Towards understanding the perceptions of warmth and competence in synthetic speech

vorgelegt von
M.Sc.
Sai Sirisha Rallabandi

an der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften

-Dr.-Ing.-

genehmigte Dissertation

Promotionsausschuss:
Vorsitzende: Prof. Dr. Marianne Maertens
Gutachter: Prof. Dr.-Ing. Sebastian Möller
Gutachter: Prof. Dr. Simon King
Gutachter: Prof. Dr. Yannis Stylianou
Gutachter: Prof. Dr. Oliver Niebuhr

Tag der wissenschaftlichen Aussprache: 29. März 2023

Berlin 2023

Abstract

Artificial Intelligence (AI) can already supersede humans in many tasks like playing ATARI, GO, chess, and many more. Apart from playing games, these AI models can also be used in Natural Language Processing (NLP) tasks such as language modeling (BERT). However, the applications of AI should also be directed toward solving much harder problems that would benefit mankind in situations like COVID. In my thesis, I describe the usage of AI for social good through its applications of synthetic speech. Artificial speech generation has achieved human-like-sounding speech by leveraging Neural models such as Tacotron and Wavenet. Even though the speech generation has evolved so much in the last decade, the evaluation of synthetic voices is still in its infancy. Much of the Text-to-Speech (TTS) and Voice Conversion (VC) research still evaluates the systems on naturalness, speech quality, and speaker similarity. In my thesis, I propose additional dimensions to be included in the evaluation of synthetic speech. I posit that these additional dimensions would aid in building socially acceptable synthetic voices. These additional dimensions are various speaker attributes that are relevant to different application domains.

The main contributions of my thesis are provided below.

- The perceptual dimensions representing various speaker attributes have been evaluated for synthetic speech. A factor analysis of these perceptual dimensions has provided 2 social speaker characteristics namely, warmth, and competence; and a personality trait, Extraversion.
- The acoustic analysis of the synthetic voices has provided the vocal cues of warmth (spectral flux, F1 mean, F2 mean for female speakers; F1 mean, loudness, the slope for male TTS voices) and competence (slope, flux for female synthetic voices; F0, voiced segment length for male TTS voices) in synthetic speech.
- Various VC and TTS experiments were carried out to enable the positive perceptions (highly warm/competent) of synthetic voices. The results of subjective tests display that achieving socially acceptable synthetic voices is possible.

Zusammenfassung

Künstliche Intelligenz (KI) kann den Menschen bereits bei vielen Aufgaben ersetzen, z. B. beim Spielen von ATARI, GO, Schach und vielen anderen. Abgesehen vom Spielen können diese KI-Modelle auch bei der Verarbeitung natürlicher Sprache (NLP) eingesetzt werden, etwa bei der Sprachmodellierung (BERT). Die Anwendungen der KI sollten jedoch auch auf die Lösung viel schwierigerer Probleme ausgerichtet sein, die der Menschheit in Situationen wie COVID zugute kommen würden. In meiner Dissertation beschreibe ich den Einsatz von KI für soziale Zwecke durch die Anwendung von synthetischer Sprache. Die künstliche Spracherzeugung hat durch den Einsatz von neuronalen Modellen wie Tacotron und Wavenet eine menschlich klingende Sprache hervorgebracht. Obwohl sich die Spracherzeugung in den letzten zehn Jahren stark weiterentwickelt hat, steckt die Bewertung synthetischer Stimmen noch in den Kinderschuhen. Ein Großteil der Forschung im Bereich Text-to-Speech (TTS) und Sprachumwandlung (VC) bewertet die Systeme immer noch nach Natürlichkeit, Sprachqualität und Sprecherähnlichkeit. In meiner Dissertation schlage ich zusätzliche Dimensionen vor, die in die Bewertung von synthetischer Sprache einbezogen werden sollten. Ich gehe davon aus, dass diese zusätzlichen Dimensionen bei der Entwicklung sozial akzeptabler synthetischer Stimmen helfen würden. Bei diesen zusätzlichen Dimensionen handelt es sich um verschiedene Sprechereigenschaften, die für verschiedene Anwendungsbereiche relevant sind.

Die wichtigsten Beiträge meiner Dissertation sind im Folgenden aufgeführt.

- Die Wahrnehmungsdimensionen, die verschiedene Sprechereigenschaften darstellen, wurden für synthetische Sprache ausgewertet. Eine Faktorenanalyse dieser Wahrnehmungsdimensionen ergab 2 soziale Sprechereigenschaften, nämlich Wärme und Kompetenz, sowie eine Persönlichkeitseigenschaft, Extraversion.
- Die akustische Analyse der synthetischen Stimmen hat die stimmlichen Anhaltspunkte für Wärme (spektraler Fluss, F1-Mittelwert, F2-Mittelwert für Frauen; F1-Mittelwert, Lautheit, Steigung für Männer) und Kompetenz (Steigung, Fluss für Frauen; F0, Länge des stimmhaften Segments für Männer) in synthetischer Sprache geliefert.

- Es wurden verschiedene VC- und TTS-Experimente durchgeführt, um die positive Wahrnehmung (sehr warm/kompetent) von synthetischen Stimmen zu ermöglichen. Die Ergebnisse der subjektiven Tests zeigen, dass es möglich ist, sozialverträgliche synthetische Stimmen zu erzeugen.

Acknowledgements

Firstly, I would like to express my sincere thankfulness to my Professor, Prof. Dr.-Ing. Sebastian Möller, for his continuous support, time, patience, and encouragement throughout my Ph.D. journey. His expertise and knowledge have greatly motivated me at various stages of my Ph.D. Thank you so much for constantly supporting me with my ideas and letting me experiment with various aspects of Speech Technology.

I will be grateful to David Suendermann-Oeft for introducing me to my Ph.D. supervisor at a conference back in 2018 (after a small conversation that did not even last for 10 minutes). You are absolutely correct in describing the research lab and the work environment. I loved everything about it and I am also planning to settle in Germany in the future.

I extend my gratitude to my thesis committee, Prof. Dr. Marianne Maertens, Prof. Dr.-Ing. Sebastian Möller, Prof. Dr. Simon King, Prof. Dr. Yannis Stylianou, and Prof. Dr. Oliver Niebuhr for all your time and encouragement. I have been inspired a lot by all your contributions to the field since I started speech research.

Special thanks to Dr.-Ing Babak Naderi, Dr.-Ing. Gabriel Mittag, Dr. Sai Krishna Rallabandi, Dr. Shrimai Prabhumoye, Avashna Govender, and Dr. Srikanth Ronanki for your time and valuable discussions at the initial stages of my Ph.D. Your insights have helped me a lot in shaping my thesis.

I would like to extend my thanks to all my colleagues at the Quality and Usability Lab, Dr.-Ing. Tanja Kojic, Vera Schmitt, Salar Mohtaj, Dr.-Ing. Steven Schmidt, Dr.-Ing. Saman Zadtootaghaj, Dr.-Ing.Thilo Michael, Wafaa Wardah, Prof. Dr.Ing. Michael Wagner, Robert Philipp Spang, and many more. Especially, for helping me with your feedback and guidance for my final presentation. I have also received a lot of support with respect to the teaching activities from you all.

