice Conversion Algorithms 37
2.1.4 Metrics and Methodology for Evaluating Voice Conversion Algorithms . . 43
2.1.5 Section summary 45
2.2 Speech Attitude Recognition 47
2.2.1 General Scheme of Speech Attitude Recognition 47
2.2.2 Speech Emotion Recognition Algorithms 48
2.2.3 Methodology for Evaluating Speech Emotion Recognition Algorithms . . 49
2.2.4 Section Summary 51

2.1 Voice Conversion

The main objective of this work is to design an algorithm that converts attitude in speech signal while preserving other speech attributes such as speaker identity and linguistic content. There are various research paths that might be explored to develop such an algorithm. All of them are more or less related to the broad research area of speech synthesis, i.e. the study of speech artificial production. Speech synthesis is declined in many sub domains such as Text-to-Speech (TTS) synthesis, in which speech is synthesized from text, or Voice Conversion (VC), in which every aspects of speech are kept unchanged except from one that is converted. In this research we will mainly focus on this specific synthesis sub-genre of voice conversion.

2.1.1 General Scheme for Voice Conversion

In this subsection, we describe the typical speech conversion flow in three steps: speech analysis, determination of a conversion function that maps source and target representations of speech and final recovering of speech signal.

Speech Analysis

Here, we will restrict our discussion to speech signal representations intended for speech conversion or transformation. Reconstructibility, or the ability to synthesize a signal from its representation, is one of the fundamental characteristics that must considered when selecting a speech signal representation. The interpretability of this representation, or whether it makes sense from the perspective of the mechanisms of production or perception of the speech signal, is another crucial consideration. Finally, in the context of speech conversion, the representation used must be compact - that is, it must reflect all aspects of speech and allow for efficient processing by an algorithm.

As a time-varying signal, the speech signal is frequently represented in its time-frequency form, for instance, by using a Short-Term Fourier Transform (STFT). STFT provides the time-localized frequency information of the speech signal, as its frequency components vary over time. The STFT of a speech signal is a 2D-matrix of complex numbers whose temporal and frequency components are called frames and bins respectively. More precisely, each frame is obtained by applying the Fourier Transform (FT) to a temporal region of the speech signal segmented through windowing. The real part of this STFT - i.e. the amplitude spectrogram - is commonly used to represent the speech signal, the imaginary part representing its phase. The STFT is invertible, thus a speech signal can be accurately reconstructed from its STFT. The mel-spectrogram has become the dominant representation in speech conversion thanks to its advantages of being highly compressed and reflecting all aspects of speech. It is obtained through frequency filtering by a mel filter bank of the magnitude spectrogram. In addition to these signal-based representations, which make few or no assumptions about the mechanisms that govern speech production, model-based representations of speech signal have been proposed. In this representations, it is assumed that the production of the speech signal is based on a physical model, e.g. the filter source model (Fant, 1970; Markel and Gray, 1982), then each frame of the speech signal is described by a set of parameters of this model.

In the following, we introduce a formal framework that will be used throughout this document. We denote $\mathcal{A}$ the function which, to a speech signal $\mathbf{x} \in \mathbb{R}^T$, associates its representation $\mathbf{X} \in \mathbb{R}^{T' \times D}$ such as

$$\mathbf{X} = \mathcal{A}(\mathbf{x}) \tag{2.1}$$

where T' and D are respectively the number frames (not necessarily different from T) and the number of features of the representation yielded by $\mathcal{A}$. In the following we use T for T' for the sake of clarity.

Conversion or Mapping Function

Once the source and target speech signals $\mathbf{x}^s \in \mathbb{R}^{T_s}$ and $\mathbf{x}^t \in \mathbb{R}^{T_t}$ are represented by $\mathbf{X}^s \in \mathbb{R}^{T_s \times D}$ and $\mathbf{X}^t \in \mathbb{R}^{T_t \times D}$ respectively, a conversion function $\mathcal{C}$ - also referred to as mapping function - can be introduced such that its application to the source speech representation yields a converted speech representation $\mathbf{X}^{s \leftarrow t} \in \mathbb{R}^{T_c \times D}$ in which every aspects of the source utterance are preserved except from the converted attribute of which instance in $\mathbf{X}^{s \leftarrow t}$ must match the target utterance. The conversion can thus be formulated as

$$\mathbf{X}^{s \leftarrow t} = \mathcal{C}(\mathbf{X}^s) \tag{2.2}$$

We will see in the next section that this mapping function can be approximated in various ways ranging from statistical models to deep neural networks. In particular, those networks can be either deterministic, i.e. the model's output is entirely determined by input utterance, or probabilistic, i.e. the model's output matches a probability distribution which is determined by input utterance. Since deep learning seems to have outperformed all other methods for the speech conversion task (Desai et al., 2009; Nakashika et al., 2013; Luo et al., 2016), we will almost exclusively focus on this latest approach in this work. When considering such deep learning approaches, the conversion or mapping function is referred to as a neural network. Then, as depicted on Figure 2.1, two phases must be distinguished.

- The **training phase** : the model $\mathcal{C}$ is being fed with pairs of representations $\{\mathbf{X}^s, \mathbf{X}^t\}$ which it uses to learn a mapping between the source and target utterances.
- The **inference phase** : the trained model $\mathcal{C}$ is applied to the source representation $\mathbf{X}^s$ and yields a conversion $\mathbf{X}^{s \leftarrow t}$.

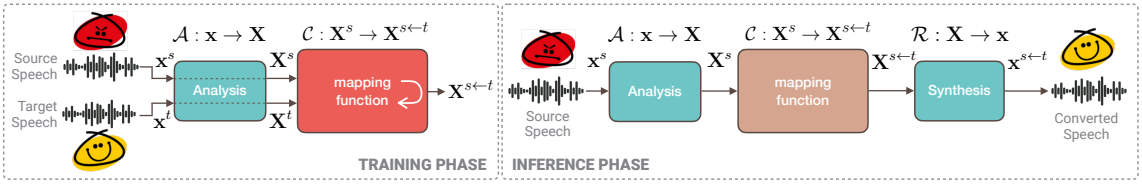

Figure 2.1: General VC scheme with training (left) and inference (right) phase. The red box represents the training of the mapping function, while the pink one applies the mapping function at the inference, in a 3-step pipeline process $\mathbf{X}^{s \leftarrow t} = (\mathcal{R} \circ \mathcal{C} \circ \mathcal{A})(\mathbf{X}^s)$

Depending on the paradigm chosen, the conversion may or may not apply to the temporal dimension, thus modifying or not the duration of the source speech signal. Ideally, the conversion model is able to change this duration such that $T_c = T_t$.

Speech Synthesis

Generating the speech signal from the converted source representation is the final step of this conversion process. Basically, the function that allows to retrieve the original audio signal from its representation is the inverse function of the speech analysis function $\mathcal{A}$. Nevertheless, a distinction must be made between complete representations such as the STFT - from which the signal can be perfectly reconstructed - and incomplete representations such as the amplitude spectrogram or the mel-spectrogram - which are not formally invertible. In the case of incomplete representations, methods allowing a more or less accurate reconstruction of the speech signal have been proposed - such as the Griffin-Lim algorithm (Griffin and Lim, 1984) for the amplitude spectrogram or neural vocoders (Airaksinen et al., 2018) for the mel-spectrogam. It should be noted that in our case the synthesis is performed after the conversion of a signal representation. Synthesis is therefore performed from a representation $\mathbf{X}^{s\leftarrow t}$ which is itself an approximation of the target representation $\mathbf{X}^t$. As a result, there are two potential sources of inaccuracy for the converted signal: conversion and synthesis. Denoting $\mathcal{R}$ the reconstruction function, then any speech representation $\mathbf{X}$ derived from $\mathcal{A}$ can be - more or less approximately - recovered as

$$\hat{\mathbf{x}} = \mathcal{R}(\mathbf{X}) \tag{2.3}$$

The general scheme for voice conversion can be formulated in the composition of these three functions ($\mathcal{A}$, $\mathcal{C}$ and $\mathcal{R}$) in a three-block sequential process

$$\mathbf{X}^{s\leftarrow t} = (\mathcal{R} \circ \mathcal{C} \circ \mathcal{A})(\mathbf{X}^s) \tag{2.4}$$

In the following of this section, we mean to provide a state-of-the art for each of those functions. The analysis and synthesis functions are first introduced. Then we introduce the conversion functions.

2.1.2 Representing Speech Signals

This part provides a state of the art of speech signal analysis-synthesis ($\mathcal{A}$ and $\mathcal{R}$ in 2.1.1) methods involved in the speech conversion context. The vast majority of these methods are derived from signal-based representations, physical model-based representations being hardly used today.

The Vocoder. Whether based simply on the signal or derived from a physical model of speech production, speech representations are yielded through the use of what is called a vocoder. A vocoder, i.e. a voice encoder, is a model for encoding a speech signal into a temporal parametric representation from which it can be reconstructed more or less accurately. Research into vocoders has a long history, while originally designed using electrical circuits, the majority of the known vocoders are now implemented in software. The following provides an overview of the methods proposed from the early 90' to the present day.

Model-Based Representations

The model-based approach assumes that a speech signal signal can be mathematically represented by a model whose parameters vary with time. In a vast majority, the speech signal representations that are based on physical models derived from the filter source model of speech production (Markel and Gray, 1982; Fant, 1981) and obtained through the use of a vocoder. The source-filter model postulates that any speech signal can be described as the filtering via the vocal

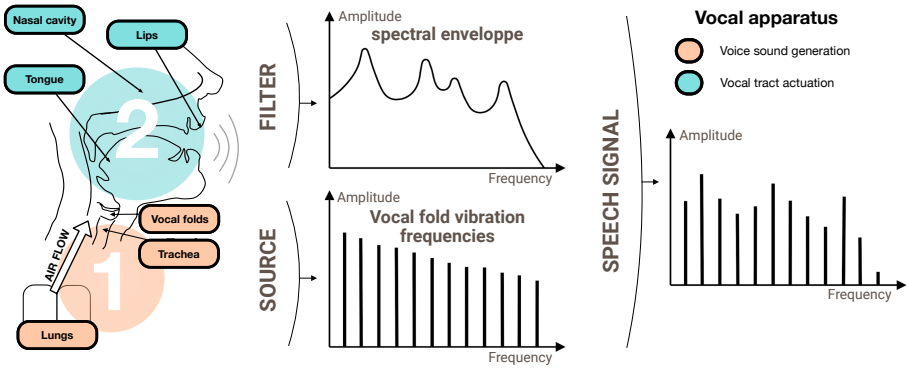

Figure 2.2: Schematic view of the source-filter model of speech production after (Fant, 1970; Markel and Gray, 1982) and adpated from

tract (filter) of an excitation produced by the larynx (source). An illustration of this model is proposed in Figure 2.3.

Although the source-filter model - on which we based our study of speech attitude production in Chapter 4 - provides a good understanding of speech production, the representations derived from it and, in general, physical model-based representations did not prove to be efficient enough in terms of both calculation and reconstruction quality.

Signal-Based Representations

Since the beginning of the 90', a large number of vocoders involving signal-based representations have been proposed. Among the first major proposals, we find the phase vocoder (Flanagan and Golden, 1966), sinusoidal models (McAulay and Quatieri, 1986), and pitch-synchronous overlap-add (PSOLA) (Moulines and Charpentier, 1990). Later many more or less elaborate systems were proposed (Quatieri and McAulay, 1992; Kawahara, 1997; Zen et al., 2009; Roebel, 2010; Degottex et al., 2013; Morise et al., 2016).

STRAIGHT or "Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum", proposed by Kawahara et al. in 1999 (Kawahara, 2006), is one of the popular vocoders in speech synthesis and voice conversion. It decomposes a speech signal into: 1) a smooth spectrogram in which any traces of interference caused by the signal periodicity are eliminated in both temporal and frequency dimensions 2) a fundamental frequency (F0) contour which is using an instantaneous-frequency-based technique; and 3) an aperiodicity map which captures the temporal and frequency characteristics of the noise. It was one of the first vocoders to allow a re-synthesis whose timbre was no longer described as artificial, thus prompting several attempts at voice conversion (Toda and Tokuda, 2005; Sisman and Li, 2018). In a similar way the WORLD vocoder (Morise et al., 2016) breaks down the signal into three parts but calculate the vocal chord vibrations on the basis of the convolution of the minimum phase response and the extracted excitation signal while STRAIGHT uses periodic and aperiodic responses independently to compute them. It offers the same high quality synthesis while limiting the number of calculations, thus reducing the synthesis computation duration.

The trend in voice conversion has long been to change the speech signal by mean of interact-

ing with the parameters. In this case, the interactions between altered and unaltered parameters needed to be handled explicitly but no method was found to achieve this. While the division of a voice model into a source and a filter allows a simple and schematic representation of the functioning of the vocal apparatus, it has the disadvantage of eluding the complex physical coupling between these two anatomical entities. For instance and as pointed out in (Roebel and Bous, 2022), in real life a change in pitch will generally be accompanied by changes in glottal pulse form, formant positions, intensity and noise level. Therefore, an important challenge of this research in the quest of a voice model formulation is the fine representation of the interaction between the source and the filter.

Neural vocoders

The arrival of deep neural networks gave rise to numerous attempts to represent such complex and non-linear coupling. The idea is then to approximate the synthesis function $\mathcal{R}$ by a neural network which, trained on a large amount of speech data, learns to reconstruct the speech signal from its representation with a minimal error. Due to the training on real world data, neural networks are expected to reflect all the complex inter-relations between different speech signal components. Recently, a deep neural network for generating raw audio waveforms called *WaveNet* has been proposed (Van Den Oord et al., 2016). The model is fully probabilistic and autoregressive, with the predictive distribution for each audio sample conditioned on all previous ones. It is composed of many residual blocks, each of which consists of 2×1 dilated causal convolutions ensuring the model cannot use future information, gated activation functions and 1×1 convolutions. This model has shown remarkable performance for the specific task of neural vocoding. When conditionned on mel-spectrogram representations of speech, the *WaveNet* yields syntheses that are almost perceptually indistinguishable from real speech signals (Shen et al., 2018). The neural reconstruction of the mel-spectrogram's related time-domain signal makes no assumptions about the model underlying speech production. Nourished by numerous examples during training, the neural vocoder itself implicitly learns the underlying acoustic structure of speech signal and the complex coupling of its parameters. Neural vocoding can be referred to as a data-driven method.

The quality of the reconstructions resulting from (Shen et al., 2018) gave rise to strong interest in the quest for a what might be called a *universal neural vocoder*. A universal neural vocoder is expected to support perceptually transparent analysis/resynthesis for arbitrary speakers, emotions, languages and voice qualities whether spoken or sung. In line with this objective and beyond the problems of computing time and costs at training and inference phases, recent research activities have started to investigate multi-speaker vocoders (Gibiansky et al., 2017; Park et al., 2019) demonstrating that multi-speaker neural vocoders can generalize not only to unseen speakers, but also to unseen languages and expressivity (Jiao et al., 2021). In the specific case of vocal attitude conversion, having a universal vocoder appears mandatory, insofar as it allows the synthesis of conversions conveying a variety of attitudes that had not necessarily been seen during at vocoder's training.

Hybrid Neural Vocoders

With the aim of designing sober vocoders - i.e. involving affordable computing resources which does not require the use of a super computer - a new trend today is to create hybrid vocoders based on both deep neural networks and elementary signal processing blocks, thus avoiding unnecessary learning of correlation already modelled by proven signal processing techniques. These systems incorporate signal processing operators implemented in such a way that they are differentiable - e.g. the DDSP python package for differentiable digital signal processing (Engel et al.,

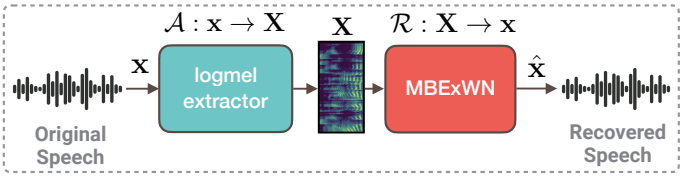

Figure 2.3: Schematic view of the analysis-synthesis framework based on the Multi-band Excited *WaveNet* proposed by (Roebel and Bous, 2022) and used in Chapter 7

2020) - and can be therefore trained in end-to-end. For instance, (Engel et al., 2020) implements an additive sinusoidal model whose parameters are derived from a deep neural network, the ensemble being trained in the manner of an auto-encoder. As pointed out in (Roebel and Bous, 2022), these hybrid models seem to provide a solution to two problems inherent to the representation of speech signals. First, it solves the issue of high-level control of signal-based speech models. Second, the filtering carried out by the signal processing blocks allows to significantly reduce the amount of data required to generate speech signals of similar quality.

Pursuing this double objective of a universal structuring of representation, (Roebel and Bous, 2022) proposed a Multi-band Excited *WaveNet* introducing an automatic and adaptive signal normalization - allowing the model to work independently from the signal energy - and based on a source filter speech model. From one hand, the excitation source generation is obtained through a first convolutional neural network which, feed with a mel-spectrogram, ouputs a F0 contour, which is turned into a quasi-periodic excitation and passed to a WaveNet - conditioned on the mel-spectrogram - that properly forms the excitation pulse. On the other hand, a second neural network generates mel-cepstral filter coefficients which are then converted into a spectral envelope through applying a Discrete Fourier Transform. With addition of a few extra steps, both source and filter outputs are multiplied to form the desired speech signal. This model has been shown to compare favorably to the state-of-the-art in its capacity to generalize to unseen voices and voice qualities.

These latest findings, especially the team-made proposal (Roebel and Bous, 2022), are a real boon to this research. We will therefore use this neural vocoder to synthesise speech signals whose conveyed attitude has been converted in the last chapter 7 of this paper.

2.1.3 Voice Conversion Algorithms

Over the past years, a variety of systems have been proposed for the specific task of voice conversion. In this section we will first present statistical modelling approaches before focusing more extensively on deep neural network based approaches, which will serve as the research's preferred paradigm.

The main determinant for the development of those algorithms is the the availability of so-called parallel data. While dealing with conversion of emotion attribute in speech, a parallel set of data would provide for each utterance $\mathcal{X}^s$ in attitude a_s, its corresponding version $\mathcal{X}^t$ in attitude a_t, i.e everything in $\mathcal{X}^t$ is identical to $\mathcal{X}^s$ at the exception of attitude, notably both utterances must be pronounced by the same speaker and convey the same linguistic content. Due of the expense and time involved in obtaining such databases, this requirement is exceedingly challenging to fulfill. This is even more true in the case of attitudes where the parallelism is very difficult to control. To

overcome this need, several methods have been proposed, we present them in a last part.

Statistical modelling for Voice Conversion

The first attempts of Voice Conversion based on statistical approaches were proposed in the 1980s (Stylianou and Cappe, 1998). The basic idea behind those systems was to represent the relationship between the source and target utterances by a statistical model. Two broad categories can be distinguished among these approaches : parametric and non-parametric approaches. The former make assumptions about the statistical distributions that underlies the considered speech features and how they are mapped when the latter make only little or none assumptions.

Among the variety of parametric approaches, Gaussian Mixture Models (GMM) have extensively used to model the relation between the source and target sets of features used for the conversion. To do so, a Gaussian Mixture Model assumes that each feature vector can be generated from a mixture of a finite number of Gaussian distributions with unknown parameters and thus model the joint probability density of the paired feature vector sequence which represents the joint distribution of source and target speech signal. The conversion parameters are estimated using Minimum Mean-Square Error (MMSE) on the source-target pairs of the training set. First introduced for speaker identity conversion in (Stylianou and Cappe, 1998), GMMs were then employed to model the relation between source and target speakers acoustic spaces by using HNM-based (Harmonic + Noise Model) representations of the speech signal. A number of improvements in this approach have been proposed. Stressing the importance of the analysis-synthesis paradigm underlying voice conversion, (Kawanami et al., 2003) proposed to use GMMs along with Dynamic Frequency Warping (DFW) directly on the features produced by STRAIGHT, thus avoiding over-smoothed and muffled converted speech. Toda (Toda et al., 2007) proposed a voice conversion algorithm based on the estimation of spectral parameter trajectory thus better considering local correlation of features between successive frames. Only few proposals focused on the specific case of speech emotion conversion (Kawanami et al., 2003; Veaux and Rodet, 2011). (Veaux and Rodet, 2011) proposed an intonation conversion system from neutral to expressive , they only focuses on F0, represented through Discrete Cosine Transform (DCT) at different temporal scales. Gaussian mixture models are used to map the prosodic features between neutral and expressive speech, and the converted F0 contour is generated under the dynamic features constraints.

Conversely, among non-parametric methods, the Vector Quantization (VQ) approach was introduced for speaker identity conversion in (Abe et al., 1988). The method consists in mapping source and target codewords, i.e. approximations of vector features by their nearest vector in a codebook, source and target being related to different codebooks. To our knowledge it has never been used for speech emotion conversion. As an examplar-based approach, Non-Negative Matrix Factorization (NMF) has been proposed for Voice Conversion. NMF is based on sparse representations, the observed magnitude spectrogram is represented by a linear combination of a small number of atoms. Successful implementation includes non-negative spectrogram deconvolution (Wu et al., 2013), locally linear embedding (LLE) (Wu et al., 2016), and unit selection (Jin et al., 2016; Obin et al., 2018). In NMF-based approaches, a target spectrogram is constructed as a linear combination of exemplars which may result in over-smoothing issues. Phonetic sparse representation (Sisman et al., 2017) is an extension to sparse representation for voice conversion. It is built on the idea of phonetic sub-dictionaries, and dictionary selection at run-time. The study shown that multiple phonetic sub-dictionaries consistently outperform single dictionary.

The quality and naturalness of the conversions yielded by these statistical approaches has re-

mained quite low and has not managed to deceive human perception. Around 2015, the advent of deep neural networks has been a game changer for the voice conversion task, as it has for many others leading to an important qualitative leap. Thanks to this new paradigm, the VC research community has been enabled to adress old and new challenges for converting speech.

Neural Voice Conversion

The use of deep neural networks in voice conversion task has not been straightforward, due to the specific constraints relative to voice conversion. In particular, deep neural networks based methods are known to require huge amounts of data, meanwhile, databases dedicated to speech conversion are generally small, rather rare when it comes to converting speaker identity (Yamagishi, 2017) and extremely rare for other speech attributes such as emotion (Burkhardt et al., 2005; Zhou et al., 2022) or accent (Kalluri et al., 2021). In addition, the almost systematic need for very costly and time consuming parallel databases has been a strong limitation for the use of deep neural networks to voice conversion tasks. These considerations initially seem to contradict the decision to use machine learning for speech conversion.

The parallel development of synthesis tasks via neural vocoders and voice conversion has contributed to the decomposition of the voice conversion issue into two sub-problems: 1) learning the conversion function $\mathcal{C}$ from compressed representations and 2) the synthesis via a neural vocoder $\mathcal{R}$ of the converted speech signal from this representation. As a result, the voice conversion issue seen through the prism of deep neural networks must be thought of as an ancillary task to the more fundamental task of speech analysis-synthesis. Indeed, the recently developed neural vocoders can now be learnt on huge amounts of data, allowing speech signals to be encoded in highly compressed representations - such as mel-spectrograms (Shen et al., 2018; Roebel and Bous, 2022) - from which they can be recovered very accurately. Therefore, Voice Conversion systems can refrain from addressing certain issues related to what constitutes the foundations of a speech signal : *What is a phoneme ?* or *How noise is distributed in the speech signal ?*. These tasks are taken over by the neural vocoder which can then be connected directly to the output of the voice conversion system. Consequently, the role of the Voice Conversion system is to model high-level attributes of the speech signal while that of the neural vocoder is to model the low-level attributes. Deep learning based methods have several advantages over standard statistical methods. They allow a non-linear mapping between source and target features that can better match the complexity of the attribute whose conversion is being modelled. This is particularly interesting when dealing with highly complex attributes such as emotions that cannot be described by basic speech features taken solely. Such deep models are also less restrictive on the size and number of features that can be modelled, for instance high-dimensional features, such as mel-spectrogams - for instance featuring 80 frequency bins.

The early attempts of neural voice conversion were deep neural networks. A speaker identity conversion based on deep neural networks has been proposed in (Desai et al., 2009) and achieved better results in transformation than the standard GMM-based methods when applied to spectral features. In (Nakashika et al., 2013), a deep belief network is employed to learn speaker identity abstractions from which a deep neural network learns conversions. (Luo et al., 2016) proposed a system for converting emotion in speech, it involves a deep belief network to learn mel-cepstral coefficients and a neural network to learn normalized segment-F0 features (NSF0) which allows for emotional prosody conversion. Those models did not allow for temporal dependencies modeling, which greatly limited their performances, especially in the case of emotion conversion. Indeed, as described in section 1.2.3, speech prosody such as the time variation of pitch (Banse and

Scherer, 1996; Bachorowski and Owren, 1995; Bänziger and Scherer, 2005) and rhythm (Mairesse et al., 2007) is of primary importance in the expression of emotions by speech.

To tackle this limitation, another kind of networks were introduced for voice conversion, namely recurrent neural networks (RNN). The advent of those recurrent networks enabled voice conversion systems to model the temporal correlation across speech frames. In particular, the long-short term memory (LSTM) cells (Hochreiter and Schmidhuber, 1997) were shown to temporally widen the contextual integration allowed by vanilla recurrent networks and achieved great improvements in naturalness and temporal consistency of the yielded conversions (Sun et al., 2015; Ming et al., 2016).

First, recurrent networks do not by themselves allow sequences of different duration to be mapped. In most cases, the mapping between source and target utterances was learned on aligned, time-stretched representations. The alignment between the source and target utterances can be done linearly or with respect to sentence related linguistics e.g. phoneme wise. The need for alignment as well as the inability of mapping sequences of different duration is a very limiting factor, even more when dealing with emotion conversion. The alignment step being error-prone and requiring linguistic knowledge of the data to be efficient.

Second, the newly proposed recurrent networks gave rise to a specific kind of architecture, referred to as sequence-to-sequence (Seq-2-Seq) modelling (Sutskever et al., 2014). Initially used for the task of translation, it was later introduced for voice conversion (Obin et al., 2019), the latter of which could finally be seen as a translation. This architecture maps the complete temporal sequences rather than just map source and target utterances frame by frame. An encoder produces fixed-size code representing the aspects of the source representation that are meant to be preserved. This code, usually the last internal state of the encoder's last recurrent layer, is then passed to a decoder that yields the converted representation. The classic problem with this architecture is that the temporally propagated information tends to vanish for long sequences, i.e. the last internal state of the encoder's last recurrent layer no longer contains information about the first temporal temporal frames.

Third, solving this issue led to the introduction of a second generation of Seq-2-Seq architectures with attention mechanism. The attention mechanism aims to implicitly learn an alignment between the source and target sentences. No assumptions are made about how the linguistic units are mapped, the model is left to learn the alignment that will maximize the conversion performance from source to target domain. Most attentive algorithms are based on Seq-to-Seq architecture: an auto-encoder mediated by an attention mechanism. Within this context, attention is used to learn a combination of the encoder's recurrent internal states thus helping to predict each decoder's temporal step. By learning such a combination, the attention mechanism focuses on salient temporal information from the encoder and past decoder time steps. With this mechanism, the code produced by encoder is an actual temporal summary of the whole source sequence, thus improving the model's conversion performance. This kind of architecture has been used in (Obin et al., 2019) for emotion pitch contours conversion, interpolated pitch sequences of different duration are mapped and conversion is learnt from one emotion to another. In (Zhang et al., 2019), an analog architecture is used to learn a mapping of between source and target mel-spectrograms for speaker identity conversion. The converted signal are recovered from converted mel-spectrograms using a WaveNet based neural vocoder as introduced in (Shen et al., 2018). Improvements were made by adding constraints during training (Tanaka et al., 2019), notably on the shape of the attention matrix with the guided attention loss (Tachibana et al., 2018).

Fourth, while recurrent neural networks represent an effective implementation for voice conversion, recent studies have shown that convolutional neural networks (CNN) with gating mechanisms also learn well the long-term dependencies (Gehring et al., 2017; Kameoka et al., 2020). The main constraint being to force the decoder's convolutions to be causal so as each frame can be predicted

only using the already predicted frames (and not the ones in the future) at inference. Moreover, using dilated convolutions broadens the temporal receptive range of the model and thus captures the context at the sentence level (Kameoka et al., 2020).

Finally, the transformer networks (Vaswani et al., 2017), of which principle is a generalization of attention mechanism that basically replaces all recurrent layers by self attention mechanisms in a network, have been proposed for Voice conversion yielding noteworthy performance in terms of sound quality (Kameoka et al., 2021). A variational version has been proposed in (Chen and Zhang, 2021) along with a connectionist temporal classification (CTC) loss used to encode each utterance phoneme per phoneme. In (Lee et al., 2022) a transformer learns to encode a mel-spectrogram into a sequence of phoneme-related vectors. From each vector a sequence of converted mel-spectrogram frames is decoded with controlling each phoneme duration throughout conversion.

Towards Parallel Data Free Voice Conversion Algorithms

Facing the difficulty of obtaining parallel data dedicated to voice conversion, many researchers have sought to circumvent this constraint. The strategies proposed to get rid of parallel data were inspired by research on image translation from unmatched images.

Inspired by the field of image-to-image translation, (Kaneko and Kameoka, 2018; Fang et al., 2018) proposed the first studies with a Cycle-GAN for parallel-free data Voice Conversion. Cycle-GAN is based on the concept of adversarial learning (Goodfellow et al., 2014), which is to train a generative model to find a solution in a min-max game between two neural networks, called as generator (G) and discriminator (D). The adversarial loss measures how distinguishable between the data distribution of converted features and source features $\mathbf{X}^s$ or target features $\mathbf{X}^t$. The closer the distribution of converted data with that of target data, the smaller the loss becomes. The adversarial loss only tells us whether $G_{s\to t}(\mathbf{X}^s)$ follows the distribution of target data and does not ensure that the contextual information, that represents the general sentence structure we would like to carry over from source to target, is preserved. To encourage that we maintain the consistent contextual information between $\mathbf{X}^s$ and $G_{s\to t}(\mathbf{X}^s)$, the cycle-consistency loss is introduced. This loss encourages $G_{s\to t}$ and $G_{t\to s}$ to find an optimal pseudo pair of $\{\mathbf{X}^s, \mathbf{X}^t\}$ through circular conversion. Experimental results shown that, with non-parallel training data, Cycle-GAN achieves comparable performance to that of GMM-based system that is trained on twice amount of parallel data (Kaneko and Kameoka, 2018).

Preserving linguistic content while converting a source utterance is one of the key challenge of voice conversion and a fortiori the main reason for the inescapable need for parallel data. A text-to-speech system allows to produce speech from text, i.e. linguistic content, it is generally trained using large speech databases that provides a high-quality speech reconstruction method given the linguistic content. The databases used for speech conversion are generally much smaller and thus do not allow VC systems to learn fundamentals of speech signal, e.g. how the phonemes are formed within the signal, they focus on what they are made to transform. Therefore, it is often a good strategy to leverage text-to-speech systems to improve a voice conversion algorithm's performance. Encoder-decoder models with attention have lately demonstrated notable success in modeling a range of complicated sequence-to-sequence challenges. Tacotron (Wang et al., 2017; Shen et al., 2018) represents one of the successful text-to-speech implementations, that has been extended to voice conversion (Zhang et al., 2019; Park et al., 2020). The interplay between text-to-speech and voice conversion has been made easier by the neural representation of deep learning. By using text-to-speech systems to adhere to linguistic content, we mean to enhance the training and run-time inference of voice conversion. Such methods, however, typically require sizable

training corpus. A framework for developing limited-data voice conversion algorithms by bootstrapping from a speaker-adaptive text-to-speech model was recently described in studies (Huang et al., 2020; Luong and Yamagishi, 2019).

A different tactic is to use the Automatic Speech Recognition (ASR) - or speech-to-text - as a side task and give it the burden of learning latent linguistic unit representations. The voice conversion algorithm may for instance use the context posterior probability sequence produced by the automatic speech recognition model to generate a target speech feature sequence (Miyoshi et al., 2017). For each input utterance, the auto system produces a latent code from which text prediction is made possible. Trained on a large speech corpus, the speech-to-text model learns fine representations of linguistic units (for instance phonemes). Using the linguistic latent representation of the source utterance, a voice conversion system is more likely to preserve the source's linguistic content in the conversion. It also become free from parallel data as non-parallel pair of utterances can be easily plunged into the speech-to-text's latent space thus yielding representations that can be furtherly used to condition the conversion to the matching text.

A last trend referred to as neural disentanglement has also been used for parallel data free voice conversion. The idea behind voice disentanglement is to consider any speech signal as a composition of several types of information related to different speech attributes. A speech signal can be conceptualized as the combination of linguistic content, speaker identity, and expressive content in the broadest sense (emotions, attitudes, etc.) that are entangled in a complex manner during the realization of speech. This is obviously a schematic view and these different types of information are not independent of each other. For example, the expression of emotions is strongly dependent on the speaker who communicates them. This view, although not based on an actual partition either of the signal or in terms of anatomical mechanisms, allows for effective manipulation of the various attributes of speech. However, assuming those different attributes are independent from one another, a model can be learnt to disentangle one attribute from the others, i.e. produce a representation that isolates this attribute from the others, while still being able to reconstruct the original signal from these representations. To convert a speech attribute, one just has to manipulate its disentangled representation and then pass it to the decoder along with preserved information. Auto-encoder (Larsen et al., 2016) represents one of the common techniques for speech disentanglement, and reconstruction. There are other techniques such as instance normalization (Chou et al., 2019) and vector quantization (Tang et al., 2022), (Wu et al., 2020), that are effective in disentangling speaker from the content. In the auto-encoder paradigm first proposed in (Qian et al., 2019), the model means to disentangle the information to be converted from the rest. To do so a first encoder produces a code containing the information that is to be manipulated while a second encoder yields a code containing the information that allows for the reconstruction of the speech signal. For the conversion to be effective, the latter code must not contain any information about the attribute being converted. This is ensured simply by reducing this code's dimension so that it cannot contains any other information than what is necessary to reconstruct speech signal. This both simple and convincing bottleneck principle has shown very good results for identity conversion (Qian et al., 2019). This idea has also been applied for speech emotion conversion. In that case, the first encoder is similar to an emotion recognition (SER) system deprived from its last fully connected layer (endorsing classification task properly). This system can thus be trained separately as proposed in (Zhou et al., 2021).

Although these attempts are promising, the best performing algorithms (Kameoka et al., 2020, 2021), in terms of the sound quality of the conversions yielded, still require learning on parallel data. There are still many challenges to be met in the context of speech conversion and specifically for

the conversion of emotions. In particular, the issues of generalisation to new speakers and to all linguistic content remain to be addressed and are currently the subject of much research.

2.1.4 Metrics and Methodology for Evaluating Voice Conversion Algorithms

An efficient voice quality assessment is necessary to validate the algorithms, evaluate technical advancement, and compare a newly proposed algorithm to the best one available. We would like to quantify and qualify how natural and intelligible the conversions yielded by the model are, to what extent the algorithm actually does what it was made for. Usual quality assessment involve both objective and subjective measurements.

Objective Evaluation

In order to provide an objective evaluation of the model performance, each conversion must be compared to a reference speech which thus required. In the case of a deterministic model learned on so-called parallel pairs, the reference sound will be the target sentence of a given pair. The closer the conversion is from the target, the better the model performs. However, depending on the model architecture, there is a guarantee that the conversion will have the same length than the target utterance. In that case, before computing metrics, target and conversion must be aligned frame-wise using an aligner, through dynamic time warping (DTW) for instance. Mel-cepstral distortion (MCD) is used for spectral enveloppe while RMSE and PCC are used for prosodic features such as pitch and energy contours.

Mel Cepstral Distortion (MCD). This metric is often used for objective evaluation on spectral features. It is calculated between the converted and target mel-cepstral coefficients, or MCEPs, [240], [241], $\mathbf{X}^t$ and $\mathbf{Y}^{s\leftarrow t}$ and expressed in dB as follows

$$\text{MCD}(\mathbf{X}^t, \mathbf{Y}^{s\leftarrow t}) = \frac{10}{T_t \ln 100} \sum_{n=1}^{T} \sqrt{2 \sum_{k=1}^{D} (X_{n,k}^t - Y_{n,k}^{s\leftarrow t})^2} \tag{2.5}$$

where $X_{n,k}^t$ and $Y_{n,k}^{s\leftarrow t}$ are the k^{th} coefficients of the n^{th} frame of $\mathbf{X}^t$ and $\mathbf{Y}^{s\leftarrow t}$ respectively.

The prosody of a speech utterance, of main importance in the production of vocal attitudes, is characterized by many speech parameters such as pitch and energy contours, speech rate and phoneme durations. Those speech parameters can be either part of the input of the conversion systems or extracted from the representations or re-synthesized signals. To effectively measure how close the prosody patterns of converted speech is to the reference speech, objective measurements are provided.

Root Mean Square Error (RMSE). This metric can be used to measure differences between target and converted sequences for all the parameters mentioned above. For instance, if we consider both target and converted F0 sequences $\mathbf{f}^t$ and $\mathbf{f}^{s\leftarrow t}$, RMSE can be computed as follows

$$\text{RMSE}(\mathbf{f}^t, \mathbf{f}^{s\leftarrow t}) = \sqrt{\sum_{n=1}^{T_t} \frac{(f_n^t - f_n^{s\leftarrow t})^2}{T_t}} \tag{2.6}$$

Pearson Correlation Coefficient (PCC). This metric can also be used to quantify differences between target and converted sequences of speech prosodic parameters. If we consider F0 sequences again, the PPC can be formulated as follows

$$\text{PCC}(\mathbf{f}^t, \mathbf{f}^{s\leftarrow t}) = \frac{cov(\mathbf{f}^t, \mathbf{f}^{s\leftarrow t})}{\sigma_{\mathbf{f}^t}\sigma_{\mathbf{f}^{s\leftarrow t}}} \tag{2.7}$$

These metrics measure quantitative differences in acoustic parameters and are not necessarily correlated with human perception. In particular, there is not just one way to pronounce a sentence or convey an emotion or attitude. Consequently, quantitative measurements from a single reference only give a rough idea of the quality of the conversion. Since the ultimate criterion for judging a conversion is whether it sounds convincing to human subjects in general, those measures are of limited interest.

Subjective Evaluation

Subjective measures are based on the judgements by individuals towards the conversions yielded by the model. In that sense, subjective measures anchor the evaluation of models in people's actual experience of the resulting applications. These measures are therefore complementary to objective measures.

The most popular method, widely used in listening test (Kameoka et al., 2021; Qian et al., 2019; Zhou et al., 2021), is mean opinion score (MOS). In MOS experiments, listeners use a 5-point scale to assess the converted voice's quality: "5" for excellent, "4" for good, "3" for fair, "2" for poor, and "1" for bad. This method involves absolute judgements, which may be demanding for the subjects of the experiment, but which provide great insights on the conversions obtained. Several methods were derived from MOS such as DMOS (Tamura et al., 2001), which is a "degradation" or "differential" MOS test, requiring listeners to rate the sample with respect to this reference. MUSHRA (Zielinski et al., 2007), which stands for MUltiple Stimuli with Hidden Reference and Anchor, and requires fewer participants than MOS to obtain statistically significant results. Another standard method of subjective assessment is the preference test, also called the AB test (Toda et al., 2007). In AB tests, participants listen to a of sounds and are asked to give a preference with respect to a specific property; for instance in terms of naturalness, or similarity to a reference sound. This method involves relative judgements that are often easier for the participant to make. However it only provides an evaluation of the conversions with regards to presented data and not in general. A last method called best-worst-scaling (BWS) (Louviere et al., 2015) has gained interest over the past years. In such an experiment, participants listen to batches of a small number of sounds (4 or 5 usually) and are asked to chose which sound is the best and which is the worst with respect to a property.

Whatever method is used to collect subjective judgements, we will need to measure how significant the observed effects or trends are. To this end, we compute a 95% confidence interval $I_{95\%}$ than can be used, along with the average score obtained, for interpreting the test results. Denoting N the number of participants, $\bar{x}$ and $\sigma(x)$, the average and standard deviation of their judgements, this interval can be computed as

$$I_{95\%} =]\bar{x} - 1.96\frac{\sigma(x)}{\sqrt{N}}, \bar{x} + 1.96\frac{\sigma(x)}{\sqrt{N}}[\tag{2.8}$$

Evaluation with Deep Learning approaches

Common subjective metrics to rate the effectiveness of synthesized or converted speech - such as MOS test - often uses multiple human judges to analyze each speech utterance. Numerous

approaches to automatically anticipate MOS test results have been put forth in an effort to lower labor costs. In order to adress this issue, numerous approaches have been put out to automatically predict the MOS of an utterance: AutoMOS (Patton et al., 2016) predicted MOS with a recurrent network based on long short term memory (LSTM) cells. Quality Net (Fu et al., 2018) employed a frame-level quality constraint thus providing the training procedure with stability. After comparing various architectures, (Lo et al., 2019) showed that the MOSNet based on stacked convolutional and bi-directional recurrent network was a better architecture for MOS prediction. Later, (Choi et al., 2020) incorporated MOSNet with multi-task learning to improve performance. Finally, (Choi et al., 2021) used global quality token and encoding layer to achieve better prediction accuracy.

2.1.5 Section summary

In this research we focus on the voice conversion paradigm that we intend to apply to speech attitude conversion - i.e. we mean to convert attitude in a source speech signal with respect to a target one while preserving other speech attributes such as speaker identity and linguistic content. The typical speech conversion flow is composed with three steps: speech analysis - i.e. choice of speech signal representation, determination of a conversion function that maps source and target representations of speech and synthesis of the converted speech signal from the converted representation.

The question of how to represent speech signal for conversion must considered closely. Whether based directly on the signal or derived from a physical model of speech production, speech representations are yielded through the use of what a vocoder. The three main characteristics we need to consider when choosing a speech representation for conversion are: completeness, i.e. whether the signal can be reconstructed perfectly from its representation or not, interpretability - i.e. whether it makes sense from the perspective of the mechanisms of production or perception of the speech signal - and compactness - i.e. whether it allows for efficient processing by an algorithm while reflecting all aspects of speech. Vocoders such as STRAIGHT (Kawahara, 2006) or WORLD (Morise et al., 2016) have proven to yield good quality sounding syntheses. However, the trend in voice conversion based on those representations was to change the speech signal by mean of interacting with the parameters. Therefore, an important challenge of this research in the quest of a voice model formulation is the fine representation of the interaction between the components of speech. The arrival of deep neural networks gave rise to numerous attempts to represent such complex and non-linear coupling. When conditionned on mel-spectrogram representations of speech, the WaveNet has been the first vocoder to yield syntheses that are almost perceptually indistinguishable from real speech signals (Shen et al., 2018). This breakthrough gave rise to strong interest in the quest for a *universal neural vocoder*. Computationally sober vocoders based on both deep neural networks and elementary signal processing blocks - e.g. (Engel et al., 2020) - have been recently proposed, thus avoiding unnecessary learning of correlation already modelled by proven signal processing techniques. We will therefore use the team-made proposal (Roebel and Bous, 2022) to synthesise speech signals whose conveyed attitude has been converted in the last chapter 7 of this paper.

Once the source and target speech signals are represented, a conversion function can be introduced. Applied to the source speech representation, this function must yield a converted speech representation in which every aspects of the source utterance are preserved except from the converted attribute of which instance in the conversion must match the target utterance. Formerly approximated by statistical models, this mapping function is now by deep neural networks thus yielding way better sounding speech conversions (Desai et al., 2009; Nakashika et al., 2013; Luo

et al., 2016). We exclusively focus on this latest approach in this work. The main determinant for the development of such algorithms is the the availability of parallel data, e.g. a dataset that features pairs of utterances that only differ in terms of the attitude conveyed. The early attempts of neural voice conversion were deep neural networks that did not allow for temporal dependencies modeling (Desai et al., 2009; Nakashika et al., 2013), which greatly limited their performances. To tackle this limitation, Recurrent Neural Networks (RNN) were introduced for voice conversion and achieved great improvements in naturalness and temporal consistency of the yielded conversions (Sun et al., 2015; Ming et al., 2016). Since recurrent networks do not by themselves allow sequences of different duration to be mapped, the sequence-to-sequence (Seq-to-Seq) architecture (Sutskever et al., 2014) has been introduced posing the conversion task as a translation problem. This architecture maps the complete temporal sequences rather than just map source and target utterances frame by frame. The classic problem with this architecture is that the temporally propagated information tends to vanish for long sequences. Addressing this issue, the attention mechanism aims to implicitly learn an alignment between the source and target sentences without any assumption about how the linguistic units are mapped (Obin et al., 2019; Zhang et al., 2019; Tanaka et al., 2019; Tachibana et al., 2018). Although recurrent networks represent an efficient implementation for sequence-to-sequence, recent studies have shown that Convolutional Neural Networks (CNN) with gating mechanisms also learn well the long-term dependencies (Gehring et al., 2017; Kameoka et al., 2020). Finally, the transformer networks (Vaswani et al., 2017), of which principle is a generalization of attention mechanism that basically replaces all recurrent layers by self attention mechanisms in a network, have been proposed for voice conversion yielding noteworthy performance in terms of sound quality (Kameoka et al., 2021; Chen and Zhang, 2021; Lee et al., 2022). Facing the challenge of gathering parallel data dedicated to voice conversion, many researchers have sought to circumvent this constraint. Experimental results shown that, with non-parallel training data, Cycle-GAN achieves comparable performance to that of GMM-based system that is trained on twice amount of parallel data (Kaneko and Kameoka, 2018). Other efficient strategies was to leverage text-to-speech (Zhang et al., 2019; Park et al., 2020) or speech-to-text (Miyoshi et al., 2017) models to improve a voice conversion algorithm's performance. A last trend referred to as neural disentanglement has also been used for parallel data free voice conversion. It conceptualizes speech signal as the combination of linguistic content, speaker identity, and expressive content in the broadest sense (emotions, attitudes, etc.) that are entangled in a complex manner during the production of speech. Auto-encoder (Larsen et al., 2016; Qian et al., 2019; Zhou et al., 2021) represents one of the common techniques for speech disentanglement, and reconstruction.

An efficient voice quality assessment is necessary to validate the algorithms, evaluate technical advancement, and compare a newly proposed algorithm to the best one available. Algorithms are evaluated objectively by comparing conversions to reference sounds using metrics such Mel-Cepstral Distorsion (MCD) for spectral envelope or Root Mean Squared Error (RMSE) for pitch or energy contours. The smaller these errors, the closer the conversion is from the reference and the better the model performs. Nevertheless, these metrics measure quantitative differences in acoustic parameters and are not necessarily correlated with human perception. For this reason, algorithms are also evaluated subjectively by asking individuals to judge the conversions they yield. The most popular method, widely used in listening test (Kameoka et al., 2021; Qian et al., 2019; Zhou et al., 2021), is mean opinion score (MOS). Another standard method is the preference test, also called the AB test (Toda et al., 2007), involving relative judgements that are often easier for the participant to make. A last method, also involving relative judgements but among batches of several speech samples, called best-worst-scaling (BWS) (Louviere et al., 2015) has gained interest over the past years. Common subjective metrics to rate the effectiveness of synthesized or converted speech often uses multiple human judges to analyze each speech utterance. Numerous

approaches to automatically anticipate MOS test results have been put forth in an effort to lower labor costs (Patton et al., 2016; Fu et al., 2018; Lo et al., 2019; Choi et al., 2020, 2021).

2.2 Speech Attitude Recognition

Human have an outstanding ability to communicate social signals conveying rich and useful information, such as emotions or attitudes. This information can be communicated through different channels, among which speech plays an important role. To perceive emotions, humans process each modality separately and then recompose units of meaning from each piece of information. For speech modality, they uses discriminative acoustic features from which emotions are inferred. Analogously, the speech emotion recognition task aims to mimic this decoding process of vocally communicated emotions. The key to this task is thus to obtain discriminating features from which emotions can be distinguished.

2.2.1 General Scheme of Speech Attitude Recognition

In this subsection, we describe the typical speech attitude - or emotion - recognition flow in two steps: speech analysis and determination of a classification function that enables speech attitude - or emotion - recognition from a representation of the speech signal.

Speech analysis for speech attitude recognition

Recognizing the attitudes, or emotions, requires identifying the characteristics of the speech signal that convey and make them distinct. First, a parametric representation of the speech signal can be used. The fundamental frequency is known to convey emotions in speech, so it can be used to represent the speech signal for emotion recognition purpose. However, emotions and attitudes are produced and perceived through complex mechanisms involving many aspects of the speech signal. In particular, the mechanism by which individuals decode attitudes and emotions cannot be reduced to the capture in the speech signal of variations in independent parameters, many of these aspects are still unknown. To address this limitation, many approaches have attempted to produce hand-crafted features adapted to the emotion recognition task (Badshah, 2017). Since the advent of neural networks, the task of identifying the signal's salient features has been combined with the determination of an emotion prediction function. Most contemporary approaches no longer make assumptions about what conveys emotions in the signal and therefore choose to use complete and compact mel-spectrogram representations (Chen et al., 2018; Meng et al., 2019).

Finally, the analysis of the speech signal can be formalized in the same way as it is for speech conversion, i.e. according to the equation 2.1. We denote $\mathcal{A}$ the function which, to a speech signal $\mathbf{x}^a\mathbb{R}^T$ conveying an attitude labelled a associates its representation $\mathbf{X}^a\mathbb{R}^{T'\times D}$ such as

$$\mathbf{X}^a = \mathcal{A}(\mathbf{x}^a) \tag{2.9}$$

Classification or prediction function

Once the speech signal $\mathbf{x}^a \in \mathbb{R}^T$ is represented by $\mathbf{X}^a \in \mathbb{R}^{T\times D}$, a prediction function $\mathcal{P}$ - also referred to as classification function - can be introduced such that its application to the input speech

representation yields a label a representing the attitude conveyed. The classification can thus be formulated as

$$a = (\mathcal{P} \circ \mathcal{A})(\mathbf{X}^a) \tag{2.10}$$

We will see in the next section that this mapping function can be approximated in various ways ranging from statistical models to deep neural networks.

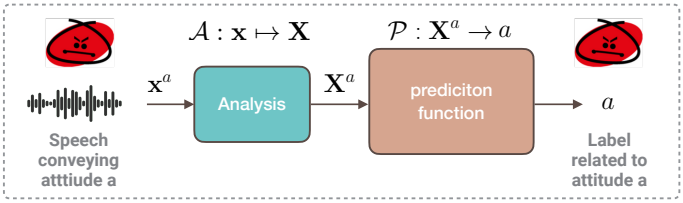

Figure 2.4: General speech emotion - or attitude - recognition scheme. The two-step pipeline process begin formulated as $a = (\mathcal{P} \circ \mathcal{A})(\mathbf{X}^a)$

2.2.2 Speech Emotion Recognition Algorithms

Formerly addressed using statistical methods and traditional learning techniques such as Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs) and Support Vector Machines (SVMs), the SER task has known significant improvements over the past years with the advent of Deep Neural Networks (DNNs). Indeed such deep networks have shown excellent abilities to model more complex patterns within speech utterances by extracting high-level features from speech signal for better recognition of the emotional state of the speakers.

Deep learning based approaches

Mao et al. (Mao et al., 2014) firstly introduced Convolutional Neural Networks (CNNs) for the speech emotion recognition task and obtained remarkable results on various datasets by learning affective-salient features. Recurrent neural networks has also been introduced for speech emotion recognition purpose with a deep Bidirectional Long Short-Term Memory (BLTSM) network proposed by Lee et al. (Lee and Tashev, 2015). Several papers have then presented convolutional neural networks in combination with recurrent networks based on long-short term memory cells to improve speech emotion recognition, based on log Mel filter-banks (logMel) (Keren and Schuller, 2016) or raw signal in an end-to-end manner (Trigeorgis et al., 2016).

Recently, attention mechanisms have raised great interest in the SER research area for their ability to focus on specific parts of an utterance that characterize emotions. (Mirsamadi et al., 2017) approached the problem with a recurrent neural network and a local attention model used to learn weighted time-pooling strategies. Neumann et al. (Neumann and Vu, 2017) used an attentive convolution neural network (ACNN) and showed the importance of the model architecture choice against the features choice. Ramet et al. (Ramet et al., 2018) presented a review of attention models on top of recurrent networks based long-short term memory cells and proposed a new attention computed from the outputs of an added bidirectional long-short term memory layer. Chen et al. (Chen et al., 2018) proposed a 3-D Attention-based Convolutional Recurrent Neural Networks (ACRNN) for speech emotion recognition with 3-D log-Mel spectrograms (static, deltas and delta-deltas) as input features. They showed 3-D convolution can better capture more effective

information for speech emotion recognition compared with 2-D convolution. Recently, Meng et al. (Meng et al., 2019) outperformed this method by using dilated convolutions in place of a pooling layer and skip connection.

Attempts to inform emotion classification networks with extra-information involved in the description of emotions were proposed in the past years. Based on previous works (Ververidis and Kotropoulos, 2004; Vogt and André, 2006; Zhang et al., 2018), Li et al. (Li et al., 2019) proposed a multitask learning framework that involves gender classification as an auxiliary task to provide emotion-relevant information leading to significant improvements in speech emotion recognition. Analogously, speaker identity has been used to inform emotion classification networks. The problem was approached by Sidorov et al. (Sidorov et al., 2014) with speaker dependent models for emotion recognition. Recently, a method for speaker aware SER was introduced by Assunção et al. (Assunção et al., 2020), a convolutional neural network model VGGVox (Nagrani et al., 2017) is trained for speaker identification but is instead used as a front-end for extracting robust features from emotional speech. These first attempts have shown that teaching speech emotion recognition systems with additional signal-based information can greatly improve performances.

Speech emotion recognition as a voice conversion side task

For several reasons, the specific task of speech emotion recognition can be considered as inherent to the conversion of emotions in speech. First, it provides information on the ability of a system to capture emotional information in a specific dataset, it is complementary to a perceptual study conducted on the same data which would provide information on the ability of humans to distinguish these emotions. In a sense, if there is no system capable of predicting the emotions of a dataset, there cannot be one that allows their conversion either. Second, learning such SER models results in the implicit learning of a definition of the emotions present in the dataset. The latent space resulting from this learning is expected to be emotionally structured, each utterance can be represented by a code that reflects the emotion it contains. These *emotion embeddings* can then be used to inform conversion systems as recently proposed by Zhou et al. in (Zhou et al., 2021).

2.2.3 Methodology for Evaluating Speech Emotion Recognition Algorithms

There are two main ways of evaluating emotion recognition systems. The first involves comparing the model's predictions to the actual ground truth, while the second involves examining the implicitly learnt latent space structure. One or the other may be preferred depending on the model's architecture and optimisation options. However in most cases, these two methods ought to be applied in concert.

Assessing Prediction Performance

Accuracy. One of the most used metrics to evaluate classification tasks in machine learning. It represents the proportion of correct predictions to all examples. Although, this metric is easy to interpret and compute, it has limitations. For this reason, it is often replaced by another objective measure called Unweighted Average Recall (UAR). In the following, we set out to explain what this measure is about.

Unweighted Average Recall (UAR). Depending on its relation with ground truth, a prediction is given a status of true positive, true negative, false positive, or false negative. When the ground thruth category is predicted, it is known as a true positive. When the ground truth category is not predicted, it is known as a true negative. When a category that differs from ground truth is predicted,

it is known as a false positive. When the ground truth category is not predicted, it is known as a false negative. Thus if we denote those status tp, fp, tn and fn respectively, thus with $p = tp + fp$ and $n = tn + fn$ the accuracy can be re-write as follows

$$accuracy = \frac{tp + tn}{p + n} = \frac{tp}{p} \times \frac{p}{p + n} + \frac{tn}{n} \times \frac{n}{p + n} \tag{2.11}$$

The factors $\frac{tp}{p}$ and $\frac{tn}{n}$ are known as recalls on the positive and negative classes respectively, also known as sensitivity and specificity. Accuracy can thus be seen as a weighted sum of those factors. The final score is more affected by classes with more samples than by classes with fewer. Due to this, accuracy is often unfit to determine how well the model performs for class unbalanced datasets. To tackle this, UAR was introduced as

$$UAR = \frac{1}{n_c} \left(\frac{tp}{p} + \frac{tn}{n} \right) \tag{2.12}$$

Assessing the Latent Space Structure

Analyzing a latent space's structure mainly involves examining how its many categories are organized within this space. To do so, there are several objective measures.