Special thanks to Irene Hube-Achter, Yasmin Hillebrenner, and Tobias Jettkowski for all the administrative and technical support. I cannot thank you enough for all the timely help, support, and encouragement I have received from each one of you in this journey.

I extend my thanks to all the Bachelor's and Masters students who did their thesis with me during this journey. I have loved the collaboration, discussions, ideas, and approaches you have proposed in shaping your thesis. It gives me immense

pleasure that some of your works have also been published in various conferences or workshops. The datasets and the data analysis you have provided through your research projects will be very useful to the scientific community in the future.

Thanks to all my crazy friends I have met in Berlin. My life would have been very different if I had not met each of you. Thank you so much for everything!

Finally, I would like to thank my family for all the love, support, and encouragement. My brother, Dr. Sai Krishna Rallabandi, is my mentor in both my personal and professional life. Thank you so much for handling all my mood swings all these years. Thank you so much Suchi (sister-in-law) for being a very nice, friendly, kind-hearted, and a perfect match for my brother. Love you both for all the motivation and inspiration I get from you. Thanks to my parents, Seshamma (mother), and Seshacharyulu (father) for constantly instilling confidence in me, especially in the last days of my Ph.D.

Contents

Abbreviations

ANN	Artificial Neural Networks
AMT	Amazon Mechanical Turk
AGI	Artificial General Intelligence
AP	Aperiodicities
BFI	Big Five Inventory
CART	Classification and Regression Tree
CNN	Convolutional Neural Network
CFA	Confirmatory Factor Analysis
CBHG	Convolutional bank + highway network + GRU
DNN	Deep Neural Networks
dB	Decibels
EAR	Electronically Activated Recorded
EFA	Exploratory Factor Analysis
eGeMAPS	Extended Geneva Minimalistic Acoustic Parameters Set
F1,2,3	Formant frequncies 1,2,3
FC	Fully Connected layer
F0	Fundamental Frequency
GMM	Gaussian Mixture Model
GRU	Gated Recurrent Units
GAN	Generative Adversarial Networks
GPU	Graphics Processing Unit
GST	Global Style Tokens
H1-H4	Harmonic difference between 1 and 4
HNR	Harmonics-to-Noise Ratio
HI	Hammerberg Index
HMM	Hidden Markov Model
HCI	Human Computer Interaction
ITU-T	International Telecommunication Union - Telecommunication Standardization Sector
ICC	Intraclass Correlation
KL loss	Kullback-Leibler loss

LC	Linear Classifier
LDA	Linear Discriminant Analysis
LSTM	Long Short-Term Memory
LLD	Low-Level Descriptors
ML	Machine Learning
MLPG	Maximum Likelihood Parameter Generation
MOS	Mean Opinion Scores
MSE	Mean Squared Error
MCD	Mel Cepstral Distortion
MFCC	Mel Frequency Cepstral Coefficients
MCC	Mel Cesptral Coefficients
OCEAN	Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism
PCA	Principal Component Analysis
QoE	Quality of Experience
RNN	Recurrent Neural Networks
RMSE	Root Mean Square Error
RBF	Radial Basis Function
Relu	Rectified Linear Unit
SL	Significance level
SIRI	Speech Interpretation and Recognition Interface
SSC	Social Speaker Characteristics
SVM	Support Vector Machine
SVR	Support Vector Regressor
SPC	Speaker Personality Corpus
TTS	Text-to-Speech Synthesis
USS	Unit Selection Synthesis
UPM	USS with prosody modification
VC	Voice Conversion
V, UV	Voiced, Unvoiced
VAE	Variational Autoencoder
WC	Warmth Competence dataset

Chapter 1
Introduction

> ”AGI speaks to something deeply human—the idea that we can become more than we are, by building tools that propel us to greatness. And that’s really nice, except it also is a way to distract us from the fact that we have real problems that face us today that we should be trying to address using AI.”
>
> *Vilas Dhar, president of the Patrick J. McGovern Foundation*.

Artificial General Intelligence (AGI) or “general” AI is a branch of AI that focuses on the intellectual capabilities of humans. AI targets outperforming humans in every task. AGI aims at the human level of interpreting and solving problems even under uncertain scenarios. For instance, a machine translation task requires the machine to, a) understand both languages equally, b) possess the ability to translate the content and the intentions of the writer, and c) have adequate knowledge of the topic being discussed. All these tasks need to be performed simultaneously by the machine in order to achieve human-level intelligence or general intelligence. GATO - A generalist AI model has been recently developed by Deepmind[1][1] that could achieve this human-level performance (in executing multiple tasks). The model was designed to handle 604 distinctive tasks without requiring a change in the network and its weights. The tasks performed by the network are language modeling, playing ATARI, GO, chess, image captioning, chatting, and many more. Some of these tasks have already been performed by AI algorithms like *MuZero* [2], and *AlphaGo Zero* [3]. *AlphaGo Zero* solely depends on reinforcement learning without any human interventions in playing and mastering *tabula rasa*. [2] is designed and investigated in real-world scenarios without defining game rules and any human intervention. The algorithm achieves comparable performance to that of the *AlphaGo Zero* that was trained with rules (in a simulated environment). Other than vision and text, these

[1] https://venturebeat.com/datadecisionmakers/is-deepminds-gato-the-worlds-first-agi/

AI algorithms have also excelled in generating speech that is as close as possible to that of human speech [4, 5, 6, 7, 8, 9].

1.1 Motivation

With the improvements in AI, once prescient Human Computer Interactions (HCI) have now become part and the parcel of our lives. We are in the wonderful phase of AI, where most of the tasks are accomplished by just voice commands. For example, from, "Alexa! set the temperature in the living room to 27 degrees" to "Alexa! order a pizza from XX restaurant", voice assistants like Amazon's Echo, Apple's SIRI, and Google home are available to make our lives much easier. Additionally, these conversational agents are also used for various customer needs such as booking appointments over a phone call [7], navigation [8], language learning applications [9], and many more. Google's Duplex [7] can handle conversations over a phone call for *specific tasks* such as booking appointments, reserving a table at a restaurant, or ordering food. The model also inserts the fillers such as "hmm", and "uh" which makes the conversations even more lively and natural. Nevertheless, general human intelligence is yet to be achieved by these conversational agents in different application domains. This is because telling a personal assistant that "I am sad!" would result in the following,

Human: *Alexa! I am sad!*
Alexa: *I am sorry to hear that. Please try talking to a friend or listening to music or taking a walk. Hope you will be fine soon.*

On the other hand, saying the same to a human would result in a completely different response. The COVID-19 pandemic has affected most of our lives in the last couple of years. The employees/industries that were high in demand during the situation were the front-line workers in health care and customer service. The motivation of my thesis is derived from the predominant need of these two industries during the pandemic situation. The best way to reduce the overwork on the employees was through employing conversational agents (chatbots or personal assistants) in the loop. However, while doing so, it is indispensable that these agents express empathy and compassion during the conversations. Even though the application domains health care and customer service, require much more than just warmth and competence, in my thesis, I focus only on these two aspects of social perception.

1.1.1 Why warmth and competence?

Framing impressions of others (animals or humans) have been performed effortlessly and involuntarily by humans for ages. Since, the era of the hunter-gatherer, (when humans feared wild predators like lions and tigers) till today of human-computer interactions (when we fear the dangerous effects of AI in the future), we have been making social judgments in our day-to-day lives. However, these judgments are mostly based on our first impressions and they may or may not be true. For example, we love babies and tend to like and trust people with cute faces more than someone with a heavily built body [10]. Similarly, women for ages are considered as caregivers, and men were seen as the breadwinners [2] [3] [11].

Decades of research on these social judgments through interpersonal relations have provided that the social perceptions of humans are based on two criteria: a) is the person warm/friendly enough? (are his/her intentions good?), b) is the person competent enough? (can he/she achieve those intentions?) [12, 13, 14]. These studies depict that both dimensions (warmth and competence) are equally important. The person with good intentions and incapable of achieving those is of no use. Accordingly, one who is capable of achieving anything but does not have good intentions is a threat to society. Therefore, finally, these two criteria were termed as ***"universal dimensions of social perception"*** [15, 16, 17].

Human beings unintentionally associate these dimensions of social perception with trustworthiness. Individuals who are warm and competent are found to be more trustworthy than others [10, 18]. In the domains like health care and customer service, the agents would require the patients/customers to trust them with their problems/queries. An individual who is in severe pain (either physical or mental) would benefit from a caring response from a trustworthy person. Similarly, a customer vexed with the poor service of a product would need a problem solver who is also reliable and understanding. In [19], the researchers posit that healthcare workers should not only possess expertise in the field but also exhibit compassion and empathy towards the patients they are dealing with. Further, empathetic agents are considered to be both warm and competent [20]. Humans tend to be comfortable in articulating their thoughts with others who exhibit similar levels of empathy and behavior as theirs [21]. Alternatively, people tend to "like" the individuals who are warm and "respect" those who are competent [10]. [22] presents that the universal dimension, competence has a significant impact in driving the attention of the customers towards the product or the company. Correspondingly, the dimension, warmth is essential in maintaining relationships with customers over longer periods of time. This thesis aims at studying the perceptions of these universal dimensions from various Text-to-Speech (TTS) voices.

---

[2] https://arborsassistedliving.com/the-gender-gap-women-predominate-among-caregivers-of-the-elderly/

[3] https://blogs.koolkanya.com/we-need-more-breadwinning-women-and-caregiving-men/

1.1.2 Evaluation of synthetic voices

The advent of end-to-end neural models facilitated the fidelity of human-like speech generation [5, 4, 6]. With this development, the current TTS research is focused on the generation of expressive speech [23, 24, 25] and the generation speed of the existing models [26, 27]. However, the parallel evolution of the evaluation of these synthetic voices is also crucial. Much of the TTS evaluations were found similar to the measures proposed in the Blizzard challenges [28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44]. Contrary to these evaluation setups, authors in [45] investigate the effect of speech quality and naturalness on the intelligibility of synthetic voices through the use of pupillometry. They study the cognitive load on the listeners perceiving human speech vs different model-based TTS voices obtained from the Blizzard Challenges. Not until recently have Neural Network-based evaluation metrics been introduced for synthetic speech [46, 47, 48]. [46] presents an automatic evaluation of speaker similarity in an unsupervised setting (human ratings are not known during the training of the models). Authors report that the automatically predicted similarity scores are correlated with the subjective responses while displaying accuracy of 96%. The speaker classification is carried out using a 2-dimensional Convolutional Neural Network (CNN) followed by a Gated Recurrent Unit (GRU) combined with two fully connected linear layers. Similarly, [47] employs a Convolutional Net and a Long-Short Term-Memory Units (CNN-LSTM) network for the prediction of naturalness on the synthetic speech and the Voice Conversion (VC) data. Also, they report better predictions through the transfer of knowledge from the speech quality ratings collected from humans. [48] discusses the VoiceMOS challenge introduced this year. The challenge encourages the automatic evaluation of Mean Opinion Scores (MOS) of naturalness from Blizzard and Voice Conversion challenges collected over a decade. [49] presents an objective evaluation of naturalness and speaker similarity in VC voices using a CNN-Bidirectional LSTM (CNN-BiLTSM). Further, the research team developed an end-to-end evaluation setup for speaker similarity in [50] which leverages the attention mechanism. Additionally, perceptions of various traits and the likability of synthetic speech by humans have also been investigated [51, 52, 53]. [51] presents the studies on the paralinguistic traits such as age (scales provided were as follows, 1-10, 11-20, 21-30, and so on till 91-100), gender (the choices were: male, female, both, neither), accents (listeners can choose from 249 options the origin of the speaker), and the human-likeness (on a 5-point Likert scale) from the voices derived from the IBM Watson's Text-to-Speech (TTS) synthesizer. Correspondingly, [52] reports the evaluation of likability ("how much does the listener like the voice on a 5-point Likert scale") and the human-likeness ("how close is the voice to that of a human voice"). These studies were carried out on German speech generated from male TTS voices. [53] discuss the importance of forming first impressions (warmth and competence) of virtual agents by humans. Their studies also detail the change of first impressions by humans of the agents in a prolonged human-agent interaction. The study presents two visual conversational agents (one similar to a robot and one similar to a human). The participants of the study could rate their impressions of the agent based on per-

ceived speech, non-verbal behaviors, and interaction time. However, these studies were carried out on intelligent virtual agents, and in the current work, we plan to investigate these social aspects from the synthetic voices (speech-alone scenarios) alone. In the course of this work, these aspects of social perception are termed as the desired Social Speaker Characteristics (SSC).

1.2 Thesis objectives and Research questions

In this section, I provide the objectives of my work and the corresponding research questions that are addressed in this thesis.

- **Objective 1:** Postulate the significance of investigating the social perceptions of synthetic speech.

  Through this thesis, we emphasize the need for studying and analyzing the perceptual dimensions contributing to socially acceptable synthetic voices. The characteristics of interest are the fundamental dimensions of social perception, warmth, and competence. These first impressions are essential in the case of both HCI and also human-human interactions. These two dimensions aid in building long-term relationships between the agents and the users along with their trust.

- **Objective 2:** Transform the negatively perceived synthetic voices to positive ones.

  The analysis of the subjective evaluations of the synthetic voices provides three different clusters: socially acceptable voices (positively perceived), neutral voices, and socially unacceptable (negatively perceived) voices. The goal is to employ various modeling techniques to alter the acoustic features of negatively perceived voices so as to manifest them as socially acceptable (positively perceived) voices.

- **Research question 1:** What social speaker characteristics do people perceive in synthetic speech?

- **Research question 2:** Which acoustic features of synthetic speech affect the subjective perceptions of social speaker characteristics?