Silhouette coefficient. Typically used to evaluate how well clustering algorithms perform, shows how separate the categories in the latent space are. Given a distance $d : \mathbb{R}^{d_{latent}} \to \mathbb{R}$, typically euclidean, it is defined for a latent vector $\mathbf{h}_i$ in the cluster of index $\mathcal{C}_k$ as follows

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{2.13}$$

$$a_i = \frac{1}{|\mathcal{C}_k| - 1} \sum_{j \in \mathcal{C}_k, j \neq i} d(\mathbf{h}_i, \mathbf{h}_j) \tag{2.14}$$

$$b_i = \min_{k' \neq k} \frac{1}{|\mathcal{C}_{k'}|} \sum_{j \in \mathcal{C}_{k'}} d(\mathbf{x}_i, \mathbf{x}_j) \tag{2.15}$$

The silhouette coefficient lies between -1 and +1, indicating respectively the limits of a bad and a good clustering of the data. An overall measure can be obtained through computing the average silhouette coefficient over all data samples.

Davies-Bouldin score. Defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances. Thus, clusters which are farther apart and less dispersed will result in a better score.

Mutual Information (MI). Computed between two random variables, MI is a non-negative value, which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency. The function relies on non parametric methods based on entropy estimation from k-nearest neighbors distances as described in (Ross, 2014).

2.2.4 Section Summary

Human have an outstanding ability to communicate social signals conveying rich and useful information, such as emotions or attitudes. To perceive those conveyed through speech, humans use discriminative acoustic features from which they are inferred. Analogously, the speech emotion recognition task aims to mimic this decoding process of vocally communicated emotions. The typical speech emotion recognition flow can be formalized in two steps: speech analysis - i.e. choice of speech signal representation - and determination of a classification function that enables speech emotion recognition from its representation.

Emotions and attitudes are produced and perceived through complex mechanisms involving many aspects of the speech signal. In particular, the mechanism by which individuals decode attitudes and emotions cannot be reduced to the capture of variations in independent speech parameters, many of these aspects are still unknown. In view of this, many approaches have attempted to produce hand-crafted features adapted to the emotion recognition task (Badshah, 2017). Since the advent of neural networks, the task of identifying the signal's salient features has been combined with the determination of an emotion prediction function. Most contemporary approaches no longer make assumptions about what conveys emotions in the signal and therefore choose to use complete and compact mel-spectrogram representations (Chen et al., 2018; Meng et al., 2019).

Formerly addressed using statistical methods and traditional learning techniques such as Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs) and Support Vector Machines (SVMs), the SER task has known significant improvements over the past years with the advent of Deep Neural Networks (DNNs). Convolutional Neural Networks (CNNs) were first introduced for the speech emotion recognition task and obtained remarkable results on various datasets by learning affective-salient features (Mao et al., 2014). Recurrent neural were then proposed, thus better exploiting temporal dependances (Lee and Tashev, 2015). Combinations between convolutional and recurrent neural networks have shown even better performances (Keren and Schuller, 2016; Trigeorgis et al., 2016). Attention mechanisms have raised great interest in the SER research area for their ability to focus on specific parts of an utterance that characterize emotions (Mirsamadi et al., 2017; Neumann and Vu, 2017; Ramet et al., 2018; Chen et al., 2018; Meng et al., 2019). Finally, attempts to provide emotion classification networks with extra-information involved in the description of emotions such as speaker identity (Sidorov et al., 2014; Assunção et al., 2020) or gender (Li et al., 2019) were proposed in the past years. There are two main ways of evaluating emotion recognition systems. The first involves comparing the model's predictions to the actual ground truth using accuracy or unweighted average recall, while the second involves examining the implicitly learnt latent space structure using silhouette coefficient, davies boudlin score or mutual information. One or the other may be preferred depending on the model's architecture and optimisation options. However in most cases, these two methods ought to be applied in concert.

Chapter 3

ATT-HACK : A DATASET FOR SPEECH SOCIAL ATTITUDES

Contents

3.1 Genesis of Att-HACK 53
3.1.1 A Need for Expressive Speech Data in French 53
3.1.2 In Defense for a Research on Vocal Communication of Social Attitudes . . 54
3.1.3 Choice of Speech Attitudes 54
3.2 Design, Methodology and Recording 56
3.2.1 Attitude Communication Context 56
3.2.2 Set of Sentences 57
3.2.3 Recording Sessions 59
3.3 Processing and Preliminary Analyses 60
3.3.1 Data Cleaning and Formatting 60
3.3.2 Att-HACK metadata statistics 61
3.3.3 Investigating pitch patterns underlying the production of vocal attitudes . 62
3.4 Discussion 64
3.4.1 Balanced vs Imbalanced Data 64
3.4.2 On Uncovering Data Biases 64
3.5 Chapter Summary 64

3.1 Genesis of Att-HACK

This section aims to explain the guiding principle behind the creation of a new dataset for expressive voice. Therefore, we will discuss existing databases, their uses, and the issues that make it necessary to gather more diverse speech data.

3.1.1 A Need for Expressive Speech Data in French

Though the linguistic functions of speech prosody are nowadays well documented in a large number of languages (phonology, syntax/prosody interface, etc...), its expressive or para-linguistic functions, such as speaking style or speech emotions, does not benefit from the same amount of attention from the speech community. Meanwhile, speech engineers have realized spectacular advances in the past decade creating extremely realistic synthetic voices (Wang et al., 2017) which are now integrated into voice interfaces that are increasingly present in our everyday lives, such as the voice assistants and conversational/virtual agents. However, these intelligible and natural voices still clearly lack expressiveness and adaptability which greatly limits the interaction between humans and machines (see for instance (G. Castellano, 2012)). Expressivity is the next frontier of speech research at the interface of cognitive science, linguistics and technology, as shown by the recent increase of research in this domain (Wang et al., 2018b; Zhou et al., 2021; Liu et al., 2022). Consequently, there is a clear need to better understand how humans produce and decode emotions, attitudes and all aspects that enrich vocal communication.

Available Expressive Speech Datasets

Speech expressivity is generally equated to speech emotions though the scope of expressivity includes but is not restricted to the primary emotions as denoted by Ekman (Ekman, 1992). This limitation is probably due to the difficulty of converging to an agreement on the terminology used to describe the various and subtle forms of expressivity in speech. Accordingly, the study of speech expressivity is generally limited to dedicated speech emotion databases as interpreted by actors or to audio books read by professional readers - mostly in English (Busso et al., 2008; Zhou et al., 2022), and sometimes in French (Sini et al., 2018), German (Chen et al., 2015) or Chinese (Zhou et al., 2022). In the past decade, speech emotion research has mainly focused on acted emotional speech: from its original form in which an actor is asked to interpret a short text with a given emotion (Burkhardt et al., 2005) to more open and spontaneous forms in which two actors freely improvise based on a given scenario and then asked to rate their own speech emotions with categories or on valence/arousal continuous scales (Busso et al., 2008; McKeown et al., 2010). These databases have been designed only for the purpose of speech emotion recognition, but not for emotional speech synthesis and voice emotion conversion. They moreover miss an essential aspect of speech prosody: its *variety* (Obin et al., 2012). There is always only one realization of each utterance, while any utterance can be produced with various prosodies, some being functionally equivalent, some presenting different degrees of expressivity (Gerazov et al., 2018). Only a few databases dedicated to emotions can be used for speech conversion (Burkhardt et al., 2005; Busso et al., 2008; Zhou et al., 2022).

On Gathering More and More Diverse Data in French

Not only, but also specifically for French - a langage for which there is very few resources regarding emotion in speech - there is a great need for expressive speech databases. Such databases must

allow a diversification of the research in speech prosody and expressivity, be large enough and of a sufficiently high quality to allow learning generative models.

3.1.2 In Defense for a Research on Vocal Communication of Social Attitudes

The basic emotions as described by Ekman (Ekman, 1992) have been and still are the subject of much studies, particularly in identifying the mechanisms underlying their production. The emergence and growth of research in social interaction and social psychology have made it possible to highlight the main dimensions of social cognition - namely warmth and competence - (Fiske et al., 2007) opening up new areas of research in understanding the communication of social signals. In light of these advances in social cognition, social attitudes — long studied in the shadow of emotions — are proven relevant in describing many of our social interactions (Wichmann, 2000). There is currently no database devoted to social attitudes in speech, and even less so in French.

Emotion vs Attitude

Human interactions are governed by a fantastic interplay of social signals. Although central in this mechanism, the primary emotions described by Ekman (Ekman, 1992) are undoubtedly not sufficient to describe the entire expressive spectrum of human speech. For example, a person can be friendly, distant, seductive or dominant with a stranger they just met, depending on the outcome they expect from the interaction. Such attitudes differ from emotions, because they do not only hint at the speakers' affective state, but are the expression of their social intention (Wichmann, 2000).

Defining Social Attitudes

Many works have attempted to provide a precise definition of these different expressive variants. For instance, to deal with subtle forms derived from primary emotions, the circumplex model, in which emotions are categorized in a bi-dimensional space, was proposed by psychologists (Russell, 1980). Attitude was firstly equated to the first dimension - i.e. valence - of this model (Ajzen and Fishbein, 1980). A distinction between emotion and attitude was done in (Couper-Kuhlen, 1986) by defining emotion as a speaker state and attitude as some kind of behaviour. This distinction was later refined in (Wichmann, 2000), by defining attitude as a predictor of social behaviour. In this respect, the attitudinal aspect of expressivity of course differs from the primary emotions. Although it must be partly determined by the speaker's affects (Bodenhausen, 1993). A distinction must be made between the propositional attitude - towards an utterance: irony, doubt, etc... (Leech, 1983) - and the social attitudes - towards a person: dominant, friendly, seductive, distant for instance. This last dimension has been recently investigated in the study of the role of speech prosody in neurosciences (Ponsot et al., 2018b).

3.1.3 Choice of Speech Attitudes

Following the initial definition in (Wichmann, 2000), we sought to select several vocal attitude labels - or categories - for designing our dataset.

How to represent Social Attitudes communicated vocally?

Prior to selecting particular attitudes to be featured in our database, the issue of how these vocal attitudes should be represented must be addressed. Two approaches can be employed: the

categorical approach - the whole range of attitude instances is divided into a number of formally distinct categories - or the dimensional approach - attitude expressions are represented in a multi-dimensional space, each of whose dimensions describes some aspect of what distinguishes the attitude instances from each other.

First, social cognition represents interactions between individuals along two dimensions, namely warmth and competence (Fiske et al., 2007). Indeed, as shortly as an interaction begins, individuals must evaluate if the other is a friend or foe - i.e., whether they have good or bad intentions - and, then, whether the other has the ability to act according to those intentions. This dimensional representation constitutes a first approach for us to attempt representing attitudes. Second, the Leary's rose model (Leary, 1957) provides a representation of attitudes in a bi-dimensional space, the first dimension reflects the inclusiveness - hostility or friendliness - towards the other and the second dimension reflect the position within a social hierarchy - subordination or dominance. This model served as the basis for an original representation of the musicians performing in duo (Aucouturier and Canonne, 2017). Since Leary's and Friske's models use similar dimensions to describe interactions between individuals, they will be considered as a starting point for our representation of the speech social attitudes.

Selection of Attitudes

The labels we selected for representing speech social attitudes are directly derived from the two precursory works (Leary, 1957; Fiske et al., 2007) mentioned above. In this paper, four social attitudes were defined by sampling the warmth and competence - or inclusiveness and dominance - dimensions during a speech interaction: friendly, seductive, dominant and distant, as represented in Figure 3.1. This includes one exclusive (distant) and three inclusive (friendly, seductive, dominant) attitudes sampled in the semi-space of neutral to high-hierarchy in the bi-dimensional space. Indeed, this selection was motivated by the final application of an inboard vocal assistant. In the scope of specific context, no attitude label related to submission was selected. We therefore assume that we do not encompass the whole spectrum of attitudinal expression. Let us note that, when sampling such attitude labels in the two-dimensional space depicted in Figure 3.1, we make no assumptions about the categories that arise from these selected labels. In particular, these categories may not be distinct from each other or some may be significantly broader than others. The uncovering of the production and perception mechanisms that underlie these vocal attitudes may help to establish the shape of these categories and their interactions.

In line with their dimensional representation, we attempted to provide an informative description for each of the selected attitudes as follows:

- **friendly**: you are pleasant and benevolent, you care about others' preferences, you act towards the others independently from your own situation.
- **seductive**: everything in your behaviour aims at charming the others, to make them love you, you do not care about others' preferences but you are ready for anything to seduce them even if you have to fake benevolence.
- **dominant**: you are self confident, sure of your own superiority, you do not care about others' preferences, everything in your behaviour is dedicated to make the others obey and listen to you without imposing anything explicitly.
- **distant**: you (barely) do not care about the others, you are uncommunicative, you do not care about others' preferences.

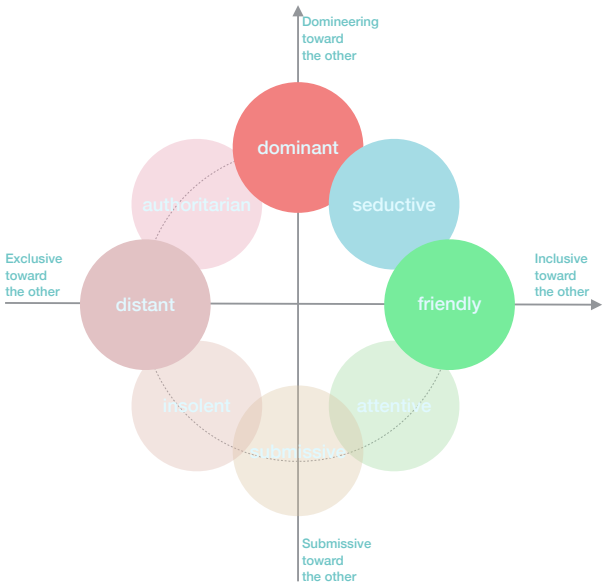

Figure 3.1: Social attitudes represented in a standard warmth/competence - or inclusiveness/dominance - bi-dimensional space following early works in psychology (Leary, 1957) and social cognition (Fiske et al., 2007).

3.2 Design, Methodology and Recording

A proper design of a dataset implies to make assumptions and methodological choices which affects both the material itself and the applications that are built from it. Must the speech attitudes in the dataset be picked up in real life or portrayed by actors? Should their linguistic support be pre-scripted or left up to improvisation by the actors if they are performed? In which case, what requirements should be followed while choosing the sentences that the actors will perform? This section is dedicated to answering all those questions.

3.2.1 Attitude Communication Context

The most crucial decision we had to make was whether to ask actors to portray attitudes or to pick up real life spontaneous utterances.

Portrayed vs Spontaneous Attitudes

First, it is important to highlight that portrayed emotional vocalisations, i.e. produced by actors, may involve different production and perception mechanisms than those involved in the case of genuine emotions. In fact, actors' vocalisations are known to be less authentic than spontaneous ones (Anikin and Lima, 2017) — which, in the case of e.g. facial expressions of emotions, even seem to rely on different neural bases (Valente et al., 2017). It has also been revealed that this authenticity standard differs depending on the emotion considered. In certain circumstances, particularly for high-stakes emotions like anger or fear, individuals would put greater importance on the criterion of emotional authenticity (Anikin and Lima, 2016). The explanation would be based on human evolution; people need to be able to rely on their own and other people's emotions when they are

in danger. Having said that, there is little doubt that choosing portrayed attitudes is not the best strategy for understanding and modeling what occurs in real life. However, no study have revealed the significance of authenticity in regard to the particular category of attitudes we have chosen to investigate. In addition, in the scope of voice conversion, exaggerated or stereotypical expression can facilitate learning by better marking the difference between the targeted categories. On the other hand, we can also hypothesize that exaggerated expressions may be more relevant for voice conversion applications, as they might be communicated more easily - notably better decoded by human users.

Practical feasibility

The other point to think about is whether it would be possible in practice to gather utterances that have an attitudinal component in real life situations. For a number of reasons, it is now very difficult to do so. Even while such records can be retrieved using internet or embedded programs on smartphones, from anywhere in the globe, it is nearly impossible to manage what is contained in them. The task of voice conversion, which is the main application for this dataset, necessitates data annotation, i.e. to know which attitude is communicated for each utterance. To obtain such annotation, one option is to carry a perceptual test in which subjects are asked to judge the attitude conveyed by a particular sentence. This preliminary step is a very expensive and time-consuming process. Conversely, asking actors to portray attitudes gives much more control on what is actually conveyed through speech utterances.

Final decision towards Portrayed Attitudes

In light of all those arguments, we decided to build a dataset featuring portrayed attitudes. The main cause for this decision is the fundamental requirement for parallel data, related to the majority of current voice conversion models, which is essentially incompatible with the spontaneous alternative. In fact, there is no guarantee that speakers would pronounce the same utterance - in the sense of linguistic content - in at least two different attitudes. Our goal is then to create a gender-balanced multi-speaker database which features several repetitions/variants per linguistic content, speaker and attitude. Provided such a dataset, we would like to be in capacity to study intra and inter-speaker variability in the production and perception of vocal attitudes.

3.2.2 Set of Sentences

The second decision to make dealt with controlling the linguistic variability in the dataset. In this regard, several options can be considered. From one hand, actors can be asked to improvise, from scratch or from given scenario or character description. On the other hand, we can provide them with a text, in respect to which they are asked to play faithfully. Designing a parallel database involves selecting a series of sentences - distinct linguistic contents - with respect to which each attitude must be portrayed. Since it would have been impossible to identify linguistically paired utterances among improvised - linguistically uncontrolled - material, this - strong - requirement prompted us to chose the second approach. The set of sentences used for the creation of the database has been designed in French as inspired by the corpus of propositional attitudes proposed by Morlec in (Morlec, 1997). The decision to choose French was quite obvious given the dearth of French expressive databases.

Criteria for the Design of Sentences

The proposed sentences have been designed according to the following criteria:

- to avoid introducing a bias due to the affective content suggested by the text - i.e. the sentences have to avoid connoting a particular expressive colouration so that only the vocal expression carries information.
- to reduce the prosodic variability due to the text structure, the sentences were designed with a limited and controlled linguistic complexity, by using simple syntactic structures and short sentences;
- to remain plausible in each social attitude.

Let us note that the first of these criteria cannot be satisfied by any sentence, since no sentence is absolutely neutral with regards to the emotional content it conveys, a study of sentiment analysis is therefore carried out on all 100 sentences in the chapter 4.

Design of Sentences

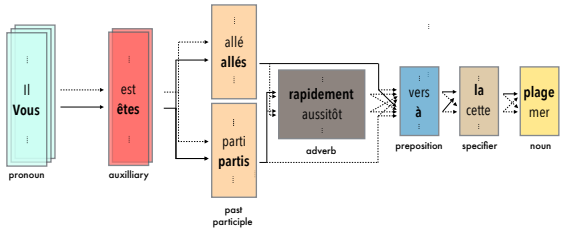

Figure 3.2: Phrase generator functioning for the above quoted phrases, chosen words are in bold, dotted lines represent possible choices for the algorithm

Accordingly, we constructed a set of 100 sentences - from 2 to 8 syllables - corresponding to simple everyday life situation - situations that are likely to occur in socialization places like home, restaurant or workplace. For this purpose, we designed a phrase generator that builds phrases from semantic nucleus ({pronoun/noun + verb} or {pronoun/noun + auxilliary}) by randomly picking words in dictionaries in order to guarantee the phrases are always conceived the same way. We randomly kept 100 phrases among the 10,000 generated ones to build our set of sentences, a sample of 10 sentences is listed below.

Oui - Yes
Bonjour - Hello
C'est vrai - That's true
A demain Paul - See you tomorrow Paul
Bonne journée Marie - Have a good day Mary
Il est tard à Londres - It's late in London
Vous êtes allés à la plage - You went to the beach
Vous êtes partis rapidement - You left quickly
Impossible, attendons un peu - Impossible, let's wait a bit
C'est vrai, allons prendre un café - That's right, let's get a coffee

Figure 3.2 depicts the functioning of the automatic sentence generator considering the example of two sentences created from the nucleus “Vous êtes ..." (“You are ...”).

3.2.3 Recording Sessions

To feed this database, twenty actors - coming from different practices of the professional voice community (dubbing, theatre, advertising) - were recorded in professional studios at Ircam. The recruitment of actors was carried out via a dedicated website [1] on which various actors are listed. The website gives access to a number of information such as the actor’s gender and age as well as recording samples of their voice. It should be noted here that the actors were not asked any questions about their sexual orientation, nor were they asked to complete a personality questionnaire. The only important information we kept was their gender, which they themselves declared on the website where we recruited them.

Recording Procedure

The recordings consisted of 4 hours sessions during which one actor had to play 100 sentences in the four different social attitudes, proposing from 3 to 6 different versions of each sentence in each attitude. At the beginning of each session, the four attitudes were shortly described as stipulated above. Those descriptions have been used as acting options, actors were told to act according to their own understanding of each attitude label and not necessary in regards to those descriptions. The actors were told to be as natural as possible, no other information was given during the session. At the end of a session, we had at least $2,500$ audio files for each actor.

The recording sessions went the following way : an attitude was picked randomly, the actor played the sentences in a random order, offering several versions. Once all the phrases had been played, another attitude was randomly picked and so on until all four attitudes were completed. I personally took care of half of the recordings, the other half was handled by Darick Lean who was a trainee in the team at the time.

Technical Details

For the recording, we used a Neumann U87 static microphone plugged into a RME fireface interface synchronized with the ProTools software [2]. All audio recordings were made with a sampling rate frequency of 44.1 kHz and a quantization of 16 bits per sample. The recording were made in two different studios depending on their availability. A patch implemented with the Max for Live software [3] was used to provide a visual interface to the actor, displaying on a screen in front of the actor the sentence to be read and the expected attitude, this display being monitored by a sound engineer. The patch also allowed to store the time codes corresponding to each sentence, which were used after the session to segment and name automatically the continuous recording made with the ProTools software.

The experimenter must press the space bar on the computer keyboard to capture the time codes related to a sentence’s start and finish. Therefore, it is crucial for the experimenter to maintain focus in order to avoid pronunciation (text) or interpretation (attitude) errors, direct or correct the playing, and eventually do retakes. Sessions used to drag on for a while, and it has been typical for actors

[1] https://www.castingmachine.com/
[2] https://www.avid.com/fr/pro-tools
[3] https://www.ableton.com/en/live/max-for-live/

to become sidetracked and forget what they had to say or the attitude they had to portray. In a way, the experimenter has to direct the actor he is recording.

3.3 Processing and Preliminary Analyses

Once collected, these data just waited to be used. Nonetheless, it was essential to first ask: what is actually contained in these data? This subsection offers a first attempt at answering that question.

3.3.1 Data Cleaning and Formatting

Once the recordings were completed, the collected data needed to be processed in order for it to be usable effectively. Essentially, this involves doing two things: first, cleaning up the data by removing any fraudulent examples, and second, formatting it.

Detecting fraudulent samples

At the end of the recording sessions, the audios were run through a click detector called the Hook-Net. This novel adaptation of the Wave-U-Net for the task of detecting artefacts in audio signals has been developed in the Analysis Synthesis Team by Daniel Wolff (Wolff et al., 2021). Fed with raw waveform segments, it can return time series of artefact classification results. The yielded corrupted audios were either removed from the database or fixed for less serious cases. This method was still not able to catch all of the fraudulent examples, despite being effective. As a result, a significant portion of the database has been processed manually, and some new false samples have been discarded. The database, however, is too big to be fully handled manually. The database has got cleaner and cleaner as it has been used for different purposes, each use enabling for the detection of remaining faulty samples. It should be highlighted that this process was solely aiming at deleting the samples of poor sound quality but has nothing to do with judging the quality of the actors' performances.

Data formatting

First, each sound sample in Att-HACK has been stored in wav format in $48kHz$ and $24b$. Furthermore, we employed a neural aligner (Teytaut and Roebel, 2021) proposed by Yann TEYTAUT - also PhD candidate in the team - to perform phoneme-to-audio alignment on the Att-HACK samples. Since there is no way to assess the proper performance of this aligner on the Att-HACK samples, we rely on this aligner's performance on TIMIT dataset (Garofolo et al., 1992) - in English - for which this algorithm holds the state-of-the-art. Using this alignement, we have trimmed the examples to prevent too large silences at the start and end of phrases. Finally, the data formatting and manipulation was done within the pandas data frame framework [4] throughout the entire course of this research. This format for data manipulation makes it simple to select a subset of the data using a key, which in our instance could possibly be a speaker, an attitude, or even a sentence.

[4] `https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html`

3.3.2 Att-HACK metadata statistics

A first analysis consisted in making basic statistics on Att-HACK by quantifying some of its facets such as attitude, gender or linguistic content. Such statistics are displayed in Table 3.1. In total the database represents just over 27 hours of expressive speech in French, which makes it unique. However, the question of its intrinsic balance must be asked: is there the same number of sound samples for each attitude? And, is this distribution also balanced between male and female speakers ?

Att-HACK	**friendly**		**distant**		**dominant**		**seductive**	
	F	**M**	**F**	**M**	**F**	**M**	**F**	**M**
total duration *(h)*	3.67	2.97	3.82	3.12	3.64	3.01	3.83	3.05
sample mean duration *(s)*	2.67	2.42	2.74	2.75	2.61	2.52	2.78	2.83
number of samples	4953	4432	5015	4094	5007	4290	4955	3887

Table 3.1: Att-HACK metadata (duration and number of samples) statistics by attitude and speaker's gender.

Firstly, it can be seen that there is a roughly equivalent number of speech examples for each of the attitude categories. There are 9385 samples for friendliness, 9109 samples for distance, 9297 for dominance and 8842 for seduction. It can therefore be said that the database is rather balanced with regard to vocal attitudes. What about the distribution with respect to the gender of speakers? First of all, we have to mention that the database is constituted from the recordings of 12 females and 8 males. As a result, the number of samples associated with male gender (16703) is significantly lower than the number of samples associated with female gender (19930). Despite our best efforts, we were unable to attain perfect parity because only male speakers repeatedly withdrew. In this regard, the left part of the Figure 3.3 provides a clear account. We can therefore speak of an over-representation of the female gender in Att-HACK. If we now consider the distribution in regards of both attitude and speaker's gender, we see that the balance between attitudes is almost perfect for females. Conversely, there are differences in representation between attitudes for males. We acknowledge that creating a balanced database across all of its facets is exceedingly challenging. The challenge resides in both the logistics — being able to find the speakers we require — and the method used for recording and database post-processing.

The right part of the Figure 3.3 represents the distribution of duration for each attitude and each speaker's gender. As intended, the Att-HACK samples are rather short, thus learning conversion models from entire, unsegmented sound samples will be tractable. In particular, we found an average duration between 2 and 3 seconds for both female and male speakers. We can already see that the distributions of duration vary depending on attitudes, indicating that there may different production strategies in the speech rate, rhythm, and duration of phonemes depending on the portrayed attitude. These issues will be discussed further in Chapter 3 of this document.

Future efforts may be made to grow this database, especially by attempting to balance out the numerical disparity between male and female speakers. Such a balance is crucial in the quest for a comprehensive understanding of both the mechanisms underlying the production of vocal attitudes and those governing the perception of these attitudes. It is all the more necessary as we expect to see differences between male and female speakers in both of these factors.

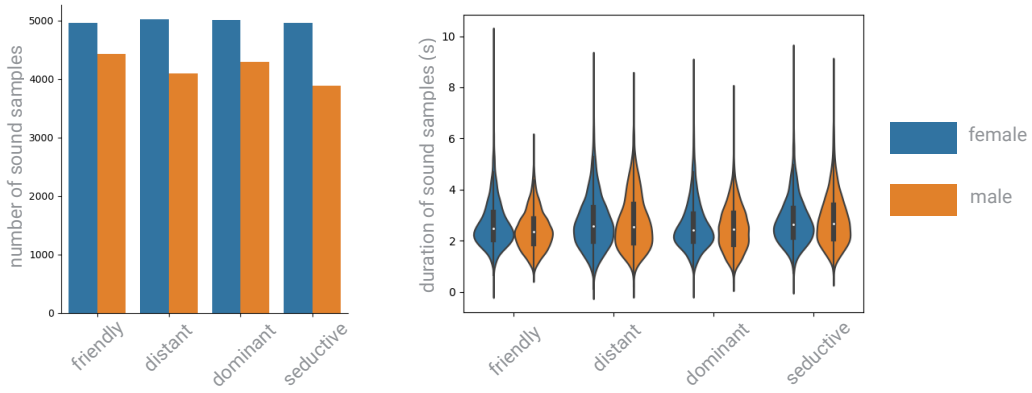

Figure 3.3: Representation of the sound samples's number (left) and duration (right) in Att-HACK depending on the portrayed attitude and the presumed speaker's gender.

3.3.3 Investigating pitch patterns underlying the production of vocal attitudes

Prior to making a first attempt to learn to convert vocal attitudes from Att-HACK data, it seems crucial to wonder what aspects of the signal convey these attitudes. Two main study areas underpin this question. On the one hand, the analysis of the vocal attitude production asks how speakers employ their vocal apparatus to produce one attitude or another. On the other hand, the analysis of attitude's human perception asks how do listeners decode the attitude conveyed by a speech signal. These two key areas enable us to understand the entire vocal attitude communication chain, from the intention to alter one's own voice by modifying one's body configuration to the capacity to capture specific elements of the speech signal in order to understand the social signal being communicated. A parametric model of vocal attitudes will describe the temporal variations of vocal parameters involved in the production and decoding of vocal attitudes. Such a model is difficult to develop in that it involves understanding both ends of the communication process.

We saw in the introduction that both attitudes and emotions, as they belong to the para-linguistic domain, are mainly encoded in the speech signal via the prosody. In addition, speech prosody can be broken down into four main phenomena: intonation, rhythm, intensity and voice quality. In particular, intonation - which is the most prominent component in the communication of para-linguistic content - is encoded from a signal point of view by the variations of the fundamental frequency - i.e. pitch or F0. The idea here is to observe pitch variations, the speech parameter which is commonly the most prominent in the communication of para-linguistic content.

Pitch contours

The prominence of pitch contours in the communication of emotions has been shown many times (Chuenwattanapranithi et al., 2007; Rodero, 2011; Amir and Globerson, 2014), while only few studies have dealt with attitudes. For instance, it has been shown that raising one's pitch helps to produce a friendly voice in English, Dutch, Chinese and Swedish (Chen et al., 2004; Li and Wang, 2004; House, 2005). The same strategy has been proven to convey dominance for English speakers in (Puts et al., 2006, 2007). Conversely, findings on vocal attractiveness report a lowered pitch for male speakers (Collins, 2000; Feinberg et al., 2005; Xu et al., 2013). Note that the attractiveness of a voice differs from what we call the seductive attitude. An attractive voice may not be seductive. It is even conceivable that a voice showing shyness or reserve may be attractive to some people.

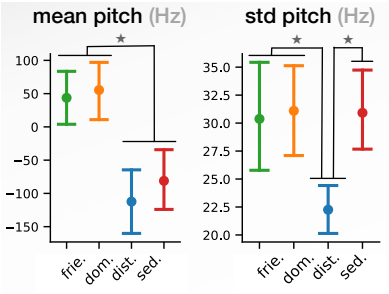

Figure 3.4: F0 analyses for friendliness (green), dominance (orange), distance (blue) and seductiveness (red). '' : statistically significant difference (p<0.05), '•' : marginally significant difference (p<0.1); paired t-tests. Error bars represent 95% confidence intervals on the mean.

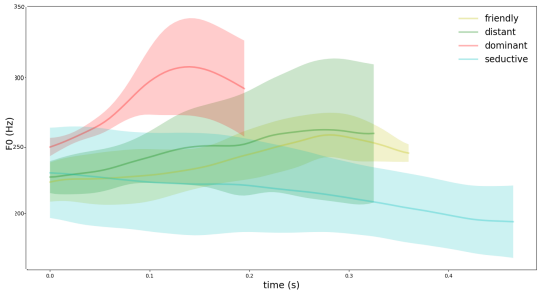

Figure 3.5: F0 contours mean (solid line) and standard deviation (filled with color) for the phrase *"Oui"* for a given female speaker

F0 estimation

The fundamental frequency of the speakers was estimated by using the SWIPEP algorithm (Camacho, 2007) with a minimum pitch value of 75 Hz, a maximum pitch value of 450 Hz, and a hop size of 5 ms, without any post-processing for correcting or smoothing the raw pitch values. The voiced/unvoiced decision was computed from the pitch strength associated with the pitch value estimate with a threshold of 0.25 (the pitch strength being a value between 0 and 1 corresponding to the periodicity of the speech frame).

A F0 contour was extracted for each syllable in order to illustrate the F0 patterns realized by the actors for the different social attitudes. The F0 contour of a syllable was identified as the one corresponding the longest sequence of F0 values considered as voiced over the syllable, i.e. the longest F0 segment for which each F0 value corresponds to a pitch strength value which is above a given threshold (in this work, 0.3).

F0 statistics on Att-HACK

A preliminary investigation was conducted to compare the F0 statistics of the actors across the social attitudes. We computed mean and standard deviation statistics on F0 segments extracted as stipulated above, for each speaker and attitude. We found a main effect of attitude on mean pitch (χ^2(3)=560, p$<$.001), std pitch (χ^2(3)=396, p$<$.001).

Figure 3.5 illustrates F0 countours distributions obtained for a given sentence with the four attitudes, each attitude being represented by a dedicated color. In each color, the solid line represents the mean F0 contour obtained by averaging the variations realized by the actor, the area filled with color represents the corresponding standard deviation around the pattern, and the length of the pattern the corresponding mean duration. This illustration reveals that distinctive F0 patterns are associated with the social attitudes, and also highlights the diversity of strategies employed by actors to communicate a social attitude.

3.4 Discussion

In line with the above outlined Att-HACK's metadata statistics, this section provides an insights about the design of such a dataset and the potential impact of the biases it introduces - or reproduces.

3.4.1 Balanced vs Imbalanced Data

A statistical analysis of the Att-HACK metadata allowed to establish the balance of the database with respect to the vocal attitudes that it intends to represent. On the other hand, there is a slight imbalance between female speakers and male speakers, which can prove to be problematic in bringing to light the potential gender effects encountered in both production and perception mechanisms that underlie vocal attitudes. This can also cause a slight bias of our conversion algorithm, in the sense that the algorithm would see more female speakers than male during training. The implicit definition of vocal attitudes learned by such an algorithm could then be biased in gender, it would better represent what females express than what males do. Through this conjecture, we see how the constitution of a database can influence the performance of the algorithm it feeds. In general, the biases observed when using different systems based on machine learning are almost always related to the dataset on which they were trained.

3.4.2 On Uncovering Data Biases

Two types of bias should be distinguished here. On the one hand there are those which are introduced by the constitution of a given database, because it does not reflect reality or because it is partial. We may think in particular of the non-authentic characteristic of the attitudes in Att-HACK - that have portrayed by actors - which thus potentially distances us from the reality of the vocal attitudes in the daily interactions between individuals. On the other hand, there are the biases that pre-exist the constitution of any database - i.e., the structural biases that are inherent in a society, a culture in a given time frame. We may wonder to what extent algorithms, such as those that this research proposes to design, should reproduce these biases in their uses. One could imagine using these algorithms as tools for social or even anthropological change by training them to either limit biases or even compensate for them. This perspective, as interesting as it is terrifying, deserves to be discussed a lot more broadly, gathering insights from a variety of disciplines such as machine learning, cognitive science and psychology. This is the purpose of the collaboration I started with clinical psychologist and PhD candidate Nadia GEROUAOU of the Perception and Sound Design (PDS) team at Ircam (Guerouaou et al., 2021).

In order to specifically uncover these biases, we intend to carefully examine the data gathered in Att-HACK, both from the perspective of the production of vocal attitudes and the perception of these attitudes by individuals. This is what the following Chapter 4 of this document aims to accomplish.

3.5 Chapter Summary

Att-HACK constitutes a first attempt to widen the scope of expressivity in speech, by providing a database of acted speech with social attitudes: friendly, seductive, dominant, and distant. The proposed database, recorded in French, comprises 20 speakers interpreting 100 utterances in 4

social attitudes, with 3-5 repetitions each per attitude for a total of around 27 hours of expressive speech. The Att-HACK is freely available for academic research under Creative Commons Licence. A publication at Speech Prosody 2021 - in Tokyo - (Le Moine and Obin, 2020) covers the elements presented in this chapter. A preliminary analysis showed that the pitch contours were involved in the production of the attitudes represented in Att-HACK. A first attempt at converting speech attitudes and based on such pitch contours is presented in Chapter 5. In addition, it must be noted that the attitude labels assigned to all Att-HACK's utterances only reflect the instructions given to the actors at recording. We thus do not know what the attitudes in Att-HACK *are*. In order to assess those attitudes, a proper uncovering of their production and perception mechanisms is initiated in the next Chapter 4, thus providing a global understanding of speech attitude communication as well as useful criteria to clean data for conversion learning.

Chapter 4

PRODUCTION STRATEGIES AND PERCEPTION OF SPEECH ATTITUDES

Contents

4.1 On the need for questioning the term attitude 67
4.2 Anatomical division of the vocal apparatus 68
4.2.1 Vocal fold behaviour 68
4.2.2 Vocal tract actuation 69
4.2.3 Phonetic structure 69
4.3 First Study - Uncovering the production strategies of vocal attitudes 70
4.3.1 Experiment 70
4.3.2 Results 71
4.3.3 Discussion 73
4.4 Second Study - Understanding the perception of vocal attitudes . . . 75
4.4.1 A Perceptual Validation Method Based on Best-Worst-Scaling (BWS) . . 75
4.4.2 Experiment 77
4.4.3 Preliminary Analysis 79
4.4.4 Understanding the human perception of speech attitudes 84
4.5 Chapter Summary 85

4.1 On the need for questioning the term attitude

In this chapter we aim to question the ontology of speech attitudes, i.e. what IS a speech attitude. A first step would be to question how we feel it, produce it, perceive it... People's various stances towards an attitude provide it very different definitions. Schematically, there are four ways to qualify the attitude related to a speech utterance in the scope of Att-hACK. We may speak of either

A PRIORI ATTITUDE. We refer to the instructions given to the actors, for instance "play it seductive", as a priori attitude. This first aspect is not particularly intriguing insofar as it may be reduced to an instruction. The direction may or may not have been followed, as more importantly, actors gave it a specific interpretation. However, at the end of the recording process, this was the only way to distinguish utterances from one another. We notably used these a priori attitudes for our first attempt of voice conversion described in chapter 5.

FELT ATTITUDE. One could feel seductive but not be able to infuse seductiveness in one's vocalizations. Conversely, one could feel friendly and be perceived as dominant. Therefore there is a gap between felt attitude and produced/perceived attitude. In this work, we do not access nor focus on this aspect of attitude, which is related to psychology, as its investigation would require a lot more information than recordings of the portrayed attitudes. However, it remains important to acknowledge the existence of that a apriori.

PRODUCED ATTITUDE. This aspect of attitude might be of more interest since it refers to how people *intend* to behave. A person will specifically set up their vocal apparatus to produce an attitude when they want to communicate it vocally. This intentional use, although not necessarily conscious, is referred to as a production strategy. In a first study (4.3), we attempted to uncover those strategies through an anatomically based acoustic analysis of the utterances in Att-HACK.

PERCEIVED ATTITUDE. The last definition of attitude deals with how people perceive it. It is noteworthy that it emphasizes the function that attitudes play in communication. From this perspective, unsuccessfully communicated attitudes must be, if not discarded, at least disregarded. In a second study 4.4, we try to understand how people decode - or perceive - vocal attitudes.

None of these elements can be excluded from a thorough account of what a vocal attitude is. Before delving deeper into these two studies, a first section 4.2 provides a schematic description of the vocal of how the vocal apparatus works. In order to investigate these different aspects, I decided to initiate a collaborative research thus surrounding myself with researcher and doctoral students who had different skills than mine and whose access to speech - as a research object - was also different than mine. While the layout of this chapter is my responsibility, its content is the result of this collaboration between myself and

- **Léane SALAIS**, doctoral student in machine learning the Analysis-Synthesis Team at IRCAM.
- **Pablo ARIAS**, postdoctoral researcher in cognitive (neuro)science at Glasgow University.
- **Victor ROSI**, doctor in psycho-acoustics formerly in the Perception Team at IRCAM.

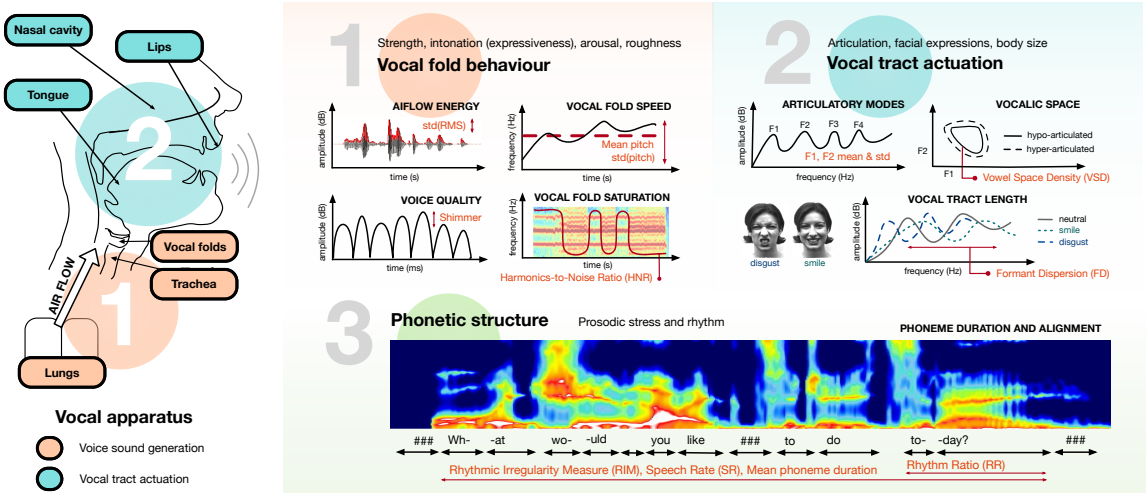

Figure 4.1: Anatomical voice production mechanisms and corresponding acoustic features. To describe the production strategies of vocal attitudes, we analyse three main categories of speech descriptors, relating to (1) vocal fold behaviour, (2) vocal tract actuation and (3) phonetic structure.

4.2 Anatomical division of the vocal apparatus

We built our analysis on a list of selected acoustic variables, following previous research on emotional speech (Arias et al., 2021). We split them into three clusters that reflect how the studied attitudes impact the speakers' control of their (1) vocal fold behaviour, (2) vocal tract actuation and (3) phonetic speech structure (Fig. 4.1).

4.2.1 Vocal fold behaviour

To quantify speakers' control over their vocal folds, we use acoustic descriptors of their vibration amplitude and rate.

Root Mean Square (RMS). First, we estimate the voice signal's RMS (dB) and its standard deviation (window size=2048). Although it reflects the general airflow energy which is partly affected by the activation of the vocal tract, we only look at it for its impact on vocal folds vibration strength. Indeed our anatomical division is schematic and therefore involves some simplifications.

Harmonics-to-Noise Ratio (HNR). It serves as an indicator of vocal fold saturation in vocalisations. In speech, HNR measures the energy ratio between harmonics produced by the vibrating vocal folds and the glottal noise in the spectrum. A sustained and subtle airflow produces harmonic vocal fold vibrations with a high HNR; in contrast, strong airflow from the lungs makes the vocal folds oscillate in non-linear or chaotic regimes, resulting in a rough voice (Erhard et al., 1999). This feature has been previously linked with e.g., aversiveness, arousal, negative valence and to some extent, emotion intensity (Anikin et al., 2020).

Shimmer. As a complementary measure of HNR, shimmer (dB) is associated with voice quality. Shimmer corresponds to the voice's amplitude variation over glottis cycles: high shimmer is often

associated with a breathy voice.

Pitch. Finally, we measure vocal pitch (Hz, mean and standard deviation), which reflects the vocal fold's vibration speed, i.e. the count of glottis cycles. Its variations summarise the modulations in intonation, a key feature to communicate vocal intentions, attitudes and emotions (Ponsot et al., 2018b; Rachman et al., 2018; Piazza et al., 2017).

4.2.2 Vocal tract actuation

Here, we pay particular attention to spectral parameters, which are traditionally less prevalent than parameters derived from vocal cord behavior in the acoustic analysis of affective signals. By extracting those parameters, we aim to investigate the speakers' strategies in terms of articulation.

Formants (F1 and F2). We measured the first and second formant frequencies (F1, F2, Hz, mean and standard deviation (**?**)), which represent the articulatory resonances of the vocal tract: they are impacted by the lips, mouth and tongue positions. Formants are not only essential to convey phonetic information, but also key to convey emotional information such as facial expressions (Arias et al., 2018).

Formant Dispersion (FD). To estimate the speakers' dynamic vocal tract elongation, we measure Formant Dispersion (FD) (Hz) i.e. the averaged difference between successive formant frequencies (F1 to F4). FD reflects the vocal tract length — which is also closely tied to body size (Anikin et al., 2022). Speakers can extend (lower FD) or shorten (higher FD) their vocal tract (Belyk et al., 2022) through facial expressions: previous results have reported an association between FD and expressions of emotions such as smiles (Drahota et al., 2008) and disgust (Chong et al., 2018). However, FD does not allow for an exhaustive description of the underlying articulatory strategies, e.g. switching from one articulatory mode to another by shifting only one formant (Pisanski et al., 2022).

Vocalic Space (VS). To accurately account for those strategies, we examine the vocalic space (VS), i.e. the space formed by F1 and F2 formants). We consider each vowel-related time frame in the dataset — extracted using the phoneme-to-audio alignments described in Chapter 3. To study the topology of this space, we compute the Vowel Space Density (VSD) (Story and Bunton, 2017). We first estimate a probability density function for the count of time frames located in the neighbourhood of each point in the space, and normalise the density to [0, 1] for each speaker and attitude. To account for prototypical strategies, we only keep samples located in high density areas (above a threshold of 0.5). VSD offers a holistic understanding of vocal articulatory strategies. Positions of attitude clusters in the vocalic space reflect how speakers articulate to convey attitudes (articulatory modes, e.g., closed/open mouth), while the surface covering all samples in the VS shows how much they articulate: the broader the surface, the easier it is to discriminate the vowels pronounced (Story and Bunton, 2017).

4.2.3 Phonetic structure

Finally, to investigate speech's phonetic structure, we take advantage of the phoneme-to-audio alignments and estimate several time related speech descriptors.

Speech Rate (SR). We estimate the Speech Rate (SR) — i.e. the mean number of phonemes per second in a speech utterance.

Rhythmic Irregularity Measure (RIM). The rhythmic irregularity measure quantifies the mean duration difference between all segments in a sentence (Gibbon and Gut, 2001). This feature yields indices on the global stability of speech rate. For a given utterance x composed with p phonemes of durations $\mathcal{D}_x = \{d_1, ..., d_p\}$, the RIM is formulated as follows

$$RIM = \sum_{d \in \mathcal{D}_x} \sum_{\substack{d' \in \mathcal{D}_x \\ d' \neq d}} \log \frac{d}{d'} \tag{4.1}$$

Rhythm Ratio (RR). The rhythm ratio is the mean duration difference between contiguous speech segments (Gibbon and Gut, 2001). Conversely to the rhythmic irregularity measure, the rythm ratio yields indices on the local stability of speech rate and is formulated as follows

$$RR = \frac{100}{p-1} \sum_{i=1}^{p-1} \frac{|d_{i+1} - d_i|}{(d_{i+1} + d_i)} \tag{4.2}$$

4.3 First Study - Uncovering the production strategies of vocal attitudes

In the light of this anatomical division the vocal apparatus functioning and the related measures we mentioned in the previous section - thus providing a general method for speech production assessment - this section presents an attempt to uncover the strategies that underlie the production of vocal attitudes. This study has been published at Interspeech 2022 (Salais et al., 2022).

4.3.1 Experiment

In this experiment, we exclusively focus on our Att-HACK dataset in which 20 actors are portraying 4 speech attitudes : friendly, distant, dominant and seductive.

A subset of Att-HACK

As the extraction of all the voice parameters (notably formants) involved in the study was very costly in terms of time and computational resources, we chose to use a subset of the of the Att-HACK database. We randomly sampled two recordings per speaker and per attitude for 62 sentences, thus obtaining 2400 recordings per attitude. The 62 sentences were selected to maximise semantic diversity, i.e. achieve an optimal coverage of the semantic space yielded by the CamemBERT (Martin et al., 2020) French language model. The second study on the perception of voice attitudes also uses the same subset — considered to be representative of the database.

Features extraction and normalization

We have extracted the features using Parselmouth [1], a Python library for the Praat [2]software. Praat (**?**) is a free scientific software package for the manipulation, processing and synthesis of speech sounds. It was developed at the Institute of Phonetic Sciences of the University of Amsterdam by

[1] `https://parselmouth.readthedocs.io/en/stable/`
[2] `https://www.fon.hum.uva.nl/praat/`

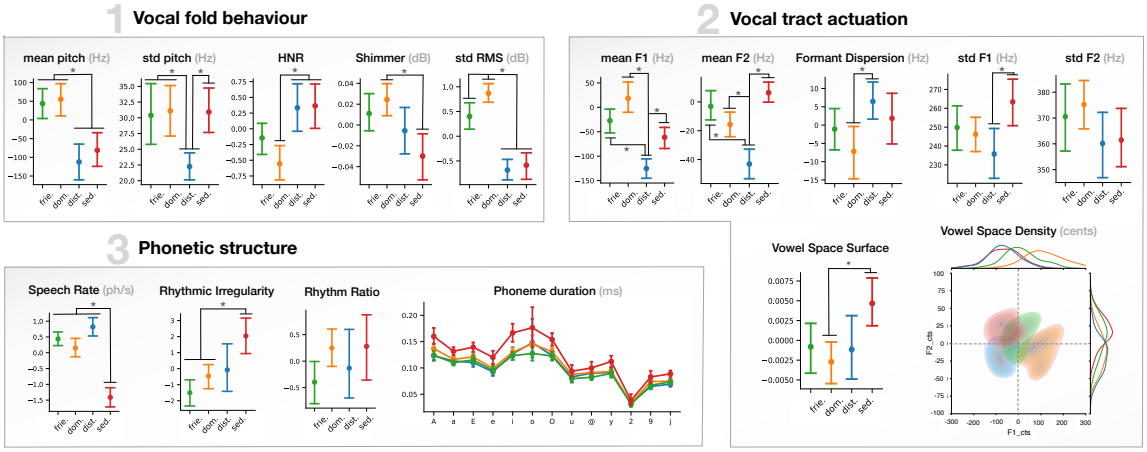

Figure 4.2: Feature analyses for (1) vocal fold behaviour, (2) vocal fold actuation and (3) phonetic structure on friendliness (green), dominance (orange), distance (blue) and seductiveness (red). '⋆' : statistically significant difference ($p<0.05$), '•' : marginally significant difference ($p<0.1$); paired t-tests. Error bars represent 95% confidence intervals on the mean.

Paul Boersma and David Weenink. It can run on a wide range of platforms. Praat is written in C++.

In particular, we used the following parameters for extraction: a time step of $10ms$ and a window size of $20ms$, a pitch floor and ceilling of respectively $45Hz$ and $600Hz$, a silence threshold of 0.1. We extracted 5 formants and used until 4 formants to compute formant dispersion. All the features are averaged over the whole utterance for each utterance - considering voiced parts - except for the calculation of the VSD which implies considering the temporal segments associated with vowels, for this metric only we keep temporal series of formants.

So as to observe intra-speaker variations of those features, we applied speaker normalization. For each speaker s and each acoustic feature x in Hz, we denote $\bar{x}^s$ the mean value of x over this speaker's utterances. Then the speaker normalized feature x^s is obtained through computing the distance - in cents - to the average value $\bar{x}^s$ such as

$$x^s = 1200 \log 2 \left(\frac{x}{\bar{x}^s} \right) \tag{4.3}$$

For all other features not expresses in Hz, we computed the speaker normalized feature x^s as

$$x^s = x - \bar{x}^s \tag{4.4}$$

To investigate articulatory strategies and temporal structure, we need to access the segmental information in speech (i.e. temporal information at the phoneme level). To infer it from Att-HACK recordings, we generate phoneme-to-audio alignments using a recent deep learning-based phonetic aligner (Teytaut and Roebel, 2021).

4.3.2 Results

In this subsection, we present the results obtained throughout the statistical analysis conducted.

Statistical analysis

To statistically evaluate the differences in vocal production strategies, we analysed acoustic features with GLMMs (Generalised Linear Mixed Models). We report p-values, estimated from hierarchical model comparisons using likelihood ratio tests (Gelman and Hill, 2006), and only present models that satisfy the assumption of normality (validated by visually inspecting the plots of residuals against fitted values) and statistical validation (significant difference with the nested null model). To test for main effects, we compared models with and without the fixed effect of interest. We performed post-hoc comparisons with paired t-tests, and applied Bonferroni corrections to correct for multiple comparisons. We report Cohen-d as a measure of effect size.

For each attitude, we present mean values of acoustic descriptors over full utterances. Because we are not investigating inter-speaker variability, but the speakers' own production strategies, we normalise features by speaker and get zero-centred values. Thus, variations between the conditions below reflect intra-speaker variations. In consequence, the statistical differences between attitudes bring out the shared part of the attitude production strategies among the speakers.

Vocal fold behaviour

We found a main effect of attitude on mean pitch (χ^2(3)=560, p$<$.001), std pitch (χ^2(3)=396, p$<$.001), HNR (χ^2(3)=83, p$<$.001), shimmer (χ^2(3)=59, p$<$.001) and std RMS (χ^2(3)=905, p$<$.001). Post-hoc analyses revealed that mean pitch was higher for dominance and friendliness as compared to distance and seductiveness (paired t-tests, p$<$.001, d$>$0.75). Speakers' pitch variability was also smaller for distance than for other attitudes (p=.05, d$>$0.9). On another line, friendliness and dominance seemed to be opposed to seductiveness and distance in terms of dynamics and roughness. HNR was significantly higher in dominant speech than in distant (p=.002, d=0.68) and seductive speech (p=.001, d=0.7); similarly, we found higher RMS variability for friendliness and dominance as compared to distance and seductiveness (p$<$.001, d$>$1.1). In addition, shimmer was significantly higher for dominant utterances as compared to seductive ones (p=.001, d=0.7).

Vocal tract actuation

We found a main effect of attitude for Formant Dispersion (FD) (χ^2(3)=61, p$<$.001), F1 (χ^2(3)=99, p$<$.001), F2 (χ^2(3)=37, p$<$.001), std F1 (χ^2(3)=73, p$<$.001) and std F2 (χ^2(3)=24, p$<$.001). Post-hoc analyses revealed that speakers significantly decreased their FD when producing dominance as compared to distance (p=.02, d=0.6). In line with this result, we found significantly lower F1 (p=.005, d$>$0.8) and F2 (p$<$.001, d$>$0.8) frequencies for distance, compared to all other attitudes. On another line, we found that distant utterances were produced with significantly more F1 variability as compared to seductiveness (p=.01, d=1.2), but found no significant differences for std F2. Finally, we only found a marginal difference between the surfaces of VSDs of the attitudes between seductive and dominant attitudes (p=0.06; see VSD plot in Fig. 4.2).

Phonetic structure

We found a main effect of attitude for Speech Rate (SR) (χ^2(3)=81, p$<$.001) and Rhythmic Irregularity (RIM) (χ^2(3)=18, p$<$.001), but no significant effect on Rhythm Ratio (RR) (χ^2(3)=4.2, p=.23). That is, attitudes influence global rhythmic patterns rather than local ones. Post-hoc analyses revealed that seductive samples had a significantly lower SR (p$<$.001, d$>$1.24) and higher RIM as compared

to friendliness and dominance (p=.02, d$>$0.7). The duration of all vowels was also extended accordingly.

4.3.3 Discussion

In the present study, we investigated how speakers modulate their voice to communicate vocal attitudes. To do this, we analysed the vocal production of dominant, friendly, seductive and distant attitudes in a multi-speaker and multi-attitude French database. For each attitude, we reported the changes in the speakers' vocal fold behaviour, vocal tract actuation, and phonetic speech structure.

Production strategies of vocal attitudes

In the following we discuss the findings for each attitude in comparison with others when statistically relevant. We obtained two statistically strong prototypes for dominance and seductiveness and two weaker ones for friendliness and distance.

Friendliness. Friendliness was produced with a raised and dynamic voice (high pitch, high std RMS). The speed and regularity of friendly versus seductive speech (higher SR, lower RIM) may hint at an uncomplicated and extraverted persona (Mairesse et al., 2007). These results are in line with cross-lingual literature for English, Dutch, Chinese and Swedish (Chen et al., 2004; Li and Wang, 2004; House, 2005).

Distance. The production strategies were of particular interest for distant speech. Indeed, distance was conveyed by fast speech that lacks expressiveness (low and steady pitch, high SR vs. seductiveness), pronounced with a low mouth aperture and a shortened vocal tract (low F1 and F2; high FD compared to dominance). In light of these results, it seems that when producing distance, speakers do not put much effort into being understood. Their calmness (e.g. high HNR when compared to dominance) suggests that their rendition of distance is close to indifference. Distance is hence distinct from neutrality, and could be interpreted as a marker of dissent, mistrust, or disgust.

Dominance. In line with previous findings, we found that dominance was expressed through a vocal tract elongation (lower formant dispersion) (Feinberg et al., 2005; Puts et al., 2007) as well as a rough and dynamic voice (low HNR, shimmer, high std RMS). However, contrary to previous findings, speakers raised their pitch in comparison with other attitudes (Puts et al., 2006, 2007). This discordance may be explained by the language setting, culturally learned vocal associations, or more simply by the fact that previous studies contrasted dominance with neutral speech and not other vocal attitudes (Ponsot et al., 2018b).

Seductiveness. We also found strong prototypical strategies for seductiveness, which was produced with low pitch, low dynamics (low std RMS), and a relatively high harmonic content (high HNR, low shimmer) in comparison to other attitudes. Importantly, we also found a strong effect of seduction on speech's phonetic structure. Specifically, seductive utterances were produced with a slow and irregular rhythm, as if speakers took time to expose their intentions. Previous findings on vocal attractiveness report a lowered pitch for male speakers (Feinberg et al., 2005; Xu et al., 2013). We complement these findings by studying seduction as a modulated vocal attitude, rather than an intrinsic vocal trait, and highlight the specific modulations that all speakers use to convey seductiveness.

Communicative signal of vocal intent

To our knowledge, this is the first study to reveal diverging voice production strategies at the articulatory level. Specifically, we found that speakers' productions were distributed across specific clusters in the vowel space (Fig. 4.2-2). For example, we found that distance had a lower F1 than other attitudes, suggesting that distance is produced with a more closed mouth. Similarly, analysing the Vowel Space Density surface revealed that some attitudes span more articulatory modes than others. For instance, the vowel space for seductiveness was marginally wider than for dominance, which, in complement with formant dispersion findings, suggests that speakers switched between articulatory modes to produce vocal attitudes, by e.g, restraining or modulating their articulatory range. This result suggests that subtle cues in speech articulation can convey a communicative signal of vocal intent.

Overall, these results shed light on the social intentions behind the production of social attitudes. For example, speakers limited their vocal expressivity to sound distant and hinted at a larger body size to sound dominant. Such behaviours may be closely interpreted from a social perspective, revealing the links between attitude-specific vocal behaviours and higher-order cognitive mechanisms (Goupil et al., 2021b). However, it is important to highlight that the vocalisations analysed herein were produced by actors, and actors' vocalisations are known to be less authentic than spontaneous ones (Anikin and Lima, 2017) — which, in the case of e.g. facial expressions of emotions, even seem to rely on different neural bases (Valente et al., 2017). In any case, these results uncover the shared strategies used by speakers to volitionally produce vocal attitudes.

4.4 Second Study - Understanding the perception of vocal attitudes

The study presented in the previous section has shed light on the strategies used by French speakers to produce vocal attitudes. On the opposite end of the communication channel lies the issue of how people *perceive* these attitudes. To address it, we conducted a listening experiment to question the perception of vocal attitude in a group of 100 participants.

4.4.1 A Perceptual Validation Method Based on Best-Worst-Scaling (BWS)

A variety of techniques could help us to assess the perception of sound attributes such as Rating Scale (RS) (Friedman and Friedman, 1997) and Best-Worst Scaling (BWS) (Louviere et al., 2015) methods. We chose to use BWS because it has proven to be very effective for similar tasks such as the perception of sound attributes like brightness, roughness, roundness and warmth (Rosi et al., 2022). Applying it to the perception of vocal attitudes is also an opportunity to test it again and thus to encourage other uses.

Best-Worst-Scaling (BWS)

The BWS technique involves a discrete choice experiment repeated on trials. For each trial, the subject listens to n_t - usually 4 or 5 - sounds and evaluates their perception with respect to a studied term - in our case a vocal attitude. If we use friendliness as an example, then for each trial, the subject will have to judge which of these sounds he or she perceives as the friendliest and as the least friendly. The former will be referred to as the trial's best, the latter as the trial's worst. Each term (attitude) is evaluated independently. Once the BWS experiment is complete, we rank all the assessed sounds on a scale ranging from 0 to 1. If we get back to using friendliness as an example, the sounds assessed would sit on a scale going from least perceived friendly to most perceived friendly. A fundamental point is that the evaluation is relative, e.g. friendliness would only be questioned with respect to the other in the dataset. The BWS differs from the rating scale in this regard because the latter entails making *absolute* judgments, thus making the task more difficult. In fact, although not experimentally proven, it appears to be fairly harder to determine the friendliness of a speech utterance by rating it on a scale than by comparing it with another utterance in a binary way. Therefore, by opting for BWS, we cannot answer the question of how vocal attitudes are perceived generally. The question we ask is undoubtedly less ambitious, but it is more likely to yield interesting results about the data we have at our disposal. Furthermore, as highlighted in section 4.3, Att-HACK provides a great variety of vocal attitude production strategies (20 speakers, several versions with fixed sentence, attitude and speaker); this method, at the very least, allows to distinguish between sounds whose a priori attitude is well perceived (at the top of the BWS scale) and those whose related attitude is poorly perceived (at the bottom of the BWS scale). Such a distinction might enable us to clean up the database by discarding the samples with miscommunicated attitude. Furthermore, we hope to see prototypical attitude instances emerge.

Designing the trials

Denoting n the number of sounds to assess and n_t the number of sounds per trial (here we chose $n_t = 4$), then the BWS method is likely to converge if $2n$ 4-tuples are designed with respect of the

following constraints

- Each 4-tuple cannot contain the same sound twice
- Each sound appears in 8 distinct 4-tuples

Provided that it would be impossible, both in terms of time and financial resources, to validate the complete database by assessing every sound in it for each attitude, we came up with a two-stage agenda depicted in Figure 4.3.

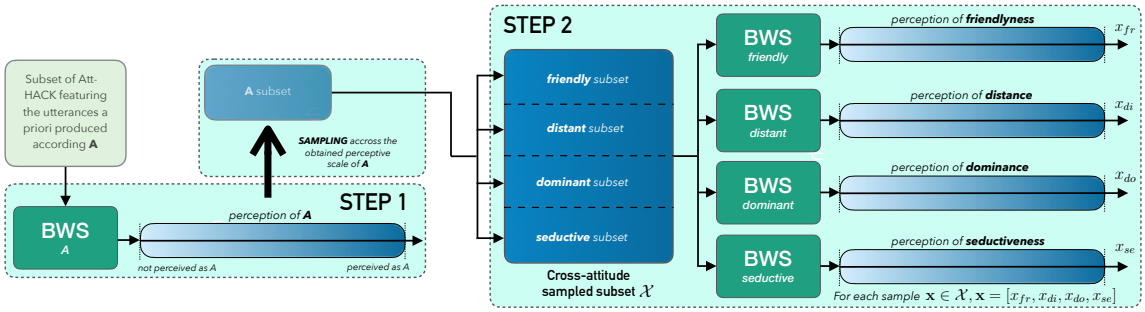

Figure 4.3: Two-step perceptual validation based on BWS experiments. At the time of writing, only the first step has been completed.

STEP 1 : A priori attitude perceptual assessment

The first phase involves, analyzing solely the perception of its related sounds for each a priori attitude, i.e. the ones already produced with aim of conveying this specific attitude. At the end of this first stage, we obtain four scales ranking the sounds of each a priori attitude. It is expected that this will bring out the clearest prototypical examples of attitudinal expression as well as the poorest productions overall.

STEP 2 : Assessment of the interaction between attitudes

We then study the interaction of attitudes with one another, i.e. whether a sound could be perceived with a different attitude than the one for which it was produced. Sounds from the four a priori attitudes are selected for each questioned attitude and submitted for judgment. Again, from a time and budget viewpoint, interrogating every sound for every attitude is unfeasible. As a result, we must choose a subsample that still properly reflects the dataset's diversity. To do this, we suggest sampling the obtained BWS scales uniformly over the whole range. Sampling can be adjusted to the distribution of sounds within each scale, i.e. for a given sampled BWS value, a number of sounds would be collected proportionally to the density of sounds present in the area around this value. After sampling, we obtain a subspace $\mathcal{X}$ of the Att-HACK in which the four a priori attitudes are equally distributed. It remains to test this subspace for each attitude using BWS. At the end of the experiment, each sound in $\mathcal{X}$ would be represented in a quadri-dimensional space.

4.4.2 Experiment

At the time of writing, only the first step was carried out. The second step dealing with interactions between attitudes remains to be done due to lack of time and money. The completion of Att-HACK's perceptual validation procedure hence depends on the strategy adopted by the researchers that will pursue this work. Thus, we only describe the first step experiment here.

On preparing the experiment

We used the same Att-HACK subset as used in the first study. We sampled each audio at 16kHz and stored in wav format. For each attitude, we had 120 sets of 4-trials designed from the Att-HACK subset. These series are indexed in groups of 3 so that each participant examines only one third of the subset associated with a given attitude.

In order to save time, we retrieved Victor ROSI's BWS experimental assessment interface for the investigation of sound attributes (bright, round, warm and rough) (Rosi et al., 2022). This interface, based on the graphical programming tool Max [3] is depicted in Figure 4.4. At each trial, the participant is asked to judge four sounds. It is mandatory to listen to the four sounds before moving on the next trial. If the participant spends too much time on a single trial, a red light flashes. It is also mandatory to choose distinct best and worst sounds at each trial.

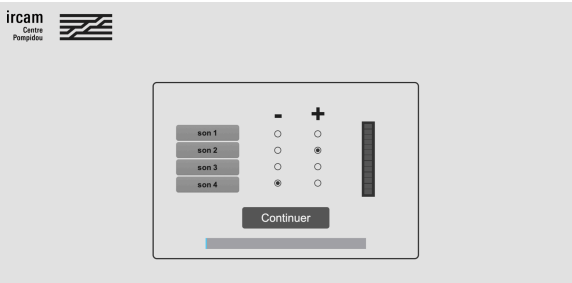

Figure 4.4: Max interface for BWS experiment based on work by Rosi (Rosi et al., 2022).

On conducting the experiment

We conducted the experiment ourselves, so that we could control the experimental setup, meet the subjects and discuss the their experience of the test.

Centre for Behavioural Sciences at INSEAD-Sorbonne. To do this we collaborated with the INSEAD-Sorbonne behavioural lab, which provides facilities (isolated cabins, rooms, etc.), equipment (computers, headphones, etc.) and recruitment logistics for scientific experiments. The disciplines covered range from experimental ethics to neuroscience. Working with this organisation has the huge advantage of not having to worry with subject recruitment or overall planning.

[3] https://cycling74.com/products/max

We had $N = 100$ subjects recruited. The experiment spanned a full week with 5 to 6 test slots a day and 3 to 6 booths hosting the experiment in parallel. This experimental set up required us to be very rigorous and focused, particularly in assigning trial sets to subjects. In fact, as mentioned above, each sound must be judged eight times in order to be able to rank the sounds, which means that it is not sufficient to provide each participant with a random set of trials; the entire set of trials given across participant must match this requirement. Throughout the week, Léane SALAIS provided me with invaluable assistance, without which it would not have been feasible to conduct the experiment.

As a basic guideline, we operated as follows. At a given time slot, we first welcomed the scheduled participants, seated them in one of the booths and invited them to read and sign a consent form. They had to watch a short explanatory video, after which we answered any subsidiary questions and launched the experiment. Halfway through the experiment, the participants had the opportunity to go out, perhaps have a drink and stretch their legs. At the end of the experiment - which lasted for about one hour in average, the participants went out for a debrief. We described the objectives of the experiment and answered any remaining questions.

On assessing the relevance of the experiment

Before any attempt at analyzing the results of this BWS experiment, we need to answer some preliminary questions. Firstly, we need to check that the participants could actually complete the task. Thanks to the debriefing sessions at the end of the tests, we were already able to affirm that no participant was overwhelmed by the task to be accomplished. However, the difficulty level varied depending on whether the participant had to evaluate one or another attitude. In particular, distance elicited many reactions describing the task as difficult and tiring, which was not the case for the other attitudes. In general, the participants found the task feasible and rather interesting. This matches our observation in terms of the average time taken by the participants to complete the experiment, which varies according to the attitude considered, but is close to one hour, as presented in Table 4.1. They often pointed out the difficulty of judging *only* the attitude, as the interaction with the linguistic content seemed strong to them. Some also mentioned the potential polysemy of the terms used to describe an attitude. In particular, distance could mean both a form of reserve and indifference.

	friendly	**distant**	**dominant**	**seductive**	**across attitudes**
Average duration (min)	61	60	56	53	**57**
Compliance	0.81	0.73	0.77	0.79	**0.78**

Table 4.1: Several indicators related to the conducted BWS experiment for each attitude.

Measuring participant compliance, i.e. the degree to which each participant agrees with the median opinion, is necessary to determine how well the experience worked from an objective standpoint. The compliance is calculated for each participant as follows. For each trial t^a given to participant for an attitude a, we consider all sound pairs (x, y) such as $x, y \in t^a$ and at least one sound in the pair has been judged best or worst. This ensures both sounds in the pair are linked by an order relation. Let us assume x has been perceived less a than y then $x \prec_a y$ denotes the relation induced by the participant's judgement between x and y. The idea behind compliance is to compare this relation to the mean relation across all participants. To do so we use the BWS scores assigned to each of these sounds s_x^a and s_y^a, the compliance for this pair can be computed as follows

$$c^a_{(x,y)} = \begin{cases} 1 & \text{if } s^a_x \leq s^a_y \\ 0 & \text{else.} \end{cases} \tag{4.5}$$

The full compliance score is computed for a participant evaluating attitude a as the sum of $c^a_{(x,y)}$ across all pairs satisfying the conditions mentioned above. It is then averaged across participants as shown in the last row of Table 4.1. Note that a participant giving random answers would receive a compliance score of 50%. In our case, the outcomes show that participants generally agree with one another on whether they perceive someone's voice as friendly, distant, dominant, or seductive. However, there are differences depending on the attitude being considered. For example, distance is associated with the lowest compliance ($C = 0.73$) and, as a result, is the least consensual attitude which was already noticed when talking to the participants. On the other hand, friendliness seems to be the most consensual attitude ($C = 0.81$).

4.4.3 Preliminary Analysis

The perceptual data collected potentially reflects other speech attributes such as linguistic content or gender that influence the communication of speech attitudes. Before seeking to understand the perception of speech attitudes, we need to question these potential biases by analyzing the interaction between the scores obtained and these different speech attributes.

Perceptual scores x linguistic content

First, we aim to investigate the interaction between the linguistic content of the evaluated sentences and the individual perception of attitudes. Indeed the linguistic content can be found to interact with attitudes through their suggested emotion. In order to investigate the perception of attitudes, we first need to evaluate this suggested emotion.

SENTIMENT ANALYSIS. Sentiment analysis captures the dominant emotional opinion in an input text through a score. This score reflects the emotion suggested by the sentence, i.e. the emotion felt, on average, by the person reading it. In our case, the vocal attitudes overlay the emotion lying in the linguistic. To do so, we designed our proper sentiment analysis study asking 65 participants to rate all Att-HACK sentences's emotional valence on a 7-point Likert scale with 0 meaning neutral valence.

The results are depicted on Figure 4.5 in which attitudes are considered independently. For each attitude, we plot the BWS scores averaged sentence-wise (blue line) and the sentiment analysis scores (red line). Provided this Figure, we may acknowledge that there is an interaction between the perception of attitude and the emotion lying in the linguistic content, at least for some specific attitude. It nonetheless comforting (though rather obvious) to note that this interaction does not fully account for the obtained perceptual scores. This figure effectively conveys the idea that attitudes are perceived *conditionally* on the linguistic context that *partly* determines how they are expressed (encoded) and perceived (decoded).

That being said, certain attitudes seem to be more vulnerable to this interaction with linguistics than others. To assess this point, we computed the correlation between sentiment analysis and BWS scores for each of the attitudes. The graphs obtained are shown in Figure 4.6. The statistical results showed significant effects for friendly - with a Pearson coefficient of 0.47 ($p < 0.001$) - and dominant - with a Pearson coefficient of -0.41 ($p < 0.001$).

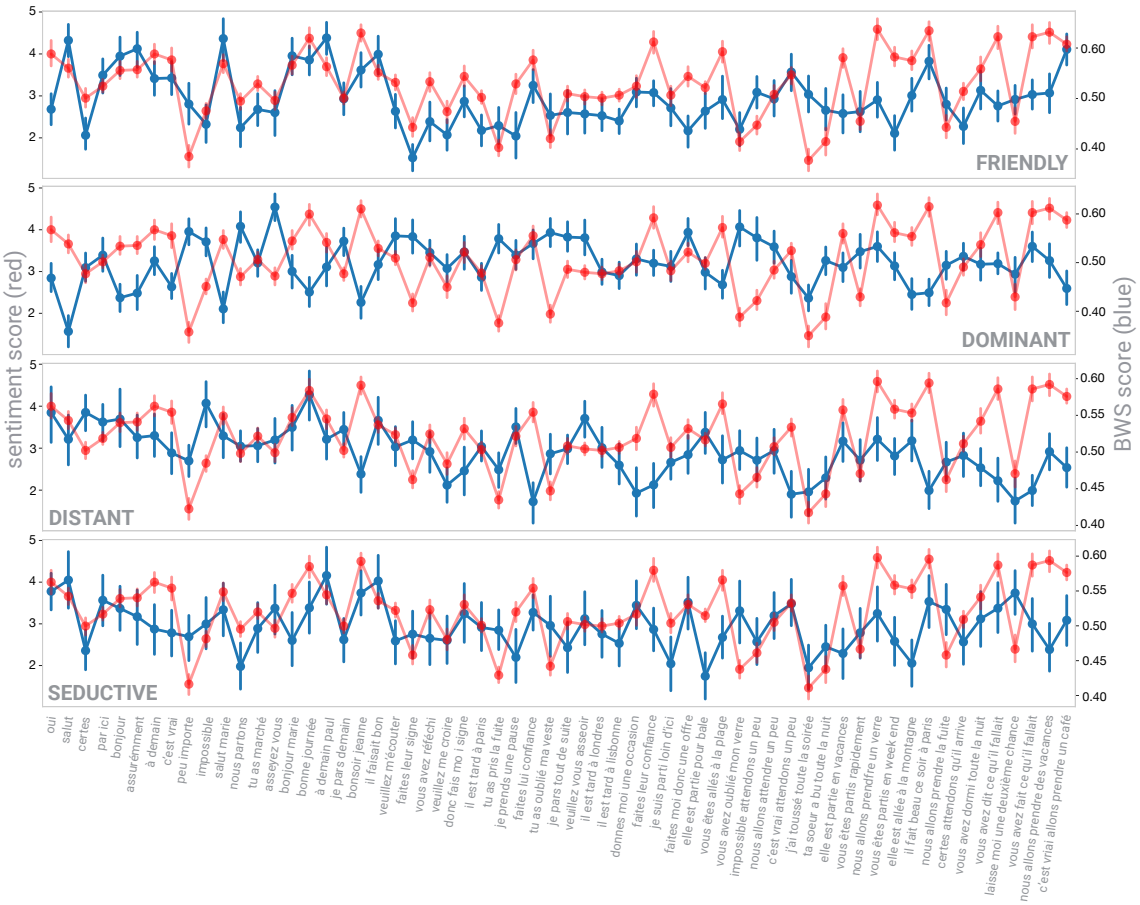

Figure 4.5: Concurrent display of the scores yielded by the sentiment analysis (red) and the BWS experiment (blue), for each considered sentence (x-axis)

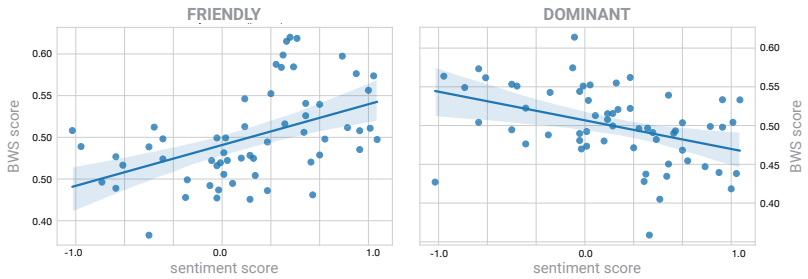

Figure 4.6: Correlation between the scores yielded by the sentiment analysis and the perceptual scores from the BWS experiment for $friendly$ and $dominant$ attitudes.

To conclude on this, the sentences in the database cannot be regarded as neutral, they have a meaning that denotes either a rather positive or negative sentiment. Reflecting this evoked sentiment, an emotional valence score is assigned to each sentence through sentiment analysis. Significant effects for friendliness and dominance were found when investigating this score in interaction

with the BWS's perceptual score, thus showing that attitude perception is influenced by linguistic content. In particular, perception of friendliness is significantly correlated to the emotional valence carried by said linguistic content, i.e. the more the sentence evokes a positive emotion, the more the utterance will be perceived as friendly. Conversely, dominance seems to be negatively correlated with the evoked emotional valence, i.e. the more the sentence evokes a negative emotion, the more the utterance will be perceived as dominant. We did not find any significant effect for the other two attitudes, namely distance and seduction.

Perceptual scores x speakers' vocal gender

We investigated the interaction between the perceptual score obtained and the gender of the speaker who produced the evaluated utterance. It is important to note that the actors were not asked any questions about their gender. The gender category mentioned here is therefore what we, as experimenters, can infer from perceived vocal characteristics (pitch, length of vocal tract, ...) as being classically associated with a female or a male. This category might be referred to as vocal gender. Given the significant research conducted on gender recently, the association of vocal masculinity and femininity with particular traits must be challenged. However, this is not the focus of this work; rather, we question the interaction between the degree to which a vocal attitude is perceived, i.e., how effectively it has been communicated, and the presence of specific vocal characteristics that are usually associated with either masculinity or femininity in the speaker's voice.

In essence, demonstrating a gender influence on the perception of vocal attitudes entails bringing out the intra-individual perception variations for male and female speakers. There are several ways to observe this interaction, two of these are depicted in Figure 4.7. A first approach consists in using raw data, i.e. unnormalized data. This allows to observe intra-individual variations, however such differences can be either amplified or compensated by inter-individual ones. To account for this, we adopted another approach (on the right in Figure 4.7) that involves normalizing scores so that the data related to each speaker, has a zero mean and a unit standard deviation for each attitude.

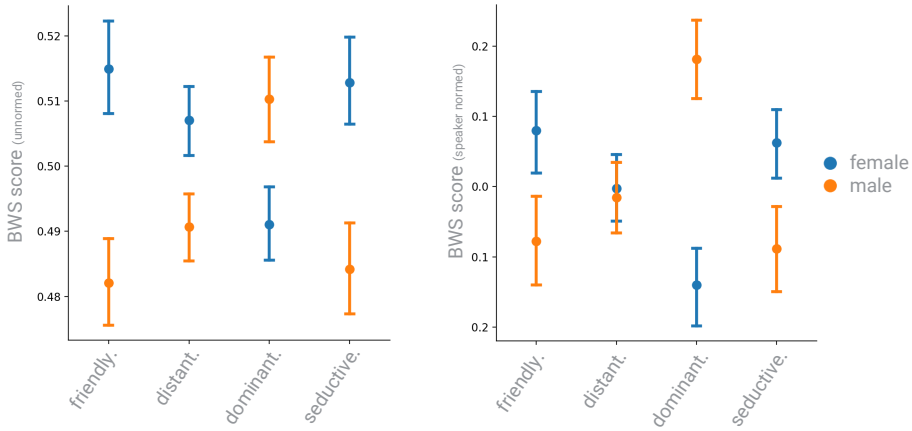

Figure 4.7: BWS scores (left) and speaker-wise normalized BWS scores (right) depending on attitude and speaker's vocal gender.

When looking at the graph on the left, for raw data, we observe a significant interaction between vocal gender and attitude perception, i.e. it seems that certain attitudes are better communicated depending on whether the speakers' vocal characteristics are classically associated with males or females. It is also observed that the effect is not the same depending on the attitude considered. Friendliness, distance and seduction seem to be best communicated by female speakers. Conversely, dominance seems to best communicated by male speakers.

As explained above, this representation does not allow one to discriminate the differences related to intra-individual variations from those related to inter-individual variations. To do this, we need to consider the graph representing the normalized data per speaker. Firstly, we notice that the gender effect for distance disappears. This suggests that the effect observed in the left-hand graph (unnormalized data) is due to inter-individual variation rather than to real differences in intra-individual perception. Then, it appears that the effects observed for friendliness and seductiveness persist when looking at the standardised data but are weaker, so some of the effects observed were due to inter-individual variations in the perception of friendliness and seductiveness. Conversely, the effect observed for dominance appears to be even stronger when considering speaker-normalised data. This suggests that inter-individual variations tended to mask intra-individual ones.

The question we must then ask is: how can we explain this gender effect? We provide two possible explanations, which are not mutually exclusive. Firstly, it may be that the better communication of certain attitudes by speakers of one gender is due to a *production* advantage, i.e. a better physiological capacity to produce and thus communicate these attitudes. For example, we found in section 4.3 that dominance was produced by speakers lengthening their vocal tract, and men usually have a longer vocal tract than women, which can be seen as an advantage of men over women in producing vocal dominance. Similarly, we observed that friendliness is produced by speakers increasing their pitch, but it is known that female speakers have a higher average pitch than male speakers, which can be seen as an advantage of women over men in the production of friendliness. The second potential cause of this interaction between vocal gender and the attitude perception is a decoding bias, i.e. a culturally constructed difference in the *perception* of each of these attitudes depending on whether it is expressed by a male or female speaker. It can be hypothesised that certain sexist prejudices, culturally constructed, learned and therefore relative to a community of individuals sharing this culture, imply that we globally perceive female speakers as being more seductive than male speakers. Similarly, the dominant position of men in a patriarchal society (or at least one with patriarchal structures) can potentially imply that we perceive male speakers as more dominant overall than female speakers. Attributing one or the other of these causes to the different effects observed is very difficult and would require looking at the mental representations that individuals attribute to different attitudes.

One last question remains: would these gendered differences in the perception of vocal attitudes exist outside the experimental framework in which we observe them? The task given to participants in the BWS experiment may force them to make a decision that they might not normally make in real life. By requiring participants to choose between sounds, we potentially force categories to appear. We should mitigate our conclusion by pointing out that gender differences do emerge in the perception of vocal attitudes within our experimental paradigm, which involves participants choosing between sounds.

Attitude perception x participant presumed genre x speaker's vocal gender

We could even go further and question the interaction between the presumed gender of the participant and the vocal gender of the speaker being judged. Prior to anything else, it is critical to note that participants in the BWS study were not questioned about their gender. As a result, the gender category that we assign them matches our perception of their gender as experimenters. Depending on the attitude considered, is a participant more likely to perceive a sentence whose vocal gender is the same as his or her own as conveying it the best? Or is it otherwise?

The scores that represent the overall average judgment of the participants do not provide an answer; we therefore need to examine the specifics of the trials. For a particular attitude, each trial was judged by a single participant. Our idea is to assess the participant's presumed gender versus the vocal gender of the utterance that, in his or her opinion, best conveys the attitude considered within this trial. By collecting this information for all trials and across the four attitudes, we produce the graph shown in Figure 4.8.

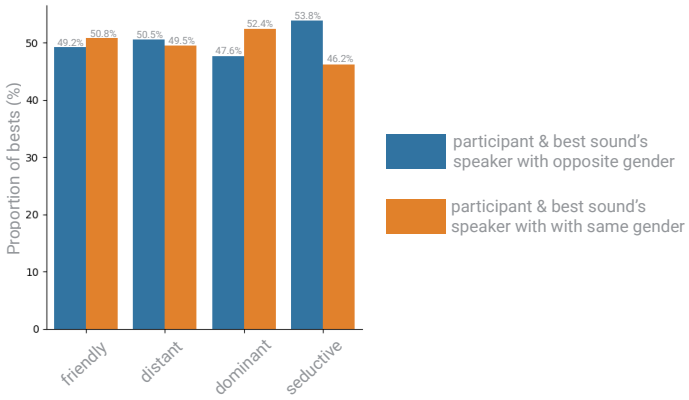

Figure 4.8: Match and mismatch between the participant's presumed gender and the vocal gender associated with the sentence he/she considers best in a trial. The graph represents the proportion of gender match and mismatch per attitude.

We observe a notable difference between the proportion of matches and mismatches for all attitudes. Nevertheless, the effect appears to be much more significant for two attitudes. Firstly, in 53.8% of the trials, participants considered sentences whose associated vocal gender was the opposite of their own as conveying seduction, compared to only 46.2% if it was the same as their own. Conversely, in 52.4% of the trials, participants considered sentences whose associated vocal genre was the same as their own as conveying dominance, compared to only 47.6% if it was the opposite of their own. Smaller effects are observed for the other two attitudes. In 50.8% of the trials, participants considered sentences whose associated vocal gender was the same as their own as conveying friendliness, compared to only 49.2% if it was the opposite of their own. Finally, in 50.5% of the trials, participants considered sentences whose associated vocal genre was the opposite of their own as conveying distance, compared to only 49.5% if it was the same as their own.

4.4.4 Understanding the human perception of speech attitudes

In the previous section, we asked how speakers use their vocal apparatus to convey different attitudes such as friendliness, distance, dominance and seduction. We showed that speakers share common strategies for producing these attitudes, which can be observed through statistical analyses of variations in certain speech parameters. Analogously, the question here is about decoding attitudes communicated through speech signals. What, specifically, are the vocal characteristics whose variations enable listeners to perceive the attitude being conveyed?

A simplistic response would be to claim that the same signal properties that enable the decoding of the conveyed attitude are also those that encode attitudes, i.e., the properties that make up the production strategies uncovered in the previous section. The process that enables the decoding of an attitude would thus be the mathematical inverse of the process by which the attitude was acoustically encoded. This conception of speech communication draws upon the legacy of first communication thinkers like Shannon.

In section 4.3, we tried to describe the differences between attitudes, in the way speakers used their vocal apparatus so as to convey them. The resulting attitude production profiles are therefore based on a statistical analysis of inter-attitude differences. In contrast, the BWS experiment conducted on Att-HACK consists of asking participants to make a preference judgment between different instances of the same attitude. Therefore, the attitude decoding strategies that we expect to find are based on intra-attitude differences. On the one hand, we wonder what distinguishes the production of two different attitudes. On the other hand, we wonder what distinguishes the perception of two instances of the same attitude.

Assessing the participants' perceptual strategy for each attitude

In this part we try to understand how participants decode attitudes. In particular, we want to determine which speech parameters allow individuals to perceive an attitude? To do so we mean make the assumption that, for any utterance, the higher the perceptual score, the more it conveys signal aspects that are salient for individual attitude decoding. We used a standard XGBoost (Chen and Guestrin, 2016) - a scalable tree boosting regression model - to learn to predict BWS scores from various speech features - the same than those used to uncover the production strategies of attitudes. Each utterance is thus represented by a vector of features $\mathbf{x}$ whose each element is a static feature - i.e. averaged over the temporal dimension. Once the model has learned to predict the BWS scores - with more or less precision - for a given attitude, we must attempt to interpret its predictions. We thus sought to determine which feature - and in what extent - the model uses to predict the BWS scores obtained for each attitude. To do so, we used SHAP-values [4].

SHapley Additive exPlanations (SHAP). SHAP is a method for explaining the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions (Lipovetsky and Conklin, 2001; Ribeiro et al., 2016).

The SHAP-values results are depicted in Figure 4.9 for the four attitudes. Two standard regression metrics are indicated in the up-left corner of each sub-figure: the coefficient of determination $r2$ - that measures the variance explained by the model - and the mean absolute prediction error

[4] `https://shap.readthedocs.io/en/latest`

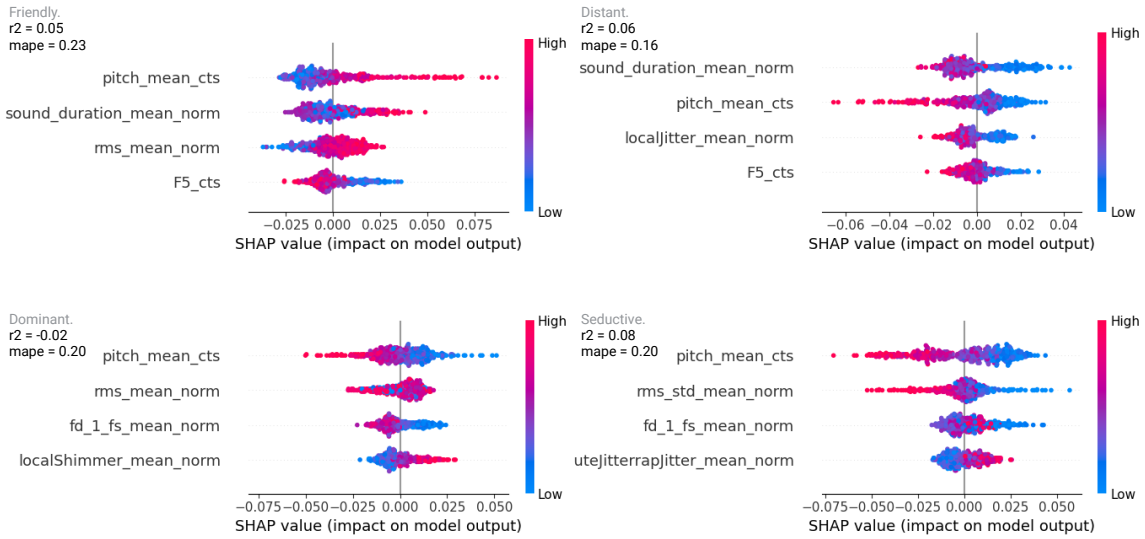

Figure 4.9: SHAP-values for friendly, distant, dominant and seductive.

$mape$ - that measures differences between actual and predicted scores. In addition, the four features that helps the model the most for predicting the BWS score are shown. Their value can be positively - red dots - or negatively - blue dots - correlated to the prediction performance of the model. Firstly, we find that the model does not succeed - or only slightly - in predicting BWS scores from the static features it takes as input. In one hand, the coefficient of determination $r2$ is slightly above 0, which means that the share of variance in BWS scores that is explained by the regression model is very small. In the other hand, $mape$ is close to 0.2 but - as most of the utterances are ranked with average scores (around 0.5) - this does not indicate that the model is able to predict the BWS score accurately. If some features seem to have significant impact on the model's predictions, its poor performance prevent from drawing any relevant conclusion out of it.

In conclusion, it is clear that static features - i.e. averaged over utterances' duration - are not predictive of BWS scores for speech attitudes as they are for sound attributes such as studied in (Rosi, 2022). It is likely that individuals use much more complex cues to decode attitudes. In particular, it is reasonable to hypothesize that temporal variations in different speech parameters play a crucial role in the perception of speech attitudes. In order to assess this hypothesis, we plan to adapt this principle of explained regression to the case of temporal sequences. By doing so, we expect to better understand the mechanisms underlying the perception of vocal attitudes.

4.5 Chapter Summary

Two studies - one uncovering the production strategies of speech attitudes and the other mainly hinting at biases involved in the perception such vocal attitudes - provided a first twofold account for how speech attitudes are communicated by individuals. The first study - which have been published at Interspeech 2022 (Salais et al., 2022) - allowed for the identification of two strong attitude production profiles - dominance and seductiveness - and two weaker ones - friendliness and distance. The second study - still unpublished at the time of writing - led to reveal that certain speech attributes such as linguistic content or gender do influence speech attitude perception. In partic-

ular, we found that some attitudes were better communicated by speakers of a given gender. We also found the sentences of Att-HACK to have strong influence on how people perceive attitude in speech. Given those findings, we attempted to understand how individuals decode attitudes without reaching a satisfying answer. It is likely that they process speech parameters - or aspects - temporally. We look forward to assessing this assumption in future works. We also plan to complement with studies on felt attitudes - investigating mental representations - to achieve a full-stack understanding of vocal social attitudes.

Chapter 5

NEURAL F0 CONVERSION OF SPEECH ATTITUDES

Contents

5.1 Related Works on F0 Contours Modelling 88
5.1.1 Pitch Contours (F0) - A First Parametric Approach 88
5.1.2 Multi-level modelling by applying CWT to F0 signals 89
5.1.3 CWT Adaptive Scales 90
5.2 Related Works on Dual-GAN-Based Voice Conversion 91
5.2.1 Generative Adversarial Network 91
5.2.2 Dual implementation of GAN 92
5.3 Contribution 92
5.3.1 Wavelet Kernel Convolutional Encoder 92
5.3.2 Framework overview 93
5.4 Speech Attitude Conversion Experiment 94
5.4.1 Implementation Details 95
5.4.2 A One-to-One Speaker Dependant Conversion Experiment 96
5.4.3 Training Procedure 97
5.5 Results & Discussions 98
5.5.1 Objective Evaluation 98
5.5.2 Subjective Evaluation 100
5.5.3 General discussion 102
5.6 Chapter Summary 104

The chapter 4 provides a substantial understanding of the communication of vocal attitudes. The choice of how to model the speech signal for the purpose of speech attitude conversion as well as how to learn to convert attitude in speech signal should be made in the light of this analysis. In this chapter we propose a first algorithm that aims at converting the speech attitude by changing the F0 of the speech signal. This research has been published at EUSIPCO 2021 (Le Moine et al., 2021b).

5.1 Related Works on F0 Contours Modelling

As mentioned above, we chose to focus on pitch contours - acoustically correlated to F0 variations - as a first - parametric - approach for the modelling of speech attitudes. In this section we review different methods for modelling pitch contours focusing on multi-level modelling such as provide by the application of the Continuous Wavelet Transform (CWT) to F0 contours, on which our proposal is based.

5.1.1 Pitch Contours (F0) - A First Parametric Approach

According to the study of both production and perception of vocal attitudes carried out in Chapter 4, pitch contours play a prominent role in the communication of speech attitudes. On the one hand, we found that distance is produced with low pitch standard deviation when compared with other attitudes, friendliness and dominance are produced with a raised pitch while speakers lower their pitch to produce seductiveness. Therefore, in order to convert attitudes, changing F0 is mandatory even though it is not enough. On the other hand, we have not yet gone far enough in understanding the individual perception of attitudes to say precisely how the pitch is used by individuals to decode attitudes. However, the experiment in learning a regression model on BWS scores from static features revealed that the mean pitch was an important static parameter for predicting scores.

Modelling speech intonation and related F0 contours is a challenging task that has been faced in the past decades for a variety of speech applications: from text-to-speech, voice identity conversion, and speech emotion conversion among others. The representation of such pitch variations is a challenging task for at least two main reasons. First, the F0 sequence corresponding to a speech signal is discontinuous by nature - F0 values are only over speech segments that are voiced, and undefined otherwise. Second, the F0 varies over multiple time scales associated with pre-defined linguistic units - e.g., syllable, phrase - or with latent units. Each of these temporal scales being associated with specific functions: linguitic - F0 helps to clarify the syntactic structure of an utterance or is used for semantic emphasis - and para-linguistic - F0 is used to encode emotional or attitudinal information. Accordingly, a number of models have been proposed to model F0 variations. First, basically as a linear sequence of F0 values defined at each time step, either from discontinuous raw F0 values or from continuous interpolated F0 values over voiced instants. Second, as a parametric stylization of the defined F0 values over linguistic units, based on the decomposition of the F0 values over a set of slow time-varying functions, pre-defined as the Discrete cosine transform (DCT) Teutenberg et al. (2008) or learned from speech datasets Obin and Belião (2018). Third, using multi-scale modelling, from multi-linear modelsGerazov et al. (2018) to more complex models such as the continuous wavelet transform (CWT) decomposition of F0 variations over multiple time scales Luo et al. (2017). These representations have been largely designed and exploited for generative modelling tasks, such as text-to-speech synthesis (TTS) and voice conversion Black et al. (2007); Latorre and Akamine (2008); Obin et al. (2011); Obin (2011); Veaux and Rodet (2011); Yin et al. (2016).

5.1.2 Multi-level modelling by applying CWT to F0 signals

Following the approach presented in (Luo et al., 2017; Luo et al., 2019), we decided to learn conversions from appropriate representations of pitch contours rather than do it directly from raw - or interpolated - temporal F0 sequences. We expect that by feeding the conversion algorithm with such salient representation of speech attitudes the conversion task will be made simpler to learn.

Multi-level aspect of F0

The speech prosody, of which intonation - represented by the F0 parameter - is one of the most important aspects, is characterized by subtle variations at multiple temporal levels. Micro prosody refers to as little - potentially uncontrolled - variations in speech due to vocal apparatus that appear at the phoneme level while macro prosody designates the global variations at the sentence contour level. These different levels of temporal variation encode different speech aspects. For example, the linguistic functions of prosody are rather encoded through global macro-prosodic variations. Thus, selecting certain levels rather than others can help to focus on certain information while rejecting others. The CWT computes a decomposition of the F0 signal over wavelet kernels which allows a representation of F0 over different temporal scales (Ming et al., 2015), with various application in expressive voice conversion (Ming et al., 2015, 2016; Luo et al., 2017; Luo et al., 2019; Zhou et al., 2020).

Continuous Wavelet Transform (CWT)

As a multi-scale modelling method, the Continuous Wavelet Transform (CWT) is entirely fitting when trying to represent both long and short-term dependencies, prosody is influenced by. In particular, the parameters that define the CWT correspond to the temporal levels whose variations the CWT intends to models. As CWT can only be applied to continuous functions, a simple linear interpolation between voiced F0 segments is needed to obtain a continuous phrase-related F0 function, as depicted in Figure 5.1, which can then be sampled in a vector $\mathbf{x} \in [0, 1]^T$.

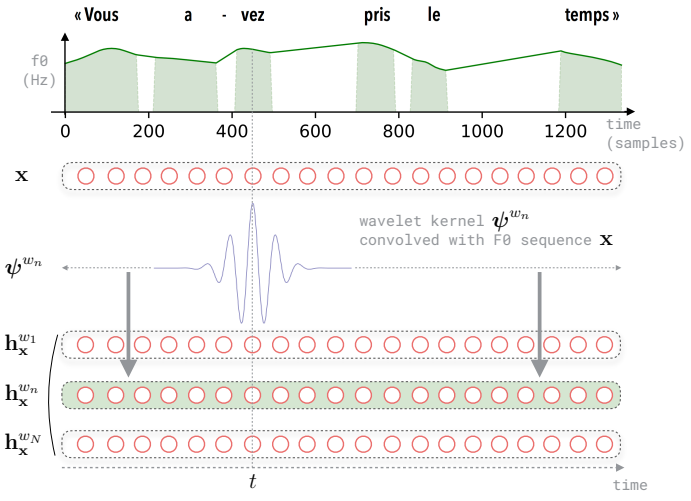

Figure 5.1: CWT applied to interpolated F0 contours

A wavelet is a wave-like function of summable square from Hilbert $\mathcal{L}^2(\mathbb{R})$ space representing an oscillation, not necessarily sinusoidal, whose envelope width and oscillation frequency are related. In this work we use the mother wavelet $\boldsymbol{\psi_w} \in \mathbb{R}^T$ defined for a time vector $\mathbf{t} \in \mathbb{R}^T$ as

$$\boldsymbol{\psi}_w(\mathbf{t}) = \frac{2\pi^{-\frac{1}{4}}}{\sqrt{3}} \left[1 - \left(\frac{\mathbf{t}}{w}\right)^2\right] e^{-\frac{1}{2}(\frac{\mathbf{t}}{w})^2} \tag{5.1}$$

For the sake of clarity, we will misuse $\boldsymbol{\psi}_w$ instead of $\boldsymbol{\psi}_w(\mathbf{t})$ in the following. The projection $\mathbf{h}_\mathbf{x}^w$ of the F0 signal $\mathbf{x}$ on the wavelet $\boldsymbol{\psi}_w$ represents the F0 variations that occur on the temporal scale w and is formulated as follows

$$\mathbf{h}_\mathbf{x}^w = \mathbf{x} * \boldsymbol{\psi}_w \tag{5.2}$$

The CWT representation of a signal $\mathbf{x}$ is its decomposition on a wavelet basis $\{\psi_w\}_{w\in\mathbb{W}}$ covering a range of N temporal scales $\mathbb{W} = \{w_1, ..., w_N\}$ and defined as $\mathbf{H}_\mathbf{x}^\mathbb{W} = [\mathbf{h}_\mathbf{x}^{w_0}, ..., \mathbf{h}_\mathbf{x}^{w_N}]$. The F0 signal is therefore represented by an 2D image showing its temporal evolution at different temporal scales, from micro-prosody (small scales) to macro-prosody (large scales). If we denote $\mathcal{A}_{cwt}^\mathbb{W}$ the operator for computing such a decomposition and $\mathcal{R}_{cwt}$ the reconstruction operation, then the analysis-reconstruction process of a signal $\mathbf{x}$ can be written as follows

$$\mathbf{H}_\mathbf{x}^\mathbb{W} = \mathcal{A}_{cwt}^\mathbb{W}(\mathbf{x}) \tag{5.3}$$

$$\hat{\mathbf{x}} = \mathcal{R}_{cwt}(\mathbf{H}_\mathbf{x}) = \frac{d_j\sqrt{d_t}}{C_d Y_0} \sum_{n=1}^{N} \mathbf{h}_\mathbf{x}^{w_n} + \bar{\mathbf{x}} \tag{5.4}$$

with $\bar{\mathbf{x}}$ the average of $\mathbf{x}$, $d_t = 1.2$, $d_j = 0.125$, $C_d = 3.541$ and $Y_0 = 0.867$ (for details, see (Torrence and Compo, 1998)).

5.1.3 CWT Adaptive Scales

Numerous approaches dedicated to speech prosody modelling are based on the use of CWT. A promising approach called CWT Adaptive Scales (CWT-AS) was proposed by Luo et al. (Luo et al., 2019).

An adaptive multi-level modelling of pitch contours

F0 modelling with CWT has been specified more recently upgraded with the possibility to compute the decomposition on arbitrary linguistic scales (e.g., phoneme, syllable, word, and utterance as described in (Luo et al., 2017)). An Adaptive-Scale (AS) algorithm (Luo et al., 2019) is described to select an optimal CWT representation for each pair of emotions, by selecting the scales that maximize in average the distance between the emotions in the CWT space. From these selected scales, the CWT decomposition of the F0 contours is computed. Finally, the transformation function between each pair of emotion is learned from those representation using a Dual-GAN.

Limitations

Though this approach appears promising, it suffers from two main limitations. First, the scale selection is only based on the maximization of the distance between the emotions, but ignores their reconstruction ability of the F0 signal. This may lead to poor F0 reconstruction which in turn would

degrade the quality and the naturalness of the transformation. Second, the CWT-AS decomposition of the F0 signal and the dual-GAN are optimized independently which constitutes a bottleneck for training. Consequently, the CWT decomposition may not be optimal in the sense of the dual-GAN objective.

5.2 Related Works on Dual-GAN-Based Voice Conversion

In (Luo et al., 2019), the function that maps source and target CWTs is learnt by a Dual-Generative Adversarial Network - or Dual-GAN. This section provides a general description of the GAN paradigm and its dual implementation.

5.2.1 Generative Adversarial Network

First introduced by Goodfellow in (Goodfellow et al., 2014), Generative Adversarial Network designates both a global architecture consisting of two main blocks - a generator G and a discriminator D - and a learning mode - i.e. a way to optimise the learning of these two modules. To learn the generator's distribution p_G over data $\mathbf{x}$ we first define a prior on input noise variables $p_{\mathbf{z}}(\mathbf{z})$. The generator is then defined through $G(\mathbf{z}, \theta_G)$ where G is a differentiable function represented by a multilayer neural network whose θ_G are the learnable weights. The discriminator is defined through $D(\theta_D)$ where D is a differentiable function represented by a multilayer neural networks whose θ_D are the learnable weights. While the output of the generator has the same shape as input data, the output of the discriminator is a scalar such as $D(\mathbf{x})$ represents the probability that $\mathbf{x}$ came from input data distribution rather than generator's output distribution. The discriminator D is trained to maximize the probability of assigning ones to the input samples and zeros to the samples yielded by G, thus considering the former as true samples and the latter as false ones. Simultaneously, the generator G is trained to trap the discriminator so that it can no longer distinguish the generated samples from the real ones. The whole architecture is depicted in Figure 5.2. Denoting $\mathbb{E}$, the mathematical expectation, D and G play the following two-player min-max game

$$\min_G \max_D \mathcal{L}_{adv}(D, G) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}\left[\log D(\mathbf{x})\right] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}\left[\log(1 - D(G(\mathbf{x})))\right] \tag{5.5}$$

Since the adversarial component as a supplementary constraint favorize the convergence to the true distribution of the original data - conversely to the standard auto-encoding paradigm - GANs are proven to yield realistic conversion. They also have the advantage of being non-deterministic, with a fixed input, the addition of noise makes it possible to generate different conversions each time whose adversarial loss guarantees the quality.

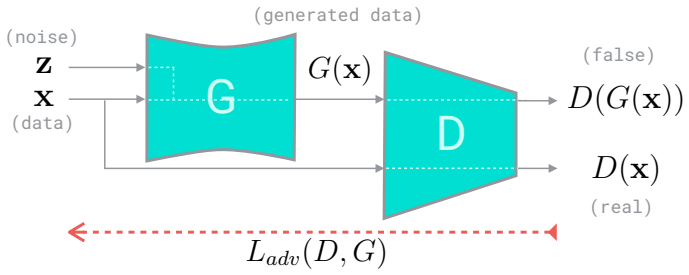

Figure 5.2: Schematic view on the Generative Adversarial Network

5.2.2 Dual implementation of GAN

Inspired by dual learning from natural language translation (He et al., 2016), (Yi et al., 2017) proposed a novel dual-GAN mechanism, which enables image translators to be trained from two sets of unlabeled images from two domains. In our architecture, the primal GAN learns to translate images from domain U to those in domain V , while the dual GAN learns to invert the task. In 2019, (Luo et al., 2019) proposed this same architecture to address the problem of voice conversion. Their system, referred to as dual Supervised Adversarial Network (dual-SAN) is trained to convert the emotion in the Mel Cepstral Coefficients (MCC) and CWT-F0 features. The duality here is not to be compared with the mathematical concept but refers to the joint learning of a forward conversion from emotion a_s to emotion a_t and a backward conversion from emotion a_t to emotion a_s, depicted in Figure 5.3. Therefore, the basic GAN functioning described through equation 5.5 can be extrapolated in a dual version as

$$\min_{G_{s\to t},G_{t\to s}} \max_{D_s,D_t} L_{d-adv}(D_s, D_t, G_{s\to t}, G_{t\to s}) = L_{adv}(D_t, G_{s\to t}) + L_{adv}(D_s, G_{t\to s}) \tag{5.6}$$

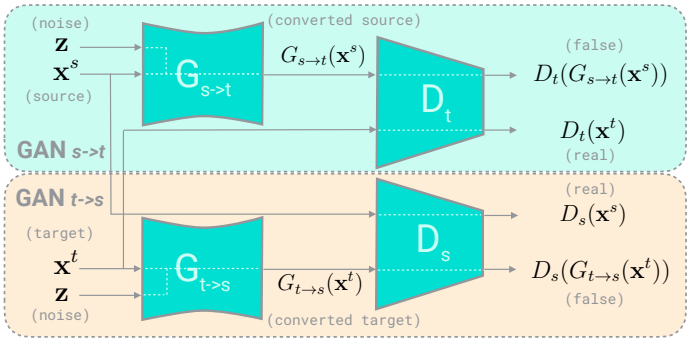

Figure 5.3: Schematic view on the dual Generative Adversarial Network (dual-GAN)

This dual mechanism is shown to improve the training performances by leveraging the probabilistic connection between both tasks to enhance the learning.

5.3 Contribution

To overcome the limitations of (Luo et al., 2019), we propose an end-to-end architecture to learn efficiently F0 transformation between attitudes. The proposed neural architecture brings together the F0 decomposition and the dual-GAN into a single network, so that the CWT decomposition is optimized in the sense of the dual-GAN objective, and combining separation and reconstruction losses of the resulting decomposition.

5.3.1 Wavelet Kernel Convolutional Encoder

This work's first contribution is our newly proposed Wavelet Kernel Convolutional Encoder (WKCE), which is a custom layer for learning CWT representations of 1-D sequences and that we use as a pre-network to encode source and target F0 sequences. As our proposed voice conversion algorithm requires parallel data, sets of utterances $\mathcal{X}^s$ and $\mathcal{X}^t$ respectively relative to attitudes s and t

are considered. A pair of utterances is then sampled and source $\mathbf{x}^s$ and target $\mathbf{x}^t$ F0 sequences are extracted. Aside from the attitude, each utterance in a pair has the same content (linguistic content, speaker identity).

If we consider $\mathbf{x}$ sampled from data distribution $P(\mathbf{x})$, this module can be trained for reconstruction objective with respect to L_1 loss formulated as

$$L_{rec}(\mathbf{x}) = \mathbb{E}_{\mathbf{x}\sim P(\mathbf{x})}\left[||\mathcal{R}_{cwt}(\mathcal{A}_{cwt}(\mathbf{x})) - \mathbf{x}||_1\right] \tag{5.7}$$

A constraint of classification on the CWTs latent space can be added, $\mathcal{A}_{cwt}$ and C are trained with respect to L_{cl}, the cross-entropy (CE) loss between actual emotion $\mathbf{a}$ of an utterance $\mathbf{x}$ and its prediction by the classifier $\mathcal{C}$ applied to its representation $\mathcal{A}_{cwt}(\mathbf{x})$. The classification loss L_{cl} is formulated as

$$\mathcal{L}_{cl}(\mathbf{x}) = \mathbb{E}_{\mathbf{x}\sim P(\mathbf{x})}\left[\mathbf{a}\log(\mathcal{C}(\mathcal{A}_{cwt}(\mathbf{x}))) + (1-\mathbf{a})\log(1-\mathcal{C}(\mathcal{A}_{cwt}(\mathbf{x})))\right] \tag{5.8}$$

5.3.2 Framework overview

Here we present the complete proposed system bringing together this specific F0 encoding module (WKCE) with a Dual-GAN architecture, thus forming an end-to-end system for F0 conversion.

CWT-based Pre-Network

The source and target F0 are given to what we called a Wavelet Kernel Convolutional Encoder (WKCE) denoted $\mathcal{A}_{cwt}$. A classifier, denoted C, whose objective is to predict the expressivity is fed with WKCE outputs. As shown in Figure 5.4, these two modules must be seen as a pre-network (pN) for Dual-GAN (DG) that can be pre-trained as well as trained along with Dual-GAN forming an end-to-end system for f0 conversion.