- **Research question 3:** Which alterations of the synthetic voices or synthetic procedure would lead to positive perceptions of speakers?

  The focus of this thesis is not to propose new frameworks for TTS/VC but rather to examine if the existing methods enable the positive perceptions of synthetic voices.

1.3 Publications

Most of the important scientific contributions of this thesis were published in conferences (or workshops). The details of the publications are provided below.

- Sai Sirisha Rallabandi, Abhinav Bharadwaj, Babak Naderi, Sebastian Möller. Perception of social speaker characteristics in synthetic speech. *In Proc. Interspeech*, Brno, Czechia, 2021.

This paper studies the social perceptions of synthetic voices. The study was carried out on two commercial Text-to-Speech (TTS) systems namely, Google TTS and Amazon Polly. The social perceptions of these synthetic voices were studied using continuous 100-point scales with adjective-antonym pairs at the extremes of the scales. The factor analysis on the subjective responses has provided three factors among which two were the social speaker characteristics, warmth, and competence, and the third factor was the personality trait, extraversion.

The design of the experiments and the organization of the paper were prepared by Sai Sirisha Rallabandi. The discussion of the dataset (text) to be used for speech generation was done between Sai Sirisha Rallabandi, Benjamin Weiss, Babak Naderi, and Sebastian Möller. The integration of the subjective evaluation setup on Amazon Mechanical Turk (AMT) was done by Babak Naderi. The pre-processing of the subjective data was carried out by Abhinav Bharadwaj.

Chapter 4 provides the basis for this work. A brief description of the scientific details contained in this paper is also provided in the chapter.

- Sai Sirisha Rallabandi, Babak Naderi and Sebastian Möller. Identifying the vocal cues of likeability, friendliness, and skillfulness in synthetic speech. *In Proc. Speech Synthesis Workshop (SSW11)*, Gárdony, Hungary, 2021.

This paper aims at deriving the vocal cues of social speaker characteristics (SSC), warmth, and competence from synthetic speech. The study consists of two parts, a) prediction of acoustic correlates of SSC, and b) automatic prediction of SSC from the vocal cues of SSC. The acoustic features used in these studies were 88-dimensional OpenSMILE features extracted using the eGeMAPS configuration. Further, the automatic prediction (regression) of SSC is performed using two models, Linear regressor and Support Vector regressor.

The background and motivation for this work, experimental setup, and the organization of the paper were designed by Sai Sirisha Rallabandi. A discussion on the limitations of the study was done between Sai Sirisha Rallabandi and Babak Naderi.

The motivation and the experimental procedure employed in this work were derived from the studies provided in chapter 5.

- Sai Sirisha Rallabandi, and Sebastian Möller. On incorporating social speaker characteristics in synthetic speech. *arXiv*, 2022.

In this paper, we investigate the modification of the synthesis procedure for the generation of positively perceived synthetic voices. This was enabled using the conditioning of the end-to-end TTS model, Tacotron using quantized acoustic correlates of SSC. Further, the evaluation of the generated speech was carried out using the adjectives derived from studies on social perceptions of synthetic voices (from the work presented in "Perception of social speaker characteristics in synthetic speech").

The motivation for this work and the organization of the paper were planned by Sai Sirisha Rallabandi. The suggestions on the subjective evaluation setup were delivered by Sebastian Möller. The discussion on the experimental setup was done between Sai Sirisha Rallabandi and Sai Krishna Rallabandi (Carnegie Mellon University).

This paper presents the alterations employed in the synthesis procedure for the perception of warmth and competence in the generated speech. Some of the experimental details of the study are provided in chapter 7.

1.4 Thesis outline

- **Chapter 2** provides the distinction between speaker characteristics, emotions, and personality traits. It also presents the works previously carried out on understanding the perceptions of various human behaviors from speech. Further, a brief description of machine learning approaches employed in various studies in this thesis is provided followed by the literature review on the conventional approaches for Voice Conversion and Text-to-Speech synthesis. Finally, the chapter concludes with an overview of perceptual studies prevalent through in-lab and crowd-sourcing experimental setups.
- In **chapter 3**, the challenges encountered in deciding on the datasets and the questionnaire to be included during the subjective evaluations of synthetic speech are discussed. Later on, the datasets and the adjectives used in different studies of this thesis are presented.
- **Chapter 4** aids in answering the first research question discussed in this chapter. The chapter details the studies conducted on social perceptions of synthetic voices. Two different studies were carried out on multiple TTS voices for this analysis namely, a) studies on a wide variety of TTS systems, and b) studies on 2 commercial TTS systems. In this chapter, we can also observe the extraction of ground truth information from the current studies. This ground truth information is used throughout the thesis for the positive perceptions of synthetic voices (discussed in chapter 6 and 7).

- **Chapter 5** addresses the second research question discussed in this chapter. This chapter is two-fold, a) analysis of acoustic correlates of SSC in synthetic speech, b) automatic prediction of SSC from synthetic speech. The first part of the chapter discusses the analysis of vocal cues responsible for various speaker attributes (or SSC) in synthetic speech using the OpenSMILE features. The second part presents the automatic prediction of those speaker attributes using the relevant vocal cues (derived from the previous step). Finally, the chapter concludes with a comparison of the current results with the results of similar studies provided in the literature (studies on both natural and synthetic speech).
- **Chapter 6** is designed in such a way that it can address the first part of the third question (which alterations of synthetic speech would lead to positive perceptions of synthetic speech?) and also addresses the second objective presented in this chapter. In chapter 6, we find the alteration of negatively perceived TTS voices for their positive perceptions using Voice Conversion. Further, a detailed description of the subjective evaluations of the converted speech with different evaluation setups that were not utilized before is presented (scales/adjectives used in the evaluation setup were not used before to the best of my knowledge). Finally, a comparison of the subjective responses derived from different evaluation setups is presented.
- **Chapter 7** deals with the second part of the third research question (which alterations of synthetic procedure would lead to positive perceptions of synthetic speech?). This chapter presents multiple TTS experiments carried out (separately for warmth and competence) using the Tacotron model. The acoustic correlates of SSC derived using the procedure described in chapter 5 were utilized in order to condition the Tacotron model in each of those experiments. Finally, the subjective evaluation of the generated speech is performed using the adjectives derived from the ground truth information defined in chapter 4.
- **Chapter 8** This chapter presents the synopsis of the thesis. Here, we find the summary of various scientific studies and the evidence gathered to achieve the objectives and address the research questions discussed in chapter 1. Further, the follow-up work for this thesis is discussed by detailing the challenges encountered during the course of this work.

Chapter 2
Background

This chapter presents the difference between speaker characteristics, emotions, and personality traits followed by the literature review on the perception of each of these from speech. Later on, an overview of the Machine Learning algorithms utilized in various studies of this thesis is presented. Finally, the chapter concludes with the details of the evaluation approaches employed in the analysis of various perceptual dimensions of speech.