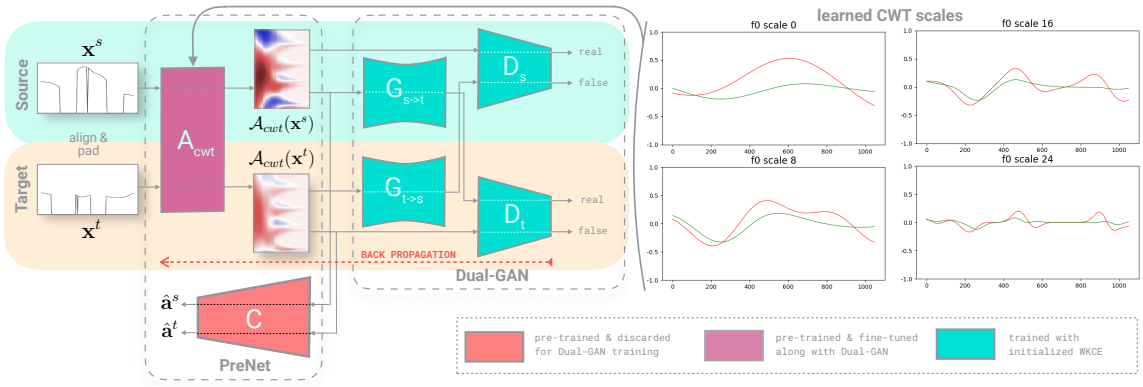

Figure 5.4: End-to-end neural architecture for F0 voice conversion. On the right, F0 decomposition over 4 of the learned scales obtained for the source (red) and target (green) expressivities.

Dual-GAN for conversion learning

The algorithm we used for conversion - referred to as Dual-GAN - is based on two concepts: Adversarial learning (Goodfellow et al., 2014) - that involves training a generative model to find a solution in a min-max game between two neural networks, called as generator G and discriminator D - and Dual supervised learning (Xia et al., 2017) - that involves training the models for two dual tasks simultaneously, thus exploiting the probabilistic correlation between them to regularize the training process. Combining those aspects allows to take advantage of the GAN's ability to produce realistic transformations as well as the significant improvements due to dual supervised learning.

This last point implies that both forward and inverse transformations, respectively $\mathcal{G}_{s\to t} : (\mathcal{A}_{cwt}(\mathbf{x}^s), z^s) \to \mathbf{x}^t$ and $\mathcal{G}_{t\to s} : (\mathcal{A}_{cwt}(\mathbf{x}^t), z^t) \to \mathbf{x}^s$, are learned jointly, where z^s and z^t are random independent noises provided in the form of dropout at each layer of $\mathcal{G}_s$ and $\mathcal{G}_t$. Two losses $L_{s\to t}$ and $L_{t\to s}$ are required to train $\mathcal{G}_{s\to t}$ and $\mathcal{G}_{t\to s}$ respectively

$$\mathcal{L}_{s\to t}(\mathbf{x}^s, \mathbf{x}^t) = \mathbb{E}_{(\mathbf{x}^s,\mathbf{x}^t)\sim P(\mathbf{x}^s,\mathbf{x}^t)}(||\mathcal{R}_{cwt}(\mathcal{G}_{s\to t}(\mathcal{A}_{cwt}(\mathbf{x}^s))) - \mathbf{x}^t||_1) \tag{5.9}$$

$$\mathcal{L}_{t\to s}(\mathbf{x}^s, \mathbf{x}^t) = \mathbb{E}_{(\mathbf{x}^s,\mathbf{x}^t)\sim P(\mathbf{x}^s,\mathbf{x}^t)}(||\mathcal{R}_{cwt}(\mathcal{G}_{t\to s}(\mathcal{A}_{cwt}(\mathbf{x}^t))) - \mathbf{x}^s||_1) \tag{5.10}$$

At the same time, $\mathcal{D}_s$ has to discriminate between converted outputs $\mathcal{G}_{t\to s}(\mathcal{A}_{cwt}(\mathbf{x}^t))$ - considered false - and actual input CWT representations $\mathcal{A}_{cwt}(\mathbf{x}^s)$ - considered true. Analogously - to complete the adversarial mechanism - $\mathcal{D}_t$ has to discriminate between converted outputs $\mathcal{G}_{s\to t}(\mathcal{A}_{cwt}(\mathbf{x}^s))$ - considered false - and actual input CWT representations $\mathcal{A}_{cwt}(\mathbf{x}^t)$ - considered true. Two adversarial losses $\mathcal{L}^s_{adv}$ and $\mathcal{L}^t_{adv}$ are required to train $\mathcal{G}_{s\to t}$, $\mathcal{G}_{t\to s}$, $\mathcal{D}_s$, $\mathcal{D}_t$ and $\mathcal{A}_{cwt}$

$$\mathcal{L}^s_{adv}(\mathbf{x}^s, \mathbf{x}^t) = \mathbb{E}_{\mathbf{x}^s\sim P(\mathbf{x}^s)}[\mathcal{D}_s(\mathcal{A}_{cwt}(\mathbf{x}^s))] + \mathbb{E}_{\mathbf{x}^t\sim P(\mathbf{x}^t)}[1 - log(\mathcal{D}_s(\mathcal{G}_{t\to s}(\mathcal{A}_{cwt}(\mathbf{x}^t))))] \tag{5.11}$$

$$\mathcal{L}^t_{adv}(\mathbf{x}^s, \mathbf{x}^t) = \mathbb{E}_{\mathbf{x}^t\sim P(\mathbf{x}^t)}[\mathcal{D}_t(\mathcal{A}_{cwt}(\mathbf{x}^t))] + \mathbb{E}_{\mathbf{x}^s\sim P(\mathbf{x}^s)}[1 - log(\mathcal{D}_t(\mathcal{G}_{s\to t}(\mathcal{A}_{cwt}(\mathbf{x}^s))))] \tag{5.12}$$

A third constraint called Dual loss is added so as to strengthen the intrinsic connection between $G_{s\to t}$ and $G_{t\to s}$, it can be understood as a regularization of the process.

$$\begin{aligned}\mathcal{L}_{dual}(\mathbf{x}^s, \mathbf{x}^t) = \mathbb{E}_{(\mathbf{x}^s,\mathbf{x}^t)\sim P(\mathbf{x}^s,\mathbf{x}^t)}(||&\mathcal{A}_{cwt}(\mathbf{x}^s) * \mathcal{G}_{s\to t}(\mathcal{A}_{cwt}(\mathbf{x}^s)) \\ &-\mathcal{A}_{cwt}(\mathbf{x}^t) * \mathcal{G}_{t\to s}(\mathcal{A}_{cwt}(\mathbf{x}^t))||_1)\end{aligned} \tag{5.13}$$

Therefore two final losses can be formulated for pre-Net pretraining and proper Dual-GAN training, respectively $\mathcal{L}_{pN}$ and $\mathcal{L}_{DG}$ with α, β, λ and γ respectively weighting reconstruction, classification, transformation and dual objectives.

$$\mathcal{L}_{pN} = \alpha\mathcal{L}_{rec} + \beta\mathcal{L}_{cl} \tag{5.14}$$

$$\mathcal{L}_{DG} = \lambda\mathcal{L}_{a\leftrightarrow b} + \mathcal{L}_{adv} + \gamma\mathcal{L}_{dual} \tag{5.15}$$

5.4 Speech Attitude Conversion Experiment

In this section we present an application of our proposed model to speech attitude conversion and compare it with the baseline CWT-AS approach.

5.4.1 Implementation Details

Our proposal and the CWT-AS baseline were implemented in Python 3.6 and Tensorflow 2.1. and trainings were performed on GeForce GTX 1080 Ti GPUs. In the following, we give all details about the design and implementation of our proposed architecture.

Input pipeline

We extracted fundamental frequency from the speech signal by using SWIPEP algorithm (Camacho, 2007). All F0 sequences are sampled to 1ms (as recommended in (Luo et al., 2019)), passed to log(F0) and a linear interpolation has been processed between voiced segments. For each pair, a mapping between syllables starting and ending times has been done to align source and target. Once pairs are aligned syllable-wise, F0 sequences are padded with zeros up to a value T_{max} corresponding to the longest sentence in the dataset.

Wavelet Kernel Convolutional Encoder

Our WKCE denoted $\mathcal{A}_{cwt}$ has been implemented as a Tensorflow custom layer. It is made up of a convolutional layer whose kernel - which takes the shape of a wavelet - is learnt. The temporal support of this kernel corresponds to the maximum size of a sequence of F0 in the dataset. This layer is associated with N learnable variables which are the scales from which the wavelets are derived. A constraint of growth is added on the range of scales to ensure the continuity of the learned CWTs. The output of this layer is then unpadded and sliced in segments of size T_{slice}.

Parameters. We chose to use $N = 32$ learnable scales, thus making good reconstruction of the original F0 from CWT representation possible with . The slicing length is set to $T_{slice} = 512$ what was considered suitable for Dual-GAN processing. Therefore each sliced segment has shape $[32, 512]$.

Pre-Net classifier for attitude prediction from CWTs

The classifier is built using convolutional blocks composed of $n^{\mathcal{C}}_{cnn}$ 2-D convolutional layers. Each convolution layer has $d^{\mathcal{C}}_{cnn}$ filters, $k^{\mathcal{C}}_t$ and $k^{\mathcal{C}}_f$ as temporal and feature-wise kernels and $s^{\mathcal{C}}_t$ and $s^{\mathcal{C}}_f$ as temporal and feature-wise strides. Strides are used to reduce time and features dimensions as is usually done with convolutional networks. Those convolutions are followed by a flatten layer and two fully connected layers with respectively, $d^{\mathcal{C}}_{emb}$ and $d^{\mathcal{C}}_{cl}$. An activation function $\phi^{\mathcal{C}}$ is applied to the last layer.

Parameters. We set $n^{\mathcal{C}}_{cnn} = 3$, chose filters sizes $d^{\mathcal{C}}_{cnn}$ starting from 32 up to 128, $k^{\mathcal{C}}_t = k^{\mathcal{C}}_f = 3$, $s^{\mathcal{C}}_t = 2$ and $s^{\mathcal{C}}_f = 4$. A dropout $\rho^{\mathcal{C}} = 0.2$ is applied for each layer. Padding is set to same (TF argument for convolution) so that input and output signal has the same shape. For the last fully connected layers, we set $d^{\mathcal{C}}_{emb} = 1000$ and $d^{\mathcal{C}}_{cl} = 2$. The activation function is $\phi^{\mathcal{C}} = softmax$

The architecture of the Dual-GAN itself as well as the contribution of each module, $\mathcal{G}_{s\to t}$, $\mathcal{G}_{t\to a}$, $\mathcal{D}_s$ and $\mathcal{D}_t$, are taken from (Luo et al., 2019) and described in the following.

Dual-GAN generators

The generators are made up of $n_{cnn}^{\mathcal{G}}$ convolutional and backward convolutional layers with d_{cnn} filters, $k_t^{\mathcal{G}}$ and $k_f^{\mathcal{G}}$ as temporal and feature-wise kernels, $s_t^{\mathcal{G}}$ and $s_f^{\mathcal{G}}$ as temporal and feature-wise strides (used for downsampling and upsampling respectively for convolutions and backward convolutions), and a dropout $\rho^{\mathcal{G}}$ which is used to provide the network with noise, as needed for GANs. Each layer is followed by an activation function $\phi^{\mathcal{G}}$ with the exception of the output layer. Residual blocks are employed in the middle of convolutional and backward convolutional layers.

PARAMETERS. According to (Luo et al., 2019), we have $n_{cnn}^{\mathcal{G}} = 2$ convolutional layers with filters sizes d_{cnn} of 128 and 64, $k_t^{\mathcal{G}} = k_f^{\mathcal{G}} = 3$, $s_t^{\mathcal{G}} = s_f^{\mathcal{G}} = 2$. Analogously, we have $n_{cnn}^{\mathcal{G}} = 2$ backward convolutional layers with filters sizes $d_{cnn}^{\mathcal{G}}$ of 64 and 128, $k_t^{\mathcal{G}} = k_f^{\mathcal{G}} = 3$, $s_t^{G} = s_f^{\mathcal{G}} = 1/2$. A dropout $\rho^{\mathcal{G}} = 0.5$ is applied. We chose a non linear activation function $\phi^{\mathcal{G}} = ReLU$.

Dual-GAN discriminator

Inspired by PatchGAN classifier (He et al., 2016), we designed the same architecture for the discriminator as we did for the pre-Network classifier described in 5.4.1. Only the size of its output, which is here reduced to a scalar (*real* : 1 or *false* : 0), and the activation function $\phi^{\mathcal{D}}$, stand out. The only constraint being that the output must fall within the range of 0 and 1, which must guaranteed by the activation.

PARAMETERS. Everything identical to Pre-Network Classifier, excepting from the output $d_{cl} = 1$ and the activation $\phi^{\mathcal{D}} = sigmoid$.

5.4.2 A One-to-One Speaker Dependant Conversion Experiment

The purpose of this experiment is to compare different configurations of our model with the baseline method (Luo et al., 2019) to show the relevance of our contributions in the context of speech attitudes conversion. We chose a one-to-one paradigm for conversion which means that the mapping is learned between specific pairs of attitudes. For a given pair, the source and target utterances share the same linguistic content and are spoken by the same speaker, they only differ in the attitude produced. For our experiments we used our homemade speech database Att-HACK (Le Moine and Obin, 2020). The design as well as description of Att-HACK is presented in Chapter 3.

TRAIN-TEST DATA SPLIT. Since prosodic strategies conveying emotion has been shown to be speaker dependent (Sisman and Li, 2018), we might assume that each speaker has their own way of expressing attitudes. To assess the model ability to learn a specific strategy of vocal attitude production, we chose to learn transformations on two different speakers independently. We thus selected a female (F08) and a male (M07), representing almost 400 utterances each. The *train/valid* split was 80/20 % and has been done linguistically, i.e. as Att-HACK features 100 different linguistic contents, we chose 80 sentences for training and 20 for validation.

Configurations

We selected three configurations for comparison, the baseline Dual-GAN featuring CWT Adaptive Scales (AS) proposed in (Luo et al., 2019) and two configuration of our proposed end-to-end model with variations of the pre-network pN

- ***CWT-AS*** strictly re-implemented following (Luo et al., 2019)
- ***WKCE*** : $pN = \{\mathcal{A}_{cwt}\}$ learns the CWT scales regarding the CWT reconstruction objective ($\alpha = 1$, $\beta = 0$)
- ***WKCE-C*** : $pN = \{\mathcal{A}_{cwt} + \mathcal{C}\}$ learns the CWT scales regarding both the CWT reconstruction and the CWT related attitude classification objectives ($\alpha = 10$, $\beta = 1$)

5.4.3 Training Procedure

The training procedure is divided in two phases. First we pre-train the pre-network to produce CWT representations of F0 sequences. The full pre-training process, described in Table 5.1, depends on the configuration considered. Second, we train the Dual-GAN along with the pre-network to learn transformations between attitudinal pairs of F0 sequences. The process is fully described in Table 5.2.

We chose $\lambda = 5$ and $\gamma = 15$ as a balance between transformation, adversarial, and dual objectives. For the configuration *WKCE-C*, we used the classification scores obtained at the end of pre-training as sample weights for Dual-GAN training. ADAM optimizer with 0.0001 as learning rate has been used. All codes are written in Python-Tensorflow 2.1, the baseline has been re-implemented.

PreNet pre-training process

Source input tensor is made of b source F0 sequences $\mathcal{B}^s = \{\mathbf{x}_1^s, ..., \mathbf{x}_b^s\}$
while target input tensor is made with the related target F0 sequences $\mathcal{B}^t = \{\mathbf{x}_1^t, ..., \mathbf{x}_b^t\}$
Voicing tensors $\mathcal{V}^s = \{\mathbf{v}_1^s, ..., \mathbf{v}_b^s\}$ and $\mathcal{V}^s = \{\mathbf{v}_1^t, ..., \mathbf{v}_b^t\}$ are made voicing sequences,
i.e. ones for voiced segments and zeros for unvoiced segments.

for any $\{\mathcal{B}^s, \mathcal{B}^t\}$
$\mathcal{B}^s* = \mathcal{V}^s$
$\mathcal{B}^t* = \mathcal{V}^t$
$\mathcal{A}_{cwt}(\mathcal{B}^s) = \{\mathbf{X}_1^s, ..., \mathbf{X}_b^s\}$
$\mathcal{A}_{cwt}(\mathcal{B}^t) = \{\mathbf{X}_1^t, ..., \mathbf{X}_b^t\}$
$\mathcal{A}_{cwt}(\mathcal{B}^s)* = \mathcal{V}^s$
$\mathcal{A}_{cwt}(\mathcal{B}^t)* = \mathcal{V}^t$
$\mathcal{L}_{\mathcal{A}_{cwt}} = \frac{1}{b}\left[\sum_{\mathbf{x}\in\mathcal{B}^s} \mathcal{L}_{rec}(\mathbf{x}) + \sum_{\mathbf{x}\in\mathcal{B}^t} \mathcal{L}_{rec}(\mathbf{x})\right]$

if the configuration is *WKCE-C*:
$\mathcal{L}_{\mathcal{A}_{cwt}} += \frac{1}{b}\left[\sum_{\mathbf{x}\in\mathcal{B}^s} \mathcal{L}_{cl}(\mathbf{x}) + \sum_{\mathbf{x}\in\mathcal{B}^t} \mathcal{L}_{cl}(\mathbf{x})\right]$

$\mathcal{A}_{cwt}$ is optimized to minimize $\mathcal{L}_{\mathcal{A}_{cwt}}$

until convergence

Table 5.1: Algorithm1. pretraining process of the pre-network pN

Dual-GAN training process

Source input tensor is made of b source F0 sequences $\mathcal{B}^s = \{\mathbf{x}_1^s, ..., \mathbf{x}_b^s\}$
while target input tensor is made with the related target F0 sequences $\mathcal{B}^t = \{\mathbf{x}_1^t, ..., \mathbf{x}_b^t\}$
Voicing tensors $\mathcal{V}^s = \{\mathbf{v}_1^s, ..., \mathbf{v}_b^s\}$ and $\mathcal{V}^s = \{\mathbf{v}_1^t, ..., \mathbf{v}_b^t\}$ are made voicing sequences,
i.e. ones for voiced segments and zeros for unvoiced segments.

for any $\{\mathcal{B}^s, \mathcal{B}^t\}$

$\mathcal{B}^s* = \mathcal{V}^s$
$\mathcal{B}^t* = \mathcal{V}^t$
$\mathcal{A}_{cwt}(\mathcal{B}^s) = \{\mathbf{X}_1^s, ..., \mathbf{X}_b^s\}$
$\mathcal{A}_{cwt}(\mathcal{B}^t) = \{\mathbf{X}_1^t, ..., \mathbf{X}_b^t\}$
$\mathcal{A}_{cwt}(\mathcal{B}^s)* = \mathcal{V}^s$
$\mathcal{A}_{cwt}(\mathcal{B}^t)* = \mathcal{V}^t$

$$\mathcal{L}_R^s = \frac{1}{b} \sum_{\mathbf{x}^s \in \mathcal{B}^s, \mathbf{x}^t \in \mathcal{B}^t} (\mathcal{L}_{t \to s}(\mathbf{x}^s, \mathbf{x}^t) + \mathcal{L}_{rec}(\mathbf{x}^s))$$

$$\mathcal{L}_R^t = \frac{1}{b} \sum_{\mathbf{x}^s \in \mathcal{B}^s, \mathbf{x}^t \in \mathcal{B}^t} (\mathcal{L}_{s \to t}(\mathbf{x}^s, \mathbf{x}^t) + \mathcal{L}_{rec}(\mathbf{x}^t))$$

$$\mathcal{L}_{ADV}^s = \frac{1}{b} \sum_{\mathbf{x}^s \in \mathcal{B}^s, \mathbf{x}^t \in \mathcal{B}^t} \mathcal{L}_{adv}^s(\mathbf{x}^s, \mathbf{x}^t)$$

$$\mathcal{L}_{ADV}^t = \frac{1}{b} \sum_{\mathbf{x}^s \in \mathcal{B}^s, \mathbf{x}^t \in \mathcal{B}^t} L_{adv}^t(\mathbf{x}^s, \mathbf{x}^t)$$

$$\mathcal{L}_{DUAL} = \frac{1}{b} \sum_{\mathbf{x}^s \in \mathcal{B}^s, \mathbf{x}^t \in \mathcal{B}^t} \mathcal{L}_{dual}(\mathbf{x}^s, \mathbf{x}^t)$$

$\mathcal{A}_{cwt}, \mathcal{G}^{s \to t}$ and $\mathcal{D}^t$ are optimized to minimize $\lambda \mathcal{L}_R^t$.
$\mathcal{A}_{cwt}, \mathcal{G}^{t \to s}$ and $\mathcal{D}^s$ are optimized to minimize $\lambda \mathcal{L}_R^s$.

$\mathcal{A}_{cwt}, \mathcal{G}^{s \to t}$ and $\mathcal{D}^t$ are optimized to maximize L_{ADV}^t.
$\mathcal{A}_{cwt}, \mathcal{G}^{t \to s}$ and $\mathcal{D}^s$ are optimized to maximize L_{ADV}^s.

$\mathcal{A}_{cwt}, \mathcal{G}^{s \to t}$ and $\mathcal{G}^{t \to s}$ are optimized to minimize $\gamma \mathcal{L}_{DUAL}$

until convergence

Table 5.2: Algorithm2. Dual-GAN training process

5.5 Results & Discussions

In this section, we assess the performance of the different configurations and baseline to convert speech attitude by changing pitch contours. We used both objective and subjective criterion and considered each pair of attitudes (12 pairs) as well as overall results across all the pairs.

5.5.1 Objective Evaluation

In order to validate the outputs of our system, we define two objective measures : 1) ϵ_R reflects the reconstruction performance of the pre-network in both configurations *WKCE* and *WKCE-C* and yielded by the *CWT-AS* scales optimization. 2) ϵ_T reflects the transformation performance of the three different models.

ROOT MEAN SQUARED ERROR (RMSE-F0). RMSE is used to compute both measures. A first measure ϵ_R is computed between the target F0 sequences and their reconstructions from the CWTs $\mathbf{x}^t$ and $\hat{\mathbf{x}}^t$ respectively. Thus ϵ_R is computed as follows

$$\epsilon_R = \sqrt{\sum_{t=1}^{N} \frac{(\hat{x} - x)^2}{N}} \tag{5.16}$$

A second measure ϵ_T is computed between the target and converted F0 sequences $\mathbf{x}^t$ and $\mathbf{x}^{s\leftarrow t}$ respectively. Thus ϵ_T is computed as follows

$$\epsilon_T = \sqrt{\sum_{t=1}^{N} \frac{(\hat{x}^{s\leftarrow t} - x^t)^2}{N}} \tag{5.17}$$

Performance in transformation

Let us begin by looking at the overall results shown in Table 5.4, i.e. the results in RMSE between the target and converted F0 sequences averaged across all attitudinal pairs. First of all, we observe that the proposed end-to-end system for F0 voice conversion outperforms the traditional **CWT-AS**, for both configurations. The best performance is obtained with the most elaborated configuration **WKCE-C**. On the other hand, the differences in RMSE remain small and do not allow us to conclude on the real performance of the two configurations of the proposed model in comparison with the baseline.

If we now look at the attitude pair-wise results shown in Table 5.3, we observe that our proposal **WKCE-C** achieves the best transformation performance four 8 pairs out of 12. For the worst pairs ($do. \rightarrow fr.$, $di. \rightarrow do.$, $do. \rightarrow di.$ and $se. \rightarrow do.$), this configuration achieves comparable or better performance than the baseline **CWT-AS** excepting from two pairs ($di. \rightarrow do.$, $do. \rightarrow di.$). These results make **WKCE-C** a rather reliable configuration that well captures the attitudinal conveyed through pitch contours and transform it accurately, when compared to the baseline **CWT-AS**.

Models		fr. → di.	di. → fr.	fr. → do.	do. → fr.	fr. → se.	se. → fr.
CWT-AS	ϵ_R	15.5	17.1	17.3	19.1	17.2	16.2
	ϵ_T	25.8	19.4	20.2	23.8	19.6	20.7
WKCE	ϵ_R	**8.2**	**8.2**	**12.3**	**8.0**	9.5	**9.0**
	ϵ_T	15.8	19.9	23.4	**12.3**	19.3	18.9
WKCE-C	ϵ_R	11.4	12.2	13.3	15.2	**9.1**	12.0
	ϵ_T	**13.9**	**14.5**	**16.2**	17.9	**16.1**	**16.3**
		di. → do.	do. → di.	di. → se	se. → di.	do. → se	se. → do.
CWT-AS	ϵ_R	18.3	20.1	15.7	17.2	16.0	18.2
	ϵ_T	**19.1**	20.1	23.5	21.2	22.1	22.5
WKCE	ϵ_R	**10.3**	**7.4**	10.2	**10.5**	**9.1**	**7.3**
	ϵ_T	22.2	**16.3**	21.9	22.8	22.2	**15.8**
WKCE-C	ϵ_R	16.0	16.3	**10.1**	15.1	14.3	16.1
	ϵ_T	23.5	25.6	**17.1**	**20.3**	**21.3**	23.2

Table 5.3: Performance results in reconstruction ϵ_R and transformation ϵ_T for all considered pairs of attitudes based on three models : **CWT-AS**, **WKCE** and **WKCE-C**. Friendly, distant, dominant and seductive being respectively denoted $fr.$, $di.$, $do.$ and $se.$

Performance in reconstruction

Across attitude pairs, the configuration **WKCE** achieves the best results in reconstruction by far, logically since it is optimised for the reconstruction criterion. The other configuration **WKCE-C** is slightly worse but substantially outperforms the **CWT-AS** baseline.

Examining further into details considering each pair of attitudes, we observe that the configuration **WKCE** achieves the best results for all pairs excepting from $fr. \rightarrow se.$ and $di. \rightarrow se.$. Thus it interesting to note that, in most cases, the best results in transformation are not obtained with the same configuration than those in reconstruction. This supports the idea that the representation given to the network must above all be specifically tailored to the transformation task. Moreover, our two configurations always yield better reconstruction results than the **CWT-AS** baseline which validates the ability of our wavelet kernel convolutional encoder to produce representations from which F0 sequences can be accurately reconstructed.

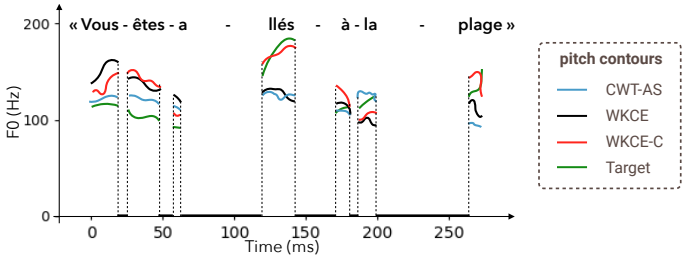

Figure 5.5: Example of F0 conversion from distant to dominant for speaker M07 for **CWT-AS** and ours **WKCE** and **WKCE-C**.

Those objective results tend to show an advantage of our proposal in the ability to produce accurate transformations for most considered pairs of attitudes. An example of conversion is depicted in Figure 5.5. As mentioned, since such objective measurements do not necessarily reflect human perception of those conversions, a subjective study is proposed in the following section to confirm these trends.

5.5.2 Subjective Evaluation

To assess the perception of converted speech, we conducted a listening experiment.

Test Design

We decided to perform a AB preference test to assess the relative performance of different models by comparing the average subjective perception of their yielded conversions. The test consists in repeated trials, for each of which two questions were asked. At each trial, participants started by listening to a reference speech sample (a target), then they listened to a pair of sounds - two different conversions supposed to match the target. First, they had to choose the speech sample from that pair that, in their opinion, most closely resembles the reference in terms of conveyed attitude. Second they had to judge which of the four attitudes was conveyed in the reference speech

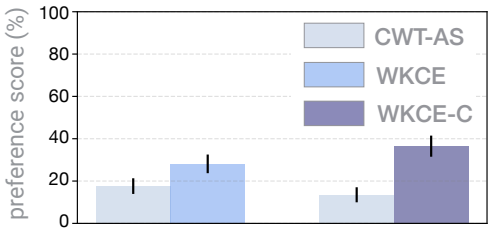

Figure 5.6: The XAB preference results with 95% confidence interval between the **CWT-AS** and ours **WKCE-C** and **WKCE** regarding attitude similarity.

Models	ϵ_R	ϵ_T
CWT-AS	17.32	21.71
WKCE	9.16	19.15
WKCE-C	13.4	**18.83**

Table 5.4: A comparison of the RMSE results of the **CWT-AS**, **WKCE** and **WKCE-C** for reconstruction and transformation

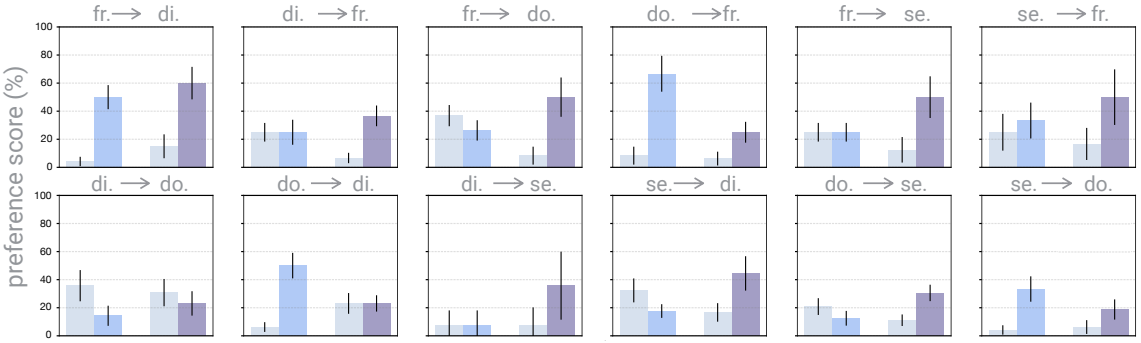

Figure 5.7: The XAB preference results with 95% confidence interval between the **CWT-AS** and ours **WKCE** and **WKCE-C** regarding attitude similarity.

sample [1]. We built the speech pairs by choosing a conversion yielded by the **CWT-AS** configuration and one of our configurations' - **WKCE-C** and **WKCE** - conversion .

Results - Similarity to the reference attitude

We start by looking at the overall results reported in Figure 5.6, thus assessing general performance of our proposals. Compared to baseline **CWT-AS**, our configurations **WKCE** and **WKCE-C** are respectively preferred by 9% and 23% in terms of similarity to the target attitude. We can thus confidently state that our approach, particularly the setup that features a classifier **WKCE-C**, is better at converting attitude, in general, than the baseline **CWT-AS**.

This preliminary conclusion needs to be refined by examining the pair-wise outcomes. Indeed, our configuration **WKCE-C** outperforms significantly the baseline for most attitude pairs at the exception of $do. \rightarrow di.$ and $di. \rightarrow do.$. **WKCE** is preferred to baseline by listeners in four pairs out of twelve. This difference in performance between **WKCE-C** and **WKCE** tends to highlight the role played by the classifier in the production of CWT representations of F0 sequences. Since the classifier forces the attitudinal salience of the representations, the model seems to more easily capture the information to convert. This information being more or less reachable depending on the attitude pair considered.

[1] When conducting this experiment, the Att-HACK database has not been perceptually validated, i.e. there were no evidence that the attitude categories are actually perceived as such by individuals.

Results - Attitude recognition

The results in attitude recognition are shown in the form of a confusion matrix in Figure 5.8. The best recognized attitude is seduction (71%) followed by distance (60%), while dominance (48%) and friendliness (42%) are recognized as such in less than half of the cases. Friendliness and distance are often confused, as are friendliness and dominance. An important point should be made here, what is being measured is a recognition score. This score reflects globally the identification confusion by the subject. This however may be decomposed into various sources of confusion: the actors' mental representations of the attitudes, their ability to convey them acoustically, the ability of listening test subjects to decode them. In particular, in the case of friendliness, it is currently impossible to say whether friendliness is simply more difficult to recognize than seduction (which is unlikely) or whether some occurrences of the database have been badly portrayed by the actors, leading to confusion. Questioning the recognition of attitudes through forced choices allows us to acknowledge how difficult it is for individuals to recognize the a priori attitude labels attached to Att-HACK samples. The BWS study conducted in Chapter 4 helped to identify utterances for which the attitude was miscommunicated. We expect that by removing these samples from Att-HACK, individuals' recognition of attitudes will be improved.

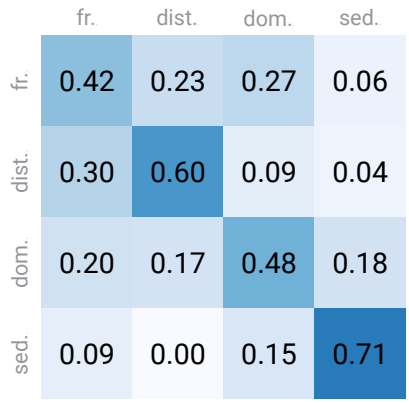

Figure 5.8: Normalized confusion matrix where each row presents the class dependant recognition performance across all participants.

5.5.3 General discussion

Our model's validity has been assessed both objectively, using RMSE measures in transformation and reconstruction, and subjectively through a listening experiment in which participants gave their preference between pairs of conversions in terms of similarity to a reference speech sample. If both of our proposed configurations achieve better performance than the baseline in average, looking at pair-wise results showed that the addition of the classifier in the pre-network led to better adaptation across different attitude pairs. The best results in both objective and subjective terms are thus obtained with the **WKCE-C** configuration. In the following we discuss the choices made regarding the modelling of speech attitudes as well as the paradigm employed to learn transformations between source and target utterances. First, the scale distributions that underlie the learned representations of pitch contours are closely examined in order to question the relevance of CWT representations for modeling pitch contours.

Examining learned pitch contours' scales distributions

This part provides an a posteriori comparison of the F0 scales distribution that underlies the F0-CWT representations used for the speech attitude conversions. Each transformation (forward and backward) between a pair of attitudes is associated with a set of temporal scales that are used to compute the CWT representations used for the conversion. Consequently, each transformation can be described by a distribution of the temporal scales that are used to convert the F0 optimally. Figure 5.9 presents the distribution selected by the baseline CWT-AS algorithm and learned by our proposed contribution, as obtained for a specific speaker (F08) for the six pairs of attitudes.

As stated above, the best performance has been obtained with **WKCE-C**, but what does it mean in regards with the underlying F0 scales distribution? First, one can clearly observe that the temporal distribution of the **WKCE-C** is wider than the others, the transformation covering a wide range of temporal scales from the micro variations over the phonemes to the global contours of the sentence. Additionally, the distributions associated with the **CWT-AS** and the **WKCE** appear mostly independent with respect to the transformation pair, while the distribution associated with the **WKCE-C** tend to be more varied depending on the transformation pair. This suggests that the **WKCE-C** may adapt more efficiently to the singularity of each pair.

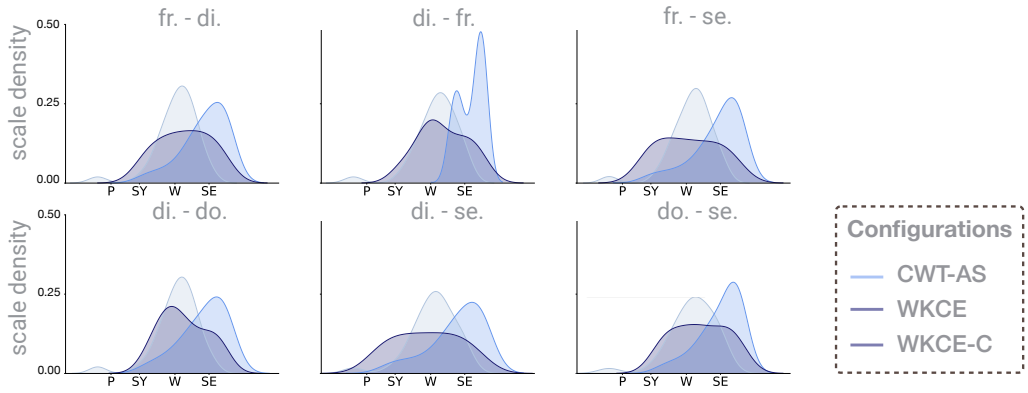

Figure 5.9: CWT scale distributions for the three considered algorithms: **CWT-AS**, **WKCE** and **WKCE-C**. For interpretation, P, SY, W and SE markers respectively denote average duration of the phone, syllable, word and sentence linguistic units.

The scale distributions supporting the encoding of the F0 contours can help to explain transformation performance, at least in part. Our proposal for modeling pitch contours through learnt CWT representations therefore appears relevant. It allows a multi-level representation of F0 contours, thus modeling a wide prosodic spectrum from micro to macro, while enabling a focus on certain levels rather than others depending on the attitude pair considered. This adaptive method can be further investigated and applied to other multi-level time varying signals.

On the limitations of F0 modelling for voice conversion

Listening to the generated conversions clearly shows the limitations of using isolated parameters as modelling for expressive conversion. First, since the conversion only applies to the F0, artefacts appear due to the inconsistency between the altered F0 and the rest of the parameters of the

speech signal which have remained unchanged. Second, our Chapter 4 shows that F0 is not the only parameter involved in the communication of vocal attitudes, both in terms of production strategy and of perception. Ignoring the other parameters for conversion leads to less realistic conversion. Voice conversion thus requires modelling and converting all the parameters of the speech signal - as well as their inter-correlations - to ensure the consistency and quality of the converted speech signal. In order to better convert speech attitudes, we decide to abandon the parametric representation - based on F0 - and attempt to model the conversion directly from compact representations like mel-spectrograms - that prove to encompass every component of the speech signal.

Rethinking the conversion paradigm

With the Dual-GAN framework, a transformation and its inverse are learnt jointly. This goes beyond the classic case of a one-to-one learning: only six learnings were necessary against twelve in a standard GAN configuration thus offering a time saving as well as a potential gain in performance. Another interesting point is the probabilistic nature of the conversions yielded by the Dual-GAN, i.e. for one source utterance several different conversions can be produced. This is typically attributable to GAN, which operates by filtering random noise. It offers a diversity in conversions that may be desired, for example to avoid the monotony of responses from a voice assistant. The paradigm chosen for transformation thus has many qualities, yet some limitations need to be mentioned.

First, the training of a GAN is made difficult by its intrinsic instability due to the addition of noise. In practice, the best optimization of the model from the point of view of conversions is not necessarily the global minimum of the cost function employed. It is necessary to save the weights of the model at different checkpoints and compare the conversions from these different sets of weights. We would like the training criterion, the cost function, to be correlated with what we hear when we listen to the conversions and therefore be sufficient to determine at what point in the training the model achieves its best performance. Second, this model's fundamental drawback is that it disallows the mapping of time sequences with various duration. This point is fundamental, specifically for the case of the conversion of emotions (close to attitudes in terms of acoustic realisation). The main way to achieve such a sequence-to-sequence mapping is to employ an attention mechanism that learns an implicit alignment between the source and target sequences. Finally, we would like to learn to convert any attitude into any other attitude through a single training procedure, this kind of learning being referred to as many-to-many. Since learning a conversion and its inverse simultaneously appears to improve the overall conversion performance, we might expect to gain even more by feeding a conversion algorithm with the four attitudes simultaneously.

5.6 Chapter Summary

In line with the findings of Chapter 4, we chose to focus on pitch contours - acoustically correlated to F0 variations - to parametrically model the speech attitudes. To highlight the multi-level nature of F0 variations, we model it through using the Continuous wavelet Transform (CWT) - that computes a decomposition of the F0 signal over wavelet kernels which allows a representation of F0 over different temporal scales. In order to adapt this representation for conversion purpose, an Adaptive-Scale (AS) algorithm (Luo et al., 2019) has been proposed. It finds an optimal CWT representation for each pair of emotions, by selecting the scales that maximize in average the distance between the emotions in the CWT space. Though this approach appears promising, it suffers

from two main limitations. First, the scale selection is only based on the maximization of the distance between the emotions, but ignores their reconstruction ability of the F0 signal - which may affect the quality and the naturalness of the conversions. Second, the CWT-AS F0 decomposition is optimized independantly from the conversion model - a dual-GAN - which may cause the CWT decomposition not to be optimal for conversion objective.

To overcome the limitations of (Luo et al., 2019), we propose an end-to-end architecture to learn efficiently F0 conversions between attitudes. The proposed neural architecture brings together the F0 decomposition and the dual-GAN into a single network. We compared two setups of our proposal with the baseline CWT-AS in an speech attitude conversion experiment. The experiment is speaker dependant, two speakers were selected and subjects of independent trainings for converting from any attitude to any other. Both objective and subjective evaluation tends to show our proposal achieves better conversions than the baseline. However, listening to the generated conversions clearly shows the limitations of using isolated parameters as modelling for attitude conversion. First, since the conversion only applies to the F0, artefacts appear due to the inconsistency between the altered F0 and the rest of the parameters of the speech signal which have remained unchanged. Second, our Chapter 4 shows that F0 is not the only parameter involved in the communication of vocal attitudes, both in terms of production strategy and of perception. Ignoring the other parameters for conversion leads to less realistic conversion. Voice conversion thus requires modelling and converting all the parameters of the speech signal - as well as their inter-correlations - to ensure the consistency and quality of the converted speech signal. In the Chapter 7, we chose to use mel-spectrogram as speech representation for attitude conversion.

PARADIGM SHIFT

At this stage of our research, the various findings we made led us to a paradigm shift both in the way we model the speech signal and in the way we learn attitude conversion. In the following, we discuss this paradigm shift, its motivations and implications.

Speech Attitude Modelling for Voice Conversion

Listening to the conversions yielded by our first proposal - presented in Chapter 5 - clearly shows the limitations of using isolated parameters - in this instance F0 contours - as modelling for attitude conversion. First, because the conversion only affects the F0, artefacts show up as a result of inconsistencies between the altered F0 and the other speech signal parameters that have stayed unchanged. Second, our Chapter 4 confirmed that F0 is not the only parameter involved in the communication of vocal attitudes, in terms of both production strategy and perception. Instead, all aspects of the speech signal appear to play a role in the speech communication of attitudes. In the perspective of speech attitude conversion, as every aspect in the signal needed to be changed, we decided to move from partial parametric representation to complete representation of the speech signal using mel-spectrogram. The mel-spectrogram constitutes a compelling alternative to standard multi-parametric representations - e.g. spectral envelope, F0 and aperiodicity - such as used in STRAIGHT (Kawahara, 2006) or WORLD (Morise et al., 2016) vocoders. First it is perceptually relevant thanks to representing frequency in mel scale and amplitude dBs that both reflect human auditory perception. Second, it constitutes a compact representation that can be efficiently used by conversion algorithm. In addition, we chose to use the neural vocoder proposed in (Roebel and Bous, 2022) - at different stages of its development - to reconstruct speech signal from mel-spectrograms. This vocoder has particularly shown to achieve near transparent speech quality even for out of domain data (Roebel and Bous, 2022).

Speech Attitude Conversion

After the modeling issue has been addressed leading to adopt the mel-spectrogram for representing speech signals, the question of *how to convert vocal attitudes using this representation?* must be considered closely.

Towards Changing Any Aspect of Speech Signals

By choosing mel-spectrogram as speech signal representation, we account for all the aspects of the speech signal that convey attitudes. Converting attitudes in proper way involves changing those aspects and thus requires a specific algorithm's architecture. The main point is about learning a mapping between speech signal representations of different duration. This strong constraint excludes frame-aligned mapping algorithms (Ming et al., 2016; Zhou et al., 2021) that does not allow for temporal integration. We thus focused on sequence-to-sequence architecture introduced in (Sutskever et al., 2014). At the heart of most of those architectures lies an attention mechanism that aims to implicitly learn an alignment between the source and target sentences. While recurrent neural networks represent an effective implementation for Seq-to-Seq voice conversion (Shen et al., 2018; Tanaka et al., 2019; Tachibana et al., 2018), recent studies have shown that convolutional neural networks (CNN) with gating mechanisms also learn well the long-term dependencies (Gehring et al., 2017; Kameoka et al., 2020). Finally, the transformer networks (Vaswani et al., 2017), lately proposed for Seq-to-Seq conversion, have shown noteworthy performance in terms of sound quality (Kameoka et al., 2021; Chen and Zhang, 2021; Lee et al., 2022). The main advantage of such transformers being that it basically replaces all recurrent - or dilated convolution - layers by self attention mechanisms in a network. This allows for more efficient learning in terms of both time and computing resources.

Conversion Learning Mode

Another point in this paradigm shift deals with the mode of conversion learning. In our first attempt at converting speech attitudes - chapter 5 - we learned for a given attitude pair, the conversion and its inverse simultaneously. Here, we intend to learn conversions in a many-to-many fashion, i.e. we would learn to convert any attitude into any other through a single training procedure. Such a paradigm would allow to learn an implicit definition of attitudes - as represented in Att-HACK. (Kameoka et al., 2021) proposed a many-to-many extension to transformer-based speaker identity conversion allowing for learning conversions between four speakers at once.

Towards Perceptually Conditionned Attitude Conversion

Finally, a question emerged at the end of the BWS experiment on Att-HACK: *how can we use the perceptual data collected to improve the quality of attitude conversions?* A first answer comes with the perceptual validation of Att-HACK data, in other words, it is crucial to clean the Att-HACK data so that it actually represents vocal attitudes as perceived by individuals before feeding our conversion model. To do so, we mean to extend the judgements made by participants during the experiment to un-evaluated utterances. Going further, one might even ask whether the gathered perceptual data can be used to control certain aspects of the conversion such as attitudinal intensity which appears to be one of the aspects effectively reflected in the participants' judgements. Finally, we would like to be able to evaluate the subjective perception of converted utterances by inferring from the evaluated ones in order to avoid the time-consuming and logistically demanding task of a listening test. These three objectives led us to design algorithms capable of automatically mimicking the average judgment made by BWS participants. This is the purpose of the next chapter 6.

Chapter 6

SPEECH ATTITUDE RECOGNITION

Contents

6.1 Towards Speech Attitude Recognition 110
6.1.1 Mel-spectrogram as Speech Signal Representation 110
6.1.2 Model Architecture 110
6.1.3 Preliminary experiment 112
6.1.4 Results & Discussion 114
6.2 Perceptual Regression Based on BWS Scores 116
6.2.1 About the Possible Uses of Gathered Perceptual Data 117
6.2.2 Proposal for a Perceptual Regressor 117
6.2.3 Experiments with the Perceptual Regressor 119
6.2.4 Results & Discussion 121
6.3 Perceptual Classification based on BWS scores 124
6.3.1 Proposal for a Perceptual Classifier 124
6.3.2 Experiment with the Perceptual Classifier 126
6.3.3 Results & Discussion 127
6.4 Perceptual Metric Learning Based on BWS Raw Judgements 129
6.4.1 Proposal for a Perceptual Arranger 130
6.4.2 Experiments with the Perceptual Arranger 134
6.4.3 Results & Discussion 137
6.5 Chapter Summary 139

The process by which individuals decode attitudes communicated vocally is complex - as shown by the second study of Chapter 4. In particular, it was found that static speech parameter values alone are insufficient to predict whether or not an attitude can be decoded by individuals - in average. As a result, it is currently impossible to perceptually validate Att-HACK, in the sense that we do not have a reliable criterion to extend our knowledge about attitude perception from what has been judged in the BWS experiment - i.e. barely a third of Att-HACK. From the perspective of attitude conversion learning - which is the main objective of this research - this lack of proper validation is a substantial issue as no conversion should be learnt on biased data. In order to tackle this issue, we propose to design a BWS-Net - i.e. an algorithm that artificially mimic the process by which individuals decode attitudes by learning on the gathered perceptual data. With such an algorithm, we intend to enhance the quality of our speech conversion by:

- **Validating Att-HACK** - and especially the part that has not been evaluated by participants during the BWS experiment (chapter 4) - and thus provide clean data for speech attitude conversion learning.
- **Conditioning our attitude conversion algorithm on perceptual data** by incorporating such a BWS-Net as depicted in Figure 6.1 and thus providing control on the attitudinal intensity of the conversions. Doing so, we expect to propose an effective and perceptually relevant algorithm for speech attitude conversion.
- **Validating the conversions** - yielded by our conversion algorithms - by application of a BWS-Net, thus assessing their subjective perception without the need for a listening experiment.

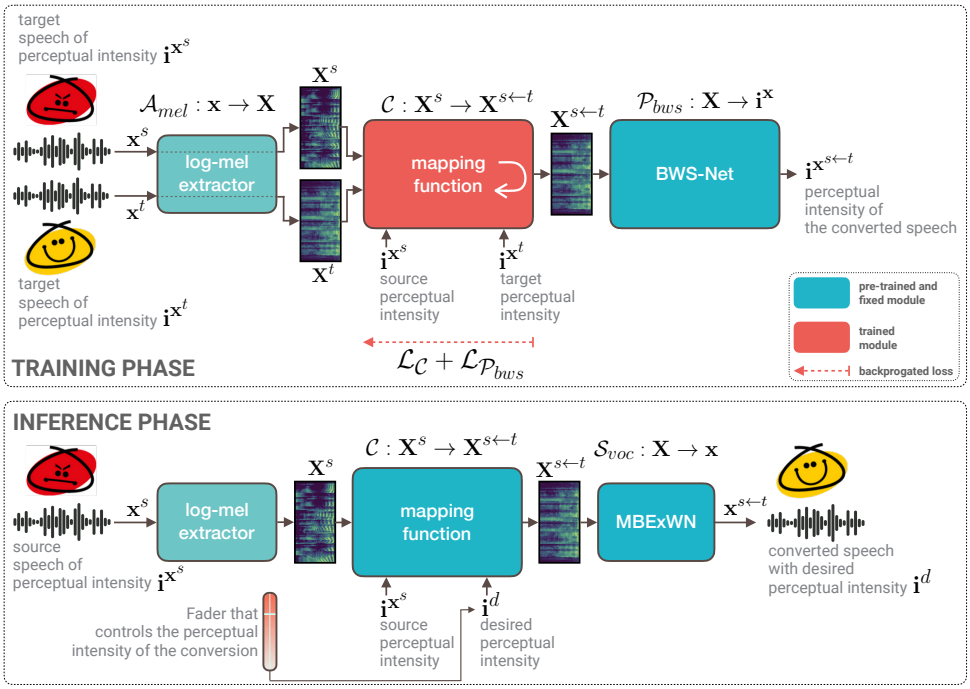

Figure 6.1: General scheme of the perceptual conditioning of our speech attitude conversion algorithm with attitudinal intensity control. At the time of writing, we conduct experiments with such an algorithm testing the different BWS-Nets designed in this Chapter.

6.1 Towards Speech Attitude Recognition

Before addressing the issue of perceptual regression/classification, we start by focusing on a more standard task of speech attitude recognition. In this section we will use the apriori attitude labels as ground truth for recognizing the Att-HACK attitudes. Through this recognition task, we introduce a baseline architecture from which we will experiment in the other sections to model subjective perceptual judgements. This architecture we used for speech attitude recognition is inspired from (Li et al., 2019) which proposed to learn to predict emotion and gender jointly. We only kept the emotion part to build our proposal.

6.1.1 Mel-spectrogram as Speech Signal Representation

We chose to use mel-spectrograms to represent speech signals. Indeed, we saw in Chapter 4 that many aspects of speech are involved in the communication of attitudes. Reflecting all those aspects is mandatory to properly representing speech for both purpose or speech attitude recognition and conversion. In particular, mel-spectrograms have the benefit of condensing the information in the speech signal while keeping its fundamental components, including the paralinguistic aspects. Thus, in the following - including in Chapter 7 - we always use this representation of speech signal extracted with the exact same parameters as given this section's experiment part.

Parameters. The mel-spectrograms are obtained through computing Short-Term-Fourier-Transform (STFT) of parameters N_{ft}, R_{ft} and M_{ft} corresponding to the size of the FFT, the hop and window sizes respectively. The number of mel channels is set to D. Melspectrograms are then padded batch-wise up to T_b which corresponds to the longest utterance in the batch. This way, computation can be performed in an tensorial way.

6.1.2 Model Architecture

This part provides a comprehensive description of the standard Attentive Convolutional Recurrent Neural Network (ACRNN) architecture - proposed in (Li et al., 2019) - as well as an interpretation of the role of each block that compose it in relation to the speech attitude recognition task.

A Convolutional Neural Network (CNN)

First, a block of convolutional layers is applied to the melspectrograms in order to extract high-level features that we expect to be useful for classifying the considered speech attribute. We choose to use 2D convolutions that incorporate both the time and frequency dimensions. At this stage, capturing the utterance's temporal context accurately is a key challenge. There are several approaches to achieve this, such as staking several convolutional layers, expanding the temporal kernel's size or using dilated convolutions, thus widening the temporal receptive field. Here, the first and second options are kept, while the third is omitted to allow for striding over the frequency-dimension, thus reducing its shape layer after layer. Calculations can be sped up and overall performance improved by employing temporal pooling to reduce the number of frames after the first layer.

Parameters. The number of convolutions in the block is n_{cnn}. Each convolution layer has identical d_{cnn}-dimensional output filters, kernels of size k_t and k_f and strides s_t and s_f, respectively for time and frequency dimensions. A pooling factor r_t is applied on temporal dimension. The padding is set to *same*. Then for a batch b, each melspectrogram $\mathbf{X}$ of shape $[T_b, D]$ processed by the CNN

block yields a tensor of high-level features of shape $[T_b/r_t, D/(s_f n_{cnn}), d_{cnn}]$ which is then resahped to $[T', D']$ where $T' = T_b/r_t$ and $D' = d_{cnn}D/(s_f n_{cnn})$.

A Bi-Long Short Term Memory (LSTM) Network

After being processed by the convolutional block, the features are passed to a recurrent network whose objective is to produce a temporal summary of those features. Here we use special recurring cells called Long-Short-Term-Memory (LSTM) cells . LSTMs employ a number of gates that regulate how data in a sequence enters, is stored in, and leaves the network. The forget gate, input gate, and output gate are the three gates that make up a conventional LSTM. These gates each represent a separate neural network and can be viewed as filters. They are well-known for their capacity to spread temporal information across extended sequences while avoiding the pitfalls of vanishing gradients. We use bidirectional LSTMs to collect temporal information processing in both forward and backward directions. Outputs from both directions are concatenated. This recurrent block is potentially composed of several layers, however it should be noted that recurrent layers are very costly in terms of time and computing resources.

Parameters. The number of LSTM layers is n_{blstm}. For each layer, each of the T' different LSTM states have d_{blstm} features produced with a dropout rate ρ_{blstm}. Thus the output of the Bi-LSTM block is a tensor $\mathbf{H}^{blstm}$ of shape $[T', 2 * d_{blstm}]$.

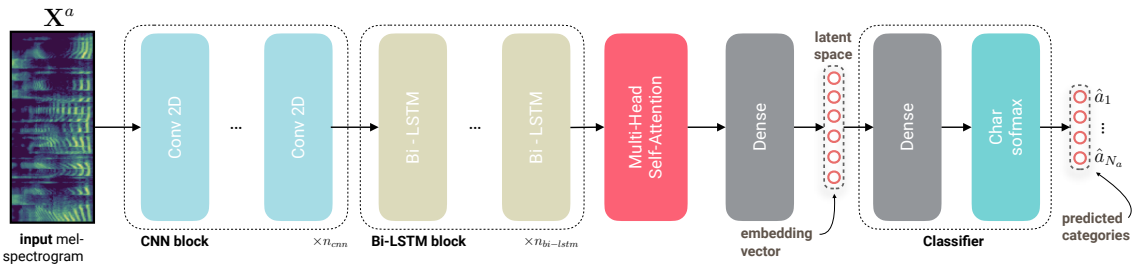

Figure 6.2: Schematic view of the ACRNN neural architecture

The attention block can be configured in a variety of ways, including basic self-attention and multi-head attention.