2.1 Distinguishing characteristics, emotions and personality traits

The study of human behavior was not so easy and has been performed for decades. These studies were carried out by different research groups throughout the years [54, 55, 12, 56, 57, 58]. Nevertheless, the analysis of these behaviors has been studied in a similar fashion [58, 59]. Most of the studies on human behavior employ a semantic differential scaling test with the bi-polar adjectives at the extremes of the scale (kind vs unkind, attractive vs unattractive, etc.,). The subjective evaluations are carried out through peer-rating or self-rating of a list of adjectives. Correspondingly, the factor analysis of the subjective data is performed using any of the dimensionality reduction techniques such as Principal Component Analysis (PCA) (reduces the high dimensional data while retaining the most essential information), Exploratory Factor Analysis (EFA) (used when the number of factors to be determined is unknown), Confirmatory Factor Analysis (CFA) (used to confirm the number and type of the factors underlying the data) [59]. Each of the derived factors is named after the term that would best represent its factor loadings.

[54] presents the studies evaluated by undergraduates in psychology. The students were provided with two lists of adjectives in a specific order and were asked to form impressions of the person. Among the two lists, all the adjectives were the same except for one. The first list consisted of the characteristic, “warm” and the second one had “cold”. The study shows that the impressions formed based on the first

list were more positive and socially acceptable than those from the second. Through these studies, the social dimension, warm was proposed as the central trait in forming impressions of a person. However, a subsequent study involved a multidimensional scaling for the clustering of 64 adjectives [55]. This study has provided two orthogonal dimensions namely, social good/bad and intellectual good/bad. Using these two dimensions, a hypothesis was carried out to verify the effect of one dimension on another [12]. The results showed that the social dimension, warm as previously thought to be the central dimension [54] is not completely true as the intellectual dimensions are not affected by the manipulations in the social dimension and vice versa. Only after several years of research by Bales and his students at Harvard University, it has been confirmed that the two primary dimensions of social perception are based on social interactions (warmth) and task accomplishment (competent) [56, 57]. Thus, these two dimensions are considered the fundamental dimensions of social perception that are long-lasting and constant [15]. Further, the researchers propose that a different combination of these characteristics/dimensions (upward or downward movements in the dimensions/longitudinal study of social perceptions) would contribute to different emotions and behaviors of people [60].

I present an example to provide the slight difference between emotions and characteristics (social speaker characteristics). We all remember the dialogue of Wanda Maximoff from Doctor Strange, Multiverse of Madness.

> "You break the rules and become a hero, I do it and I become the enemy!!! ***That does not seem fair.***"
>
> *Wanda Maximoff, Doctor Strange, Multiverse of Madness, MCU.*

In the first part of the dialogue, the actor seems angry and disgusted (emotions). Emotions are feelings that are controlled by external stimuli and can be either temporary (fear or anger) or long-lived (grief) [61, 13]. The basic emotions are anger, fear, shame, contempt, disgust, surprise, joy, etc., [62]. As suggested by Plato in his philosophy, the human soul can be categorized into three partitions: emotions, cognition, and motivation. Emotions can affect motivations and cognitive abilities such as perception, memory, and decision-making. Accordingly, humans with positive emotions exhibit decisiveness, and responsibility in their actions and can lead a peaceful life. On the other hand, when negative emotions are dominant, humans tend to be depressed and make bad choices in their lives. This theory of emotion-cognition interdependence was further supported by numerous studies throughout the literature [13, 14, 63]. Such varying human behaviors based on their emotions are also associated with the term, *instincts* [63]. Authors in [63] detail such associations in the case of four basic emotions namely, anger, happy, sad and fear. Those associations are as follows: happiness is tied to the human actions of rewarding (either others or oneself), sadness is linked to punishment, and stress is associated with two emotions, fear and anger. These human actions (reward, punishment and stress) are termed as *core effects* and are also associated with different colors, blue,

red and yellow. A different combinations of these core effects can further give rise to complex emotions such as love and other feelings in humans.

The last part of the sentence (in bold) shows the character's real characteristics (cynical). "Cynical" is a negative characteristic and its antonym/positive attribute is trust or belief. People with this characteristic (cynicism) are less likely to participate in group activities or collaborate with others [64] as they fear of being deceived. Literature shows that due to this disbelief, people with cynical behavior may also lose a lot of opportunities. Researchers have also explored the relationship between cynical behavior and competence in individuals [65]. The study presents that cynicism has a negative impact on social well-being, task accomplishment, and cognitive abilities. Thus, cynicism does not contribute to the success of an individual.

In the example, we also observe a sudden difference in the facial expressions and also the tone of the actor. Emotions vary based on different situations but as mentioned before, the characteristics remain constant in humans. Hence, this property of the social characteristics would facilitate the data collection (without the need for enacted speech).

Apart from forming first impressions (warmth, competence) and temporary feelings (emotions), there are also other stable characteristics that human beings possess (or perceive) [66]. All those adjectives (or characteristics) that describe an individual were grouped together (into 5 orthogonal dimensions) and were termed personality traits. Robert R. McCrae proposes the most commonly used personality traits and Paul T. Costa Jr in their FIVE FACTOR Theory of personality [67]. The universal dimensions of personality that they proposed were Openness to Experience (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N). Over the years, these are termed the BIG FIVE personality traits and are employed in various personality assessment tests. Among these five dimensions of human personality, we can further categorize them based on their perceptions of humans. All the traits can be referred to as interpersonal (obtained through peer ratings and self-ratings). However, the traits, Conscientiousness, Neuroticism, and Openness are not only identified through the interpersonal relationships of humans but also through different attitudes exhibited by humans under different and difficult circumstances (similar to varied emotions and behaviors in humans, these traits can also vary with the external stimuli). But, the traits Extraversion and Agreeableness are perceived only through social interactions [68]. Initial behavioral studies utilized high-dimensional (128-item Interpersonal Adjective Scales (IAS)) circumplex models (circular ordering of dimensions) for various behavioral studies on humans. However, [69] propose a reduced version termed IAR-R with 64 items (64 adjectives) for various personality assessments. Some of the adjectives used in this thesis are derived from the ones proposed in [69, 70]. These studies include the questions (adjectives) required for the personality judgments accumulated from studies carried out on personality (prior to this work) along with the adjectives in general (which can define various human qualities).