Self-Attention (SA)

With a sequence of high-level representations, an attention layer is employed to focus on relevant features and produces discriminative utterance-level representations for classification, since not all frame-level CRNN features contribute equally to the representation of the attributes to recognize.

Specifically, with the model's BLSTM output $\mathbf{H}^{blstm} \in \mathbb{R}^{T' \times 2 * d_{blstm}}$, a temporal vector $\alpha^{att} \in \mathbb{R}^{T'}$, representing the per-frame contribution to the target attribute, is computed depending on learnt weights vector $W^{att} \in \mathbb{R}^{d_{blstm}}$. Then $\boldsymbol{\alpha}^{att}$ is used to obtain an utterance-level representation by computing the weighted sum of temporal BLSTM internal states $\mathbf{c}^{att}$ often called context vector.

$$\boldsymbol{\alpha}^{att} = softmax(\mathbf{W}^{att}{\mathbf{H}^{blstm}}^{T}) \tag{6.1}$$

$$\mathbf{c}^{att} = \boldsymbol{\alpha}^{att}\mathbf{H}^{blstm} \tag{6.2}$$

Multi-Head Self Attention

When using the fundamental self-attention approach, the combination derived from α^{att} may only focus on one specific aspect of the input information, thus leaving other important aspects aside. In order to obtain a representation that is both compressed and exhaustive, multiple combinations of the Bi-LSTM hidden states $\mathbf{H}^{blstm}$ can be computed. This refinement of the basic attentional principle, known as Multi-Head-Attention, can be obtained through projecting inputs in n_{att} sub-spaces generated by as many weight matrices $\{\mathbf{W}_h^2\}_{h\in[\![1,n_{att}]\!]}$. Once obtained for each sub-space, the weighted sums are concatenated to get the final encoding vector $\mathbf{c}^{att}$ often referred to as the context vector.

$$\boldsymbol{\alpha}_h^{att} = softmax(\mathbf{W}_h^2 \tanh(\mathbf{W}^1{\mathbf{H}^{blstm}}^{T})) \tag{6.3}$$

$$\mathbf{h}_h^{att} = \boldsymbol{\alpha}_h^{att}\mathbf{H}^{blstm} \tag{6.4}$$

$$\mathbf{c}^{att} = \mathbf{h}_1^{att} \oplus ... \oplus \mathbf{h}_{n_{att}}^{att} \tag{6.5}$$

Parameters. The multi-head self attention network has n_{att} heads of latent dimension d_{att}. Provided with an input tensor $\mathbf{H}^{blstm} \in \mathbb{R}^{T'\times 2*d_{blstm}}$, the attention outputs a vector $\mathbf{c}^{att} \in \mathbb{R}^{2n_{att}d_{blstm}}$.

Embedding space projection and classification layer

Through the use of a fully connected layer, a final projection is made to the output $\mathbf{c}^{att}$ of the attentional network. It results in a so-called embedding vector $\mathbf{h}^{emb}$ on which classification can be performed properly through another fully connected layer. Finally, an activation function is applied to this final layer to produce a prediction vector $\mathbf{a}^{attr}$. Depending on the goal being pursued, such as multi-categorical or multi-label classification or even regression, this vector may take several shapes. In every instance, it should contain the information that our system is designed to predict.

Parameters. The first and second fully connected layers have d_{emb} and d_{pred} as outputs sizes. The chosen activation function depends on the task performed.

6.1.3 Preliminary experiment

In this part, we use the architecture previously described to tackle the task of speech attitude recognition as a preliminary - but fundamental - experiment.

Input data

We chose $N_{ft} = 2048$, $R_{ft} = 200$ and $M_{ft} = 800$ respectively for the FFT size, hop size and window size. The number of mel channels is set to $D = 80$. Note that those same parameters are used for mel-specteograms extraction in the following of this document - including the Chapter 7.

Ablation Study

Various SER studies have already been conducted, allowing for the emergence of a prototypical classification architecture : a convolutional block used to extract high-level features is followed by a recurrent network which captures temporal dependencies, an attention helps to select salient features, finally two fully connected layers are used to reduce data dimensionality, the last properly performs classification. However, the vocal attitude prediction task has not been specifically addressed in any studies. Thus, this initial investigation intends to highlight the relevance of the various blocks employed in regard to the model's capacity to predict vocal attitudes. Arguments in favor of doing such a study include the database's size in comparison to other databases often used for the SER, its multi-speaker nature, and the fact that it is in French.

Here, we compare several settings in order to develop a fundamental architecture from which to derive subsequent experiments. In the following, each configuration uses $n_{cnn} = 2$ convolutional layers with $d_{cnn} = 64$ filters, a temporal kernel $k_t = 5$ and a feature kernel $k_f = 3$, the variation of these parameters having little impact compared to the others.

- **CRNN**: The model is deprived of its attention mechanism, thus $n_{att} = 0$.
- **SA-CRNN**: The model features a Self-Attention (SA) layer implemented as described in 6.1.2, thus $n_{att} = 1$
- **MSA-CRNN**: The model features a Multi-Head-Self-Attention (MHSA) layer implemented as described in 6.1.2, following (Li et al., 2019) we chose $n_{att} = 8$.

Those configurations are examined for three kinds of *train-valid* splits, if **C** denotes a given configuration, it is paired with a suffix that indicates the split employed.

- **C-r** : The dataset is divided randomly with the single restriction of having all the sentences and speakers present in both training and validation sets.
- **C-l** : The dataset is divided linguistically, a part of the dataset sentences is never seen by the model. This division enables assessment of the model's generalizability to new sentences.
- **C-s** : The dataset is divided speaker-wise. Also referred to as Leave-One-Speaker-Out (LOSO) approach, this division enables assessment of the model's generalizability to new speakers.
- **C-oa** : This configuration does not refer to a specific split but designs the average over all three prior splits, reflecting split-independent performances.

Evaluation Metrics

In order to assess how well these different setups perform, we employ various objective measures.

Unweighted Average Recall (UAR). This metric measures the recognition performance of the different attitude classes, furtherly taking into account the imbalance between classes in terms of the number of samples.

Silhouette coefficient (Sil). This metric reflects the extent to which the different attitude instances are clustered together in the latent space generated by the model.

Davies Bouldin score (DBs). This metric reflects here how similar an attitude cluster is to its most similar cluster.

Mutual Information (MI). This metric reflects the extent to which the embeddings yielded for a given attitude are independent of the ones of other attitudes.

6.1.4 Results & Discussion

In this part, the results obtained are first discussed with the aim of providing a baseline architecture dedicated to speech attitude recognition. This architecture will constitute a starting point from which we will experiment towards perceptual attitude recognition.

Speech Attitude Recognizer's Performance

The results for all considered configurations and splits are displayed in Table 6.1. We can start by observing the significant interaction between the model's performance and the type of split that is performed on the data. The last few rows at the bottom of the table - denoted MAg for model agnostic - show model's performance averaged over configurations. We note that when the split is performed randomly on speech data, the model reaches its best performance. In particular it consistently outperforms the linguistic split by 0.02 and the speaker split by 0.12, in terms of UAR.

Configurations	UAR	Sil	DBs	MI
CRNN-r	0.60	0.15	2.22	0.13
SA-CRNN-r	0.67	0.20	1.62	0.06
MHSA-CRNN-r	**0.78**	**0.24**	**1.41**	**0.21**
CRNN-l	0.59	0.14	2.14	0.08
SA-CRNN-l	0.65	0.21	1.61	0.12
MHSA-CRNN-l	**0.75**	**0.22**	**1.46**	**0.17**
CRNN-s	0.55	0.11	2.46	0.07
SA-CRNN-s	0.55	0.19	1.56	0.09
MHSA-CRNN-s	**0.58**	**0.20**	**1.58**	**0.10**
CRNN-oa	0.56	0.13	2.27	0.09
SA-CRNN-oa	0.62	0.20	1.59	0.09
MHSA-CRNN-oa	**0.70**	**0.22**	**1.48**	**0.14**
MAg-r	**0.68**	**0.20**	1.75	**0.13**
MAg-l	0.66	0.19	**1.70**	0.12
MAg-sp	0.56	0.17	1.87	0.09

Table 6.1: Speeh attitude recognizer's performance in terms of UAR, silouhette coefficient, Davies Bouldin score and MI for all considered configurations and splits.

Let us then focus on the elements that make up the model architecture and their impact on the performance of the model. This experiment aims, among other goals, to assess the relevance of standard neural attention blocks, namely self-attention and multi-head self attention, for the specific challenge of speech attitude recognition. To do so, we focused on the penultimate group of rows in the Table 6.1, considering the average performance independent of the split achieved. The addition of a basic attention mechanism (SA) to the core CRNN architecture leads to a considerable gain of 0.15 in terms of UAR. Turning this block into a Multi Head Self Attention (MHSA), thus allowing the model to observe several different aspects of the speech signal to compute attention

scores, yields even better performance ($+0.08$ in UAR). It should be noted here that the best performance in terms of UAR was obtained for a number of attention heads n_{att} equal to 8. We do not detail here the numerous attempts to further improve the results by changing the hyper-parameters of the model.

In a second analysis, let us look at the model's performance for each of the split types. We observe, for instance, that the addition of a self-attention block has no impact on the performance in UAR for the case where the data are split by speaker while predictions are improved by 0.07 and 0.06 for random and linguistic split respectively. Although the addition of a multi-head attention block has a positive impact on UAR, it remains moderate compared to the other two splits : 0.3 for speaker split against 0.08 and 0.06 for random and linguistic split respectively. It can be hypothesised that the signal's aspects captured by the attention mechanism in addition are speaker-dependent. This hypothesis explains why the addition of such a mechanism has little or no effect when the speakers in the validation set have not been seen by the model.

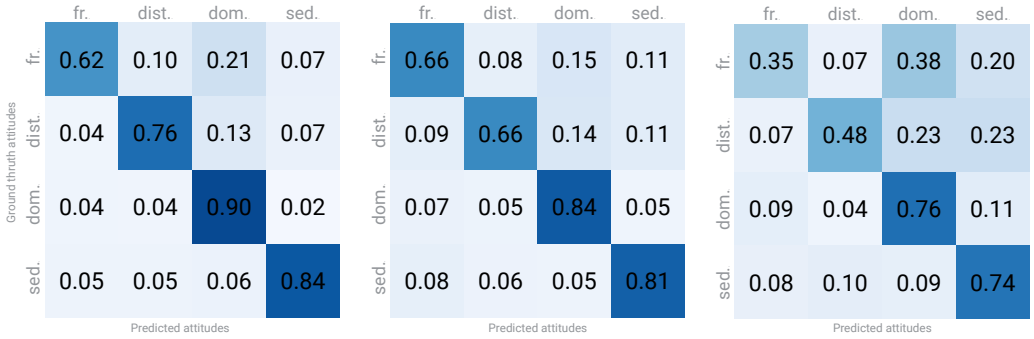

Figure 6.3: Speech attitude recognizer's confusion matrices for three split types, random (left), linguistic (middle) and speaker (right) in its best configuration MHSA-CRNN.

The confusion matrices displayed for what seems to be the best configuration, namely MHSA-CRNN, in Figure 6.3 enable us to go deeper in the analysis by looking at class dependant performance. It can be observed that two of the attitudes - dominance and seduction - are clearly better recognised than the other two - friendliness and distance. It can be seen that depending on the split considered, the performance gap between attitudes is not the same. Therefore, the effect caused by the choice of data split - random, linguistic or speaker wise - is dependent on the attitude considered. That suggests that the attribute according to which the split is performed does not have the same role in attitudinal recognition whether we consider one attitude or the other. While the performance gap remains moderate for the random and linguistic splits, it becomes huge for the speaker split. In comparison with friendliness, the best recognized attitude - dominant - is better recognized by 0.28, 0.18 and 0.41 in UAR for random, linguistic and speaker splits respectively. Analogously, in comparison with distance, dominant is better by 0.14, 0.18 and 0.28 in UAR for those same split types. It seems, therefore, that the model is not able to recognize friendliness when it is communicated by a speaker it has not seen. Even if the effect is less important, the same trend is also observed for distance.

Towards Speaker Informed Speech Attitude Recognition

As pointed out by Scherer in (Scherer, 1986), there is an undeniable contradiction between the apparent ease with which listeners judge emotions from speech and the intricacy of finding discriminative features in speech signal for emotion recognition. Indeed, the emotion recognition scores obtained in the case where the speaker is unknown, i.e. when the speaker has been seen by the model, are quite low compared to other split settings (random and linguistic). A plausible first explanation is that the model's learnt definition of what is an attitude is insufficient, in the sense that it is not always transferable to other speakers.

Let us reconsider how vocal emotion is processed by humans. When we meet someone we did not meet before, we risk misinterpreting their emotions, notably when vocalized. For instance, some people's subtle vocal traits may convey the same emotion as someone who would produce it much more emphatically. This brings up the issue of personal emotional granularity (Barrett, 2017) and individual production strategy of emotional - or attitudinal - content. In most cases, this problem is solved by learning. We come to know the person in front of us and their emotional expression patterns. The process of mastering this decoding skill takes time, it may take years to master it. If we use a schema : we inform our decoding of emotions with knowledge about the identity of the one who expresses them, in order to improve it.

A first step in enhancing the performance of speech emotion and attitude recognition models might to mimic this human functioning thus informing prediction with speaker identity. We have conducted many experiments to try to inform attitude recognition by speaker identity without really solving the problem at this date. Nevertheless, in (Le Moine et al., 2021a) we show that it is possible to improve the prediction performances of emotions - through the use of the IEMOCAP database (Busso et al., 2008) - and of attitudes - through the use of the (Le Moine and Obin, 2020) database - by informing the classifier with speakers embeddings. However, we have only managed to show this in the context of a speaker dependent experiment. When the speaker is not seen by the classification model during training, the model appears uncapable of exploiting the speaker embedding by which it is informed. Other experiments have shown that learning these speaker representations from a large multi-speaker database (Yamagishi, 2017) does not solve the problem either. It can be assumed that the way in which speaker identity and attitudes - or emotions - interact depends on the speaker that communicates them. In addition, the speaker identity informed attitude recognition algorithm outlined in (Le Moine et al., 2021a) did not reach the performance of the neural classifier introduced in this section. Although its architecture and functioning was interesting, we decided not to present it in this document.

6.2 Perceptual Regression Based on BWS Scores

In this section - and the two next ones - we mean to model the perceptual latent spaces that underlie the decoding of speech attitudes by individuals, thus having direct access to their mental representations of such attitudes. This marks a shift of paradigm as we no longer attempt to recognize attitudes from their apriori labels neither from their acoustic correlates. To do so, we conducted experiments using the data collected during the study on the perception of vocal attitudes (Section 4.4 of Chapter 4) using the BWS method. Indeed, this data may potentially be used to train models for different purposes such as perceptual regression, i.e. a model is trained to predict BWS perceptual scores which might allow to assess the perception of new sounds' attitude. The perceptual structuring of latent spaces is an another goal that might be pursued, i.e. a model is trained to

produce utterance embeddings that make sense with regards to the perception of attitudes.

6.2.1 About the Possible Uses of Gathered Perceptual Data

The data collected can be viewed in two ways. The raw data are the judgements made by participants, i.e relations between sounds within a trial, and the processed data, i.e. the BWS scores yielded by the BWS post-processing algorithm described in (Louviere et al., 2015).

Raw Judgements and How They Can Be Interpreted

Each utterance can be represented as a point in a perceptual space. Each sound is compared to a finite number of other sounds in the BWS paradigm. These comparisons are represented in the perceptual space by relations between the distances separating the compared point from the other points. As points cannot be ordered in a space of dimension less than one, the perceptual space has dimension greater than or equal to one. The data collected do not allow to access this space whose dimensionality is unknown. However, building such a space could be beneficial in many ways. For example, to improve our understanding of how attitudes are perceived through its analysis or even to provide perception-based information for an attitude conversion algorithm. Such a space could also be used for attitude perception assessment, i.e. news sounds could be plunged in the space, their perception being inferred from their relative position and distance to yet assessed utterances.

BWS Scores - A Projection of the Raw Perceptual Judgements

This second data format is a dimensional reduction of the actual perceptual space underlying the decoding of speech attitudes. In the perceptual space, the BWS scores are the projections of various utterance-related points on a 1-D space contained in this space. As a result, the information contained in these scores is: in the best case, a good representation of the participants' judgements, in the worst case - a rough summary of it. We may hypothesize that this projection, as a dimensional reduction, results in a loss of information and a simplification of the actual structure of the perceptual space of attitudes. In this section, we start by directly using the BWS scores as training data for a regression model. In particular, we will use the architecture presented in section 6.1 as a starting point for this experiment. The next section is an attempt to go beyond those scores by using raw perceptual judgements as data for model optimization.

6.2.2 Proposal for a Perceptual Regressor

Since the judgements made by participant only have a relative value which prevails only within trials, the use of raw BWS data involves a technical elaboration that is complex both to formalize and implement. Less ambitious but safer, the use of BWS scores in a classical neural regression framework is a good trade-off solution. The regression model shares the same objective as the participants in the perceptual experiment carried out on Att-HACK. In order to accurately predict the proper perceptual score, the regression model must be able to exploit the information from the speech signal that relate with the actual perception judgment. Furthermore, investigations outlined in 4 revealed significant interactions between BWS scores and both speaker identity and linguistic content. Thus it seems that individuals use both types of information to judge their own perception of attitudes. Therefore, it makes sense to explicitly input this data into the model and track how it affects the performance in regression.

Problem Positioning

In this framework, it is not necessary to take into account the experimental trials since the post-processing of the judgments, which enables the assignment of a perceptual score to each sentence, allows for the relativity issue to be avoided. Hence, each utterance x questioned with respect to an attitude a is being assigned a BWS score s_x^a that represents how a is perceived. It should be noted that we do not have access to the participants either, in the sense that each score is an average of the judgements of all participants. A regression model $\mathcal{R}^a$ which takes as input a representation of the speech signal $\mathbf{X}$ and produces a scalar p is then introduced. The model is optimized to yield outputs that match the BWS score s_x^a assigned to the input utterance. Pursuing such a goal, we expect the model to learn an implicit function that represents the underlying strategy used by individuals to decode speech attitudes. The aspects of speech signal used by receivers for attitude decoding must be captured by $\mathcal{R}^a$. Be aware that depending on the attitude decoded, these elements of the signal and, consequently, the implicit function that connects the signal's representation to the related perceptual score, may differ. So we learn as many models as there are attitudes.

Architecture of the Perceptual Regressor

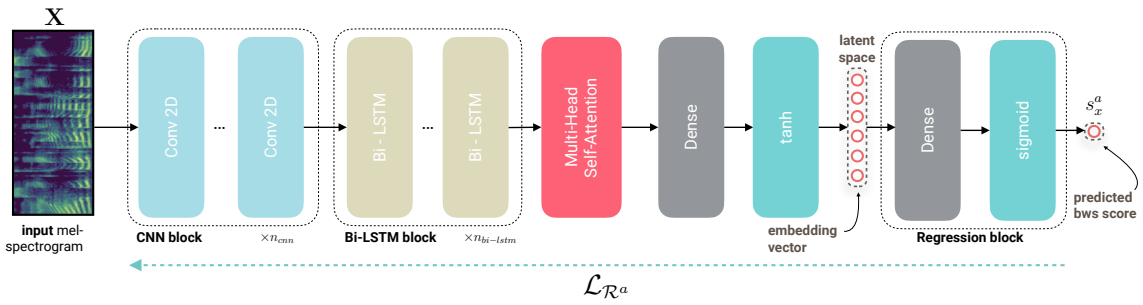

Figure 6.4: Schematic view of the perceptual regressor's neural architecture

We propose a perceptual regressor $\mathcal{P}_{reg}^a$ fully based on the ACRNN architecture - presented in previous section 6.1 - that takes mel-spectrogram representations of speech signal as inputs but only outputs a scalar in place of a multi-class vector. The selected architecture, depicted in Figure 6.4, slightly differs from the one we used for speech attitude recognition in the previous sections. As mentioned at the beginning of the chapter, the model consists of four main blocks.

A first convolutional block is used here to capture what is relevant in the signal to decode an attitude. The difference between this scenario and attitude recognition is that we are now trying to differentiate between instances of a particular attitude rather than trying to distinguish between attitudes themselves. In order to forecast a low BWS score for poor realizations, the model will need to capture the signal elements that convey the attitude being assessed. However, the model cannot stop with this fundamental job; it also needs to capture what may be described as an attitudinal intensity, i.e. it has to learn to differentiate between successful outcomes. The latter task seems much more complicated. Once these elements have been captured over time, the model produces a temporal summary of the yielded features using a recurrent block based on BLSTM layers followed by a self-attention block (potentially multi-headed). These three blocks aim to model the patterns, i.e. the temporal co-variation of various speech parameters, that underlie the perception of an

attitude. After the attention block there is a fully connected layer that allows the features obtained to be projected into a smaller space. By choosing the dimension of this space we make a hypothesis about the number of global criteria used by individuals to decode an attitude. The regression is finally performed from this space, i.e. each embedding vector is passed through a fully connected layer and converted into a scalar.

Towards Informed Perceptual Regression

In order to inform our perceptual regressor with the speaker identity and the linguistic content, we chose to represent each information by a one-hot-vector encoding. Assuming Att-HACK has finite number of speakers, each one can be assigned a number and represented as a one-hot-vector. For instance, if we denote $\mathcal{S} = \{s_1, ..., s_{Nsp}\}$ the set of N_{sp} speakers in Att-HACK, then a speaker s_i can be represented by a vector $\mathbf{h}^{sp_i} = [h_1^{sp_i}, ..., h_{N_{sp}}^{sp_i}]$ such as

$$h_n^{sp_i} = \begin{cases} 1 & \text{if } n = i \\ 0 & \text{else.} \end{cases} \tag{6.6}$$

Analogously, if we denote $\mathcal{S} = \{s_1, ..., s_{N_{sent}}\}$ the set of N_{sent} sentences, i.e. linguistic contents, in Att-HACK, then any sentence s_i in Att-HACK can also be represented by a one-hot-vector. Both types of information yield two different vectors that can be tiled along time dimension to match the melspectrogram shape. Once processed in that way, tensors can be concatenated to their related input melspectrograms and given to the model.

Model's Optimization

To train this model, we use a standard mean absolute error as cost function. Therefore, for a batch $\mathcal{B}$, the loss can formulated as follows

$$\mathcal{L}_{\mathcal{R}^a} = \sum_{x \in \mathcal{B}} |s_x^a - \mathcal{P}_{reg}^a(\mathbf{X})| \tag{6.7}$$

6.2.3 Experiments with the Perceptual Regressor

In order to obtain the best algorithm, we tested many configurations varying the number of convolutional layers (from 2 to 8), the size of their filters (from 16 to 256), the number of recurrent layers (from 0 to 2) and the size of their internal states (from 32 to 256), the size of the latent space (from 8 to 128) but also the number of attention heads (from 1 to 16) and their size (from 32 to 2048). The purpose of this section is not to report on this quest for the right architecture but rather to report on the performance of this optimal architecture depending on what is given as input. Nevertheless we give some keys on what in the model seems to impact the regression performance.

Data Processing

We used exactly the same data involved in the BWS experiment which represents nearly 2400 utterances for each attitude.

We normalized mel-spectrograms speaker wise for each of the four attitudes. We also normalized bws scores so that they properly range from 0 to 1 for each attitude. In fact, the scores were initially lying within $[0, 1]$ but values were found not to be close to 0 nor 1. Thus, denoting $\mathcal{X}^a$ the

set with all sentences questioned with respect to attitude a, for any utterance $x \in \mathcal{X}^a$, the related BWS scores were transformed as follows

$$s_x^a = \frac{s_x^a - \min\limits_{x' \in \mathcal{X}^a} s_{x'}^a}{\max\limits_{x' \in \mathcal{X}^a} s_{x'}^a - \min\limits_{x' \in \mathcal{X}^a} s_{x'}^a} \tag{6.8}$$

Criteria for Assessing Model Performance

Before launching experiments in order to select a baseline architecture, let us briefly review the metrics used to evaluate the performance of our model. We chose to use three metrics accounting three different aspects of model performance.

Mean Absolute Error (MAE). MAE is at the same time our training cost function and an interesting metric for evaluation. It measures how accurately the model predicts BWS scores in average over all validation samples. In particular it gives an idea of how far the predicted score of an unseen utterance is from its actual score.

Pearson Correlation Coefficient (r). This coefficient quantifies the linear relationship between the expected values and the actual values. The ideal situation is when the predicted values exactly match the observed values, their relation is precisely linear, and the correlation coefficient is 1. If the coefficient is close to 0, the model does not work, it does not manage to capture what in the signal leads to a rather low or rather high score. If the coefficient is close to 0.5, the model fails to make good predictions for the whole score scale. For example, it may predict an average score instead of a rather low score.

Coefficient of determination (r2). This coefficient, widely used for regression model's assessment, measures the proportion of variance of the predictions explained by the regression model. It is important to note that the amount of variability that the model can account for does not determine how significant the correlation between the predictions and the actual values is. This significance is measured through p-value. However it gives insight about how data is distributed around the linear regression line associated with the correlation coefficient. It would be close to 0 if data is far from this line and rather close to 1 if close.

Finding the Right Architecture

During this test phase we limited ourselves to the specific case of friendliness perceptual regression. Indeed, it would have taken too much time to test all the configurations (almost 40) on the four attitudes. The batch size was set to 80 and the learning rate to 10^{-5}. We trained each configuration for at least 200 epochs and stopped training once lowest validation loss was reached. We found that increasing the number of convolutional layers to the limit of 6 layers as well as decreasing the size of the filters to 32 also significantly improved performance. Changing the number of recurrent layers radically impacts the r2 score, so we kept 2 BLSTM layers. The size of the internal states showed little impact on the performance of the model, so we kept a size of 60. The number and size of the attention heads also have a great influence on the performance of the model, the chosen configuration uses 8 heads of size 1024. Finally, the size of the latent space has been drastically reduced to 8 for slightly better performance. The selected regression model is therefore rather small, using far fewer parameters than the model used for the emotion recognition task. It is also deeper, 6 convolutional layers compared to 4 for the SER.

Towards Homogeneous Prediction Performance Across the Full Range of BWS Scores

A major problem we faced when experimenting with perceptual regression is caused by the specific distribution of the BWS scores. Most of BWS values are close to 0.5 and very few are close to 0.0 and 1.0, i.e. the scores' distribution is some sort of a Gaussian. Moreover, the model's objective is to have the mean absolute error between predictions and actual values averaged over all samples the lowest as possible. To fulfill this objective the model can barely ignore those extreme values and just needs to provide accurate predictions for mean values (located around 0.5). The paradox is that such a perceptual regressor is partly dedicated to identify badly communicated instances of attitudes, i.e low BWS scores, as well as very well communicated ones, i.e. high BWS scores. Thus, it appears crucial to give more importance to those extreme valued samples during training so that the model take them into account.

To enforce this compensation, we multiply each sample's loss value by a factor γ_x^a that depends on the actual BWS score of the utterance x. We formulated this factor as follows

$$\gamma_x^a = 0.5 + \alpha * |0.5 - s_x^a| \tag{6.9}$$

where α is a positive scalar that controls how much we want to give importance to extreme valued samples during training. The greater α is the more importance extreme valued samples are being given.

6.2.4 Results & Discussion

In this subsection, we display performance results for the considered configurations of our perceptual regressor, REG basically works with melspectrograms, REG-SP is informed by speaker identity and REG-SP-LING is informed by speaker identity and linguistic content.

Performance Results for Informed Perceptual Regression Models

While the BWS task has been completed by participants without troubles - as proven by objective measures such as average experiment duration and participant consistency outlined in section 4.4, it appears difficult to artificially mimic these judgements, the main problem encountered being generalization. The performance results for the four selected configurations REG, REG-SP, REG-LING and REG-SP-LING across the four attitudes are displayed in Table 6.2.

Model	friendly			distant			dominant			seductive		
Metric	mae	r	r2	mae	r	r2	mae	r	r2	mae	r	r2
REG	0.14	0.47	0.18	0.14	0.18	0.01	0.14	0.35	0.09	0.13	0.57	0.30
REG-SP	0.14	0.51	0.24	0.13	0.32	0.06	0.13	0.43	0.18	0.13	**0.60**	**0.34**
REG-LING	0.15	0.42	0.16	0.13	0.23	0.02	0.14	0.41	0.14	0.14	0.49	0.20
REG-SP-LING	**0.12**	**0.64**	**0.39**	**0.12**	**0.37**	**0.12**	0.13	**0.54**	**0.30**	0.13	0.58	0.32

Table 6.2: Performance of the model in terms of mean absolute error (MAE) and r2 score for the four selected configurations and across attitudes

At first sight, we observe that the last configuration REG-SP-LING informed by both speaker identity and linguistic content obtains the best results in terms of MAE, pearson correlation coefficient r and coefficient of determination r2 for all attitudes except seduction that obtains better r and r2

with REG-SP. We note that observed correlations were all significative ($p < 0.001$). We found significant impact of the speaker identity information on the model's performance in terms of r2. When compared to basic REG configuration, speaker informed configuration REG-SP achieves improvement of 0.06 for friendliness, 0.08 for distance, 0.09 for dominance and 0.04 for seductiveness. We found the effect of linguistic information on the model's performance in terms of r2 to be dependant of the considered attitude. When compared to basic REG configuration, linguistically informed configuration REG-LING achieves improvement of 0.06 for distance, 0.05 for dominance. Conversely, linguistically informed configuration REG-LING is lower of 0.06 for friendliness and 0.10 for seductiveness in terms of r2. However, the two strong production profiles - namely seduction and dominance - uncovered in Section 4.3, are those for which the best results are obtained. We have $r^2 = 0.39$ and $r^2 = 0.22$ respectively for seduction and dominance while only $r^2 = 0.17$ and $r^2 = 0.11$ respectively for friendliness and distance.

By looking at the latent spaces - which here represent all the data - using a UMAP representation (McInnes et al., 2018) obtained through mapping the actual latent space with a 2D reduced space, one can see a polarisation according to the BWS score. High scores tend to be on one side of the space while low scores tend to be on the other side. We examined correlations between BWS scores and utterances' coordinate on the dimension of the reduced space in which polarisation is observed. We found $p = 0.28$ for friendliness, $p = 0.20$ for distance, $p = 0.31$ for dominance and $p = 0.42$.

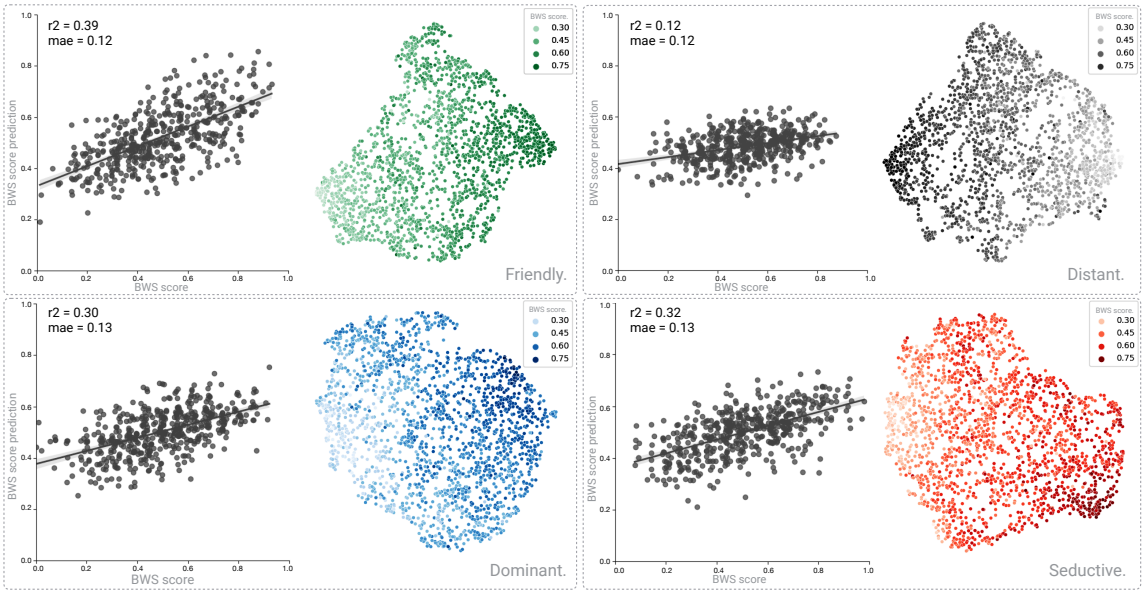

Figure 6.5: Regressor's performance in REG-SP-LING configuration for the four attitudes, friendly, distant, dominant and seductive. For each attitude, the correlation between actual and predicted bws scores (left) and a latent space UMAP vizualisation (right), in which darker dots are associated with higher BWS scores and lighter dots with lower scores, are depicted.

By looking at the graphs that display the predictions as a function of the actual BWS scores, we can observe that our regression models mostly predict average scores. Notably, our models make no predictions below 0.2 and almost none above 0.8 (except for friendliness) while some sentences have scores within these ranges. To explain this drawback, it can be hypothesised that

the model sees too few examples whose scores fall within these ranges for it to learn to yield correct predictions for these cases. Indeed the scores are approximately Gaussian distributed, the majority of the examples have scores between 0.3 and 0.7. In the following we try to take this into account in order to improve the predictions for the extreme bands of the BWS scale. Note that it seems essential that our models make good predictions for these extreme intervals, since one of their objectives is to detect poor achievements - related to low BWS scores - as well as very good achievements - related to high BWS scores.

On Improving Predictions for Low and High BWS Scores

So as to make up for the model's incapacity to predict low and high values, we intend to give the examples related to underrepresented scores more weight when the model is being optimized. To do so we weight each sample's x related loss when considering attitude a by a factor γ_x^a defined as formulated in Eq. 6.9.

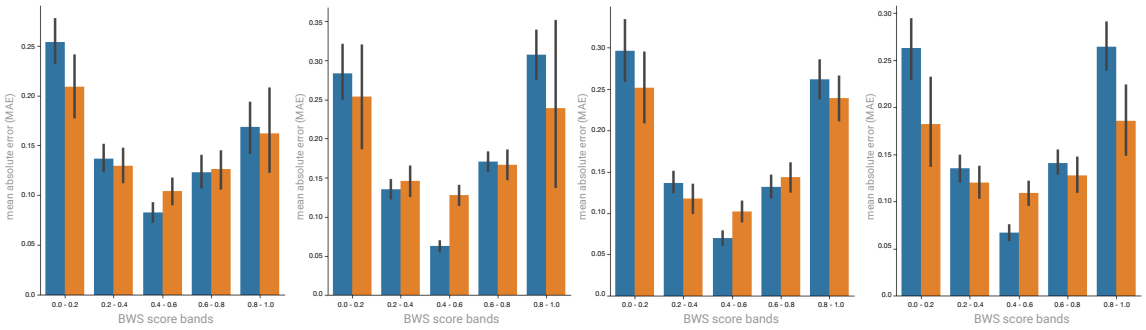

Figure 6.6: Regressor's performance in REG-SP-LING configuration with weighting constraint in terms of mean absolute error for different BWS bands for the four attitudes, friendly, distant, dominant and seductive. Error bars show 95% confidence interval.

The approach is fair in practice; as shown in Figure 6.6, this adjustment lowers the prediction error for the marginal bands. However, it also raises the error on the other bands leading to poorer global performance. As shown in Table 6.3, both mae, r and r2 are worst in the weighted case than in the basic one.

Model	friendly			distant			dominant			seductive		
Metric	mae	r	r2	mae	r	r2	mae	r	r2	mae	r	r2
REG-SP-LING	**0.12**	**0.64**	**0.39**	**0.12**	**0.37**	**0.12**	0.13	**0.54**	**0.30**	0.13	0.58	**0.32**
WEIGHTED	0.13	0.60	0.30	0.15	0.29	-0.3	0.13	0.49	0.19	0.13	**0.61**	0.30

Table 6.3: Performance of the model in terms of mean absolute error (MAE), pearson correlation coefficient r, and coefficient of determination r2, for the selected baseline and its sample weighted version across attitudes

How can this be interpreted? The assumption we make is that there are two different sub-tasks underlying this seemingly simple regression, and it is likely that the model is being asked to do something that is too complex. In order to produce accurate predictions, the model must account for what can be described as an attitude intensity, or a difference in the degree of communication

of an attitude. However, it should also be able to identify instances where the attitude has been miscommunicated, or just not perceived as so by the receivers. The model likely needs to capture a variety of quite distinct signal properties for these two objectives. In one case the attitude is effectively communicated and it is a matter of distinguishing degrees between instances of this attitude. In the other case, it is not successfully communicated, the vocal characteristics underlying its production are possibly absent. It is likely that the model captures fluctuations around a global pattern with an associated attitude for the first subtask. Conversely, for the second task, we can assume that the pattern might be completely absent. Therefore, in order to predict, the model would need to detect its absence. With regard to the high scores, it can be hypothesised that what underlies the individuals' judgements has more to do with intrinsic characteristics of the voice being judged than with the attitude being communicated. To make good predictions, the model should then capture these intrinsic features, e.g. the identity of the speaker. We must emphasize that the above statements are merely hypotheses, and that they lack any explanatory value until they are supported by evidence.

We propose to take a step back by considering ranges - or intervals - of scores rather than scores in order to, on the one hand, question these hypotheses and, on the other hand, try to obtain a reliable algorithm that can mimic the average perceptual judgment for each of the attitudes. The problem therefore no longer takes the form of a regression task but of a classification task.

6.3 Perceptual Classification based on BWS scores

It is challenging to develop a model that predicts the BWS score from a mel-spectrogram, as we have seen in the previous section. The hypothesis we have retained is that several sub-tasks are intertwined within this seemingly basic regression task. The BWS score, in particular, might be interpreted in two different ways. For instance, a very low score could signify a low intensity of the attitude conveyed, but it could also denote a realization for which no attitude related pattern can be identified, a sentence for which the target attitude is not recognized as such by receivers or for which other aspects are used by receivers to judge. Our attempts of regression were unsuccessful most likely due to the polysemy of the resulting BWS scale. An efficient model must be able to discriminate between these two interpretations in the signal as well as what ties a sentence to its score. This can be seen as a two-stage task: first a classification task that distinguishes interpretations of the BWS score and then a regression task that aims to predict the score accurately.

6.3.1 Proposal for a Perceptual Classifier

In this section, we consider implementing one simpler subtask: the recognition of perceptual domains. For this purpose we make the following assumption: depending on the BWS ranges considered, one of the aforementioned interpretations takes precedence over the others. For example, we will consider that sentences with a very low score are bad productions or poorly communicated attitude instances. Sentences with average scores will be considered to have come from the attitude's typical production. Differences within this range will be interpreted as variations in the intensity of attitudes. High-scoring sentences will form the final perceptual domain and will be regarded as non-typical production. It is considered that in these sentences, some other feature of speech like the speaker's identity takes primacy in the decoding process.

Problem Positioning

A classifier $\mathcal{C}^a$ which takes a representation of the speech signal $\mathbf{X}$ as input and produces a multi-class vector $\mathbf{d}$ is introduced. Provided BWS scores normalized with respect to Eq. 6.8, we split data in N_b categories corresponding to samples related to different perceptual bands, i.e. with scores lying in different bands of the BWS scale. Each band $\mathcal{B}^{a,i}$ is defined by a lower bound $s_l^{a,i}$ and a higher bound $s_h^{a,i}$. Thus any sample x assigned with a BWS score s_x^a is ensured to be related with exactly one band provided

$$x \in \mathcal{B}^{a,i} \text{ if } \begin{cases} s_l^{a,i} \leq s_x^a \\ s_x^a < s_h^{a,i} \end{cases} \tag{6.10}$$

Each band $\mathcal{B}^{a,i}$ is represented by a one-hot vector $\mathbf{b}^{a,i}$ of size N_b which is defined as follows

$$b_k^{a,i} = \begin{cases} 1 \text{ if } k = i \\ 0 \text{ else.} \end{cases} \tag{6.11}$$

The classifier's objective is to capture the aspects of speech signal that enable individuals to distinguish between those N_b domains. Again, we note that depending on the attitude decoded, these elements of the signal and, consequently, the implicit function that connects the signal's representation to the related BWS domain, may differ. So we learn as many models as there are attitudes.

Architecture of the Perceptual Classifier (PC)

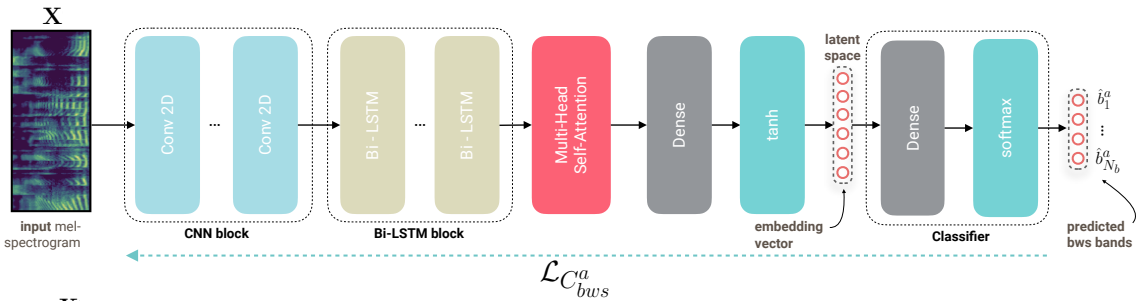

Figure 6.7: Schematic view of the perceptual band classifier's neural architecture

We propose a classifier $\mathcal{C}^a$ fully based on the ACRNN architecture that takes melspectrogram representations of speech signal as inputs and outputs a multi-class vector. The selected architecture, depicted in Figure 6.7, slightly differs from the one we used for speech attitude recognition in the previous sections. As mentioned at the beginning of the chapter, the model consists of four main blocks. Since the model's design is the same as that utilized in the previous part and can be interpreted in the same way, we won't go into great detail about it here. The only notable differences are in the last block, i.e. the classification block. Here the fully connected layer predicts a vector of size N_b and is followed by a softmax activation. Similar to the regression task presented in the previous section, we found interesting to provide additional information to the classification model such as speaker identity and linguistic content. For more details on how this information is represented and integrated into the model, please refer to the subsection 6.2.2 of the last section.

To train this model, we use a standard categorical cross entropy as cost function. Therefore, for a batch $\mathcal{B}$, the loss can formulated as follows

$$\mathcal{L}_{\mathcal{C}^a_{bws}} = -\sum_{i=1}^{N_b} d_i \log(\hat{d}_i) \tag{6.12}$$

6.3.2 Experiment with the Perceptual Classifier

The number of convolutional layers, the size of their filters, the number of recurrent layers and the size of their internal states, the size of the latent space, as well as the number and size of attention heads, were all varied within aforementioned ranges (6.2) to find the optimal algorithm. We do not think it is worthwhile to report on the influence of any of these parameters here. There are two key goals for this experiment. The first is to investigate the claims that distinct perceptual domains exist. The second entails developing a system to validate the Att-HACK database, specifically the detection of instances of poorly or well-communicated attitudes. Through this second goal, a use that is dear to us emerges: conditioning a system of attitude conversion with perceptual information. In particular, at the time of writing we work on using the information gathered in the BWS experiment to monitor the intensity of the converted attitude.

Data Processing

The data are processed and normalized is exactly as in the previous section. The only difference is that we used balanced batches, i.e. with as many examples for each BWS sub-band. We thus used nearly 2400 utterances for each attitude. Melspectrograms were normalized speaker wise for each of the four attitudes. BWS scores were also normalized following Eq. 6.8 to have values lying between 0 and 1 for each attitude.

Proving the Classifier with Additional Information

For this experiment, similar to what was done before, we propose to inform the model with two types of information: on the one hand the identity of the speakers, on the other hand the linguistic content. Since we needed to run the model on data that was not evaluated in the BWS experiment, and hence on other linguistic contents, we had to come up with a different method of providing the model with linguistic information than through the use of one-hot encoding. Even though every speaker in the database was evaluated during the perceptual experiment, only some of the sentences, i.e. linguistic contents, were evaluated.

To feed the model with relevant representations of the sentences in the Att-HACK, we used a semantic content encoding algorithm called CamenBERT. CamemBERT (Martin et al., 2020) is a state-of-the-art language model for French based on the RoBERTa architecture (Liu et al., 2019) pretrained on the French subcorpus of the newly available multilingual corpus OSCAR. Although semantic and linguistic aspects are interdependent, they should not be confused. By immersing the 100 sentences of the database in the language model's latent semantic space, we obtain a vector representation $\mathbf{h}^{sem}$ of size 768 for each sentence. Obviously, concatenating a sequence of vectors of this size to the input melspectrogram slows down the training considerably and ultimately gives too much importance to the semantic information. We therefore reduce the dimension of these semantic embeddings by projecting them into a smaller space using a fully connected layer.

Criteria for Assessing the Model's Performance

Before launching experiments in order to select a baseline architecture, let us briefly review the metrics used to evaluate the performance of our model. We chose to use two metrics accounting two different aspects of model performance.

Unweighted Average Recall (UAR). Already used in previous experiments, this metric measures the average of the recall on the positive class and recall on the negative class.

Precision (P). The precision is the ratio between the number of true positives and both true and false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative. The addition of this second criterion makes it possible, among other things, to select the configuration for which sentences with poorly communicated attitude are least misclassified.

Finding the Right Architecture

During this test phase we considered the four attitudes. We did not test as many configurations as in the previous experiment in the case of regression. We have reduced the batch size to 32 so that the different BWS bands can be evenly distributed. We kept a learning rate of 10^{-5}. We trained each configuration for at least 200 epochs and stopped training once lowest validation loss was reached. We have tested different numbers and values for the low and high limits that define the BWS bands we wish to be able to identify. We have chosen to make class prediction on four distinct bands ($N_b = 4$) listed below.

- A **lower band** with scores $s_x^a \in [0.0, 0.2[$ in which badly communicated instances of a as well as low intensity instances of a can be observed.
- A **low medium band** with scores $s_x^a \in [0.2, 0.5[$ that can be associated with normally communicated attitude instances of low intensity.
- A **high medium band** with scores $s_x^a \in [0.5, 0.8[$ that can be associated with normally communicated attitude instances of high intensity.
- A **higher band** with scores $s_x^a \in [0.8, 1.0]$ in which extremely well communicated attitude instances of a can be observed.

Towards a Lightweight Model for BWS Range Classification

It is conceivable that such a model could be connected to the output of a speech attitude conversion system in order to condition the training of the transformation on perceptual data. Note that as many models as attitudes are converted will be required, which can be problematic both in terms of memory and computing time. So, in order to make these models easily included and still effective, we must endeavor to make them smaller.

6.3.3 Results & Discussion

In this section, we present the results of our BWS perceptual band classification models. Regarding the assumptions stated at the beginning of this part, we attempt to analyze their performance.

Configurations	FRIENDLY	DISTANT	DOMINANT	SEDUCTIVE
UAR	0.51	0.43	0.47	0.50
P	0.46	0.29	0.38	0.39

Table 6.4: BWS range classifier's performance in terms of unweighted average recall and precision score for each attitude.

BWS Range Classifier's Performance

In general, the scores obtained are quite modest. In particular, it can be seen that the precision scores are rather low, especially for distant with $P = 0.29$, dominant with $P = 0.38$ and seductive with $P = 0.39$. This means that the model cannot avoid to predict false positives. The UAR scores are literally average (around 0.5) but this must be put into perspective by the variability of performance depending on the perceptual band being considered.

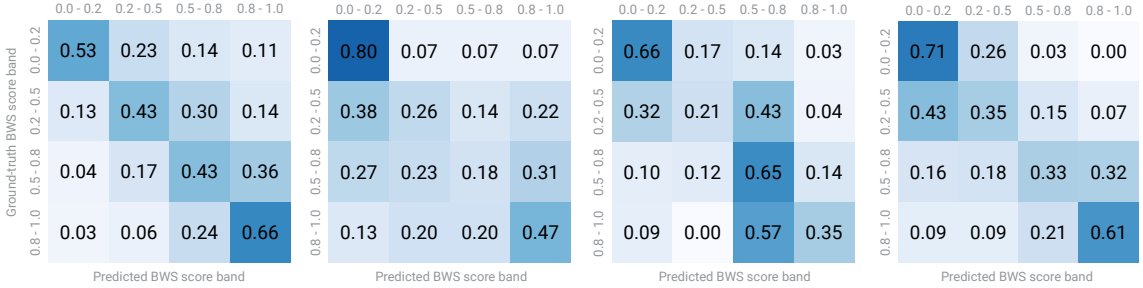

Figure 6.8: BWS score band classifier's confusion matrices for the four attitudes (from the left to the right) friendly, distant, dominant and seductive with the selected configuration.

If we look at the prediction performance per BWS score range, i.e. BWS domain, we immediately see that there are big differences depending on whether we consider one band or the other. In particular, we observe for the four attitudes a clear better prediction performance for the extreme ranges, namely the range associated with very low scores (< 0.2) and the range associated with very high scores (> 0.8). This finding supports the hypothesis that there are different perceptual or communicative domains within the resulting BWS scale. It can be hypothesized that this indicates a difference in the nature of the attitude communicated for sentences associated with very low scores and that communicated for sentences associated with scores between 0.2 and 0.8. Similarly, although to a lesser extent, it seems that the attitude communicated for sentences associated with very high scores is different in nature from that communicated for medium score sentences. For the average scores, between 0.2 and 0.8, it seems difficult to identify categories, so we can imagine that this domain is relatively homogeneous, we can think of a continuum of attitude within which differences of degree could be observed between attitude instances.

Latent Space Visualization

We represented latent spaces yielded by our perceptual classifier $\mathcal{C}^a$ using UMAP visualizations in Figure 6.9. The perceptual domains associated with the different BWS bands are represented by different colors. In line with the assessment of predictive performance, the domains are not distinct, they largely overlap. Nevertheless, we can see that the extreme bands are well separated,

with sentences associated with scores below 0.2 being spatially opposed to those associated with scores above 0.8. As for the other two categories, while there is a tendency for polarisation, with sentences with scores between 0.2 and 0.5 being closer to the sentences associated with very low scores and similarly for high scores, there is also a large part of the space that is blurred and contains points from both of these categories.

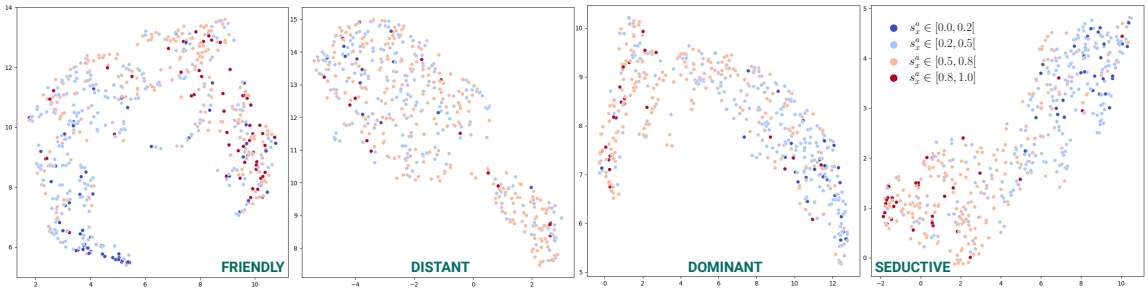

Figure 6.9: Latent spaces UMAP visualizations for the four attitudes. The four perceptual domains are represented using different colors. Note that only validation data is represented.

Preliminary Discussion

We have found that learning a regression from BWS scores is rather challenging. Using such a regression model, it was specifically found to be difficult to predict extremely low and extremely high scores. As one of the objectives of this Chapter is to identify bad realizations - i.e. instances of poorly communicated attitudes - we could not leave it at that. We then adopted a different strategy. We divided the BWS scale into four categories thus changing the issue into an apparently easier classification task. Some of these categories - the extreme bands - proved to be easier to predict than the others, suggesting that they represented radically different things, we therefore speak of perceptual domains. By identifying the utterances predicted as to be related to the lower perceptual range, this classification model will also enable a minimal validation of the entire database. As only a sub-part of the base was judged by the participants. It will be assumed that these identified samples are essentially poor vocal productions and are therefore not of interest in our attitude conversion framework.

6.4 Perceptual Metric Learning Based on BWS Raw Judgements

We have shown how challenging it is to design an algorithm - BWS-Net - that mimics human judgements artificially. In the previous two sections we have limited ourselves to training these models from the BWS scores. These scores vary on four linear scales which can be seen as 1-D sub-spaces of the actual multi-dimensional space in which the perception of attitudes occurs. From this point of view, a BWS score is a projection onto one of these sub-spaces of the average judgement about an attitude instance. The BWS score therefore potentially contains much less information about the perception of the attitude conveyed by the sentence with which it is associated than the actual judgment that was made. In the light of this observation, we propose to learn a model to mimic the process by which an individual decodes a vocal attitude, from the raw judgements of the BWS experiment.

6.4.1 Proposal for a Perceptual Arranger

In this section we attempt to develop a method for learning a model from the raw BWS judgements. This method is all the more interesting as it can be transposed to other cases of application of the BWS such as the perceptual assessment of timbre investigated in (Rosi et al., 2022).

Problem Positioning

A trial is a tuple of N sounds $t^a = \{x^1, ..., x^N\}$ on which judgements are made with respect to an attribute a. Here, a can be either friendly, distant, dominant or friendly. Within a trial, one utterance is judged best, one is judged worst and others can be considered neutral. Those judgements are denoted b, w and n respectively. We denote $\mathcal{T}^a$ the set containing all the trials considered for the BWS experiment that investigates a. Hence, in a given trial $t^a \in \mathcal{T}^a$ depicted in Figure 6.10, any utterance $x \in t^a$ is being assigned a judgement $j_x^{t^a} \in \{b, w, n\}$ in regards with the attribute a.

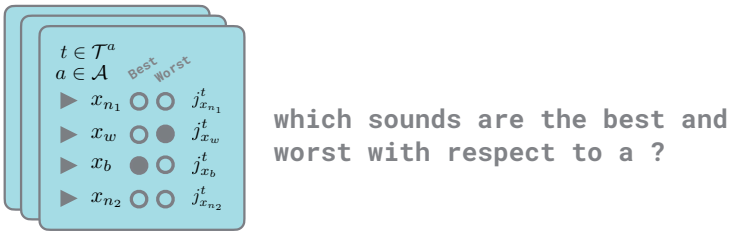

Figure 6.10: A trial $t \in \mathcal{T}^a$ of $N = 4$ sounds judged with respect to the BWS paradigm

To use utterances with regards to their BWS status, we rename them with respect to the trial they lie in and the judgement they have been assigned. The best of trial t^a can be indexed as t^a_b, i.e. as part of trial t investigating attribute a and judged best. Analogously the worst and neutrals of trial t^a can be indexed t^a_w and $t^a_{n_i}$ with $i \in [\![1, N-2]\!]$ respectively. Thus, depending on the trial considered and the judgement made, an utterance x could be denoted $x^{t^a_b}$, $x^{t^a_w}$ or $x^{t^a_{n_i}}$ with $i \in [\![1, N-2]\!]$. We would thus design different instances of a given sound within the BWS experiment. In the perceptual space related to a, i.e. the space formed by judgements with regards to a, each sound x is represented by a vector $\mathbf{h}^x$. Therefore, the vectors representing the utterances of trial t^a can be denoted for $i \in \{1, ..., N-2\}$ as follows

$$\mathbf{h}^{t^a_b} = \mathbf{h}^{x^{t^a_b}} \tag{6.13}$$

$$\mathbf{h}^{t^a_w} = \mathbf{h}^{x^{t^a_w}} \tag{6.14}$$

$$\mathbf{h}^{t^a_{n_i}} = \mathbf{h}^{x^{t^a_{n_i}}} \tag{6.15}$$

Proposal for a BWS-Based Perceptual Arranger

Following our intention to exploit the data obtained from BWS experiment, we thought of developing a model that captures what underlies the perception of speech attitudes in the signal. The model would take mel-spectrograms as inputs and generate perceptual embeddings, i.e. perceptually salient representations of input utterances. For any input mel-spectrogram $\mathbf{X}$, our perceptual arranger $\mathcal{A}^a$ outputs an embedding vector $\mathbf{h}^x$. $\mathcal{A}^a$ is basically a ACRNN, as introduced in section

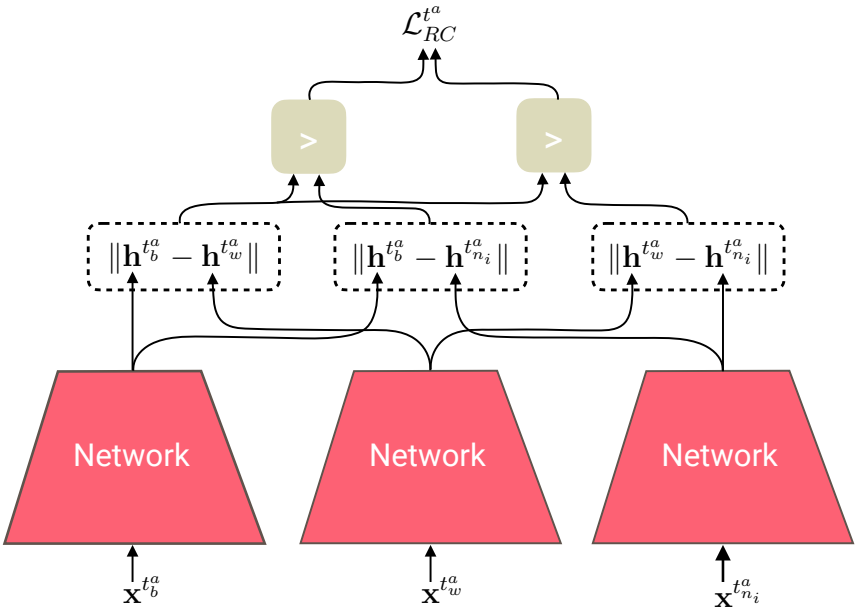

Figure 6.11: Relative Contrasting (RC) loss structure processing a batch formed with a trial t^a. The structure has to be repeated $N-2$ times depending on the size N of each trial that changes the number of neutrals in a trial. Scheme inspired from the triplet loss structure depicted in (Hoffer and Ailon, 2015).