2.2 Perception of various characteristics, emotions, and personalities from speech

Studies show that it is possible to interpret different characteristics, mental states, traits, and emotions of a person based on auditory cues [71, 72]. In [71] researchers examined the perception of the speaker's age, weight, and height from the speech samples of unknown speakers. Additionally, the study provides a comparison of the visual cues and the vocal cues in determining these characteristics of an unknown person. The study displays comparable results in person identification from both speech perception as well as visual perception (from the photographs of unknown people). Similarly, the assessment of personality has also been carried out for human voices as well as machine-generated voices [72]. The study shows that the participants were attracted to and believed the machine-generated voice when its personality matched with their own personality (extrovert Vs introvert). Additionally, akin to personality studies on humans (studies based on overall behaviors), interpreting the personality from speech-alone scenarios has also been investigated. The personality studies on speech data were carried out using the datasets, SSPNet Speaker Personality Corpus (SPC) and the Electronically Activated Recorded (EAR) corpus. A subset of the BIG FIVE Inventory (BFI-44, a questionnaire with 44 items [73]), a BFI-10 (a questionnaire with 10 items)) was prepared for the personality analysis of English and German samples [74]. Correspondingly, [75, 76] examined the speaker characteristics of 300 German speakers from semi-spontaneous conversations. The authors propose five perceptual factors that can be perceived in zero-acquaintance scenarios. The derived perceptual factors were both social (apathy, serenity, confidence, incompetence) and physical (attractiveness). There have also been some works on analyzing the social speaker characteristics from speech [77, 78, 79, 80, 81]. In [77], the speaker characteristics were examined in telephone communications through an agent-customer conversation. The speaker characteristics considered were the age, gender, emotions of the speakers, accents, etc., The study provides a three-class taxonomy for speaker characterization namely, online, mirroring, and critical. In [78], the authors investigate the perceptual dimensions (speaker characteristics) of charismatic speech. The proposed dimensions were, enthusiastic, charming, persuasive, passionate, and convincing. Additionally, they report the negation of the negative attribute, boring (not boring) to be one of the responsible speaker characteristics responsible for the perception of charisma in speech. They also report the acoustic correlates of charismatic speech to be the speaking rate, loudness, and f0, and its dynamics (mean, standard deviations). Some other works on charismatic speech include, [82, 83]. In [83] authors investigate the influence and the degree of influence of the charismatic voice in an unusual experimental setting (mock drive with a highly charismatic voice Vs less charismatic voice guiding the driver). The study displays that the participants obey the navigator and still continue to listen to the instructions even after a couple of mistakes in the case of a highly charismatic voice. [80] study the speaker characteristics, insecure, hesitant, monotonous, aggressive, accusing, agitating, objective, trustworthy, humble, expressive, powerful, and

involved using a 5-point Likert scale. The study proposes that the speakers with the highest ratings on the adjectives, expressive, powerful, involved, and trustworthy are regarded as "good speakers". Additionally, they suggest that the acoustic correlates of good speakers are the F0 peak height and F0 range. Similarly, [84] investigate the association of the vocal cues, F0, jitter, and shimmer with the perception of emotions from natural speech. The study examines the hypothesis that the acoustic parameters are the external and visible cues of internal emotional states and moods of a person. The experimental results propose that the relevant vocal parameters are triggered by the intensity of the speaker's emotional state, which in turn gives rise to various expressions in their speech. In line with the dimensional approach previously discussed in speaker characterization and personality assessment, emotions have also been analyzed on multiple dimensions such as valence, activation, potency, and emotional intensity [85]. Authors in [85] study the above-mentioned dimensions in the following emotional states: anger, sadness, disgust, happiness, and fear. The acoustic features responsible for multiple emotions and dimensions are speaking rate, pitch, voice intensity, and spectral energy distribution. Apart from emotions and various speaker characteristics, personality has also been found to be perceived from speech. In [86], authors examine the automatic prediction of the personality of the speaker on 60 different prepositions (prepositions representing different personality traits). The subjective analysis consisted of a 5-point scale in the range of 0-4. Each point on this scale is defined as follows: *0 = strongly disagree*, *1 = disagree*, *2 = neutral* , *3 = agree*, *4 = strongly agree*. The experimental results display high accuracy in the automatic classification (accuracy = 60% on a 10-class classification task) of various personality traits in enacted speech. [87] investigate the automatic perceptions of personality by non-native speakers. The study utilizes the news bulletin available in French from SSPNet Speaker Personality Corpus. The perceptual studies were carried out by 11 non-native speakers. The idea was to inspect the personality from non-verbal cues in speech. Further, the automatic classification of these personality traits through logistic regression (classification) resulted in the perception of the traits, extraversion, and consciousness with higher accuracy over others. The perceptions of fundamental dimensions, warmth, and competence were previously performed between visually impaired and sighted individuals [88]. The speech data used in the study consisted of a series of vowel pronunciations by both genders. In addition to the vowel pronunciations, the questionnaire also consisted of the varied pronunciations of these vowel samples (raise and lowering of the voice pitch). The study displays that the social perceptions in both the categories of test participants (visually impaired and sighted individuals) were similar. The voices with lowered pitch were perceived as trustworthy and competent irrespective of the gender of the speaker. Nevertheless, the raised pitch in females contributed to an improved perception of warmth.

However, to the best of my knowledge, the analysis of the social speaker characteristics such as warmth and competence in synthetic speech (alone) has not been performed before. Therefore, through my thesis, I provide a) an analysis of the social perceptions (warmth and competence) of synthetic voices, b) acoustic correlates of warmth and competence in female and male TTS voices, c) automatic prediction

of warmth and competence from synthetic speech, and d) modeling of synthetic voices and modification of synthesis procedure for the generation of warmth and competence from the TTS voices.

2.3 Machine Learning approaches

This section details various machine-learning approaches utilized in this thesis.

2.3.1 Decision Trees

A decision tree is a non-parametric supervised machine learning algorithm used for both classification and regression tasks [89]. Based on these functionalities, it is also termed a CART Tree (Classification and Regression Tree). The components of a tree are leaves and nodes. The main node is called the root node and the data split is enabled at different layers of the tree based on decision criteria (conditions). The data split follows a greedy approach and the best split is determined based on the cost function (the data split with the lowest cost or loss is chosen). The final node which does not split anymore is called the leaf and the intermediate layers are called branches. Figure 2.1 displays an example of a decision tree and data split. The example displays a classification tree to determine if a person could step out of the house. Similarly, a regression task would determine the continuous values such as population, pricing (houses, groceries, properties), etc.,

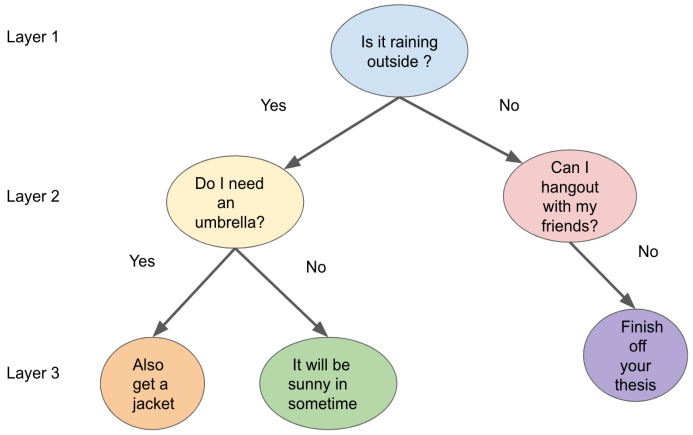

Fig. 2.1: Schematic of Decision Tree

.

The cost functions for each classification and regression task are provided below. Equation 2.1 represents the loss function calculated due to the classification using a decision tree. Here p_i represents the probability of a data point being in the class, i, and c is the total possible classes. Similarly, equation 2.2 represents the loss function calculated due to the regression costs, where y is the actual (target) outcome and $f(x)$ is the predicted outcome, N is the total number of data points. The decision trees have been widely used in unit selection based TTS for Diphone synthesis [90], [91], [92]. In this thesis, decision trees are employed in the modeling of acoustic features of the synthetic voices, in the prediction of vocal cues of warmth and competence (in chapter 5).

$$c\_loss = \sum_{i=1}^{c} p_i * (1 - (p_i)) \tag{2.1}$$

$$r\_loss = \sum_{i=1}^{N} (y - f(x_i))^2)/N \tag{2.2}$$

2.3.2 Support Vector Machines

Support Vector Machines (SVMs) were pioneered by Boser, Guyon, and Vapnik through their work in [93]. SVMs employ a supervised learning algorithm that can perform both classification and regression tasks [94, 95]. They perform well with limited training data as opposed to Neural Networks. The model predictions are carried out through a hyperplane or a decision boundary. For 2-dimensional data, the decision boundary is a line and for 3-dimensional data, it is a plane, and the data points that lie close to/or on the decision boundary are called the *support vectors*. The corresponding mathematical representations are provided below. The equation of a line for a regression task is presented in equation 2.3, where y denotes the predicted values, W and b are weights and biases respectively and x represents the input data points. The equation for the hyperplane in the case of classification is provided in the equation 2.4. x denotes the input data points, w is the weights and b represents the bias added to the model. The model optimization is carried out by calculating the loss functions such as hinge loss (classification loss) or mean squared error (regression loss function) followed by weight updates to maximize the margin (distance between the support vectors or the width of the hyperplane). SVMs can also perform non-linear tasks by moving the lower dimensional data into high dimensional space. This non-linear transformation is enabled by the functions such as the polynomial function, and radial basis function. These kernel functions calculate the relationships between the data points in the higher dimensional space without actually transforming the data into a high dimensional space. This mechanism is called the *Kernel trick*. In this thesis, SVMs are utilized for both classification and regression tasks in chapter 5. These models are used for a) the automatic prediction

of acoustic correlates of warmth and competence, followed by b) the automatic prediction of social speaker characteristics provided the derived acoustic correlates.

$$y = Wx + b \tag{2.3}$$

$$y = \begin{cases} +1, & \text{if } x * w + b \geq 0 \\ -1, & \text{if } x * w + b < 0 \end{cases} \tag{2.4}$$

2.3.3 Feed-forward neural networks

A feed-forward NN is a network in which the data/feature processing is directed from the input layer to the output layer through the hidden layers (layers between the input layer and output layer) without any feedback connections (the output of the network is not fed to itself during training). The basic architecture of a feedforward network is provided in figure 2.2. A typical neural network comprises of multiple processing units called nodes organized in a sequence of layers. The nodes in each layer of the NN are fed with the outputs of the previous layer (except the input layer). The objective of a NN is to approximate the function, $y' = fn(x;w)$, where y' is the output, fn is the activation function, x is the input, w represents weights and biases. Further, we can find the mathematical representations of the layer-wise forward propagation. Equation 2.5 represents the output of the first hidden layer when mapped with the input x at time t. In the equation, h_1 stands for the first hidden layer of the network, a is the activation function, W_1 is the weight matrix between the input and the first hidden layer and b_1 is the bias for the first hidden layer. Correspondingly, the equation for the other hidden layers can be represented as shown in equation 2.6. h_n is the n^{th} hidden layer, h_{n-1} is the previous hidden layer or $n - 1^{th}$ hidden layer, W_n is the weight matrix between the n^{th} and $n - 1^{th}$ hidden layer, b_n is the bias vector for the n^{th} hidden layer. The final predictions of the network at the output layer are mathematically represented as presented in equation 2.7. In the equation, y' is the output of the final layer/output layer of the network, W_o represents the weight matrix between the output layer and the previous hidden layer, b_o is the bias vector corresponding to the output layer. The choice of activation function varies depending on the task the model is being used for. For instance, a linear classification or regression is performed by a linear activation function. Similarly, non-linear tasks are handled by non-linear functions such as sigmoid (generally used for binary classification), softmax (performs multi-class classification), tanh (both classification and regression), ReLU (rectified linear unit), and SeLU (scaled exponential linear unit). Feedforward NNs are powerful machine-learning models that can perform frame-wise modeling. However, they have also been explored for sequential tasks such as TTS and VC [96, 97, 98]. This thesis uses basic NN architecture in chapter 5 for the classification task.

$$h_1 = a(W_1 x_i + b_1) \tag{2.5}$$

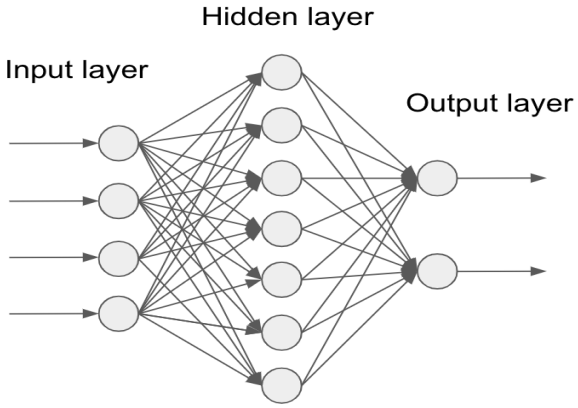

Fig. 2.2: Schematic of Neural Network

.

$$h_n = a(W_n h_{n-1} + b_n) \tag{2.6}$$

$$y' = a(W_o h_n + b_o) \tag{2.7}$$

2.3.4 Recurrent Neural Networks

Artificial speech generation requires the temporal modeling of the speech data. Recurrent neural networks (RNNs) can effectively model the sequential information [99]. In a recurrent neural network, the output at each timestep (hidden state) is fed as feedback to the network. This feedback loop helps the network to model the information sequentially while using the memory (hidden state). However, conventional RNNs suffer from vanishing gradient [100]. Therefore, the long-term dependency in sequential data is studied using the RNN cells called, Long Short-Term Memory units (LSTM) [101, 102]. LSTM consists of 3 inputs and 3 outputs. The three inputs are X_t (current input), h_{t-1} (previous hidden state) and C_{t-1} (previous cell state). The three outputs are O_t (current output), h_t (current hidden state) and C_t (current cell state). Additionally, it consists of four fully connected layers and four gates namely, input gate *(I/p)*, forget gate *(F_t)*, cell state gate *(C')* and output gate *(O/p)*.