6.1 and depicted in Figure 6.2, from which end classifier block is discarded. As a first experiment, we propose to use the raw judgements made by participants as model training criterion. Thus the model will explicitly learn a latent space in which representations of utterances are arranged with respect to the BWS judgements. The BWS judgments are not appropriate for use as such to train a model. Judgments, however, can be understood as relations between sounds within a trial. For a given trial, $2(N-1)$ relations can be inferred from raw judgements. Denoting $\prec_a$ the order relation, i.e. if $x \prec_a y$ then x is more perceived as a than y. We can write for $i \in [\![1, ..., N-2]\!]$

$$x^{t_b^a} \prec_a x^{t_w^a} \tag{6.16}$$

$$x^{t_{n_i}^a} \prec_a x^{t_w^a} \tag{6.17}$$

$$x^{t_b^a} \prec_a x^{t_{n_i}^a} \tag{6.18}$$

To turn those relations understandable by a neural model we can translate them in terms of distances. Thus, introducing a distance $\|.\|: \mathbb{R}^{d_{emb}} \to \mathbb{R}$, we can write for $i \in [\![1, ..., N-2]\!]$

$$\|\mathbf{h}^{t_b^a} - \mathbf{h}^{t_w^a}\| \geq \|\mathbf{h}^{t_b^a} - \mathbf{h}^{t_{n_i}^a}\| \tag{6.19}$$

$$\|\mathbf{h}^{t_b^a} - \mathbf{h}^{t_w^a}\| \geq \|\mathbf{h}^{t_w^a} - \mathbf{h}^{t_{n_i}^a}\| \tag{6.20}$$

To train the model to match these relations between distances in the latent space, we imagined a training criterion, a cost function inspired by the metric learning literature and notably the famous triplet loss (Hoffer and Ailon, 2015).

RELATIVE CONTRASTING (RC) LOSS. The peculiarity of this case is the relative nature of relations that must be imposed within the latent space. Analogously to the context of triplet loss, the relations here involve at least three points in the space, i.e. each of the inequalities written above involves three points as depicted in Figure 6.11. However there is no guarantee that three utterances, randomly picked in latent space, are linked by a relation. Here there are no absolute relations between sounds, relations can only be relative, valid within a trial. To avoid the model to collapse, i.e. turning all utterances into one single point in the latent space, we introduce a positive margin α. The RC loss $\mathcal{L}_{t^a}$ can be defined for all trial $t^a \in \mathcal{T}^a$ as

$$\begin{aligned}\mathcal{L}_{rc}^{t^a} = &\frac{1}{n_v^{t^a}} \sum_{i=1}^{N-2} \max\left(\|\mathbf{h}^{t_b^a} - \mathbf{h}^{t_{n_i}^a}\| - \|\mathbf{h}^{t_b^a} - \mathbf{h}^{t_w^a}\| + \alpha, 0\right) + \\ &\frac{1}{n_v^{t^a}} \sum_{i=1}^{N-2} \max\left(\|\mathbf{h}^{t_w^a} - \mathbf{h}^{t_{n_i}^a}\| - \|\mathbf{h}^{t_b^a} - \mathbf{h}^{t_w^a}\| + \alpha, 0\right)\end{aligned} \tag{6.21}$$

where $n_v^{t^a}$ is the number of valid relation, i.e. unfulfilled relation, with the trial t^a.

Imposing a fixed margin that is not dependent on the trial under consideration appears problematic. The best and worst outcomes of one trial may in fact be perceived as being closer than the best and neutral outcomes of another. We run into the relativity of the judgments made once more. To tackle this issue, we introduce another network $\mathcal{M}$ dedicated to margin learning. This model takes two parameters as arguments, a mean value μ and an amplitude δ such that any learnt margin lies between $\mu - \delta$ and $\mu + \delta$. Taking all the embeddings in the batch as input, it produces $N-2$ distinct margins $\{\alpha_{b,n_i}, \alpha_{w,n_i}\}_{i \in [\![1, N-2]\!]}$ related to each of the trial's relations respectively. Details about the implementation are given in the experiment the next section.

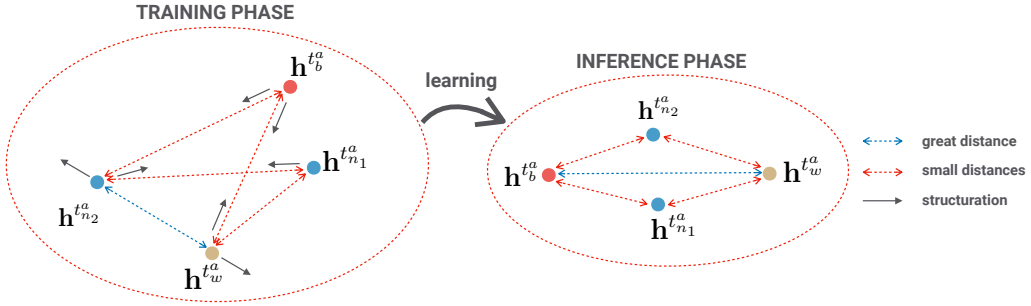

Figure 6.12: Configurations of the latent space at training (left) and inference phase (right) for a given trial t^a composed with four utterances $x^{t_b^a}$, $x^{t_w^a}$, $x^{t_{n_1}^a}$ and $x^{t_{n_2}^a}$ respectively judged best, worst, neutral and neutral. The model explicitly learns a latent space structure that reflects the participants' judgements.

DYNAMIC MARGIN RELATIVE CONSTRASTING (DM-RC) LOSS. With this refinement on margins, the RC loss formulated in equation 6.25 slightly changes as not does it take embeddings as inputs but also outputs of the margin network. Thus for a given trial t^a, the DM-RC loss can be expressed as follows

$$\mathcal{L}_{dmrc}^{t^a} = \frac{1}{n_v^{t^a}} \sum_{i=1}^{N-2} \max\left(\|\mathbf{h}^{t_b^a} - \mathbf{h}^{t_{n_i}^a}\| - \|\mathbf{h}^{t_b^a} - \mathbf{h}^{t_w^a}\| + \alpha_{b,n_i}, 0\right) + \\ \frac{1}{n_v^{t^a}} \sum_{i=1}^{N-2} \max\left(\|\mathbf{h}^{t_w^a} - \mathbf{h}^{t_{n_i}^a}\| - \|\mathbf{h}^{t_b^a} - \mathbf{h}^{t_w^a}\| + \alpha_{w,n_i}, 0\right) \tag{6.22}$$

We seek to explicitly learn the latent structure underlying speech attitude perception by fitting a model on raw judgments with respect to those losses. Such a model should be able to extract from speech signals the features that people use to decode speech attitudes. Using those discriminative features, the model rearranges the points in its latent space to move them closer to and farther away from one another in order to match the BWS judgments. The training and inference phases are depicted in Figure 6.12 for a given trial.

Initial experiments showed that if the model was given the freedom to learn a dynamic margin, it systematically tended to learn the smallest possible margin. In reality, however, some sounds are perceived as very distant and others as very close with regards to the attitude conveyed. Intuitively, we can assume a Gaussian distribution of distances between sounds in a perceptual space. It can therefore be assumed that the margins learned by the model also follow a Gaussian distribution, which means that few margins will be small and large and the majority will be medium.

Dynamic Margin Constraint (DMC). In order to counteract the model's propensity to learn the smallest margin possible, we decided to impose a constraint dependant on the learned margin. This constraint can be formalised through a function γ that takes the learned margins as an argument and produces a scalar that is then added to the total loss of the model. There are many ways to apply such a constraint, so we tested different functions. The DMC constraint can thus be formulated for a trial t^a as follows

$$\mathcal{L}_{dmc}^{t^a} = \sum_{i=1}^{N-2} \gamma(\alpha_{b,n_i} - \mu) + \gamma(\alpha_{w,n_i} - \mu) \tag{6.23}$$

One final point to note. The decrease in the DM-RC loss is not sufficient to guarantee an increase in the number of relationships actually satisfied. Since margins can decrease overall without affecting order relationships. In particular, we observed this behaviour at the end of convergence. To avoid this, we have added a final loss directly derived from the DM-RC loss.

Fulfilled Relations (FR) loss. This loss corresponds for a given trial t^a to the number of unsatisfied relationships within the trial divided by the number of elements in the trial N.

$$\mathcal{L}_{fr}^{t^a} = \frac{n_v^{t^a}}{N} \tag{6.24}$$

For any utterance x in the dataset whose perception is questioned with regards to a, we denote $\mathcal{T}_x^a = \{t^a \in \mathcal{T}^a \mid x \in t^a\}$ the subset of $\mathcal{T}^a$ in which each trial contains x and $\mathcal{T'}_x^a$ a subset containing N_t randomly picked elements of $\mathcal{T}_x^a$. From each sound x in the dataset, we generate batch $\mathcal{B}_x^a$ of size $N * N_b$ by sampling N_b trials within $\mathcal{T'}_x^a$. Hence, for a given batch $\mathcal{B}_x^a$, the total perceptual arranger loss can be expressed as follows

$$\mathcal{L}_{\mathcal{A}^a}^{\mathcal{B}_x^a} = \sum_{t \in \mathcal{T'}_x^a} \mathcal{L}_{rc}^{t^a} \tag{6.25}$$

In the case of dynamic margin, the loss becomes slightly more complex and is formulated as follows

$$\mathcal{L}_{\mathcal{A}^a}^{\mathcal{B}_x^a} = \sum_{t \in \mathcal{T'}_x^a} \mathcal{L}_{dmrc}^{t^a} + \lambda_{dmc}\mathcal{L}_{dmc}^{t^a} + \lambda_{fr}\mathcal{L}_{fr}^{t^a} \tag{6.26}$$

6.4.2 Experiments with the Perceptual Arranger

We carried out experiments using both models introduced in the previous part, namely the perceptual arranger and regressor, on the data gathered during the BWS study on Att-HACK described in 4.4. experiment.

Implementation Details

Implementing RC loss was a major challenge as the relationships between sounds are only valid within a trial. In order to optimize the model, we must execute the computation tensorially, we cannot just calculate each term of Eq. 6.25 one at a time. To do so, we have been inspired by the way the triplet loss (Hoffer and Ailon, 2015) is implemented. When questioning attitude a, from each utterance x a batch $\mathcal{B}_x^a$ is designed as follows

$$\mathcal{B}_x^a = \bigcup_{t^a \in \mathcal{T}_x^a} \{x' \mid x' \in t^a\} \tag{6.27}$$

We further denote n the number of elements in $\mathcal{B}_x^a$. Our custom RC loss takes three tensors as input. A first tensor $\mathbf{H} = \{\mathbf{h}_i\}_{i \in [\![1,n]\!]}$ is made of latent vectors corresponding to each utterance in the batch, a second tensor $\mathbf{t} = \{t_i\}_{i \in [\![1,n]\!]}$ is made of the corresponding trial names, finally a third tensor $\mathbf{j} = \{j_i\}_{i \in [\![1,n]\!]}$ is made of each batch element's corresponding judgement labels (b, w, n). For instance, if we consider a batch $\mathcal{B}_x^a$ formed by only two trials t^a and t'^a both containing x, then

$$\mathbf{H} = [\mathbf{h}^{t_b^a}, \mathbf{h}^{t_w^a}, \mathbf{h}^{t_{n_1}^a}, \mathbf{h}^{t_{n_2}^a}, \mathbf{h}^{t'^a_b}, \mathbf{h}^{t'^a_w}, \mathbf{h}^{t'^a_{n_1}}, \mathbf{h}^{t'^a_{n_2}}] \tag{6.28}$$

$$\mathbf{T} = [\mathrm{t}, \mathrm{t}, \mathrm{t}, \mathrm{t}, \mathrm{t'}, \mathrm{t'}, \mathrm{t'}, \mathrm{t'}] \tag{6.29}$$

$$\mathbf{J} = [b, w, n, n, b, w, n, n] \tag{6.30}$$

To compute RC loss, we need to compute batch embeddings' pairwise distances as $\mathbf{D}$. We first compute the dot product tensor $\{\mathbf{h}_i\mathbf{h}_j\}_{i,j \in [\![1,n]\!]}$ between $\mathbf{H}$ and its transposition $\mathbf{H}^T$ from which diagonal is extracted (operator detoned $diag$) thus representing the square norm vector of $\mathbf{H}$. The pairwise squared distances tensor $\mathbf{D}$ can be obtained as follows

$$\mathbf{D} = max\left(diag(\mathbf{H} \odot \mathbf{H}^T)[:, newdim] - 2\mathbf{H} \odot \mathbf{H}^T + diag(\mathbf{H} \odot \mathbf{H}^T)[newdim, :], 0\right) \tag{6.31}$$

The next step is to create appropriate masks that allow for RC loss computation as formulated in eqs. 6.25 and 6.22. A first mask $\mathbf{M}^g = \{m_{p,r}\}_{p,r \in [\![1,n]\!]}$ is dedicated to drop any element of $\mathbf{D}$

which is not involved in the computing of so called *great* distances $\|\mathbf{h}^{t_b^a} - \mathbf{h}^{t_w^a}\|$, i.e. the distances between embeddings of utterances that had been judged best and those that had been judged worst. To compute the mask we use both $\mathbf{t}$ and $\mathbf{j}$ tensors as follows

$$m_{p,r}^g = \begin{cases} 0 & \text{if } j_p = n \text{ or } j_r = n \\ 1 & \text{else} \end{cases} \tag{6.32}$$

A second mask $\mathbf{M}^s = \{m_{p,q}\}_{p,q \in [\![1,n]\!]}$ is dedicated to drop any element of $\mathbf{D}$ which is not involved in the computing of so called *small* distances $\|\mathbf{h}^{t_b^a} - \mathbf{h}^{t_{n_i}^a}\|$ and $\|\mathbf{h}^{t_w^a} - \mathbf{h}^{t_{n_i}^a}\|$ for all $i \in \{1, 2\}$, i.e. all the other terms in the loss formula. The mask is thus computed as follows

$$m_{p,q}^s = \begin{cases} 1 & \text{if } (j_p = n \text{ and } j_q \neq n) \text{ or } (j_p \neq n \text{ and } j_q = n) \\ 0 & \text{else} \end{cases} \tag{6.33}$$

A last constraint is added to ensure no distance between elements from different trials is involved in the loss computation

$$\begin{cases} m_{p,r}^g = 0 & \text{if } t_p \neq t_r \\ m_{p,q}^s = 0 & \text{if } t_p \neq t_q \end{cases} \tag{6.34}$$

Provided a first fully connected layer of weights $\mathbf{W}^s \in \mathbb{R}^{d_{emb} \times d_{emb}}$ and biases $\mathbf{b}_s \in \mathbb{R}^{d_{emb}}$, the network computes a first margin tensor $\mathbf{A}_s$. Analogously, a second fully connected layer of weights $\mathbf{W}^g \in \mathbb{R}^{d_{emb} \times d_{emb}}$ and biases $\mathbf{b}^g \in \mathbb{R}^{d_{emb}}$ yields another margin tensor $\mathbf{A}^g$. The biases are tiled as $\mathbf{B}^s = [\mathbf{b}^s, ..., \mathbf{b}^s]$ and $\mathbf{B}^g = [\mathbf{b}^g, ..., \mathbf{b}^g]$ respectively which allows to compute margin tensors as follows

$$\mathbf{A}^s = (\mathbf{H}\mathbf{W}^s + \mathbf{B}^s)\mathbf{H}^T \tag{6.35}$$

$$\mathbf{A}^g = (\mathbf{H}\mathbf{W}^g + \mathbf{B}^g)\mathbf{H}^T \tag{6.36}$$

Then $\mathbf{A}_s$ and $\mathbf{A}_g$ are extended to a new dimension such as

$$\mathbf{A}^s = \mathbf{A}^s[:, :, newdim] \tag{6.37}$$

$$\mathbf{A}^g = \mathbf{A}^g[:, newdim, :] \tag{6.38}$$

The final margin tensor $\mathbf{A}$ is obtained by addition of $\mathbf{A}^s$ and $\mathbf{A}^g$. The margins in equation 6.22 are specific elements of $\mathbf{A} = \{\alpha_{p,q,r}\}_{p,q,r \in \mathbb{R}^{|\mathcal{B}_x^a|}}$ isolated by means of masking such that $\alpha_{b,n_i} = \alpha_{p_b,q_{n_i},r_w}$ and $\alpha_{w,n_i} = \alpha_{p_w,q_{n_i},r_b}$.

Data for the Experiment

As we did not test interactions between attitudes during the BWS experiment - e.g. question whether an utterance that have been meant to convey distance conveys friendliness, dominance or seductiveness - trainings must have been conducted on the four attitudes separately. As preliminary experiments, we tested various configurations for friendliness in order to validate the general approach outlined here.

For our perceptual arranger, we split the data trial wise as we expect to learn relations between sounds. In this perspective, a trial that has not been seen by the model is a new bunch of relations that can be more or less inferred depending on model's ability to generalize well. Almost 80% of the

4800 trials are kept for training while the rest is kept for validation. As showed in the second study of Chapter 4, there is a strong interaction between perceptual judgments and emotion conveyed by linguistic content, we thus paid a specific attention to including all sentences (100 different linguistic contents) in both training and validation sets. As mentioned above, batches are designed as follows: for each sound x within both sets we build a batch with all trials in which this sound lies. As a result we have a variable batch size that depends on the number of trials related to x. This number of trials in a batch is comprised between 4 and 8 for training set and between 1 and 4 for validation set. Why is this so? When designing a batch for training set from a sound x, some of its related trials could be attributed to the validation set. Conversely, when designing a batch for validation set from a sound x, some of its related trials could be attributed to the training set.

For our perceptual regressor, we split the data randomly while paying attention to including all sentences (100 different linguistic contents) in both training and validation sets. In fact here the notion of trial is no longer important since each utterance is associated with a score whose validity is absolute across all the data.

Configurations

We select several configurations of both models, each uses $n_{cnn} = 2$ convolutional layers with $d_{cnn} = 64$ filters, a temporal kernel $k_t = 5$ and a feature kernel $k_f = 3$. We chose the embedding size to be $d_{emb} = 32$. We tested a fixed-margin configuration and several learnt-margin ones, with and without dynamic margin (DMC) constraint featuring different functions γ and with or without a fulfilled relation (FR) loss.

Towards Trial Independent Performance Criterion

In order to assess how well these setups perform, we employ various objective measures. Within a trial we distinguish two types of distances, the distance between the best and worst that we expect to be the greatest in the trial, and the other distances that we expect to be smaller. To assess the model performance, we would like to design a general criterion that does not depend on the trial considered, i.e. a score that reflects how great and small are the different distances over the dataset. To obtain *great* and *small* distributions of distances across the dataset we normalize distances trial-by-trial. To do so, we define a custom mean trial distance as follows

$$\bar{d}_{t^a} = \frac{1}{2}\left[\max_{i\in[\![1,N-2]\!],j\in\{b,w\}}\left(\|\mathbf{h}^{t^a_j} - \mathbf{h}^{t^a_{n_i}}\|\right) + \|\mathbf{h}^{t^a_b} - \mathbf{h}^{t^a_w}\|\right] \tag{6.39}$$

Defined this way, the custom trial mean distance $\bar{d}_{t^a}$ is ensured to be greater than any *small* distance in a trial t^a and smaller than its *great* distance. We thus define $\mathcal{D}_g$ and $\mathcal{D}_s$ the trial-wise normalized distributions of *great* and *small* distances as

$$\mathcal{D}_g = \{\|\mathbf{h}^{t^a_b} - \mathbf{h}^{t^a_w}\| - \bar{d}_{t^a}\}_{t^a\in\mathcal{T}^a} \tag{6.40}$$

$$\mathcal{D}_s = \{\|\mathbf{h}^{t^a_j} - \mathbf{h}^{t^a_{n_i}}\| - \bar{d}_{t^a}\}_{i\in[\![1,N-2]\!],j\in\{b,w\}} \tag{6.41}$$

Metrics for Evaluation

For quantitative evaluation of our arranger's performance, we used two metrics, reflecting the arrangement of speech samples in the latent space at two levels:

- **Well-Arranged Trials (WAT)** : the percentage of trials within the set that are well arranged, i.e. in which all relations are fulfilled.
- **Fulfilled Relations (FR)** : the percentage of relations within the set that are fulfilled.

6.4.3 Results & Discussion

This proposal is first of all a theoretical contribution, a method allowing the learning of a latent space from raw BWS judgements, i.e. from relative judgements. It is also the subject of numerous more or less successful experiments. The results of these experiments are presented and discussed in this section.

Perceptual Arranger's Performance

The objective results of the perceptual arranger $\mathcal{A}^a$ are displayed in Table 6.5 for friendliness in terms of Well Arranged Trials (WAR) and Fulfilled Relations (FR) which respectively serve as strong and weak criterion for latent space structuring assessment.

Configuration	margin	λ_{dmc}	γ	λ_{fr}	WAT	FR
A-f	*fixed*	-	-	-	1.46%	20.9%
A-l	*learnt*	0	-	0	-	-
A-l-Re	*learnt*	1	$x \to ReLU(-x)$	0	5.6%	31.1%
A-l-Re-fr	*learnt*	1	$x \to ReLU(-x)$	1	**17.7%**	**47.6%**
A-l-Re2	*learnt*	1	$x \to ReLU(-x)^2$	0	12.9%	42.1%
A-l-Re2-fr	*learnt*	1	$x \to ReLU(-x)^2$	1	17.3%	44.6%

Table 6.5: Objective results of $\mathcal{A}^a$ in terms of well arranged trials (WAT) and fulfilled relations (FR) computed on the validation set for different configurations on friendliness.

First, we observe that the fixed-margin configuration **A-f** does not achieve to generalize at all, the validation loss does not decrease. As expected, as some trials' best and worst can be either very distant in the latent space or rather close. Let us now consider the different learnt-margin configurations. First, we observed for **A-l** that if no constraint is applied on margins - i.e. if $\lambda_{fr} = 0$ - all the points in the latent space are collapsing into one single point, thus the distances between any pair of points is null. In order to make the DM-RC loss decrease, the model has two strategies, it can both seeks to fulfill more relations within trials or it can diminish the margins. Note that this last strategy does not ensure any trial relation to be fulfilled.

Let us consider the learnt-margin configurations with the dynamic margin constraint (DMC) - i.e. $\lambda_{dmc} \neq 0$. We tested different functions γ for the DMC. We observe that applying the FR loss leads to improvements in both WAR and FR which supports the idea that adding this loss prevents the model from engaging in the second strategy of lowering margins generally. The best performance is obtained with **A-l-Re-fr** - that features a function $\gamma : x \to ReLU(-x)$ - with a WAR of 17.7% and a FR of 47.6%. The analog configuration **A-l-Re2-fr** - with function $\gamma : x \to ReLU(-x)^2$ - achieves slightly worse performance with a WAR of 17.3% and a FR of 44.6%. The figure 6.13 depicts the distances distributions $\mathcal{D}_s$ and $\mathcal{D}_g$ for the best configuration **A-l-Re-fr** on both training and validation sets, thus reflecting the generalization issue.

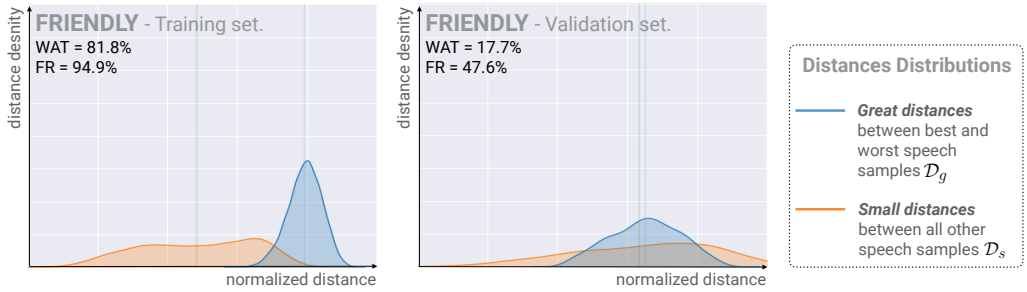

Figure 6.13: Distances distributions $\mathcal{D}_s$ and $\mathcal{D}_g$ of **A-I-Re-fr** configuration for both training (left) and validation (right) sets on friendliness.

Discussion

Two points can be made. Firstly, it appears that our model fails to generalize, on the validation set it satisfies about one out of two relations and barely one out of five trials is well ordered at the end of the training. However, if we examine the structure of the latent spaces - as depicted in Figure 6.14 through UMAP visualization - in relation to the BWS scores obtained, we can notice a certain order: the high scores are rather on one side while the low scores are rather on the other. The model thus seems to learn a certain pattern. At least, it seems to distinguish the top from the bottom of the BWS scale. Since the BWS scores are projections into a one-dimensional space of the raw judgements, it is logical that their structure appears when learning directly from the raw judgements.

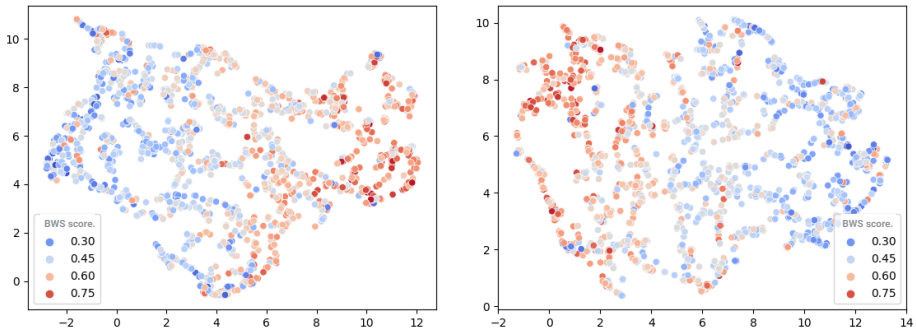

Figure 6.14: UMAP visualizations of the validation set's latent space for **A-I-Re** (left) and **A-I-Re-fr** (right), each point's color reflecting its BWS score for friendliness.

At the time of writing, we are working on a joint evaluation of these three BWS-Net proposals. While they appear to be complementary in some respects - detection of low and high scores - it is questionable which of these models is more suitable for conditioning an attitude conversion model or for the perceptual validation of converted utterances. Does the last proposal better represent the perceptual space that underlies the raw judgements made by the participants than does our regressor $\mathcal{R}^a$ or classifier $\mathcal{C}^a$? To answer this question, we could, for instance, apply the two metrics WAT and FR to the respective latent spaces of these models in order to be able to make comparison with what we have obtained for our perceptual arranger. The effectiveness of these different methods can also be assessed through their use in a voice conversion framework. At the time of writing, we are working on the integration of the perceptual classifier $\mathcal{C}^a$ into the furtherly

described - in Chapter 7 - speech attitude conversion algorithm. The idea is to add a parameter of attitudinal intensity which makes it possible to yield conversions conveying a more or less intense attitude. Using such a model to condition the conversion algorithm will hopefully allow to derive this intensity parameter from the perceptual data collected.

6.5 Chapter Summary

Before addressing the issue of perceptual regression/classification, we start by focusing on a more standard task of speech attitude recognition using the apriori attitude labels as ground truth. Although various Speech Emotion Recognition (SER) studies have already been conducted, leading for the emergence of a prototypical classification architecture - a convolutional block followed by a recurrent network, an attention mechanism and two fully connected layers - the recognition of speech attitude has not been specifically addressed in any studies, and even less so on a large multi-speaker dataset in French such as Att-HACK. Thus, an initial ablation study intends to highlight the relevance of the various blocks employed in regard to the algorithm's ability to predict vocal attitudes. To do so we tested different model's configurations and data splits. The results notably showed the importance of multi-head attention on the performance of the model but tended to moderate its impact when the evaluated utterance comes from a speaker who has not been seen by the model. In general, the performance of the model is much worse when evaluated on utterances coming from a speaker not seen by the model. This supports the idea that the production strategies of vocal attitudes are highly speaker dependent. In addition, the architecture validated here constitutes a starting point for our investigations with the aim of establishing a BWS-Net.

The rest of the Chapter marks a paradigm shift outlining three different proposals for the modelling of the perceptual latent spaces that underlie the decoding of speech attitudes by individuals. By mean of learning such space, we expect to have direct access to the individuals' mental representations of attitudes. To do so, we conducted experiments using the perceptual data collected during the BWS study we conducted on Att-HACK to understand human perception of vocal attitudes (Section 4.4). The collected data can be viewed in two ways: raw - we consider the judgements made by participants, i.e relations between sounds within a trial - or processed - we consider the BWS scores yielded by the post-processing algorithm described in (Louviere et al., 2015).

The first proposed method involves learning regression models on the BWS scores for each attitude from mel-spectrogram representations. The regressor has shown to significantly improve performance in regression when informed with speaker identities and linguistics. In particular, the two strong production profiles - namely seduction and dominance - uncovered in Section 4.3, are those for which the best results are obtained. We also observe that our regression models mostly predict average scores and fail to predict extreme ones, this drawback being likely due to the Gaussian distribution of actual scores. Without being able to enhance the overall performance of the regression, we attempted to counteract this effect by giving more weight during training to samples with extreme scores. At the end of this experiment, we hypothesized a difference in nature between the attitudes conveyed with average BWS scores and those associated with extreme scores.

In the light of these results, we changed approach by experimenting with the recognition of perceptual domains. We thus assumed that: samples with very low score are bad productions or poorly communicated attitude instances, samples with average score represent attitude's typical production and high-scoring samples non-typical production in the sense that other speech traits seem to drive the judgment of individuals. We then divided the BWS scales into four contiguous ranges and trained a classifier to predict which of these categories each sample was as-

signed to. We obtained a moderate overall performance however significant better for the extreme ranges, associated with very low and very high scores. Those results support the hypothesis of distinct perceptual domains co-existing within the BWS scales. Since only a sub-part of Att-HACK was judged by the participants, the obtained algorithm enables a minimal validation of the entire database through the detection of speech samples with poorly communicated attitude that are not of interest in our attitude conversion framework.

These first two proposals involved using the BWS scores as training material. These scores vary on four linear scales which can be seen as 1-D sub-spaces of the multi-dimensional space in which the perception of attitudes occurs. Therefore, compared to the actual judgment that was made, the BWS score may contain substantially less information about the perception of the attitude conveyed. In the light of this observation, we propose to learn a BWS-Net from the raw judgements directly by interpreting judgements within trials as distance relations between speech samples in a latent space in a metric learning fashion. The key point is that the relationships we aim to model are relative and not absolute, which entails refining standard metric learning approaches such as triplet loss (Hoffer and Ailon, 2015). We have therefore developed different variants of a cost function enabling such relative metric learning. the learning of a latent space metric that reflects those relative relations by constraining the motion of speech samples in the latent space during training. While we managed to learn the perceptual structure of participants' judgements for the data seen by the model, the model's performance remains modest for the unseen data with about one out of two fulfilled relations and barely one out of five well arranged trials at the end of the training. The examination of latent spaces through UMAP visualization revealed some structuring in relation to the BWS score on unseen data. This metric learning approach to BWS data is thus new, to our knowledge, and can be transposed to any other issue involving a BWS perceptual assessment.

At the time of writing, we are working on a joint evaluation of these three BWS-Net proposals. While they appear to be complementary in some respects - detection of low and high scores - it is questionable which of these models is more suitable for conditioning an attitude conversion model or for the perceptual validation of converted utterances.

Chapter 7

TRANSFORMER-BASED CONDITIONED VOICE CONVERSION

Contents

7.1 Related Work - Transformer-Based Voice Conversion 142
7.1.1 Voice Transformer Network's Architecture 142
7.1.2 Voice Transformer Network's Optimization 146
7.1.3 Limitations 147
7.2 Contribution - Speech Attitude Conversion 149
7.2.1 Reformulation in the Scope of Speech Attitude Conversion 149
7.2.2 Linguistic Conditioning of Speech Attitude Conversion 151
7.3 Speech Attitude Conversion Experiments 155
7.3.1 Many-to-Many Experiment for Speech Attitude Conversion 155
7.3.2 Linguistic Conditioning Experiment 156
7.3.3 Selected Configurations & Evaluation Process 157
7.3.4 Objective Evaluation 159
7.4 Perceptual Evaluation of Vocal Attitude Conversion Models 163
7.4.1 Perceptual Experiment 163
7.4.2 Results & Discussion 164
7.5 Chapter Summary 167

In this Chapter we propose an adaptation of the algorithm proposed in (Kameoka et al., 2021) for many-to-many speech attitude conversion based on mel-spectrogram representation of speech signal.

7.1 Related Work - Transformer-Based Voice Conversion

In this part, we introduce the transformer-based algorithm proposed in (Kameoka et al., 2021) for speaker identity conversion. We present its architecture, detailing the role and operation of each of its constituent parts and how the algorithm is optimised. Finally, based on initial experiments, we point out its limitations in the specific case of voice attitude conversion.

7.1.1 Voice Transformer Network's Architecture

The architecture proposed in (Kameoka et al., 2021) is depicted in Figure 7.1. The only notable difference with (Kameoka et al., 2021) is that it takes mel-spectrograms as inputs instead of WORLD features (Morise et al., 2016). In this part, we provide description for each of this architecture's components and explain their role with respect to speech attitude conversion.

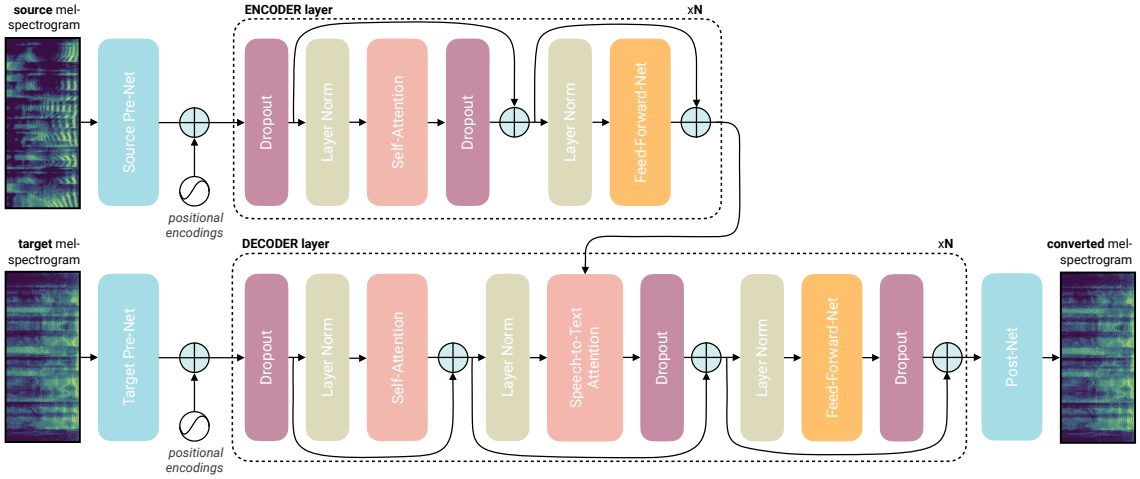

Figure 7.1: Schematic view of the Voice Transformer Network Neural Architecture according to (Kameoka et al., 2021) with mel-spectrograms as input data instead of WORLD features.

Model Inputs

In this (Kameoka et al., 2021), WOLRD vocoder is used to analyse speech signals and extract representations combining the mel-cepstral coefficients (MCCs) (spectral enveloppe), log F0, aperiodicity, and voiced/unvoiced indicator of speech all stacked into a 2-D representation. The WORLD vocoder is also used for the synthesis of converted representations. Denoting $\mathcal{A}_{world}$ the WORLD analyzer, inputs are obtained from source and target signals $\mathbf{x}^s$ and $\mathbf{x}^t$ as $\mathbf{X}^s = \mathcal{A}_{world}(\mathbf{x}^s)$ and $\mathbf{X}^t = \mathcal{A}_{world}(\mathbf{x}^t)$. We do not detail parameters for such extraction as WORLD is not used in the experiments.

Source and Target Pre-Nets

Source and target representations are passed through their respective Pre-Net so as to extract temporal features that make sense for attitude conversion. Those networks are formed by n_{cnn} convolutional blocks. Each block is composed with a 1D convolutional layer proceeding on time dimension followed by a batch normalization and an activation.

CONVOLUTIONAL LAYERS. The convolutions feature d_{cnn}-dimensional output filters, kernel of size k for time dimension, no strides are used but a dilation factor δ which allows to increase the layer's receptive field thus capturing wider context. The padding is set to *same* for the source pre-Net while it is causal for the target pre-Net.

Since time must be preserved throughout the entire network, those modules do not alter the temporal dimension. Fed with source and target mels $\mathbf{X}^s$ and $\mathbf{X}^t$ the pre-Nets would yield $\tilde{\mathbf{X}}^s \in \mathbb{R}^{T_s \times d_{cnn}}$ and $\tilde{\mathbf{X}}^t \in \mathbb{R}^{T_t \times d_{cnn}}$ respectively.

POSITIONAL ENCODINGS. As the model transformer sees each frame independently, it does not have any sense of the order of the elements in a temporal sequence. To provide it with this information we use positional encodings that is a position-dependent signal that helps the model incorporate the order of frames (Vaswani et al., 2017). Thus, once obtained, high level representations $\tilde{\mathbf{X}}^s$ and $\tilde{\mathbf{X}}^t$ are added with tensors of positional encodings $\mathbf{P}^s$ and $\mathbf{P}^t$. Positional encodings can be either learnt or fixed, here we use sinusoidal encodings $\mathbf{P} = \{p_{i,j}\}$ defined as follows

$$p_{i,j} = \begin{cases} \sin\left(\frac{1}{10000^{\frac{2j}{d_{cnn}}}}\right) & \text{if } p \text{ is pair} \\ \cos\left(\frac{1}{10000^{\frac{2j}{d_{cnn}}}}\right) & \text{else} \end{cases} \tag{7.1}$$

Transformer Encoder

The transformer encoder takes $\tilde{\mathbf{X}}^s$ as input and produces a context vector sequence $\mathbf{Z}^s \in \mathbb{R}^{T_s \times d_{model}}$ which expected to contain the linguistic content that lies in the source utterance. The transformer encoder is made of L_{enc} identical blocks. Each block is formed by self-attention (SA) and position-wise fully connected feed forward network (FFN) layers. Residual connections and layer normalizations are applied in addition to those two layers. In the following, we detail each of these elements.

MULTI-HEAD SELF ATTENTION (MHSA). Given an input tensor $\mathbf{X} \in \mathbb{R}^{T \times d_{model}}$ the MHSA layer outputs a tensor $\mathbf{Y} \in \mathbb{R}^{T \times d_{model}}$. Denoting H the MHSA's number of heads, the input tensor is projected in $3H$ different sub spaces. Thus for each head $h \in \{1, ..., H\}$, we compute a query $\mathbf{Q}_h$, a key $\mathbf{K}_h$ and a value $\mathbf{V}_h$ such as

$$\mathbf{Q}_h = \mathbf{X}\mathbf{W}_{h,q} \tag{7.2}$$
$$\mathbf{K}_h = \mathbf{X}\mathbf{W}_{h,k} \tag{7.3}$$
$$\mathbf{V}_h = \mathbf{X}\mathbf{W}_{h,v} \tag{7.4}$$

Those projections are obtained via fully connected layers with learnable weights $\mathbf{W}_{h,q} \in \mathbb{R}^{d_{model} \times d_{att}/H}$, $\mathbf{W}_{h,k} \in \mathbb{R}^{d_{model} \times d_{att}/H}$ and $\mathbf{W}_{h,v} \in \mathbb{R}^{d_{model} \times d_{att}/H}$. For each head $h \in \{1, ..., H\}$, a tensor of self attention - measuring the similarity of each frame to each other in the input sequence $\mathbf{X}$ with respect to the context observed by the considered head h - can be computed as follows

$$A_h = softmax\left(\frac{\mathbf{Q}_h \mathbf{K}_h^T}{\sqrt{d_{model}}}\right) \tag{7.5}$$

The attention tensors from different heads are multiplied through their related value tensor and concatenated along the features dimension. Thus, as depicted in Figure 7.3, the MHSA layer can be seen as multiple Self Attention (SA) layers stacked and focusing on different aspects of the input tensor. Finally, the resulting tensor $[\mathbf{A}_1\mathbf{V}_1, ..., \mathbf{A}_H\mathbf{V}_H]$ is being passed through a last fully connected layer with learnable weights $\mathbf{W} \in \mathbb{R}^{d_{att} \times d_{model}}$ thus yielding the output tensor $\mathbf{Y}$

$$\mathbf{Y} = [A_1V_1, ..., A_HV_h]\mathbf{W} \tag{7.6}$$

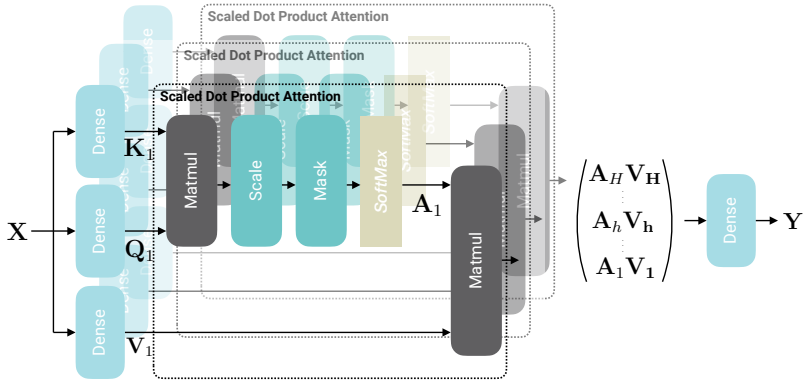

Figure 7.2: Schematic view of the Multi-Head Self Attention layer fed with input tensor $\mathbf{X}$

Position-wise Feed Forward Network (FFN). Given an input tensor $\mathbf{X}$, the FFN is formed with two fully connected layers of trainable weights $\mathbf{W}^1 \in \mathbb{R}^{d_{model} \times d_{ffn}}$, $\mathbf{W}^2 \in \mathbb{R}^{d_{ffn} \times d_{model}}$ and biases $\mathbf{b}^1 \in \mathbb{R}^{d_{ffn}}$, $\mathbf{b}^2 \in \mathbb{R}^{d_{model}}$. Thus, tiling biases tensors such as $\mathbf{B}^1 = [\mathbf{b}^1, ..., \mathbf{b}^1] \in \mathbb{R}^{T \times d_{ffn}}$ and $\mathbf{B}^2 = [\mathbf{b}^2, ..., \mathbf{b}^2] \in \mathbb{R}^{T \times d_{model}}$ the output tensor $\mathbf{Y}$ can be computed as follows

$$\mathbf{Y} = \phi(\mathbf{X}\mathbf{W}^1 + \mathbf{B}^1)\mathbf{W}^2 + \mathbf{B}^2 \tag{7.7}$$

In Equation 7.7, ϕ denotes an element-wise nonlinear activation function such as the rectified linear unit (ReLU) or gated linear unit (GLU) functions.

Layer Normalization (LN). Previous research has suggested that the position of the layer normalization in the transformer architecture has an impact on the training process' speed and stability as well as the performance of the trained model. Layer normalization is placed after the SA and FFN layers in the original transformer architecture, while the architectures shown in (Wang et al., 2019; Xiong et al., 2020) and the baseline (Kameoka et al., 2021) are designed with LN before those layers. Though, we decided to do the same, as depicted in Figure 7.1.

Let us taking a step back and consider the l^{th} encoder layer Enc_l composed with a SA layer SA_l, two normalization layers $LN_{1,l}$ and $LN_{2,l}$ and a FFN layer FFN_l. If we denote $\mathbf{X}^l$ and $\mathbf{X}^{l+1}$ the input and output of the l^{th} encoder layer, the process $\mathbf{X}^{l+1} = Enc_l(\mathbf{X}^l)$ is given by

$$\mathbf{U}^l = \mathbf{X}^l + SA_l(LN_{1,l}(\mathbf{X}^l)) \tag{7.8}$$

$$\mathbf{X}^{l+1} = \mathbf{U}^l + FFN_l(LN_{2,l}(\mathbf{U}^l)) \tag{7.9}$$

Transformer Decoder

The decoder takes $\mathbf{Z}^s$ and $\mathbf{X}^t$ as the inputs and produces a converted feature sequence $\mathbf{Y}^{s\leftarrow t} = [Y_1^{s\leftarrow t}, ..., Y_{T_t}^{s\leftarrow t}] \in \mathbb{R}^{T_t \times d_{model}}$. Similar to the encoder, the decoder consists of L_{dec} identical blocks, each of which has SA and FFN layers, residual connections and layer normalization layers. In addition to these layers, each block has a multi-head target-to-source attention (MHTSA) layer as depicted in Fig. 3, whose role is to find which frame in the source melspectrogram contextually corresponds to each frame in the target melspectrogram and convert the context vector sequence according to the predicted corresponding temporal positions. All the layers employed in the decoder have already been presented except for the attention layer between the decoder and the encoder, the MHTSA layer, which is therefore introduced in the following.

MULTI-HEAD TARGET-TO-SOURCE ATTENTION (MHTSA). Given a source input tensor $\mathbf{Z} \in \mathbb{R}^{T_s \times d_{model}}$ and a target tensor $\mathbf{X} \in \mathbb{R}^{T_t \times d_{model}}$ the MHTSA layer outputs a tensor $\mathbf{Y} \in \mathbb{R}^{T_t \times d_{model}}$. Denoting H the MHSA's number of heads, the source input tensor $\mathbf{Z}$ is projected in $2H$ different sub spaces while the target input tensor $\mathbf{X}$ is projected in H different sub spaces. Thus for each head $h \in \{1, ..., H\}$, we compute a query $\mathbf{Q}_h^{ts}$, a key $\mathbf{K}_h^{ts}$ and a value $\mathbf{V}_h^{ts}$ such as

$$\mathbf{Q}_h^{ts} = \mathbf{X}\mathbf{W}_{h,q}^{ts} \tag{7.10}$$
$$\mathbf{K}_h^{ts} = \mathbf{Z}\mathbf{W}_{h,k}^{ts} \tag{7.11}$$
$$\mathbf{V}_h^{ts} = \mathbf{Z}\mathbf{W}_{h,v}^{ts} \tag{7.12}$$

Those projections are obtained via fully connected layers with learnable weights $\mathbf{W}_{h,q}^{ts} \in \mathbb{R}^{d_{model} \times d_{att}/H}$, $\mathbf{W}_{h,k}^{ts} \in \mathbb{R}^{d_{model} \times d_{att}/H}$ and $\mathbf{W}_{h,v}^{ts} \in \mathbb{R}^{d_{model} \times d_{att}/H}$. For each head $h \in \{1, ..., H\}$, a tensor of source-to-target attention - measuring the similarity of each source frame of $\mathbf{Z}$ to each target frame of $\mathbf{X}$ with respect to the context observed by the considered head h - can be computed as follows

$$A_h^{ts} = softmax\left(\frac{\mathbf{Q}_h^{ts}\mathbf{K}_h^{ts^T}}{\sqrt{d_{att}}}\right) \tag{7.13}$$

Each produced tensor $\mathbf{V}_h^{ts}\mathbf{A}_h^{ts}$ can be understood as a time-warped version of $\mathbf{V}_h^{ts}$ with regards to the context observed by head h. All these time-warped feature sequences from different heads are concatenated along the features dimension to produce a resulting tensor $[\mathbf{A}_1^{ts}\mathbf{V}_1^{ts}, ..., \mathbf{A}_H^{ts}\mathbf{V}_H^{ts}]$ which is then passed through a last fully connected layer with learnable weights $\mathbf{W}^{ts} \in \mathbb{R}^{d_{att} \times d_{model}}$ thus yielding the output tensor $\mathbf{Y}$ as depicted in Figure 7.3

$$\mathbf{Y} = [A_1^{ts}V_1^{ts}, ..., A_H^{ts}V_h^{ts}]\mathbf{W}^{ts} \tag{7.14}$$

Let us take a step back and consider the l^{th} decoder layer Dec_l composed with a MHSA layer SA_l, three normalization layers $LN_{1,l}$, $LN_{2,l}$ and $LN_{3,l}$, a FFN layer FFN_l and a MHTSA layer TSA_l. If we denote $\mathbf{X}^l$ and $\mathbf{X}^{l+1}$ the input and output of the l^{th} decoder layer and $\mathbf{Z}^s$ the output of the encoder, the process $\mathbf{X}^{l+1} = Dec_l(\mathbf{X}^l)$ is given by

$$\mathbf{U}^{1,l} = \mathbf{X}^l + SA_l(LN_{1,l}(\mathbf{X}^l)) \tag{7.15}$$
$$\mathbf{U}^{2,l} = \mathbf{U}^{1,l} + TSA(LN_{2,l}(\mathbf{U}^{1,l}, \mathbf{Z}^s)) \tag{7.16}$$
$$\mathbf{X}^{l+1} = \mathbf{U}^{2,l} + FFN_l(LN_{3,l}(\mathbf{U}^{2,l})) \tag{7.17}$$

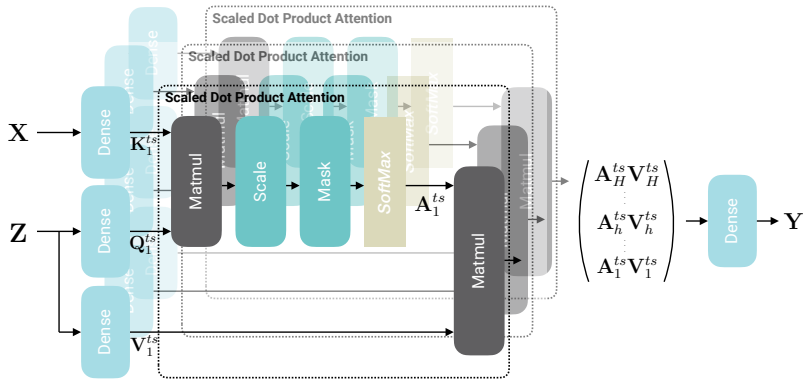

Figure 7.3: Schematic view of the Multi-Head Source-to-Target Attention layer fed with source input tensor $\mathbf{Z}$ and target input tensor $\mathbf{X}$

7.1.2 Voice Transformer Network's Optimization

At the core of transformer model lies an autoregressive structure which is here implemented as a constraint on target-to-source attention matrices. Furthermore, the model in its baseline configuration exposed in (Kameoka et al., 2021) is optimised with respect to two cost functions, a reconstruction loss that ensures the converted representation matches the target one and another loss that constrains the structure of target-to-source attention matrices.

Auto-regressive Structure

At inference phase, we want to be able to produce a conversion with only the source sentence available. To do this, an autoregressive structure is introduced into the model. The feature vector corresponding to the first time frame is produced from the source and a null vector. Subsequently, the feature vector associated with time t is produced from the source and output of the decoder at times $t < 1$. In order to enable this behavior throughout the model, we must first ensure that the decoder is not enabled to use future information about the target feature vectors while creating an output vector at each time step. This can be ensured by simply constraining the convolution layers in the target pre-Net to be causal and each self attention in the decoder to be triangular matrices. This latter constraint can be fulfilled by replacing Eq. 7.5 in all the SA layers of the decoder with

$$A_h = softmax\left(\frac{\mathbf{Q}_h\mathbf{K}_h^T}{\sqrt{d_{model}}} + \mathbf{E}\right) \tag{7.18}$$

where $\mathbf{E} = \{e_{i,j}\}_{i,j\in[\![1,T_t]\!]}$ such as

$$e_{i,j} = \begin{cases} 0 & \text{if } i \leq j \\ -\infty & \text{if } i > j \end{cases} \tag{7.19}$$

The negative values of $\mathbf{E}$ passed in the softmax are turned into zeros which make of $\mathbf{A}_h$ a triangular matrix. Meeting such constraint, the predictions for time step t can depend only on the known outputs at past instants t' such as $t' < t$.

Reconstruction Loss

So as for the model to have an autogressive structure, the output sequence $\mathbf{Y}^{s \leftarrow t}$ must correspond to the target sequence $\mathbf{X}^t$ but time-shifted by one sample, i.e. the prediction of frame t is tensorally aligned with frame $t-1$. To meet this requirement, the target mel-spectrogram $\mathbf{X}^t$ is concatenated with a null vector of size D. This way, at each time step the decoder has full access to all past information and is enabled to predict the next time step.

Thus, the model is trained with respect to a L1 reconstruction loss formulated as follows

$$\mathcal{L}_{rec} = \|[\mathbf{Y}^{s \leftarrow t}]_{1:T_t-1,:} - [\mathbf{X}^t]_{2:T_t,:}\|_1 \tag{7.20}$$

with $[\mathbf{X}]t : t', :$ the slice of tensor $\mathbf{X}$ that goes from time step t up to time step t'.

Diagonal Attention Loss (DAL)

Through attention matrices yielded by TSA layers, the tranformer model learns an implicit alignment between source and target utterances. Therefore, it is then reasonable to hypothesise that these matrices must have some structure which ensures that a given temporal segment, for instance a phoneme, in the source sentence is mapped to its counterpart in the target sentence. We will therefore try to impose a monotonic and quasi-linear diagonal on these matrices so that the learned alignment is not a basic time stretch - i.e. linear diagonal matrices - but a mapping between dilated and compressed temporal regions. Introduced in (Tachibana et al., 2018), a diagonal attention loss (DAL) is used to penalize the attention matrices for not having a diagonally dominant structure, it is formulated as follows

$$\mathcal{L}_{da} = \frac{1}{T_s T_t L_{dec} H} \sum_{l=1}^{L_{dec}} \sum_{h=1}^{H} \|\mathbf{G}_{T_s \times T_t} \odot \mathbf{A}_{l,h}\|_1 \tag{7.21}$$

where $\mathbf{A}_{l,h}$ denotes the target-to-source attention matrix of the head h in the TSA layer in the layer l of the decoder, $\odot$ denotes element-wise product, and $\mathbf{G}_{T_s \times T_t} \in \mathbb{R}^{T_s \times T_t}$ is a non-negative weight matrix whose elements $g_{n,m}$ are defined as

$$g_{i,j} = 1 - e^{-(\frac{i}{T_s} - \frac{j}{T_t})^2 / 2\nu^2} \tag{7.22}$$

7.1.3 Limitations

In this section, we point out two limitations, the first one deals with the use of the WORLD representation and vocoder in the context of speech conversion. This limitation is mainly conceptual and discussed in the light of many research works. The second limitation appeared to us as a result of preliminary experiments of speech attitude conversion based on the transform architecture presented above and deals with intelligibility of the conversions yielded.

Multi-Parametric Modelling of Speech Signals

The first limitation we point out here is the use of features associated with the WORLD vocoder in (Kameoka et al., 2021). For many years no vocoder was able to synthesize speech in a natural way, i.e. the human ear was then able to distinguish between a synthesized speech sample and an authentic one. The first vocoder to break this rule was the one proposed in (Shen et al., 2018) based on the conditioning of a WaveNet on mel-spectrogram representations of speech signals. In

addition, the superiority of the WaveNet over the WORLD vocoder was established in (Wang et al., 2018a) through a large-scale perceptual test. From the synthesis perspective, it appears that the mel-spectrogram representation is the most appropriate as it allows for high quality recovering of speech signals.

However, the use of mel-spectrogram representation in the specific context of speech conversion must also be considered. Since the optimization of the conversion algorithm is performed independently of the re-synthesis of the speech signal in most of voice conversion works, the importance of each component of the representation - in the conversion learning - must be weighted according to the impact it has on the perception of the converted utterance. In particular, in the case of conversion learning from a multi-parametric representation, one must ask how to weight the reconstruction errors of the different parameters (F0, energy and spectral envelope). One should also consider how to manage the F0 in the invoiced temporal segments of speech. In other words, this amounts to adding hyper-parameters to the conversion algorithm and to multiplying tenfold the number of configurations to be tested in order to find the best one. Conversely, the mel-spectrogram representation fundamentally differs from multi-parametric representations such as employed in WORLD vocoder in that each of its components represents the same type of information and are all equally relevant in terms of perception, in particular there is no conceptual difference between voiced and unvoiced speech temporal segments. The mel-spectrogram representation is both homogeneous and compact, which enables its direct use by a conversion algorithm without the need for any weighting in the reconstruction of its various components.

In view of this, many works in voice conversion have used the mel-spectrogram as a representation of the speech signal (Zhang et al., 2020; Qian et al., 2019, 2020; Bous et al., 2022; Bous and Roebel, 2022), we thus propose to adapt the architecture in (Kameoka et al., 2021) so that it takes mel-spectrogams as inputs for conversion learning.

Loss of Intelligibility

Early experiments of speech attitude conversion with the transformer architecture led us to identify an important issue that is indistinguishable through the standard conversion monitoring based on error and accuracy measures. Indeed, we found that in a significant number of cases, the conversions did not preserve the linguistic content of the source - some phonemes were badly formed or not formed at all. There were also cases where a word was changed into another word thus completely changing the meaning of the utterance. This issue appeared to be very disturbing in that it compromises the optimal decoding of the message conveyed by the converted utterance. Facing this issue, we intended to establish a hierarchy between the criteria that conversions generated by our model must meet. We especially place intelligibility - i.e. the ability to decode the linguistic message conveyed in a speech signal - at the top of this hierarchy, the faithfulness of the attitude conveyed to the attitude actually targeted coming in second place. Indeed, we assume that it is fundamental that the linguistic message be preserved even if it is at the expense of the attitude conveyed in the conversion which thus might be either unchanged or badly converted. Strictly speaking, there can be no conversion of attitude if the linguistic message is corrupted. In the following, we will try to quantify this loss of intelligibility and provide a solution.

7.2 Contribution - Speech Attitude Conversion

In this section, we first start by reformulating the above presented Transformer architecture in the scope of speech attitude conversion based on mel-spectrogram representation. Second we attempt to address the main problem encountered in the conversions obtained during early experiments of transformer-based attitude conversion: the occasional loss of linguistic content.

7.2.1 Reformulation in the Scope of Speech Attitude Conversion

The transposition of the proposal in (Kameoka et al., 2021) to our speech attitude conversion issue necessitates a reformulation. Moreover, as we propose to use mel-spectrograms, so we need to specify how they are obtained.

Mel-spectrograms as Model's Inputs

Unlike the proposal by Kameoka et al. (Kameoka et al., 2021) which uses WORLD vocoder features as representation of the speech signal, we chose to learn conversion between mel-spectrograms. Orginal speech signals can be retrieved from mel-spectrograms through using the team-made neural vocoder (Roebel and Bous, 2022). Denoting $\mathcal{A}_{mel}$ the melspectrogram extractor, inputs are obtained from source and target signals $\mathbf{x}^s$ and $\mathbf{x}^t$ as $\mathbf{X}^s = \mathcal{A}_{mel}(\mathbf{x}^s)$ and $\mathbf{X}^t = \mathcal{A}_{mel}(\mathbf{x}^t)$.

Parameters. The melspectrograms are obtained through computing Short-Term-Fourier-Transform (STFT) of parameters N_{ft}, R_{ft} and M_{ft} corresponding to the size of the FFT, the hop and window sizes. The number of mel channels is set to D. Source and target melspectrograms are then padded batch-wise up to $T_{s,b}$ and $T_{t,b}$ respectively the lengths of longest source and target utterances in the batch. This way, computation can be performed in an tensorial way. In the following we will use T_s and T_t for the sake of clarity.

Many-to-Many Speech Attitude Conversion

Here we propose to transpose the many-to-many paradigm, applied to speaker identity in (Kameoka et al., 2021), to vocal attitude. Let us clarify. Speaker identity conversion is learned across several speakers in (Kameoka et al., 2021) approach. That is, the same network can convert a sentence spoken by a speaker A to a speaker B and a sentence spoken by a speaker B to a speaker C. In a way that the network learns to replace a speaker's identity by capturing what their identity is generally rather than just learning to convert one specific identity into another. In contrast to a straightforward mapping of speech features, it can be claimed that the model learns an implicit definition of speaker identification. This interpretation must be tempered by the fact that the model only learns to convert 4 speakers, therefore it is impossible to pretend to any kind of generalization. Practically speaking, this approach saves time because only one training session is needed. In many regards, the many-to-many paradigm is interesting.

Similar to this, we propose employing the four pre-existing Att-HACK categories of friendliness, distance, dominance, and seduction to learn to convert the vocal attitude in general rather than just learning to convert from one attitude to another. Since it seems ambitious enough to learn such a model, we limit the problem to one speaker.

To do so, we assign each attitude a number a and consider paired utterances $\{\mathbf{X}^{s,a}, \mathbf{X}^{t,a'}\}$ with the same linguistic content but produced according to different attitudes a and a'. We use one-hot-encoding representation of the attitude. Having n_{att} attitudes in the dataset, the a^{th} attitude is represented by a vector $\mathbf{s}^a$ of size n_{att} with a one at rank a and zeros anywhere else. To provide the model with attitudinal information, each attitude one-hot-encoding is temporally tiled so that it matches the shape of its corresponding melspectrogram thus yielding a tensor $\mathbf{S}^a = [\mathbf{s}^a, ..., \mathbf{s}^a] \in \mathbb{R}^{T \times n_{att}}$. Each sub layer in the network, namely SA, TSA and FFN, as well as the pre-Nets and the post-Net are provided with attitudinal information through concatenation of their inputs with tiled one-hot attitude representations. Then, if the model is fed with a pair $\{\mathbf{X}^{s,a}, \mathbf{X}^{t,a'}\}$, thus inputs of any given sub layer, $\mathbf{X} \in \mathbb{R}^{T_s \times d_{model}}$ related to source and $\mathbf{Y} \in \mathbb{R}^{T_t \times d_{model}}$ related to target, are modified this way

$$\mathbf{X}^a = concat(\mathbf{X}, \mathbf{S}^a) \in \mathbb{R}^{T_s \times (D + n_{sp})} \tag{7.23}$$

$$\mathbf{Y}^{a'} = concat(\mathbf{Y}, \mathbf{S}^{a'}) \in \mathbb{R}^{T_t \times (D + n_{sp})} \tag{7.24}$$

$$\tag{7.25}$$

where $concat(.,.)$ denotes the concatenation of two tensors along features axis. The loss functions for such a many-to-many paradigm become

$$\mathcal{L}_{rec}^{(a,a')} = \|[\mathbf{Y}^{s,a \leftarrow t,a'}]_{1:T_t-1,:} - [\mathbf{X}^{t,a'}]_{2:T_t,:}\|_1 \tag{7.26}$$

$$\mathcal{L}_{da}^{(a,a')} = \frac{1}{T_s T_t L_{dec} H} \sum_{l=1}^{L_{dec}} \sum_{h=1}^{H} \|\mathbf{G}_{T_s \times T_t} \odot \mathbf{A}_{l,h}^{(a,a')}\|_1 \tag{7.27}$$

where $\mathbf{A}_{l,h}^{a,a'}$ denotes the target-to-source attention matrix yielded by head h of the TSA layer related to layer l of the decoder. Denoting λ_{da} the weight for controlling the influence of DA loss on training process, the full model loss can be formulated as follows

$$\mathcal{L}_{vtn}^{(a,a')} = L_{rec}^{(a,a')} + \lambda_{da} \mathcal{L}_{da}^{(a,a')} \tag{7.28}$$

So as to force the model not to alter speaker identity while source and target are from the same speaker, (Kameoka et al., 2021) introduced an identity mapping (IM) loss which is equal to $\mathcal{L}_{vtn}^{(a,a)}$. Therefore the total training loss including IM loss is

$$\mathcal{L}_{vtn} = \sum_{a,a' \neq k} \mathbb{E}_{\mathbf{X}^{s,a}, \mathbf{X}^{t,a'}} \{L_{vtn}^{(a,a')}\} + \sum_{a} \mathbb{E}_{\mathbf{X}^{(s,a)}, \mathbf{X}^{t,a}} \{\mathcal{L}_{vtn}^{(a,a)}\} \tag{7.29}$$

Speech Attitude Conversion Process

At inference phase, a source speech mel-spectrogram $\mathbf{X}^s$ can be converted to the target attitude via the following recursion

$\mathbf{Z} \leftarrow \mathbf{X}^t, \mathbf{Y} \leftarrow [0, ..., 0] \in \mathbb{R}^D$
for $l = 1$ to L_{enc} **do**
  $\mathbf{Z} = Enc_l(\mathbf{Z})$
end for
for $m = 1$ to T_t **do**
  for $l = 1$ to L_{dec} **do**
    $\mathbf{Y} \leftarrow Dec_l(\mathbf{Y}, \mathbf{Z})$
  end for
  $\mathbf{Y} \leftarrow concat([[0, ..., 0], \mathbf{Y}])$
$\mathbf{Y}^{s \leftarrow t} \leftarrow \mathbf{Y}$

where $concat$ denotes the concatenation along the temporal axis. Once converted mel-spectrogram $\mathbf{Y}^{s \leftarrow t}$ is obtained, we pass it to the neural vocoder to yield the converted speech signal such as

$$\mathbf{y}^{s \leftarrow t} = \mathcal{R}_{\text{MBExWN}}(\mathbf{Y}^{s \leftarrow t}) \tag{7.30}$$

7.2.2 Linguistic Conditioning of Speech Attitude Conversion

To tackle the issue of linguistic loss encountered in a significant number of conversions, we moved to the typical research problem of speech recognition. Automatic speech recognition (ASR) consists of transcribing audio speech segments into text. This task can be viewed as a sequence-to-sequence problem, where the audio can be represented as a sequence of feature vectors and the text as a sequence of characters, words, or subword tokens. In this case, the speech recognizer - or speech-to-text - module takes mel-spectrogram representations of speech signals as input. Its objective is to predict what is being pronounced within the input signal, i.e. the character sequence related to the linguistic message conveyed.

First, the application of such an ASR module to the conversions resulting from the models outlined in the previous section would allow for an objective evaluation of this problem of loss of linguistic content. Despite the inherent prediction error of the module, such an application would allow us to determine how much the linguistic content and, consequently, the meaning of the conversions produced by our models differ from the one of target sentences they are intended to replicate. Therefore, we first worked at implementing an efficient speech-to-text module and trained it on Att-HACK.

Second, we considered how we can use such an ASR module to enhance the performance of our conversion model. Our overall plan was to incorporate it into the conversion system and propagate its prediction error through the conversion model's layers to ensure that the transformed utterances' linguistic content was preserved. Since there are various ways to put this concept into practice, we have conducted several experiments, which are discussed in the following part.

Speech Recognizer Architecture

In this part, widely inspired by (Pham et al., 2019), the speech recognizer - or speech-to-text - architecture is presented. Most components are not examined in details as they are also used in the

voice conversion transformer exposed in the last section. However, we will focus on the part of the model that supports textual data, i.e character sequences, that have yet not been introduced in this document.

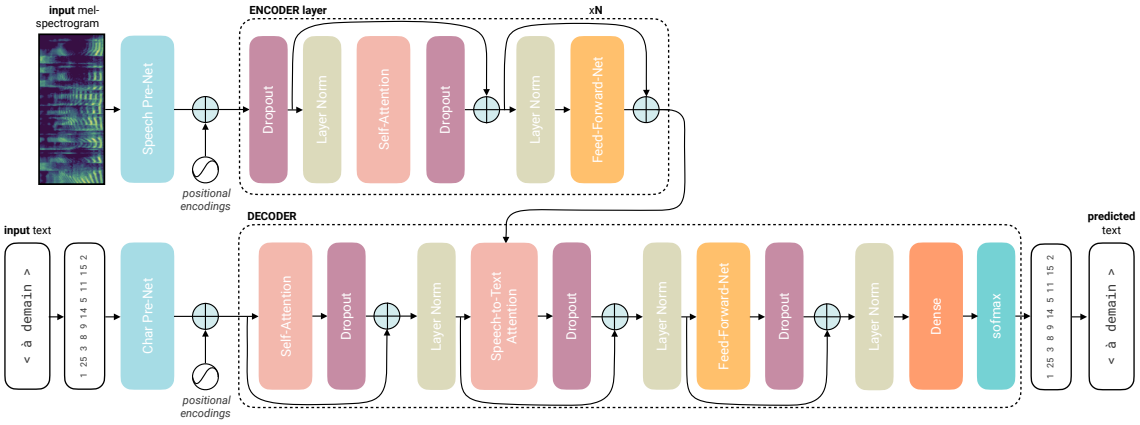

Figure 7.4: Schematic view of the Speech-to-Text Transformer Network neural architecture

MODEL INPUTS. As said before, the speech-to-text module has two types of input, firstly a representation of the speech signal, namely a mel-spectrogram computed in the same way than exposed in the last section. Secondly, the textual transcription of what is said in this signal. Those transcriptions are given to the network as sequences of characters which are encoded into one-hot-vectors.

Denoting $\mathcal{S}_{char} = \{c_1, ..., c_{n_{char}}\}$ the set of characters involved in the sentences of Att-HACK, then any sentence x - i.e. linguistic content - in Att-HACK of length - i.e. number of character - n_x, can be represented by a sequence of integers $\mathbf{y} = [y_1, ..., y_{n_x}]$ such as $y_i \in \mathcal{S}_{char}$ for all $i \in \{1, n_x\}$. Then the one-hot-encoding of y is the sequence of vectors $\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_{n_x}]$ with $\mathbf{z}_i = [z_{i,1}, ..., z_{i,n_{char}}]$ such as

$$z_{i,j} = \begin{cases} 1 & \text{if } j = y_i \\ 0 & \text{else} \end{cases} \tag{7.31}$$

SPEECH PRE-NETWORK. The speech representations - mel-spectrograms - are passed to a speech pre-Net which is exactly the same than the ones used for source and target pre-encoding in the voice conversion system descibed in the last section. Then the yielded tensors are added to positional encodings computed as described in Eq. 7.1 in the last section.

SPEECH ENCODER. The speech encoder is formed by n_{enc} standard transformer blocks, each composed of a dropout layer, a layer normalization, a multi-head self attention layer, a dropout layer after the layer's input is added (residual connections), a layer normalization and finally a feed-forward network. Since they were thoroughly covered in the previous section, the role and functioning of each of these components will not be explained here. Let us only note that this speech encoder allows to produce a representation of the speech signal in which the linguistic content is highlighted and which furtherly used by the decoder to predict the actual sequence of characters.

CHARACTER PRE-NETWORK. The one-hot-representations of linguistic content are passed to a pre-network, which consists of a fully connected layer, and whose role is to learn a linguistic space

formed by the projections of these character sequences, the idea is to adapt the linguistic representation for the specific objective of the speech-to-text module. Yielded tensors are then being added to positional encodings. Note that here the positional encodings are learnt by means of a fully connected layer and not computed as described in the last section, this choice leading to slightly better performance in prediction.

DECODER. The decoder is divided in two parts. A first part processes the characters embedding sequences through a multi-head self attention layer, a dropout layer after which the input is added, finally a layer normalization is performed. The yielded representation of the text-input is passed to a multi-head speech-to-text attention along with the representation produced by the encoder. The outputed tensor is then passed to standard dropout, layer normalizaton and feed forward layers. In the end, the tensor is passed to a fully connected layer followed by a softmax activation that both perform character classification properly.

To train this model, we use a standard categorical cross entropy as cost function. Therefore, for an utterance x of character sequence encoded as $\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_{n_x}]$ with $\mathbf{z}_i = [z_{i,1}, ..., z_{i,n_{char}}]$, the loss can formulated as follows

$$\mathcal{L}_{asr} = -\sum_{i=1}^{n_x} \sum_{j=1}^{n_{char}} z_{i,j} \log(\hat{z}_{i,j}) \tag{7.32}$$

Overall Text-Conditioned Voice Conversion Algorithm

Once we had a module for predicting text from mel-spectrogram, we wondered how to use it to preserve intelligibility when converting attitudes in speech. In order to provide an answer, we tried to find out where the problem came from, i.e. which component of the conversion model was not working well enough. To find out, we made two different assumptions with distinct technical implications and tested them.

FIRST VARIANT. Our first assumption was that the encoder of the conversion algorithm was responsible for this defect. It is, indeed, the network's component responsible for encoding the linguistic content of the source speech. In particular, it produces a representation of the source utterance in which the linguistic content - as well other characteristics such as the identity of the speaker - must lie. One explanation would be that the encoder does not always fulfil this role which prevents the decoder to render the linguistic content properly. If this assumption is correct, then it is sufficient to force the encoder to properly encode the linguistic content of the source utterance so that the conversions preserve intelligibility.

To test this assumption, we thought of a first option for the integration of the speech-to-text module in the conversion model. This option involves using the same encoder for speech-to-text and conversion while incorporating the speech-to-text's decoder as a third branch of the conversion algorithm. The representation produced by the encoder would then be sent to both the conversion algorithm's decoder and the speech-to-text's decoder. Unfortunately, early experiments with this three-branch architecture did not yield good results, not only the intelligibility did not appear to be better preserved but the model did not converge as well as it did without ASR.

SECOND VARIANT. Our second hypothesis was to make the minimal assumption of a global failure of the whole conversion algorithm. In other words, no assumptions are made about the part of the model involved in this generation failure. Then, the sufficient condition for the intelligibility of the

yielded conversions to be preserved is that the ASR model performs as well on the converted data as it does on the ground-truth data.

We thus thought of a second option for the integration of the speech-to-text module in the conversion model. We proposed to connect the speech-to-text's speech input to the output of the conversion algorithm, in this way we mean to condition the generation of a converted utterance on a specific linguistic content. Thus, the converted mel-spectrogram is given to the text-to-speech encoder, and the character sequence corresponding to the target sentence is fed into the decoder. By propagating the prediction error of the speech-to-text module through the weights of the conversion network, the latter is forced to generate mel-spectrograms from which the speech-to-text module can predict the correct character sequence. This added training criterion involves measuring the difference between the text that was actually pronounced in the source (or target) utterance and the text which is predicted from the conversion. Note that at this stage, the weights of the ASR module are frozen, only the conversion model is being trained. Indeed, we do not want the ASR module to learn to predict the actual text despite a lack of intelligibility - which it might succeed in. Conversely, we want to force intelligibility of the converted utterances. The full architecture of this ASR-upgraded voice conversion algorithm is presented in the Figure 7.5. The full model's loss $L_{vtn \times asr}$ is then the sum between the standard voice conversion transformer network's loss L_{vtn} - of which calculation is specified in the last section 7.1 - and the ASR module's prediction loss $\mathcal{L}_{asr}$, the balance between both being controlled by means of a factor λ_{asr}.

$$L_{vtn \times asr} = L_{vtn} + \lambda_{asr} \mathcal{L}_{asr} \tag{7.33}$$

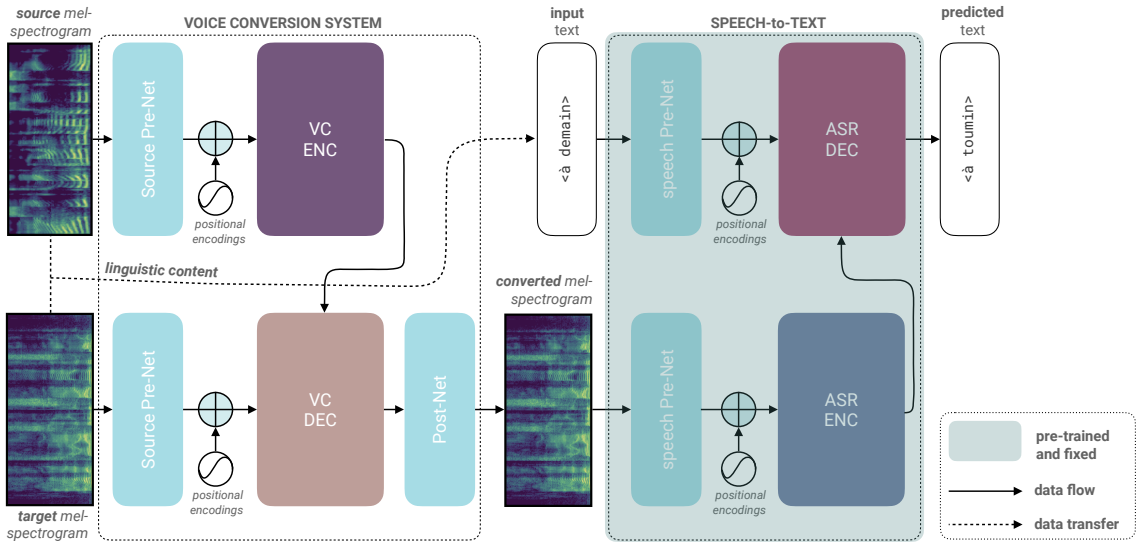

Figure 7.5: Schematic view of the proposed text-conditioned voice conversion system neural architecture

7.3 Speech Attitude Conversion Experiments

In this section, we present two experiments of many-to-many learning of voice attitude conversion. The first is a direct application of the Voice Transformer Network proposed in (Kameoka et al., 2021) with specific adjustments due to the mel-spectrogram representation of speech signals we chose to employ. The second is to evaluate the impact of incorporating an ASR module into the speech attitude conversion algorithm, particularly in terms of the intelligibility of the conversions generated.

7.3.1 Many-to-Many Experiment for Speech Attitude Conversion

The purpose of this first experiment was to bring out a baseline configuration of the voice transformer - initially proposed for speaker identity conversion (Kameoka et al., 2021) - that fits the specific task of speech attitude conversion and which involves learning on data from our database Att-HACK (Le Moine and Obin, 2020). The fundamental point here is that the speaker identity conversion does not involve modifying the same aspects of the speech signal than the speech attitude conversion. Indeed converting speaker identity mainly deals with changing the spectral envelope while we have shown in section 4.3 that almost everything in the signal, including prosodic and articulation aspects, was conveying attitudes.

Beyond this task-specific aspect, this experiment was an opportunity to validate the use of mel-spectrograms in place of multi-parametric representation of speech signal - such as WORLD vocoder features which are employed in (Kameoka et al., 2021) - for voice conversion purpose. Nevertheless, in this experiment we do not compare the results obtained from the two different representations, the arguments in favor of the use of the mel-spectrogram mentioned in section 7.1.3 being sufficient - in our opinion - to justify this choice. In addition, the choice of this representation implies slightly different choices of parameters for the Voice Transformer Network.

For these reasons, we did not simply apply the parameters specified in (Kameoka et al., 2021) directly to the conversion model but tested different configurations. We could not afford to evaluate each configuration to determine which was best, i.e. the one that would offer the best objective performance. The method we used involved radically altering the hyper-parameters to first identify those that had a major impact on the model's performance.

Data for Experiment

Since the many-to-many paradigm - i.e. learning the conversion of any attitude to any other attitude simultaneously - is an ambitious task in itself and the production of attitudes is individual (Le Moine et al., 2021a) - i.e. speakers use their own strategies to communicate attitudes - we limited this experiment to the speaker-dependent case by selecting data coming from only one female speaker ($F03$). Although this choice might appear random it is not. This actress in particular had the highest average perceptual score for friendliness which is the attitude we value most because of its numerous potential uses in daily life situations. As mentioned earlier, our perceptual domain classification model allowed us to identify utterances with poorly communicated attitude across the entire database. These samples were therefore removed and are not used in this experiment. We are fully aware that such an experiment - with only one speaker - only provides limited support for the assumptions we made. However, due to time constraint we had to keep the experimental

evaluation limited and leave a more extensive - multi-speaker - evaluation for future works.

As mentioned above, we mean here to learn attitude conversion in a many-to-many fashion. This means that each source utterance in distant yields several pairs whose targets were produced in different attitudes (including distant) but have the same linguistic content. In order to compute both terms of the model's loss as formulated in Eq. 7.29, we have to distinguish two types of pairs, the so-called intra-attitude pairs - i.e. pairs in which source and target are produced with respect to the same attitude - and the inter-attitude pairs - i.e. pairs in which source and target are produced in different attitudes. In the input pipeline we implemented, a batch is built for each type of pairs and presented to the model thus yielding both terms of the total loss.

TRAIN-VALID DATA SPLIT. For this experiment, we performed random split on data such as to have all the sentences in both training and validation set. Indeed as shown in chapter 4, attitude production can be seen as a modulation of what is already determined by the linguistic content of the utterances. Though, learning to map attitudes on a bunch of sentences is probably not sufficient to generalize to others. We selected 26,729 pairs for training and 7,923 for validation, which represents approximately 80% of the pairs kept for training.

Implementation Details

Dropouts with rate 0.1 were applied to the input sequences before being fed into the source and target pre-Nets and the post-Net only at training time. We chose to use ReLU as nonlinear activation function in each FFN sub-layer, we found GLU (gated linear unit) which used in (Kameoka et al., 2021) to yield poorer performance. The two pre-Nets and the post-Net were each designed using three 1D dilated convolution layers with kernel size $k = 5$ and dilation factor $\delta = 3^i$ for the i^{th} convolutional layer, each followed by a ReLU activation function, where weight normalization (Salimans and Kingma, 2016) was applied to each layer. The filter dimension d was set to 512 and the middle channel number d_{ffn} was set to 1024. The weighting of identity mapping loss λ_{im} was set to 1 and ν which used to compute DA loss was set at 0.3.

We tested several settings with variable number of layers in both the encoder and the decoder and different weighting of diagonal attention (DA) loss. The most salient parameter was found to be the size of the model, i.e. the number of transformer blocks in both the encoder and decoder, we thus varied L_{enc} and L_{dec} from 1 to 4. Note that beyond 2, the model became too difficult to train both in terms of memory resources and training time. While the diagonal constraint on attention matrices has a clear impact, it is difficult to find a gradation of this impact. It seems that the model simply does not converge for some values when it does for others. After testing different values for λ_{da} from 100 to 10000, we have chosen here to apply a factor $\lambda_{da} = 5000$ to this diagonal constraint on attention matrices.

7.3.2 Linguistic Conditioning Experiment

The general purpose of this experiment is not to evaluate the effectiveness of various configurations of the speech-to-text module that were previously presented, but rather to select a configuration of this module that is effective and test its impact on attitude conversion performance when incorporated in the conversion algorithm. First we intend to select this configuration.

Training the Speech-to-text Module

The speech-to-text module has first been trained to achieve its goal of predicting the appropriate texts for every sentence - i.e. a hundred different linguistic contents - in Att-HACK. For this experiment we considered all the sentences of Att-HACK, i.e. a little more than 36000 sounds. The training/validation split was done in such a way that any speaker and sentence, i.e. linguistic content, is in both the training and validation sets. predicted - achieved its best performance on the validation set. We used the same learning rate custom scheduler than for conversion learning - the architecture also being transformer-based - and batch size of 32. We selected the best configuration by saving weights at best cross-entropy performance on the validation set.

Speech-to-text Selected Architecture

The speech-to-text module itself has not even been experimented with. The configuration we started with proved to be quite efficient and we chose not to waste time trying to improve it.

The speech-to-text's encoder and speech pre-network has strictly the same architecture as those of the selected conversion system, which makes sense since the latters are partly dedicated to the encoding of the linguistic content of the source utterance. However, preliminary experiments have shown that the both modules cannot share the same encoder without compromising overall convergence. With regard to the decoder, we have chosen a dropout rate of 0.1 except for the dropout layer following the self-attention which has a higher rate of 0.5. The filter dimension d was set to 200, this also the dimension of linguistic space, i.e. the output dimension of the character pre-network. The middle channel number d_f was set to 400. The number of heads in multi-head self attention layer was set to 2.

Training the ASR-upgraded Voice Conversion Model

Once the experiment has been conducted on the speech-to-text module alone, we come to its incorporation into our conversion model. We must keep in mind that this experiment's goal is to evaluate how well our speech attitude conversion model performs after incorporating a speech recognition module. More precisely, we intend to demonstrate that such an addition enables to force the model's conversions to be intelligible. In order to solely examine the impact of the inclusion of the text-to-speech module, we used the exact same hyper-parameters as for the first experiment. As mentioned earlier, the ASR module has been pre-trained and all of its layers have been frozen to prevent it from capturing linguistic content from poor conversions. Indeed, it is conceivable that such a module would be able to decode linguistic content from a succession of poorly formed phonemes, where humans would not be able to do so without an undesired effort.

7.3.3 Selected Configurations & Evaluation Process

In the following we present the results of the two experiments described above at the same time, the main point being to evaluate the impact of incorporating a speech recognition module into the speech attitude conversion algorithm.

Selected Configurations

We selected two configurations for the first experiment: **VTN-s** and **VTN-l** respectively a small and a large version of the adapted voice transformer network proposed in (Kameoka et al., 2021). For

the second experiment, we selected three configurations each corresponding to a different level of influence of the speech-to-text module on the conversion model. For this second experiment, we used the large voice conversion model's configuration **VTN-l** - as it appeared to yield the most satisfying results. Those configurations are distinct in the impact they grant to the speech-to-text module on the overall optimisation of the conversion algorithm. We listed all those configurations bellow:

- **VTN-s**: a small model with $L_{enc} = 1$ and $L_{dec} = 1$.
- **VTN-l**: a large model with $L_{enc} = 2$ and $L_{dec} = 2$.
- **VTN-lxASR-li**: the large model with ASR $\mathcal{C}^{asr}$ of light impact on intelligibility ($\lambda_{asr} = 0.1$).
- **VTN-lxASR-me**: the large model with ASR $\mathcal{C}^{asr}$ of medium impact on intelligibility ($\lambda_{asr} = 0.5$).
- **VTN-lxASR-st**: the large model with ASR $\mathcal{C}^{asr}$ of strong impact on intelligibility ($\lambda_{asr} = 1.0$).

Evaluation Process and Metrics

To provide such an objective assessment, we chose to use the mean absolute error (MAD-mel) between the log amplitude target and converted mel-spectrograms and the root mean square error (RMSE-f0) between F0 contours of target and converted speech signals, synthesized through neural vocoding (Roebel and Bous, 2022).

Mean Absolute Difference (MAD-mel). We do not assume this metric to be correlated with human perception, i.e. a good performance in MAD does not necessarily lead to a convincing conversion from a human perception point of view. However computing differences on a mel scale, which is know to be perceptually relevant, avoid to give too much importance to frequency bands that are not for human perception. Provided with a target utterance mel-spectrogram $\mathbf{X}^t \in \mathbb{R}^{T_t \times D}$ and its related conversion $\mathbf{Y}^{s \leftarrow t} \in \mathbb{R}^{T_t \times D}$, MAD-mel denoted ϵ_{mel} is formulated as follows

$$\epsilon_{mel}(\mathbf{X}^t, \mathbf{Y}^{s \leftarrow t}) = \frac{1}{DT^t} \sum_{n=1}^{T} \sum_{k=1}^{D} |X_{n,k}^t - Y_{n,k}^{s \leftarrow t}| \tag{7.34}$$

where $X_{n,k}^t$ and $Y_{n,k}^{s \leftarrow t}$ are the n^{th} frame and k^{th} mel bin of $\mathbf{X}^t$ and $\mathbf{Y}^{s \leftarrow t}$ respectively. We also introduce a lower bound ϵ that allows avoiding a strong impact of the perceptually irrelevant small values that may arise in the noise sections. For the present evaluation, and following (Roebel and Bous, 2022), we used $\epsilon = \log 10^{-5}$ such that any values of $\mathbf{X}^t$ and $\mathbf{Y}^{s \leftarrow t}$ under ϵ are set to ϵ.

Root Mean Squared F0 Error (RMSE-f0). We use the neural vocoder (Roebel and Bous, 2022) to produce speech signals $\mathbf{x}^t$ and $\mathbf{y}^{s \leftarrow t}$ from target and source melspectrograms, $\mathbf{X}^t$ and $\mathbf{Y}^{s \leftarrow t}$ respectively. Then we employ an F0 extractor [ref] to obtain F0 contour sequences $\mathbf{f}^{\mathbf{x}^t}$ and $\mathbf{f}^{\mathbf{y}^{s \leftarrow t}}$. RMSE-f0 denoted ϵ_{f0} is computed as follows

$$\epsilon_{f0}(\mathbf{f}^{\mathbf{x}^t}, \mathbf{f}^{\mathbf{y}^{s \leftarrow t}}) = \sqrt{\frac{1}{T_t} \sum_{n=1}^{T_t} (f_n^{\mathbf{x}^t} - f_n^{\mathbf{y}^{s \leftarrow t}})^2} \tag{7.35}$$

In order to objectively measure the performance of our speech-to-text module, we employ two standard metrics. These metrics assess how well the linguistic content has been predicted by the ASR module. They also enable us to assess how well the linguistic content is preserved in the conversions our model delivers, or how well they are intelligible. In addition, we will use objective

criteria MAD-mel and RMSE-f0 for assessing the similarity of yielded conversions with target utterances, as done in the past section.

Character Error Rate (CER). This metric indicates the percentage of characters that were incorrectly predicted. The lower the value, the better the performance of the ASR system with a rate of 0 being the best performance.

Word Error Rate (WER). Analogously, this metric indicates the percentage of words that were incorrectly predicted. The lower the value, the better the performance of the ASR system with a rate of 0 being the best performance. Both metrics respectively denoted ϵ_c and ϵ_w are computed as follows

$$\epsilon_c = \frac{s_c + d_c + i_c}{n_c} \tag{7.36}$$

$$\epsilon_w = \frac{s_w + d_w + i_w}{n_w} \tag{7.37}$$

where s is the number of substitution, d the number of deletion and i the number of insertion, n being the actual number items, and indices c and w respectively denoting characters and words.

7.3.4 Objective Evaluation

In this subsection, we present the results of our conversion algorithm enhanced by a speech-to-text module. We evaluate two qualities of the resulting conversions, firstly intelligibility - through CER and WER - secondly proximity to the target sentence through MAD-mel and RMSE-f0.

Assessing the Intelligibility of the Conversions

First of all, let us specify that the objective here is to show that the addition of an ASR module makes it possible to better preserve the linguistic content in conversions. Through this we wish to ensure that it is possible for an individual to understand the meaning of what is being said. If the overall meaning of a sentence is constructed from the semantic units that are the words, we can refine this by relying on the ability of individuals to predict the meaning conveyed by an almost well-pronounced word. This is why we consider both the word error rate, which gives the true capacity of the model to ensure good semantic decoding, and the character error rate, which is less demanding.

The results in CER and WER for all the examined configurations are depicted in Figure 7.6. Looking at the average results across the four attitudes, it can be seen that the large model **VTN-l** performed significantly better than its smaller version **VTN-s** in terms of CER. However, as already mentioned, we were able to identify a linguistic loss issue that persisted after random listening to conversions, which prompted us to think about adding an ASR module to the conversion algorithm. Undoubtedly, the speech-to-text module's addition significantly lowers the prediction error at the character level. Furthermore, it is evident that the final error is lower the more important its influence on the conversion model's optimization. However, the difference between the configurations of ASR-upgraded conversion algorithm cannot be regarded as substantial. In particular, we find a CER of 3.8 for the configuration with the addition of a higher impact ASR module **VTN-lxASR-st** versus 17.3 for the baseline conversion algorithm **VTN-l**. The aforementioned trends are

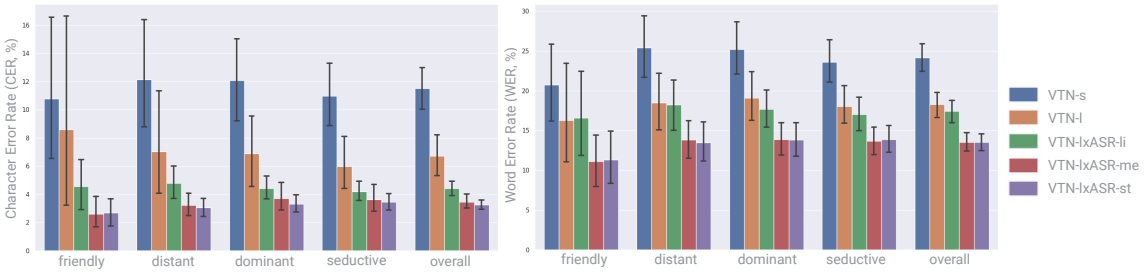

Figure 7.6: Performance of **VTN-s**, **VTN-l**, **VTN-lxASR-li**, **VTN-lxASR-me** and **VTN-lxASR-st** with respect to the character error rate (CER) on the left and to the word error rate (WER) with 95% confidence interval, on the right for friendly, distant, dominant and seductive and across attitudes.

still noticeable when we break down the results by attitude. However, there are slight differences in performance depending on the attitude considered. The performance in terms of WER, quite obviously follows the same trend. Although, the effects are less significant, especially if we consider the attitudes separately. Nevertheless, there is a substantial decrease in WER for the two higher impact configurations **VTN-lxASR-st** and **VTN-lxASR-me** compared to the baseline attitude conversion algorithm **VTN-l**.

In summary, it is apparent that adding a speech-to-text module to our conversion algorithm greatly decreases generating errors from a linguistic perspective. Nevertheless, there remains a certain proportion of utterances for which the linguistic content is not perfectly rendered by the ASR module. It should be noted that this could be a prediction error from the speech-to-text module or a generation fault from the conversion algorithm.

Assessing the Proximity of the Conversions to the Target Utterances

Here, we mean to evaluate the performance of the selected configurations of our speech attitude conversion model in regards with objective criteria. Performance results in MAD-mel and RMSE-f0 are displayed in the Table 7.1 for each conversion attitude pair.

First, we observe that the large configuration VTN-l slightly outperforms the smaller one **VTN-s** respectively by 0.2 in terms of MAD-mel and by 4.3 Hz in terms of RMSE-f0. If we now consider the performance by pair of transformations, we see that no pair escapes this trend except the distant to dominant transformation for which VTN-s performs slightly better than **VTN-l** in terms of MAD-mel, however it is probably not significant. We also find a slightly better performance of **VTN-s** than **VTN-l** in terms of RMSE-f0 for the conversion from friendly to seductive and from dominant to seductive. The correlation scores between the converted F0 contours and the target contours were also computed. With the score fluctuating somewhat around 0.9 depending on the pair of attitudes taken into consideration for the conversion, we found little difference between the two configurations.

Second we observe that the baseline **VTN-l** and light impact ASR-upgraded **VTN-lxASR-li** configurations achieve same performance with 0.76 in MAD-mel and 49.9 in RMSE-f0. Slightly better, the moderate and strong impact ASR-upgraded **VTN-lxASR-me** and **VTN-lxASR-st** configurations also achieves same performance in terms of MAD-mel with 0.74. **VTN-lxASR-st** slightly outperforms **VTN-lxASR-me** with respectively 48.2 and 48.8 in RMSE-f0. Considering pair-wise results,

it can be seen that the best performance is achieved sometimes by **VTN-lxASR-st** sometimes by **VTN-lxASR-me** depending on the transformation pair. It is impossible to determine whether the addition of the speech recognition module affects the conversion model's capacity to generate convincing conversions based solely on these results. It is however observed that the addition of such a module does not prevent the conversion algorithm from converging.

Model	Metric	fr. → di.	fr. → do.	fr. → se.	di. → fr.	di. → do.	di. → se.
VTN-s	ϵ_{mel}	0.71	0.79	0.80	0.83	**0.79**	0.80
	ϵ_{f0}	41.6	49.6	**52.7**	59.5	49.6	**52.7**
VTN-l	ϵ_{mel}	0.69	**0.75**	0.80	0.78	0.80	**0.75**
	ϵ_{f0}	40.5	48.5	54.8	55.6	48.5	54.8
VTN-lxASR-li	ϵ_{mel}	0.71	0.76	0.77	0.80	0.76	0.77
	ϵ_{f0}	42.3	48.7	**53.0**	55.4	48.7	**53.0**
VTN-lxASR-me	ϵ_{mel}	0.68	0.75	**0.76**	**0.77**	0.75	0.76
	ϵ_{f0}	40.0	47.1	54.3	54.0	47.1	54.3
VTN-lxASR-st	ϵ_{mel}	**0.67**	**0.73**	0.77	0.78	**0.74**	0.77
	ϵ_{f0}	**39.6**	**44.9**	53.7	**54.7**	**44.9**	53.7
Model	Metric	do. → fr.	do. → di.	do. → se.	se. → fr.	se. → di.	se. → do.
VTN-s	ϵ_{mel}	0.83	0.71	0.80	0.83	0.71	0.79
	ϵ_{f0}	59.5	41.6	**52.7**	59.5	41.6	49.6
VTN-l	ϵ_{mel}	0.80	0.69	0.78	0.80	0.69	0.75
	ϵ_{f0}	55.6	40.5	54.8	55.6	40.5	48.5
VTN-lxASR-li	ϵ_{mel}	0.80	0.71	0.77	0.80	0.71	0.76
	ϵ_{f0}	55.4	42.3	**53.0**	55.4	42.3	48.7
VTN-lxASR-me	ϵ_{mel}	**0.77**	0.68	**0.76**	**0.77**	0.68	0.75
	ϵ_{f0}	**54.0**	40.0	54.3	**54.0**	40.0	47.1
VTN-lxASR-st	ϵ_{mel}	0.78	**0.67**	0.77	0.78	**0.67**	**0.73**
	ϵ_{f0}	54.7	**39.6**	53.7	54.7	**39.6**	**44.9**

Table 7.1: Performance results with regards to the objective metrics MAD-mel and RMSE-f0 for **VTN-s**, **VTN-l**, **VTN-lxASR-li**, **VTN-lxASR-me** and **VTN-lxASR-st** for each conversion pair of attitudes.

Visualizing Speech-to-Text Module's Effect on Conversions

In order to illustrate the differences between the basic configuration **VTN-l**, its version augmented with a speech recognition module **VTN-lxASR-st** and what is considered as a reference - i.e. the target sentence we expect to get close to - Figure 7.7 shows two examples of conversion through representing the converted mel-spectrogram and the average attention matrix resulting from the different target-to-source multi-head attention layers that reflects the implicit alignment learnt between source and target utterances. Although it can be challenging to comprehend what a neural model does, these representations serve as a useful tool for making decisions about a model's construction and optimization. On the mel-spectrograms, the white dotted bars designate the end

and the beginning of the syllables that make up the spoken sentence, the phonetic transcription of these syllables is indicated in white at the top of each image. The question marks designate the syllables that are unintelligible, they are located at the end of the sentence and concern only the baseline configuration **VTN-I**.

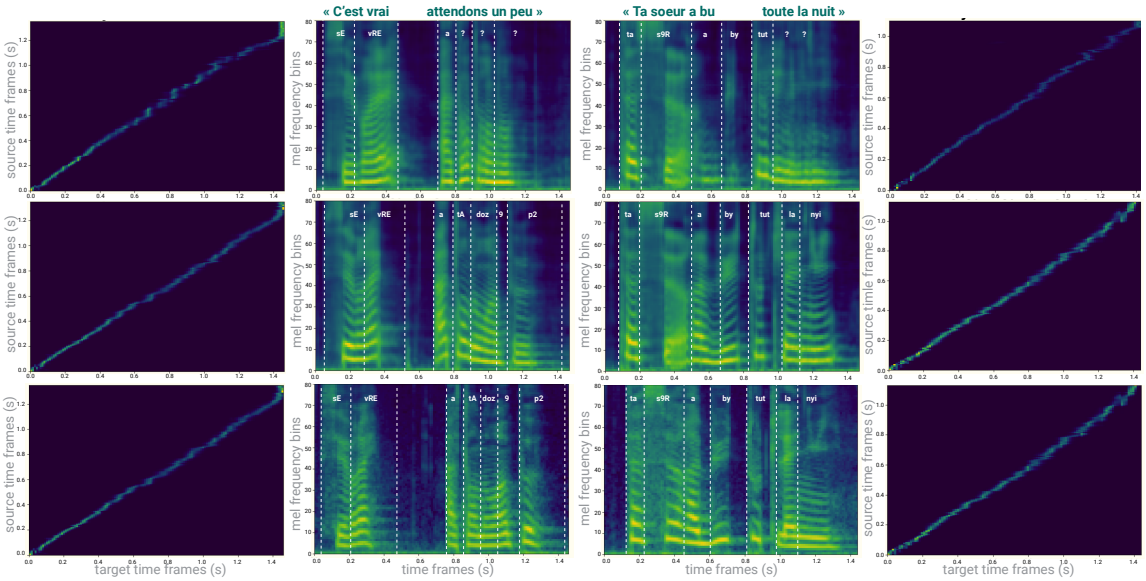

Figure 7.7: Visualization of two conversion examples. At the left, the sentence "C'est vrai attendons un peu" is converted from distant to dominant and at the right the sentence "Ta soeur a bu toute la nuit" is converted from distant to seductive. The two first lines show conversions respectively from **VTN-I** and **VTN-IxASR-st** while the last line shows the related target utterance. For each example and conversion, the mean attention matrix from the target-to-source multi-head attention layers and the converted mel-spectrogram is depicted.

By observing these two examples, we notice firstly that the addition of the speech recognition module allows, in both cases, to preserve the linguistic content. Each of the syllables underlying the formation of the global sentence is preserved by the model with ASR **VTN-IxASR** whereas the syllables at the end of the sentence are unintelligible or completely absent - this is the case of *"peu"* phonetically denoted *"p2"* - for the basic model **VTN-I**. If we leave aside this linguistic aspect, we can see that in general the mel-spectrogram converted by **VTN-IxASR-st** is much closer to the target mel-spectrogram than the one converted by **VTN-I**. In particular, **VTN-IxASR-st** seems to better render the formants' variations during the course of the sentence than **VTN-I**. Moreover, the noise which also plays an important role in speech signal seems to be modelled more finely by **VTN-IxASR-st** than by **VTN-I**.

A final observation is that the average attention estimated from the various layers of target-to-source multi-head attention is not linear. It is clear that the model captures an implicit alignment at the sentence level - including the words that make it up - as well as at lower levels - such as phonemes - reflecting micro-prosody.

7.4 Perceptual Evaluation of Vocal Attitude Conversion Models

Despite all these so-called objective evaluations of the performance of our voice attitude conversion algorithms, it seems important to us to assert the primacy of the perceptual criterion. Indeed, whatever the final application of these algorithms, the only criterion that really matters is the average individual's opinion of the conversions. This subjective judgement can obviously be broken down into different qualities such as naturalness or the similarity to a reference - or a label - regarding the attitude conveyed.

7.4.1 Perceptual Experiment

In this part, we describe the listening experiment carried out for the perceptual assessment of the various configurations' conversions.

Evaluation Data

For this experiment we have selected four different attitude conversion models among those exposed above. On the one hand, two configurations of the baseline algorithm, the small model **VTN-s** and its equivalent twice as large - in terms of number of layers - **VTN-l**. On the other hand, two configurations of the algorithm augmented with a speech-to-text module, of low impact **VTN-lxASR-li** and of high impact **VTN-lxASR-st**. Since it was obviously impossible to evaluate all the validation data perceptually, we selected ten different linguistic contents, related to short, medium and long sentences. For each of these sentences, we generated a conversion for each of the four configurations, to which we add the target sentence as a reference. This is done for the twelve transformation pairs. This amounts to 150 samples to evaluate per target attitude ($10 \times 3 \times 5$) - or a total of 600 samples.

Experimental Design

The question of the choice of experimental design used to perceptually validate the performance of our models deserves to be asked. Indeed, we successively conducted a BWS experiment to question the perception of attitudes in Att-HACK in Chapter 4 and then conducted an AB preference test to perceptually evaluate the performance of our F0-based attitude conversion model in 5. These two designs proved to be effective in revealing relative perceptual differences between sounds on the one hand and between models on the other. As a consequence, we chose to continue with this paradigm of relative assessment rather than switching to one of absolute judgment - Rating Scale (RS) kind - such as Mean Opinion Score (MOS). Since the implementation of a BWS experiment is very demanding, both from a logistical point of view and from the perceptual data post-processing perspective, we have opted to conduct an AB preference test.

Throughout this test, we mean to compare different configurations of our conversion algorithm according to two criteria: 1) **intelligibility** - i.e. being able to decode the linguistic message conveyed and 2) the **fidelity** of the attitude conveyed **to a specific attitude label**. These two criteria seem to assess common qualities, if the linguistic content cannot be decoded it seems difficult to assess the attitude conveyed. In fact, it is possible that no attitude can be conveyed if the linguistic content is not preserved. However, conversely, some conversions may preserve the linguistic content without adequately conveying the target attitude. In this case, it will be interesting to observe the answers to the two questions asked. In particular for this second criterion, we have chosen

not to compare the conversions to a reference - i.e. their related target utterance - but rather to an attitude label. Indeed, we might claim that the comparison to a reference appears too restrictive as there is no single way to communicate a vocal attitude given a linguistic message and a speaker identity. Each participant evaluates twenty randomly picked pairs of sounds that relate to a particular attitude that is also randomly picked. On each run, participants are asked to listen to two sounds, which can be either two conversions from different configurations of the speech attitude conversion algorithm or one conversion and one reference - i.e. the target sentence associated with the conversion. Participants are then asked to answer two questions: 1) *which of these sounds is more intelligible?* and 2) *which of these sounds conveys more of the attitude being studied?* To each of these questions, participants are asked to answer categorically. If we denote the sounds of a pair *A* and *B*, then participants must choose between *Mostly A*, *Fairly A*, *In between*, *Fairly B* and *Mostly B*.

7.4.2 Results & Discussion

At the end of the test we collected responses from 140 participants, which is fairly substantial considering that the test was not remunerated.

Assessing the Conversions' Intelligibility

The first question the participants were asked dealt with intelligibility. They were asked to judge, for each pair evaluated, which sound was the most intelligible. In this regard, the data collected is well represented by the confusion matrix shown on the left side of the Figure 7.8. In this matrix, each row presents the overall preference of a given configuration - or the reference - towards other configurations - and the target - with regards to intelligibility. Positive scores indicate an overall preference of the y-label over the x-label, the scale going from -2 to 2 with 0 meaning no decision. The left-hand side of Figure 7.8 is an attempt to represent the absolute performance of each of these configurations by measuring the proportion of pairs (in %) for which a given configuration is preferred over others. Note that this last representation only makes sense if all possible pairs have been evaluated by participants. However, the interpretation of the data can only be made in the light of these two representations.

First, the configuration that yields least intelligible conversions is the small model **VTN-s**, its twice as large version **VTN-l** achieving better performance. Second, it appears that the configuration with a strong impact of the speech-to-text module **VTN-lxASR-st** is preferred to any other configuration with regards to intelligibility. Note that, the configuration with light-impact of the ASR module **VTN-lxASR-li** is also better judged than our baseline algorithms **VTN-s** and **VTN-l** for intelligibility. Our large baseline algorithm **VTN-l** is preferred over the smaller one **VTN-s**. Finally, the reference is logically always found to be more intelligible whatever the conversion it is compared to. On the right-hand side of Figure 7.8, the error bars provide an indication of how significant the observed effects are - i.e. how much one configuration is preferred to the others. We note in particular the overall superiority of our strong ASR's impact algorithm **VTN-lxASR-st** over the smaller algorithm **VTN-s**.

To conclude, we can affirm that the incorporation of a speech-to-text module into our attitude conversion algorithm allows us to globally improve the intelligibility of the conversions produced. In that sense, it constitutes a solution to the issue of linguistic content loss encountered with the basic algorithms **VTN-s** and **VTN-l**.

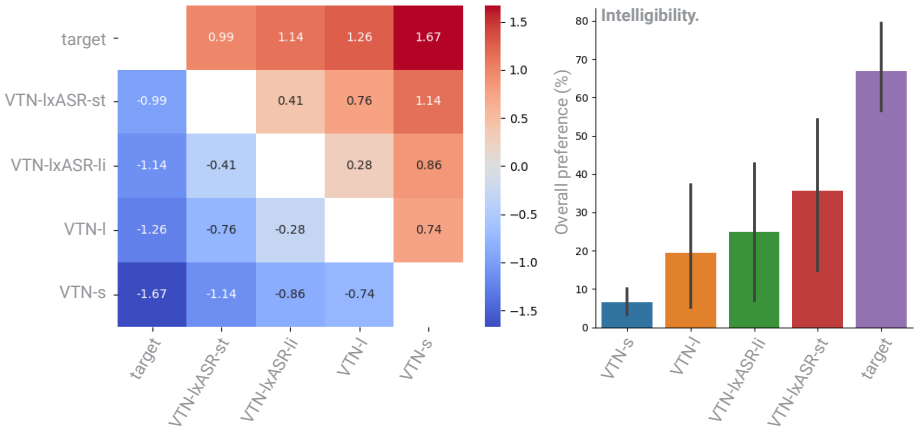

Figure 7.8: At the left, we depicted the confusion matrix where each row presents the overall preference towards other configurations and target with regards to intelligibility. Positive scores indicates an overall preference of the y-label over the x-label. At the right, we represented the proportion of pairs in which each configuration is preferred over the others with regards to intelligibility.

Assessing the Conversions' Conveyed Attitude

The second question the participants were asked dealt with the attitude conveyed. They were asked to judge, for each pair evaluated, which sound was conveying the most the attitude studied - i.e. friendly, distant, dominant or seductive depending on the attitude randomly picked by the participant. In this regard, the data collected is well represented by the confusion matrix shown on the left side of the Figure 7.9. In this matrix, each row presents the overall preference of a given configuration - or the target - towards other configurations - and the target - with regards to the attitude conveyed. As for intelligibility, positive scores indicate an overall preference of the y-label over the x-label. In a similar way, the left-hand side of Figure 7.9 represents the absolute performance of each of these configurations by measuring the proportion of pairs (in %) for which a given configuration is preferred over others.

First off, it appears clear that there are less significant perceptual differences between the various configurations than there are for intelligibility. In addition, the left-hand of Figure 7.9 shows that the target is preferred to other configurations in only 50% of the cases. It can also be seen from the left-hand side of the Figure that in the best case, the reference is only fairly preferred (mean score close to 1.0) in terms of the attitude conveyed. These points may be interpreted in two different ways. First, it is likely substantially harder to evaluate the attitude a sentence conveys than it appears to be to evaluate its intelligibility. Second, our algorithms perform rather well in regards with their ability to convert the vocal attitudes.

Let us now dive into details. Firstly, we note that **VTN-lxASR-st** is preferred to all the other configurations, that said this preference remains minor, less significant than a "fairly preferred" if we consider the judgment scale. We also notice a slight preference for **VTN-lxASR-li** over the two configurations without an ASR module, **VTN-l** and **VTN-s**. Our large baseline algorithm **VTN-l** is slightly preferred over the smaller one **VTN-s**. Note that the reference has only been slightly preferred to our best performing configuration **VTN-lxASR-st**. Although this difference is small, the right-hand side of Figure 7.9 shows that it is significant. The participants seem to agree that the

attitude conveyed through the conversions from **VTN-IxASR-st** is less faithful to the associated attitude label than the reference itself.

As a result, it appears that the participants judged the attitude conveyed in conversions from algorithms enhanced with a speech-to-text module as being even more true to the associated attitude label. This is potentially partly due to the improved intelligibility offered by these algorithms. That said, one might have feared that these algorithms, by forcing intelligibility, would inhibit conversion and produce unconvincing conversions with regards of the attitude conveyed. That is not the case, which definitely validates the incorporation of this speech-to-text module into our vocal attitude conversion algorithm. However, it should be noticed that individuals still perceive a significant difference between the real utterances and the conversions from our best model in terms of fidelity to an attitude label.

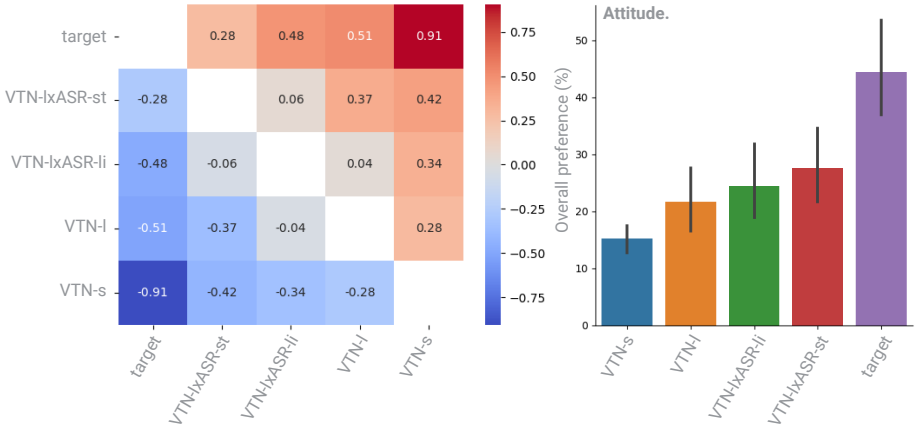

Figure 7.9: At the left, we depicted the confusion matrix where each row presents the overall preference towards other configurations and target with regards to the attitude conveyed. Positive scores indicate an overall preference of the y-label over the x-label. At the right, we represented the proportion of pairs in which each configuration is preferred over the others with regards to the attitude conveyed.

A Transformer-based Speech Attitude Converter with Improved Intelligibility

We showed that incorporating a speech-to-text module into our transformer-based conversion algorithm significantly reduced the loss of intelligibility in the conversions. The same trend was observed in terms of objective proximity to the target utterances, although the differences did not allow to draw conclusion on the naturalness of the attitudes conveyed in conversions. We supplemented this objective assessment with a listening experiment to perceptually assess the conversions in which we asked 150 participants to judge the intelligibility of the conversions and the fidelity of the attitude they convey to a specific attitude label. First, the test results proved that the incorporation of a speech-to-text module into the transformer-based attitude conversion algorithm allows to globally improve the intelligibility of the conversions. Second, it appears that the participants judged the attitude conveyed in conversions from linguistically conditioned algorithms as being even more true to the associated attitude label. However, it should be noticed that individuals still perceive a significant difference between the real utterances and the conversions from our best model in terms of fidelity to an attitude label. However, it is important here to moder-

ate these conclusions since our experiment involves only one speaker and does not allow us to validate the performance of such a model in a more general context. Nevertheless, this research marks a substantial advance in that it makes it possible to learn the conversion of any attitude to any other simultaneously, i.e. in a many-to-many fashion, which to our knowledge had never been achieved. We are currently conducting further experiments involving several speakers in order to observe whether the algorithm can learn to convert speech attitudes independently of the speaker. At the time of writing, we dispose of a powerful algorithm for the conversion of vocal attitudes and are working to improve it by extending its scope of validity.

7.5 Chapter Summary

At the end of Chapter 5, in which we outline our first attempt at converting speech attitudes by changing only F0 contours, we decided to shift paradigm. First, we chose to adopt mel-spectrogram as speech representation along with the neural vocoder proposed in (Roebel and Bous, 2022) from which speech can be recovered almost perfectly. Second, we chose to focus on the transformer architecture (Vaswani et al., 2017) for attitude conversion as it allows to learn a mapping between utterances of different duration. The main advantage of such transformers being that it basically replaces all recurrent - or dilated convolution - layers by self attention mechanisms in a network. This allows for more efficient learning in terms of both time and computing resources.

In line with this paradigm shift, we worked towards an adaptation of the algorithm proposed in (Kameoka et al., 2021) for many-to-many speech attitude conversion based on mel-spectrogram representation of speech signal. A first limitation of the approach in (Kameoka et al., 2021) was the use of WORLD vocoder features as speech multi-parametric representation, we thus proposed to employ mel-spectrogram instead. Early experiments with the transformer architecture led us to identify a loss of linguistic content in a significant number of conversions. Some phonemes were badly formed or not formed at all while in some conversions, a word was changed into another word thus completely changing the meaning of the utterance. This issue appeared to be very disturbing in that it compromises the optimal decoding of the message conveyed by the converted utterance. To face this issue, we placed intelligibility - i.e. the ability to decode the linguistic message conveyed in a speech signal - on top of the criteria our conversions must meet.

Our contribution first involved reformulating the approach of (Kameoka et al., 2021) for many-to-many speech attitude conversion based on mel-spectrogram representation. Second, we addressed the issue of linguistic loss encountered in the early experiments' conversions with the transformer. For the latter, we moved to the typical research problem of Automatic Speech Recognition (ASR) which consists in transcribing audio speech segments into a sequence of characters. First, such a speech-to-text module allows for an objective evaluation of the linguistic loss issue. Therefore, we first worked at implementing an efficient speech-to-text module and trained it on Att-HACK. Second, we considered how we can use such a speech-to-text module to enhance the performance of our conversion model. We proposed to connect the speech-to-text's speech input to the output of the conversion algorithm, in this way we mean to condition the generation of a converted utterance on a specific linguistic content. By propagating the prediction error of the speech-to-text module through the weights of the conversion network, the latter is constrained to generate mel-spectrograms from which the speech-to-text module can predict the correct character sequence, thus forcing intelligibility. We proposed several versions of this upgraded voice conversion network with different strengths of the text-to-speech module. Undoubtedly, the speech-to-text module's addition significantly reduces the loss of intelligibility in the conversions. The same trend is observed

in terms of objective proximity to the target utterances, however the differences did not allow to draw conclusion on the naturalness of the attitudes conveyed in conversions.

Indeed, whatever the final application of these algorithms, the only criterion that really matters is the average individual's opinion on the conversions. We thus conducted a listening experiment to perceptually assess the various configurations' conversions. We asked participants to judge the intelligibility of the conversions and the fidelity of the attitude they convey to a specific attitude label. First, the test results proved that the incorporation of a speech-to-text module into the transformer-based attitude conversion algorithm allows to globally improve the intelligibility of the conversions. Second, it appears that the participants judged the attitude conveyed in conversions from linguistically conditioned algorithms as being even more true to the associated attitude label. However, it is important here to moderate these conclusions since our experiment involves only one speaker and does not allow us to validate the performance of such a model in a more general context. Nevertheless, this research marks a substantial advance in that it makes it possible to learn the conversion of any attitude to any other simultaneously, i.e. in a many-to-many fashion, which to our knowledge had never been achieved. We are currently conducting further experiments involving several speakers in order to observe whether the algorithm can learn to convert speech attitudes independently of the speaker. At the time of writing, we dispose of a powerful algorithm for the conversion of vocal attitudes and are working to improve it by extending its scope of validity.

Chapter 8

GENERAL CONCLUSION & FURTHER DIRECTIONS

8.1 General Conclusions

First and foremost, our research has shed light on vocal attitudes by revealing how people use their vocal apparatus to produce them and, to a lesser extent, how they manage to decode them. This valuable knowledge has enabled us to thoroughly address the challenge of converting vocal attitudes. On the one hand, we worked at designing an algorithm that learns to convert speech attitudes with accuracy - by changing all the speech parameters involved in their description - and with efficiency - through architectural and optimisation choices. On the other hand, the knowledge provided on attitudes enabled us to feed our attitude conversion algorithm with clean and controlled data which improved the quality of conversions. The main contributions consist of: 1) the design of Att-HACK - a French database of expressive speech for social attitudes, 2) the uncovering of the production strategies and perception of vocal attitudes, 3) the design of a BWS-Net - neural predictor of perceptual judgements on attitudes and 4) the design of a sequence-to-sequence algorithm for speech attitude conversion.

The main conclusions of the present study are summarized below:

Individuals Share Common Strategies for Producing Vocal Attitudes

In the study presented in Chapter 4.3, we investigated how French speakers modulate their voice to communicate vocal attitudes. To do this, we analysed the vocal production of dominant, friendly, seductive and distant attitudes in Att-HACK. For each attitude, we reported the changes in the speakers' vocal fold behaviour, vocal tract actuation, and phonetic speech structure. We obtained two statistically strong prototypes for dominance and seductiveness and two weaker ones for friendliness and distance. To our knowledge, we conducted the first study to reveal diverging voice production strategies at the articulatory level. Specifically, we found that speakers' productions were distributed across specific clusters in the vowel space (Fig. 4.2-2). Similarly, analysing the Vowel Space Density surface revealed that some attitudes span more articulatory modes than others. This result suggests that subtle cues in speech articulation can convey a communicative signal of vocal intent.

Overall, these results shed light on the social intentions behind the production of social attitudes. Such behaviours may be closely interpreted from a social perspective, revealing the links between attitude-specific vocal behaviours and higher-order cognitive mechanisms (Goupil et al., 2021b). However, it is important to highlight that the vocalisations analysed herein were produced by actors, and actors' vocalisations are known to be less authentic than spontaneous ones (Anikin and Lima, 2017) — which, in the case of e.g. facial expressions of emotions, even seem to rely on different neural bases (Valente et al., 2017). In any case, these results uncover the shared strategies used by speakers to volitionally produce vocal attitudes.

Linguistic Content and Speaker Gender Influence the Perception of Vocal Attitudes

We conducted a Best Worst Scaling (BWS) experiment to assess the perception of attitudes in Att-HACK. As a first step, we assessed solely the perception of its related sounds for each a priori attitude, i.e. the ones already produced with aim of conveying this specific attitude. In the end, we obtained four perceptual scales ranking the sounds of each a priori attitude. The perceptual data collected potentially reflects other speech attributes such as linguistic content or gender that influence the communication of speech attitudes. Before seeking to properly understand the perception of speech attitudes, we questioned the interaction between perceptual scores obtained and other speech attributes such as linguistic content of gender.

The sentences in the database cannot be regarded as neutral, they have a meaning that denotes either a rather positive or negative sentiment. Reflecting this evoked sentiment, we assigned an emotional valence score to each sentence through a sentiment analysis conducted on 60 individuals. We found a substantial interaction between the perception of attitude and emotional valence score with significant effects for friendliness and dominance, thus revealing that those attitudes perception is influenced by linguistic content. In particular, the perception of friendliness is significantly correlated to the emotional valence carried by said linguistic content, i.e. the more the sentence evokes a positive emotion, the more the utterance will be perceived as friendly.

We found that certain attitudes are better communicated depending on the speakers' vocal gender. Friendliness and seduction seem to be best communicated by female speakers. Conversely, dominance seems to best communicated by male speakers. We provide two possible explanations for this gender effect. First, it may be that the better communication of certain attitudes by speakers of one gender is due to a *production* advantage, i.e. a better physiological capacity to produce and thus communicate these attitudes. Second, it may be caused by a decoding bias, i.e. a culturally constructed difference in the *perception* of each of these attitudes depending on whether it is expressed by a male or female speaker. Attributing one or the other of these causes to the different effects observed is very difficult and would require looking at the mental representations that individuals attribute to different attitudes. Moreover, it is unclear whether these effects would exist outside the forced-choice experiment involved in the BWS paradigm.

The Conversion of Attitudes Is Improved with Linguistic Conditioning

Early experiments at converting speech attitudes with the promising transformer approach showed issue of linguistic loss encountered in conversions. To solve this, we moved to the typical research problem of Automatic Speech Recognition (ASR) which consists in transcribing audio speech segments into a sequence of characters. We first worked at implementing an efficient speech-to-text

module and trained it on Att-HACK for proper objective assessment of the linguistic loss issue. Second, we considered how we can use such a text-to-speech module to enhance the performance of our conversion model. We proposed to connect the speech-to-text's speech input to the output of the conversion algorithm, in this way we mean to condition the generation of a converted utterance on a specific linguistic content. By propagating the prediction error of the speech-to-text module through the weights of the conversion network, the latter is constrained to generate mel-spectrograms from which the speech-to-text module can predict the correct character sequence, thus forcing intelligibility. We proposed several versions of this upgraded voice conversion network with different strengths of the speech-to-text module. Undoubtedly, the speech-to-text module's addition significantly reduced the loss of intelligibility in the conversions. The same trend was observed in terms of objective proximity to the target utterances, although the differences did not allow to draw conclusion on the naturalness of the attitudes conveyed in conversions.

Whatever the final application of these algorithms, the only criterion that really matters is the average individual's opinion on the conversions. We thus conducted a listening experiment to perceptually assess the various configurations' conversions. We asked 150 participants to judge the intelligibility of the conversions and the fidelity of the attitude they convey to a specific attitude label. First, the test results proved that the incorporation of a speech-to-text module into the transformer-based attitude conversion algorithm allows to globally improve the intelligibility of the conversions. Second, it appears that the participants judged the attitude conveyed in conversions from linguistically conditioned algorithms as being even more true to the associated attitude label. However, it should be noticed that individuals still perceive a significant difference between the real utterances and the conversions from our best model in terms of fidelity to an attitude label. Since our experiment only includes one speaker and prevents us from validating the performance of such a model in a more general context, it is crucial to mitigate these conclusions. However, this research represents a significant advancement since it allows to learn the conversion of any attitude to any other simultaneously, i.e. in a many-to-many fashion, which had previously not been achieved to our knowledge. To determine if the algorithm can learn to convert speech attitudes independently of the speaker, we are currently conducting more experiments with multiple speakers. As of this writing, we dispose of a powerful algorithm for converting speech attitudes and are attempting to improve it by broadening its scope.

8.2 Further Directions

Through pursuing the objective of converting vocal attitudes, many research directions have been explored. Although this thesis ends, the research continues: many questions remain to be addressed and many ideas are still to be implemented, the main ones are outlined below.

Understanding the Perception of Speech Attitudes

It is clear that static features - i.e. averaged over utterances' duration - are not predictive of BWS scores for speech attitudes as they are for sound attributes such as studied in (Rosi, 2022). It is likely that individuals use much more complex cues to decode attitudes. In particular, it is reasonable to hypothesize that temporal variations in different speech parameters play a crucial role in the perception of speech attitudes. In order to assess this hypothesis, we plan to adapt the principle of explained regression to the case of temporal sequences. By doing so, we expect to better understand the mechanisms underlying the perception of vocal attitudes.

Better Assessment of Conversions

In Section 4.3, we investigated how speakers modulate their voice to communicate vocal attitudes. To do, we analysed the vocal production of dominant, friendly, seductive and distant attitudes in Att-HACK by reporting the changes in the speakers' vocal fold behaviour, vocal tract actuation, and phonetic speech structure. We plan to re-apply this method, which involves a phonemes-to-audio neural alignment of the speech samples, to the conversions from our attitude conversion model. By comparing the results with the production strategies uncovered for the actual data, we could tell exactly which speech parameters were converted and which were not (and to what extent), thus providing an accurate objective assessment of the model's performance.

In Chapter 6, we worked at designing a BWS-Net that can mimic the average perceptual judgment of individuals from mel-spectrogram representations of speech signals. We plan to apply such a BWS-Net to the conversions yielded by our algorithms, thus assessing their subjective perception without the need for a listening experiment.

Perceptually Conditioned Speech Attitude Conversion

We are working on a perceptual conditioning of our speech attitude conversion algorithm that provides control on the attitudinal intensity of the conversion. At the time of writing, we conduct experiments with such an algorithm testing the different BWS-Nets designed Chapter 6. These experiments notably involve considering several speakers, in order to cover the whole perceptual spectrum for each attitude (and each sentence). In addition, each attitude requires its own BWS-Net, which makes the training considerably more demanding in terms of resources and computing time. We hope that these experiments will lead to improvements in our current attitude conversion system, particularly through attitudinal intensity control.

Bibliography

Abe, M., Nakamura, S., Shikano, K., and Kuwabara, H. (1988). Voice conversion through vector quantization. In *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, pages 655–658 vol.1.

Aden, J. (2013). Apprendre les langues par corps. In *POUR UN THÉÂTRE-MONDE. Plurilinguisme, interculturalité et transmission*. Presses Universitaires de Bordeaux.

Airaksinen, M., Juvela, L., Bollepalli, B., Yamagishi, J., and Alku, P. (2018). A comparison between straight, glottal, and sinusoidal vocoding in statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(9):1658–1670.

Ajzen, I. and Fishbein, M. (1980). *Understanding attitudes and predicting social behaviour*. Prentice-Hall, Englewood Cliffs, New Jersey.

Allport, G. (1935). *Attitudes*. Clark University Press, Worcester, MA.

Amir, N. and Globerson, E. (2014). On the role of pitch in perception of emotional speech. *Proceedings of the International Conference on Speech Prosody*, pages 154–158.

Andersen, P. (1999). *Nonverbal Communication: Forms and Functions*. McGraw-Hill.

Anikin, A. and Lima, C. (2016). Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations. *The Quarterly Journal of Experimental Psychology*, 71:1–53.

Anikin, A. and Lima, C. F. (2017). Perceptual and acoustic differences between authentic and acted nonverbal emotional vocalizations. *Q. J. Exp. Psychol. (Hove)*, page 17470218.2016.1.

Anikin, A., Pisanski, K., and Reby, D. (2020). Do nonlinear vocal phenomena signal negative valence or high emotion intensity? *Royal Society open science*, 7(12):201306.

Anikin, A., Pisanski, K., and Reby, D. (2022). Static and dynamic formant scaling conveys body size and aggression. *Royal Society Open Science*, 9(1):211496.

Arias, P., Rachman, L., Liuni, M., and Aucouturier, J.-J. (2021). Beyond correlation: Acoustic transformation methods for the experimental study of emotional voice and speech. *Emotion Review*, 13(1):12–24.

Arias, P., Soladie, C., Bouafif, O., Roebel, A., Seguier, R., and Aucouturier, J.-J. (2018). Realistic transformation of facial and vocal smiles in real-time audiovisual streams. *IEEE Transactions on Affective Computing*, 11(3):507–518.

Arias Sarah, P. (2018). *The cognition of auditory smiles : a computational approach*. Theses, Sorbonne Université.

Arias2020, P., Soladié, C., Bouafif, O., Roebel, A., Séguier, R., and Aucouturier, J.-J. (2020). Realistic transformation of facial and vocal smiles in real-time audiovisual streams. *IEEE Transactions on Affective Computing*, 11(3):507–518.

Aristotle. and Lord, C. (1984). *The politics / Aristotle ; translated, and with an introduction, notes, and glossary by Carnes Lord*. University of Chicago Press Chicago.

Arnold, M. (1960). *Emotion and personality*.

Assunção, G., Menezes, P., and Perdigão, F. (2020). Speaker awareness for speech emotion recognition. *International Journal of Online and Biomedical Engineering (iJOE)*, 16:15.

Aucouturier, J.-J. and Canonne, C. (2017). Musical friends and foes: The social cognition of affiliation and control in improvised interactions. *Cognition*, 161:94–108.

Aucouturier, J.-J., Johansson, P., Hall, L., Segnini, R., Mercadié, L., and Watanabe, K. (2016). Covert digital manipulation of vocal emotion alter speakers' emotional states in a congruent direction. *Proceedings of the National Academy of Sciences of the United States of America*, 113(4):948–953.

Bachorowski, J.-A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science*, 8(2):53–57.

Bachorowski, J.-A. and Owren, M. J. (1995). Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychological Science*, 6(4):219–224.

Badshah, A. (2017). *A Study of Hand-Crafted and Deep Features for Speech Emotion Recognition*. PhD thesis, Sejong University.

Baldwin, J. M. (1895). *The origin of motor attitudes and expressions*. Clark University Press, Worcester, MA.

Banse, R. and Scherer, K. (1996). Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70:614–36.

Barrett, L. (2011). Was darwin wrong about emotional expressions? *Current Directions in Psychological Science*, 20:400–406.

Barrett, L. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12.

Barrett, L. F. (2006). Are emotions natural kinds? *Perspectives on Psychological Science*, 1(1):28–58. PMID: 26151184.

Barrett, L. F. (2015). Ten common misconceptions about psychological construction theories of emotion. *The psychological construction of emotion*, page 45–79.

Belyk, M., Waters, S., Kanber, E., Miquel, M., and Mcgettigan, C. (2022). Individual differences in vocal size exaggeration. *Scientific Reports*, 12.

Black, A., Zen, H., and Tokuda, K. (2007). Statistical parametric speech synthesis. In *International Conference on Audio, Speech, and Signal Processing*, pages 1229–1232.

Bodenhausen, G. (1993). Emotions, arousal, and stereotype-based discrimination: a heuristic model of affect and stereotyping. *Affect, cognition, and stereotyping: Interactive processes in group perception*, pages 13–35.

Bourdieu, P. (1972). *Esquisse d'une théorie de la pratique. Paris: Seuil*.

Bous, F., Benaroya, L., Obin, N., and Roebel, A. (2022). Voice reenactment with f0 and timing constraints and adversarial learning of conversions. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 389–393.

Bous, F. and Roebel, A. (2022). A bottleneck auto-encoder for f0 transformations on speech and singing voice. *Information*, 13(3).

Bradshaw, D. (1986). *Immediate and prolonged effectiveness of negative emotion expressions in inhibiting infants' actions*. PhD thesis, University of Denver.

Brennan, S. E. and Schober, M. F. (2001). How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, 44(2):274–296.

Bryant, G., Fessler, D., Fusaroli, R., Clint, E., Amir, D., Chavez Cosamalon, B., Denton, K., Díaz, C., Duran, L., Fančovičová, J., Fux, M., Ginting, E., Hasan, Y., Hu, A., Kamble, S., Kameda, T., Kuroda, K., Li, N., Luberti, F., and Zhou, Y. (2018). The perception of spontaneous and volitional laughter across 21 societies. *Psychological Science*, 29:095679761877823.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. (2005). A database of german emotional speech. volume 5, pages 1517–1520.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower Provost, E., Kim, S., Chang, J., Lee, S., and Narayanan, S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Bänziger, T. and Scherer, K. (2005). The role of intonation in emotional expressions. *Speech Communication*, 46:252–267.

Camacho, A. (2007). *SWIPE: A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music*. PhD. thesis, University of Florida.

Campbell, N. and Erickson, D. (2004). What do people hear? a study of the perception of non-verbal affective information in conversational speech(emotion in speech). *Journal of the Phonetic Society of Japan*, 8:9–28.

Cannizzaro, M., Harel, B., Reilly, N., Chappell, P., and Snyder, P. (2004). Voice acoustical measurement of the severity of major depression. *Brain and cognition*, 56:30–5.

Chen, F., Li, A., Wang, H., Wang, T., and Fang, Q. (2004). Acoustic analysis of friendly speech. volume 1, pages I – 569.

Chen, L., Braunschweiler, N., and Gales, M. (2015). Speaker and expression factorization for audiobook data: Expressiveness and transplantation. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 23:605–618.

Chen, M., He, X., Yang, J., and Zhang, H. (2018). 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, 25(10):1440–1444.

Chen, T. and Guestrin, C. (2016). XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

Chen, Z. and Zhang, P. (2021). TVQVC: Transformer based vector quantized variational autoencoder with CTC loss for voice conversion. pages 826–830.

Choi, Y., Jung, Y., and Kim, H. (2020). Deep MOS predictor for synthetic speech using cluster-based modeling. In *Interspeech 2020*. ISCA.

Choi, Y., Jung, Y., and Kim, H. (2021). Neural MOS prediction for synthesized speech using multi-task learning with spoofing detection and spoofing type classification. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 462–469.

Chong, C. S., Kim, J., and Davis, C. (2018). Disgust expressive speech. *Speech Commun.*, 98(C):68–72.

Chou, J.-C., chieh Yeh, C., and yi Lee, H. (2019). One-shot voice conversion by separating speaker and content representations with instance normalization. In *Proc. Interspeech 2019*.

Chuenwattanapranithi, S., Xu, Y., Thipakorn, B., and Maneewongvatana, S. (2007). The roles of pitch contour in differentiating anger and joy in speech. *International journal of signal processing*, 3:129–134.

Collins, S. A. (2000). Men's voices and women's choices. *Animal Behaviour*, 60(6):773–780.

Couper-Kuhlen, E. (1986). *An Introduction to English Prosody*. Edward Arnold.

de Saussure, F. (1916). *Cours de linguistique générale*. Payot, Paris.

Degottex, G., Lanchantin, P., Röbel, A., and Rodet, X. (2013). Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis. *Speech Communication*, 55:278–294.

Depalle, P., García, G., and Rodet, X. (1994). A virtual castrato (!?). In *ICMC*.

Desai, S., Raghavendra, E. V., Yegnanarayana, B., Black, A. W., and Prahallad, K. (2009). Voice conversion using artificial neural networks. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3893–3896.

Drahota, A., Costall, A., and Reddy, V. (2008). The vocal communication of different kinds of smile. *Speech Communication*, 50(4):278–287.

Eagly, A. and Chaiken, S. (1993). *The psychology of attitudes*. Harcourt Brace Jovanovich.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Ekman, P. (1999). Basic emotions. In Dalgleish, T. and Powers, M. J., editors, *Handbook of Cognition and Emotion*, pages 4–5. Wiley.

Ekman, P., Levenson, R. W., and Friesen, W. V. (1983). Autonomic nervous system activity distinguishes among emotions. *Science*, 221(4616):1208–1210.

Elfenbein, H. A. and Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128 2:203–35.

Engel, J., Hantrakul, L. H., Gu, C., and Roberts, A. (2020). DDSP: Differentiable digital signal processing. In *International Conference on Learning Representations*.

Erhard, K. A., Kotz, S. A., Pfeifer, E., Besson, M., Friederici, A. D., and Matiasek, J. (1999). On the relations of semantic and acoustic properties of emotions. In *In: Proceedings of the XIVth International Congress of Phonetic Sciences*, pages 2121–2124.

Fang, F., Yamagishi, J., Echizen, I., and Lorenzo-Trueba, J. (2018). High-quality nonparallel voice conversion based on cycle-consistent adversarial network. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5279–5283.

Fant, G. (1970). *Acoustic theory of speech production*. Walter de Gruyter.

Fant, G. (1981). The source filter concept in voice production. *STL-QPSR*, 22:021–037.

Feinberg, D., Jones, B., Little, A., Burt, D., and Perrett, D. (2005). Manipulation of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal Behaviour*, 69:561–568.

Ferro, R., Obin, N., and Roebel, A. (2020). CycleGAN Voice Conversion of Spectral Envelopes using Adversarial Weights. In *Eusipco*, Amsterdam, Netherlands.

Fineberg, J. (2006). Lolita. http://brahms.ircam.fr/works/work/18304/.

Fiske, S., Cuddy, A., and Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, 11:77–83.

Fitch, W., Neubauer, J., and Herzel, H. (2002). Calls out of chaos: The adaptive significance of nonlinear phenomena in mammalian vocal production. *Animal Behaviour*, 63:407–418.

Flanagan, J. L. and Golden, R. M. (1966). Phase vocoder. *The Bell System Technical Journal*, 45(9):1493–1509.

Friedman, H. and Friedman, E. (1997). A comparison of six overall evaluation rating scales. *Journal of International Marketing and Marketing Research*, 22:129–138.

Frijda, N. (1999). Moods, emotion episodes, and emotions. *Handbook of emotions*, page 381–403.

Fu, S.-W., Tsao, Y., Hwang, H.-T., and Wang, H.-m. (2018). Quality-net: An end-to-end non-intrusive speech quality assessment model based on BLSTM. pages 1873–1877.

G. Castellano, M. Mancini, C. P. P. W. M. (2012). Expressive copying behavior for social agents: a perceptual analysis. *IEEE Trans Syst, Man Cybern*, 42(3).

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., and Zue, V. (1992). Timit acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium*.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1243–1252. PMLR.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press.

Gerazov, B., Bailly, G., Mohammed, O., Xu, Y., and Garner, P. (2018). Embedding context-dependent variations of prosodic contours using variational encoding for decomposing the structure of speech prosody. In *Proc. Workshop on Prosody and Meaning: Information Structure and Beyond*, Aix-en- Provence, France.

Gervasoni, S. (2007-2008). Com que voz. http://brahms.ircam.fr/works/work/19824/.

Gibbon, D. and Gut, U. (2001). Measuring speech rhythm. In *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, pages 95–98.

Gibiansky, A., Arik, S., Diamos, G., Miller, J., Peng, K., Ping, W., Raiman, J., and Zhou, Y. (2017). Deep voice 2: Multi-speaker neural text-to-speech. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Goupil, L., Johansson, P., Hall, L., and Aucouturier, J.-J. (2019). Influence of vocal feedback on emotions provides causal evidence for the self-perception theory. *bioRxiv*.

Goupil, L., Johansson, P., Hall, L., and Aucouturier, J.-J. (2021a). Vocal signals only impact speakers' own emotions when they are self-attributed. *Consciousness and Cognition*, 88:103072.

Goupil, L., Ponsot, E., Richardson, D., Reyes, G., and Aucouturier, J.-J. (2021b). Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature. *Nat. Commun.*, 12(1):861.

Griffin, D. and Lim, J. (1984). Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243.

Guerouaou, N., Vaiva, G., and Aucouturier, J.-J. (2021). The shallow of your smile: the ethics of expressive vocal deep-fakes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(20210083):1 – 5.

He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T.-Y., and Ma, W.-Y. (2016). Dual learning for machine translation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 820–828, Red Hook, NY, USA. Curran Associates Inc.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9:1735–80.

Hoffer, E. and Ailon, N. (2015). Deep metric learning using triplet network. In Feragen, A., Pelillo, M., and Loog, M., editors, *Similarity-Based Pattern Recognition*, pages 84–92, Cham. Springer International Publishing.

Holler, J. and Levinson, S. (2019). Multimodal language processing in human communication. *Trends in Cognitive Sciences*, 23.

House, D. (2005). Phrase-final rises as a prosodic feature in wh-questions in swedish human-machine dialogue. *Speech Communication*, 46:268–283.

Huang, C.-y., Lin, Y. Y., Lee, H.-y., and Lee, L.-s. (2021). Defending your voice: Adversarial attack on voice conversion. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 552–559.

Huang, W.-C., Hayashi, T., Wu, Y.-C., Kameoka, H., and Toda, T. (2020). Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining. In *Proc. Interspeech*.

Izard, C. (2008). Emotion theory and research: Highlights, unanswered questions, and emerging issues. *Annual review of psychology*, 60:1–25.

Izard, C., Youngstrom, E., Fine, S., Mostow, A., and Trentacosta, C. (2006). *Emotions and Developmental Psychopathology*, pages 244–292.

Izard, C. E. (2007). Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on Psychological Science*, 2(3):260–280. PMID: 26151969.

James, W. (1884). What is an emotion? *Mind*, 9(34):188–205.

Jiao, Y., Gabrys, A., Tinchev, G., Putrycz, B., Korzekwa, D., and Klimkov, V. (2021). Universal neural vocoding with parallel wavenet.

Jin, Z., Finkelstein, A., DiVerdi, S., Lu, J., and Mysore, G. J. (2016). Cute: A concatenative method for voice conversion using exemplar-based unit selection. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5660–5664.

Kain, A. and Macon, M. (1998). Spectral voice conversion for text-to-speech synthesis. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 1, pages 285–288 vol.1.

Kalluri, S. B., Vijayasenan, D., Ganapathy, S., M, R. R., and Krishnan, P. (2021). Nisp: A multi-lingual multi-accent dataset for speaker profiling. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6953–6957.

Kameoka, H., Huang, W.-C., Tanaka, K., Kaneko, T., Hojo, N., and Toda, T. (2021). Many-to-many voice transformer network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:656–670.

Kameoka, H., Tanaka, K., Kwaśny, D., Kaneko, T., and Hojo, N. (2020). Convs2s-vc: Fully convolutional sequence-to-sequence voice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1849–1863.

Kaneko, T. and Kameoka, H. (2018). Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2100–2104.

Kawahara, H. (1997). Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1303–1306 vol.2.

Kawahara, H. (2006). Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology*, 27(6):349–353.

Kawanami, H., Iwami, Y., Toda, T., Saruwatari, H., and Shikano, K. (2003). Gmm-based voice conversion applied to emotional speech synthesis. *8th European Conference on Speech Communication and Technology (Eurospeech 2003)*.

Keltner, D., Gruenfeld, D. H., and Anderson, C. (2003). Power, approach, and inhibition. *Psychological review*, 110 2:265–84.

Keren, G. and Schuller, B. (2016). Convolutional rnn: An enhanced model for extracting features from sequential data. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 3412–3419.

Lange, N. (1888). Neue experimente über den vorgang der einfachen reaktion auf sinneseindrücke. pages 479–510.

Lanza, M. and Pasquet, O. (2009). Häxan, la sorcellerie à travers les âges. http://brahms.ircam.fr/works/work/23986/.

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., and Winther, O. (2016). Autoencoding beyond pixels using a learned similarity metric. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1558–1566, New York, New York, USA. PMLR.

Latorre, J. and Akamine, M. (2008). Multilevel parametric-base F0 model for speech synthesis. In *Interspeech*, pages 2274–2277, Brisbane, Australia.

Lazarus, R. (1968). Emotions and adaptation: Conceptual and empirical relations. *Nebraska Symposium on Motivation*, page 175–266.

Le Moine, C. and Obin, N. (2020). Att-HACK: An Expressive Speech Database with Social Attitudes. In *Speech Prosody*, Tokyo, Japan.

Le Moine, C., Obin, N., and Roebel, A. (2021a). Speaker Attentive Speech Emotion Recognition. In *Proc. Interspeech 2021*, pages 2866–2870, Brno, Czech Republic.

Le Moine, C., Obin, N., and Roebel, A. (2021b). Towards end-to-end F0 voice conversion based on Dual-GAN with convolutional wavelet kernels. In *EUSIPCO*, Dublin (virtual), Ireland.

Leary, T. (1957). *Interpersonal diagnosis of personality; a functional theory and methodology for personality evaluation*. Ronald Press Co New York.

Lee, J. and Tashev, I. (2015). High-level feature representation using recurrent neural network for speech emotion recognition. In *Proc. Interspeech 2015*, pages 1537–1540.

Lee, S.-H., Noh, H.-R., Nam, W.-J., and Lee, S.-W. (2022). Duration controllable voice conversion via phoneme-based information bottleneck. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1173–1183.

Leech, G. (1983). *Principles of Pragmatics*. London: Sage Longman.

Levenson, R. W. (1999). The intrapersonal functions of emotion. *Cognition & Emotion*, 13:481–504.

Levenson, R. W., Carstensen, L. L., Friesen, W. V., and Ekman, P. (1991). Emotion, physiology, and expression in old age. *Psychology and aging*, 6 1:28–35.

Li, A.-j. and Wang, H. (2004). Friendly speech analysis and perception in standard chinese. In *Proc. Interspeech*.

Li, Y., Zhao, T., and Kawahara, T. (2019). Improved end-to-end speech emotion recognition using self attention mechanism and multitask learning. pages 2803–2807.

Lipovetsky, S. and Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17:319 – 330.

Liu, R., Sisman, B., Schuller, B., Gao, G., and Li, H. (2022). Accurate Emotion Strength Assessment for Seen and Unseen Speech Based on Data-Driven Deep Learning. In *Proc. Interspeech 2022*, pages 5493–5497.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Lo, C.-C., Fu, S.-W., Huang, W.-C., Wang, X., Yamagishi, J., Tsao, Y., and Wang, H.-M. (2019). MOSNet: Deep learning-based objective assessment for voice conversion. In *Interspeech 2019*. ISCA.

Louviere, J. J., Flynn, T. N., and Marley, A. A. J. (2015). *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press.

Luo, Z., Chen, J., Takiguchi, T., and Ariki, Y. (2017). Emotional voice conversion using neural networks with arbitrary scales f0 based on wavelet transform. *EURASIP Journal on Audio, Speech, and Music Processing*, 2017(1):18.

Luo, Z., Chen, J., Takiguchi, T., and Ariki, Y. (2019). Emotional voice conversion using dual supervised adversarial networks with continuous wavelet transform f0 features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(10):1535–1548.

Luo, Z., Takiguchi, T., and Ariki, Y. (2016). Emotional voice conversion using deep neural networks with MCC and f0 features. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pages 1–5.

Luong, H.-T. and Yamagishi, J. (2019). Bootstrapping non-parallel voice conversion from speaker-adaptive text-to-speech. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 200–207.

Mackie, D., Devos, T., and Smith, E. (2000). Intergroup emotions: Explaining offensive action tendencies in an intergroup context. *Journal of personality and social psychology*, 79:602–16.

Mairesse, F., Walker, M., Mehl, M., and Moore, R. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *J. Artif. Intell. Res. (JAIR)*, 30:457–500.

Mao, Q., Dong, M., Huang, Z., and Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia*, 16(8):2203–2213.

Markel, J. E. and Gray, A. H. (1982). *Linear Prediction of Speech*. Springer-Verlag, Berlin, Heidelberg.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., and Sagot, B. (2020). CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Mathews, M. V. and Pierce, J. R., editors (1989). *Current Directions in Computer Music Research*. MIT Press, Cambridge, MA, USA.

Matsumoto, D., Hirayama, S., and LeRoux, J. A. (2006). *Psychological Skills Related to Intercultural Adjustment*, pages 387–405. Springer US, Boston, MA.

McAleer, P., Todorov, A., and Belin, P. (2014). How do you say 'hello'? personality impressions from brief novel voices. *PloS one*, 9:e90779.

McAulay, R. J. and Quatieri, T. F. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoust. Speech Signal Process.*, 34:744–754.

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

McKeown, G., Valstar, M., Pantic, M., and Cowie, R. (2010). The SEMAINE Corpus of Emotionally Coloured Character Interactions. In *IEEE International Conference on Multimedia and Expo (ICME)*.

Mehrabian, A. (1972). *Nonverbal communication*. Aldine-Atherton, New York.

Meng, H., Yan, T., Yuan, F., and Wei, H. (2019). Speech emotion recognition from 3D log-mel spectrograms with deep learning network. *IEEE Access*, 7:125868–125881.

Ming, H., Huang, D., Dong, M., , Li, H., Xie, L., and Zhang, S. (2015). Fundamental frequency modeling using wavelets for emotional voice conversion. In *Affective Computuing Intell. Interact.*, page 804–809.

Ming, H., Huang, D.-Y., Xie, L., Wu, J., Dong, M., and Li, H. (2016). Deep bidirectional LSTM modeling of timbre and prosody for emotional voice conversion. In *Proc. Interspeech 2016*, pages 2453–2457.

Mirsamadi, S., Barsoum, E., and Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2227–2231.

Miyoshi, H., Saito, Y., Takamichi, S., and Saruwatari, H. (2017). Voice conversion using sequence-to-sequence learning of context posterior probabilities. *ArXiv*, abs/1704.02360.

Morise, M., Yokomori, F., and Ozawa, K. (2016). World: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Trans. Inf. Syst.*, 99-D:1877–1884.

Morlec, Y. (1997). *Génération multiparamétrique de la prosodie du français par apprentissage automatique*. PhD. thesis, Institut de la Communication Parlée, Grenoble.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5):453–467. Neuropeech '89.

Nagrani, A., Chung, J. S., and Zisserman, A. (2017). Voxceleb: A large-scale speaker identification dataset. *Interspeech 2017*.

Nakashika, T., Takashima, R., Takiguchi, T., and Ariki, Y. (2013). Voice conversion in high-order eigen space using deep belief nets. *Proc. Interspeech*, pages 369–372.

Neumann, M. and Vu, N. T. (2017). Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. In *Proc. Interspeech 2017*, pages 1263–1267. ISCA.

Obin, N. (2011). *MeLos: Analysis and Modelling of Speech Prosody and Speaking Style*. PhD. Thesis, IRCAM - UPMC.

Obin, N. and Belião, J. (2018). Sparse coding of pitch contours with deep auto-encoders. In *International Conference on Speech Prosody*, pages 799–803.

Obin, N., Lacheret, A., and Rodet, X. (2011). Stylization and Trajectory Modelling of Short and Long Term Speech Prosody Variations. In *Proc. Interspeech 2011*, pages 2029–2032, Florence, Italy.

Obin, N., Pham, P., and Roebel, A. (2018). Conversion d'Identité de la Voix Chantée par Sélection et Concaténation d'Unités Spectrales. In *Proc. XXXIIe Journées d'Études sur la Parole*, pages 1–9.

Obin, N., Robinson, C., and Roebel, A. (2019). Sequence-to-sequence modelling of f0 for speech emotion conversion. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6830–6834.

Obin, N., Veaux, C., and Lanchantin, P. (2012). Making Sense of Variations: Introducing Alternatives in Speech Synthesis. In *Speech Prosody*, Shanghai, China.

Panksepp, J. (1987). The neurochemistry of behavior. *Annu Rev Psychol.*, pages 77–107.

Panksepp, J., Nocjar, C., Burgdorf, J., Panksepp, J., and Huber, R. (2004). The role of emotional systems in addiction: A neuroethological perspective. *Nebraska Symposium on Motivation. Nebraska Symposium on Motivation*, 50:85–126.

Park, J., Zhao, K., Peng, K., and Ping, W. (2019). Multi-speaker end-to-end speech synthesis.

Park, S.-w., Kim, D.-y., and Joe, M.-c. (2020). Cotatron: Transcription-guided speech encoder for any-to-many voice conversion without parallel data. In *Proc. Interspeech 2020*.

Parra, H. (2008-2009). Hypermusic: Prologue. http://brahms.ircam.fr/works/work/23852/.

Patton, B., Agiomyrgiannakis, Y., Terry, M., Wilson, K., Saurous, R. A., and Sculley, D. (2016). Automos: Learning a non-intrusive assessor of naturalness-of-speech. In *NIPS 2016 End-to-end Learning for Speech and Audio Processing Workshop*.

Petty, R., Cacioppo, J., and Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology*, 41:847–855.

Petty, R., Cacioppo, J., and Goldman, R. (1993). Assessing the structure of prejudicial attitudes: The case of attitudes toward homosexuals. *Journal of Personality and Social Psychology*, 65:1105–1118.

Pham, Q., Nguyen, T.-S., Niehues, J., Müller, M., and Waibel, A. (2019). Very deep self-attention networks for end-to-end speech recognition. In *Proc. Interspeech 2019*, pages 66–70.

Piazza, E. A., Iordan, M. C., and Lew-Williams, C. (2017). Mothers consistently alter their unique vocal fingerprints when communicating with infants. *Current Biology*, 27(20):3162–3167.

Pisanski, K., Anikin, A., and Reby, D. (2022). Vocal size exaggeration may have contributed to the origins of vocalic complexity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377.

Ponsot, E., Arias, P., and Aucouturier, J.-J. (2018a). Uncovering mental representations of smiled speech using reverse correlation. *The Journal of the Acoustical Society of America*, 143(1):EL19–EL24.

Ponsot, E., Burred, J. J., Belin, P., and Aucouturier, J.-J. (2018b). Cracking the social code of speech prosody using reverse correlation. *Proceedings of the National Academy of Sciences*, 115(15):3972–3977.

Puts, D., Hodges-Simeon, C., Cardenas, R., and Gaulin, S. (2007). Men's voices as dominance signals: vocal fundamental and formant frequencies influence dominance attributions among men. *Evolution and Human Behavior*, 28:340–344.

Puts, D. A., Gaulin, S. J., and Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior*, 27(4):283–296.

Qian, K., Zhang, Y., Chang, S., Cox, D., and Hasegawa-Johnson, M. (2020). Unsupervised speech decomposition via triple information bottleneck. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Qian, K., Zhang, Y., Chang, S., Yang, X., and Hasegawa-Johnson, M. (2019). AutoVC: Zero-shot voice style transfer with only autoencoder loss. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5210–5219. PMLR.

Quatieri, T. and McAulay, R. (1992). Shape invariant time-scale and pitch modification of speech. *IEEE Transactions on Signal Processing*, 40(3):497–510.

Rachman, L., Liuni, M., Arias, P., Lind, A., Johansson, P., Hall, L., Richardson, D., Watanabe, K., Dubal, S., and Aucouturier, J.-J. (2018). David: An open-source platform for real-time transformation of infra-segmental emotional cues in running speech. *Behavior Research Methods*, 50(1):323–343.

Ramet, G., Garner, P. N., Baeriswyl, M., and Lazaridis, A. (2018). Context-aware attention mechanism for speech emotion recognition. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 126–131.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Röbel, A. (2003). Transient detection and preservation in the phase vocoder. In *ICMC*.

Rodero, E. (2011). Intonation and emotion: Influence of pitch levels and contour type on creating emotions. *Journal of voice : official journal of the Voice Foundation*, 25:e25–34.

Rodet, X., Potard, Y., and Barrière, J.-B. (1984). The chant project: From the synthesis of the singing voice to synthesis in general. *Computer Music Journal*, 8(3):15–31.

Roebel, A. (2010). A shape-invariant phase vocoder for speech transformation. *13th International Conference on Digital Audio Effects, DAFx 2010 Proceedings*.

Roebel, A. and Bous, F. (2022). Neural vocoding for singing and speaking voices with the multi-band excited wavenet. *Information*, 13(3).

Rohmer, E. (2007). Les Amours d'Astrée et de Céladon. http://www.imdb.fr/title/tt0823240/.

Rosi, V. (2022). *The Metaphors of Sound: from Semantics to Acoustics - A Study of Brightness, Warmth, Roundness, and Roughness*. Theses, Sorbonne Université.

Rosi, V., Houix, O., and Susini, P. (2022). Best-worst scaling, an alternative method to assess perceptual sound qualities. *The Journal of the Acoustical Society of America*, 2:064404.

Ross, B. C. (2014). Mutual information between discrete and continuous data sets. *PLOS ONE*, 9(2):1–5.

Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

Salais, L., Arias, P., Le Moine, C., Rosi, V., Teytaut, Y., Obin, N., and Roebel, A. (2022). Production strategies of vocal attitudes. In *Proc. Interspeech 2022*, pages 4985–4989.

Salimans, T. and Kingma, D. P. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 901–909, Red Hook, NY, USA. Curran Associates Inc.

Sauter, D. A., Eisner, F., Ekman, P., and Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, 107(6):2408–2412.

Scherer, K. (1986). Vocal affect expression: a review and a model for future research. *Psychological bulletin*, 99 2:143–65.

Scherer, K. (1999). Appraisal theory. *Handbook of cognition and emotion*, page 637–663.

Schwarz, D. (2003). The caterpillar system for data-driven concatenative sound synthesis. In *Proc. DAFX 2003*.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., and Wu, Y. (2018). Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783.

Sidorov, M., Ultes, S., and Schmitt, A. (2014). Emotions are a personal thing: Towards speaker-adaptive emotion recognition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4803–4807.

Sini, A., Lolive, D., Barbot, N., and Alain, P. (2022). Investigating inter- and intra-speaker voice conversion using audiobooks. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7305–7313, Marseille, France. European Language Resources Association.

Sini, A., Lolive, D., Vidal, G., Tahon, M., and Delais-Roussarie, E. (2018). SynPaFlex-Corpus: An Expressive French Audiobooks Corpus Dedicated to Expressive Speech Synthesis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Sisman, B. and Li, H. (2018). Wavelet Analysis of Speaker Dependent and Independent Prosody for Voice Conversion. In *Proc. Interspeech 2018*, pages 52–56.

Sisman, B., Li, H., and Tan, K. C. (2017). Sparse representation of phonetic features for voice conversion with and without parallel data. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 677–684.

Skinner, B. (1953). *Science and human behavior*. Free Press, New York.

Spencer, H. (1860). *First Principles*. Cambridge University Press.

Spencer, H. (1881). *Principles of psychology*. Williams and Norgate London [England], 3rd ed. edition.

Srivastava, B. L. M., Vauquier, N., Sahidullah, M., Bellet, A., Tommasi, M., and Vincent, E. (2020). Evaluating voice conversion-based privacy protection against informed attackers. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2802–2806.

Story, B. H. and Bunton, K. (2017). Vowel space density as an indicator of speech performance. *The Journal of the Acoustical Society of America*, 141(5):EL458–EL464.

Stylianou, Y. and Cappe, O. (1998). A system for voice conversion based on probabilistic classification and a harmonic plus noise model. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 1, pages 281–284 vol.1.

Sun, L., Kang, S., Li, K., and Meng, H. (2015). Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4869–4873.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Tachibana, H., Uenoyama, K., and Aihara, S. (2018). Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Tamura, M., Masuko, T., Tokuda, K., and Kobayashi, T. (2001). Adaptation of pitch and spectrum for hmm-based speech synthesis using mllr. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 2, pages 805–808 vol.2.

Tanaka, K., Kameoka, H., Kaneko, T., and Hojo, N. (2019). Atts2s-vc: Sequence-to-sequence voice conversion with attention and context preservation mechanisms. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6805–6809.

Tang, H., Zhang, X., Wang, J., Cheng, N., and Xiao, J. (2022). Avqvc: One-shot voice conversion by vector quantization with applying contrastive learning. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4613–4617.

Teutenberg, J., Watson, C., and Riddle, P. (2008). Modelling and Synthesising F0 contours with the Discrete Cosine Transform. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 3973–3976, Las Vegas, U.S.A.

Teytaut, Y. and Roebel, A. (2021). Phoneme-to-audio alignment with recurrent neural networks for speaking and singing voice. *Proceedings of Interspeech 2021*, pages 61–65.

Toda, T., Black, A., and Tokuda, K. (2007). Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15:2222 – 2235.

Toda, T. and Tokuda, K. (2005). A speech parameter generation algorithm considering global variance for hmm-based speech synthesis. *IEICE Transactions on Information and Systems*, E90D.

Tooby, J. and Cosmides, L. (2008). The evolutionary psychology of the emotions and their relationship to internal regulatory variables. In *M. Lewis, J. M. Haviland-Jones, L. F. Barrett (Eds.), Handbook of emotions*, page 114–137. The Guilford Press.

Torrence, C. and Compo, G. P. (1998). A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79:61–78.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., and Zafeiriou, S. (2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204.

Uriel, W., Labov, W., and Herzog, M. (1968). *Empirical foundations for a theory of language change*. University of Texas Press.

Valente, D., Theurel, A., and Gentaz, E. (2017). The role of visual experience in the production of emotional facial expressions by blind people: a review. *Psychonomic bulletin & review*, 25.

Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. In *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, page 125.

Varela, F. (1996). Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies*, 3(4):330–49.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Veaux, C. and Rodet, X. (2011). Intonation conversion from neutral to expressive speech. In *Proc. Interspeech 2011*, pages 2765–2768.

Veaux, C., Yamagishi, J., and King, S. (2013). Towards personalised synthesised voices for individuals with vocal disabilities: Voice banking and reconstruction. In *Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies*, pages 107–111, Grenoble, France. Association for Computational Linguistics.

Ververidis, D. and Kotropoulos, C. (2004). Automatic speech classification to five emotional states based on gender information. In *2004 12th European Signal Processing Conference*, pages 341–344.

Vettin, J. and Todt, D. (2004). Laughter in conversation: Features of occurrence and acoustic structure. *Journal of Nonverbal Behavior*, 28:93–115.

Vogt, T. and André, E. (2006). Improving automatic emotion recognition from speech via gender differentiaion. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Wang, Q., Li, B., Xiao, T., Zhu, J., Li, C., Wong, D., and Chao, L. (2019). Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822.

Wang, X., Lorenzo-Trueba, J., Takaki, S., Juvela, L., and Yamagishi, J. (2018a). A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis. In *Proc. ICASSP 2018*, pages 4804–4808.

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q. V., Agiomyrgiannakis, Y., Clark, R. A. J., and Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. In *Proc. Interspeech 2017*.

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R. J., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y., and Saurous, R. A. (2018b). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *ICML*.

Wichmann, A. (2000). The attitudinal effects of prosody, and how they relate to emotion. In *ITRW on Speech and Prosody*, Newcastle, UK.

Wolff, D., Mignot, R., and Roebel, A. (2021). Audio defect detection in music with deep networks. In Lee, J. H., Lerch, A., Duan, Z., Nam, J., Rao, P., van Kranenburg, P., and Srinivasamurthy, A., editors, *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021*, pages 762–768.

Wu, D.-Y., Chen, Y.-H., and yi Lee, H. (2020). VQVC+: One-Shot Voice Conversion by Vector Quantization and U-Net Architecture. In *Proc. Interspeech 2020*, pages 4691–4695.

Wu, Y.-C., Hwang, H.-T., Hsu, C.-C., Tsao, Y., and Wang, H.-m. (2016). Locally linear embedding for exemplar-based spectral conversion. In *Proc. Interspeech 2016*, pages 1652–1656.

Wu, Z. and Li, H. (2014). Voice conversion versus speaker verification: an overview. *APSIPA Transactions on Signal and Information Processing*, 3:e17.

Wu, Z., Virtanen, T., Kinnunen, T., and Chng, E. (2013). Exemplar-based voice conversion using non-negative spectrogram deconvolution. In *8th ISCA Speech Synthesis Workshop*, pages 201–206.

Xia, Y., Qin, T., Chen, W., Bian, J., Yu, N., and Liu, T.-Y. (2017). Dual supervised learning. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3789–3798. PMLR.

Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., and Liu, T. (2020). On layer normalization in the transformer architecture. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10524–10533. PMLR.

Xu, Y., Lee, A., Wu, W.-L., Liu, X., and Birkholz, P. (2013). Human vocal attractiveness as signaled by body size projection. *PLOS ONE*, 8(4):1–9.

Yamagishi, Junichi; Veaux, C. M. K. (2017). Cstr VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. University of Edinburgh. The Centre for Speech Technology Research (CSTR).

Yi, Z., Zhang, H., Tan, P., and Gong, M. (2017). Dualgan: Unsupervised dual learning for image-to-image translation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2868–2876.

Yin, X., Lei, M., Qian, Y., Soong, F. K., He, L., Ling, Z.-H., and Dai, L.-R. (2016). Modeling F0 trajectories in hierarchically structured deep neural networks. *Speech Communication*, 76:82–92.

Zen, H., Tokuda, K., and Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064.

Zhang, J.-X., Ling, Z.-H., and Dai, L.-R. (2020). Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:540–552.

Zhang, J.-X., Ling, Z.-H., Liu, L.-J., Jiang, Y., and Dai, L.-R. (2019). Sequence-to-sequence acoustic modeling for voice conversion. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(3):631–644.

Zhang, L., Wang, L., Dang, J., Guo, L., and Yu, Q. (2018). *Gender-Aware CNN-BLSTM for Speech Emotion Recognition: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I*, pages 782–790.

Zhou, K., Sisman, B., Liu, R., and Li, H. (2021). Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 920–924.

Zhou, K., Sisman, B., Liu, R., and Li, H. (2022). Emotional voice conversion: Theory, databases and esd. *Speech Communication*, 137:1–18.

Zhou, K., Sisman, B., Zhang, M., and Li, H. (2020). Converting anyone's emotion: Towards speaker-independent emotional voice conversion. In *Proc. Interspeech 2020*, pages 3416–3420.

Zielinski, S., Hardisty, P., Hummersone, C., and Rumsey, F. (2007). Potential biases in mushra listening tests. *Audio Engineering Society - 123rd Audio Engineering Society Convention 2007*, 2.