The mathematical expressions for each of input gate, forget gate, output gate, and the cell state gate are provided in the equations 2.8, 2.9, 2.10, 2.11 respectively where w* are weights and b* are biases, σ represents sigmoid activation function. The mathematical expressions for the current cell state *(C_t)* and the output *(ht/Ot)* of the LSTM are shown in the Equations 2.12, 2.13 respectively.

$$I/p = \sigma(w_i[h_{t-1}, x_t] + bi) \tag{2.8}$$

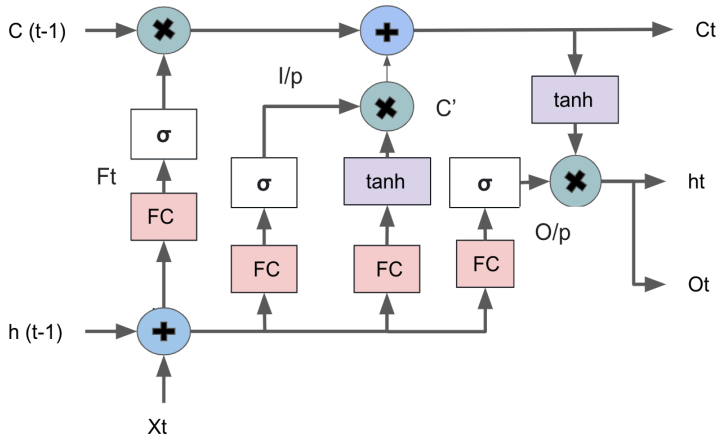

Fig. 2.3: Schematic of an LSTM cell. FC=fully connected gates, Ft = forget gate, I/p = input gate, $C^{'}$ = Cell state gate, O/p = output gate

.

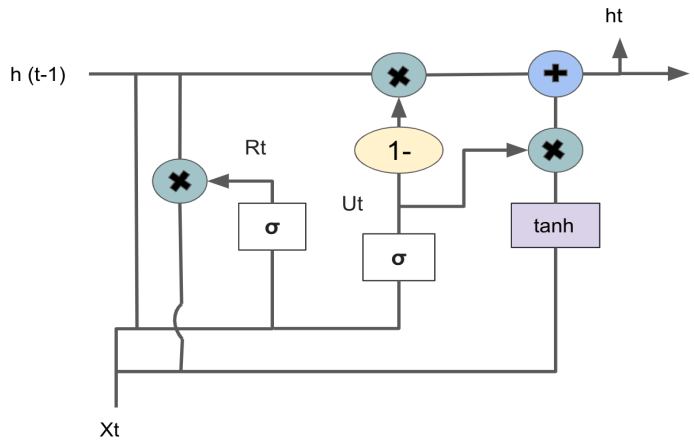

Fig. 2.4: Schematic of a GRU cell. Ut = Update gate, Rt = Reset gate, Xt = input, h (t-1) = previous hidden state, ht= output of the GRU cell

.

$$F_t = \sigma(w_f[h_{t-1}, x_t] + bf) \tag{2.9}$$

$$O/p = \sigma(w_o[h_{t-1}, x_t] + bo) \tag{2.10}$$

$$C^{'} = tanh(w_c[h_{t-1}, x_t] + bc) \tag{2.11}$$

$$C_t = Ft * C_{t-1} + I/p * C^{'} \tag{2.12}$$

$$h_t = O/p * tanh(C_t) \tag{2.13}$$

Even though LSTMs can address the memory issue seen in basic RNNs, they suffer from long training times and complex structures. As an alternative, a variation of LSTMs, called Gated Recurrent Unit (GRUs) were introduced [103]. A diagrammatic representation of GRU units is presented in figure 2.4. A GRU unit consists of 2 additional gates namely, the update gate and a reset gate. These gates would enable the network to retain long-term memory by updating (the amount of information from the past that has to be used) and resetting (which information should the network forget) the hidden state information when needed. The forget gate and the input gate in the LSTM are combined to form an update gate. The mathematical representations for the update gate (U_t), reset gate (R_{rt}), output of the current memory state $h_t^{'}$, the output of the GRU cell (h_t) are provided in the equations, 2.14, 2.15, 2.16, 2.17 respectively.

$$U_t = \sigma(w_u[h_{t-1}, x_t] + b_u) \tag{2.14}$$

$$R_{rt} = \sigma(w_{rt}[h_{t-1}, x_t] + b_{rt}) \tag{2.15}$$

$$h_t^{'} = tanh(w.[Rt * h_{t-1}, x_t]) \tag{2.16}$$

$$h_t = (1 - Ut) * h_{t-1} + Ut * h_t^{'} \tag{2.17}$$

2.3.5 Convolutional Neural Networks

Convolutional Neural Networks or ConvNets (CNNs) is a variant of artificial Neural Networks that are prevalent in computer vision and image analysis [104]. As the name suggests, the model uses the "convolution" operation to learn the complex data representations. The essential components of a CNN are a) the convolution layer, b) the pooling layer, c) the fully connected layer (FC), d) the dropout layer and e) the activation function. The convolutional layers consist of filters (kernels) that can extract the features from the input image or speech or text. The feature extraction is carried out by computing the dot product while sliding the filter (of size (N*N)), over the specific portions of the image. The pooling layer aids in reducing computational costs through dimensionality reduction. The derived features are then connected to the FC layer before deriving the predictions. The FC layer might contribute to overfitting during the training phase therefore, a dropout layer is included in a CNN. The choice of the activation function to be used depends on the task to be accomplished using the network. Due to their tremendous benefits in modeling sequential information, the convolutional layers are also employed in speech generation [4, 5] and speech quality prediction [105].

2.3.6 Training mechanism

The difference between the actual outputs and the predictions made by the NN is provided in terms of the cost functions. Different cost functions or loss functions can be utilized depending on the task. For instance, the most commonly used loss functions for classification tasks are cross-entropy loss, and the one for regression is the mean squared error (MSE). The mathematical equations for the binary classification in terms of binary cross entropy are provided in the equation 2.18, where y is the actual output, y' is the predicted output. Correspondingly, the mathematical representation of MSE is presented in equation 2.19, where N represents the total number of examples or data points, y is the actual output, and the function, $f(x,w)$ is the predicted output. Eventually, training of these networks for model optimization is carried out using the gradient descent approach. In this approach, the model weights and biases are updated iteratively in order to minimize the cost function and reach a local minimum. This process begins by calculating the partial derivatives of the cost function with respect to the network's weights and biases. Another important parameter to be considered while performing the model optimization is the learning rate. The learning rate value determines the size of the steps to be taken during the model optimization in order to reach the local minimum. Lower learning rate values can lead to slower convergence, while higher learning rates can cause overshooting (the model might miss the local minima). An optimizer such as Adam (Adaptive Moment Estimation) can be used while performing such model optimizations. Adam enables an adaptive learning rate and also provides a faster convergence. This entire optimization is carried out on multiple subsets of the input data. Each subset is called a mini-batch. The optimization finished on the entire data (once on all the mini-batches) is termed as one training epoch.

$$bce = -(ylog(y') + (1-y)log(1-y')) \tag{2.18}$$

$$mse(w) = \sum_{n=1}^{N} \frac{1}{2N} ||(y - f(x,w))||^2) \tag{2.19}$$

2.4 Text-to-Speech and Voice Conversion

In this section, I provide the conventional TTS and VC frameworks in practice.

2.4.1 Voice Conversion

Voice conversion (VC) is a technique used to convert the source speaker's voice to that of a target speaker without affecting the linguistic content in the speech sample [106, 107]. There are various applications for VC such as anonymization of

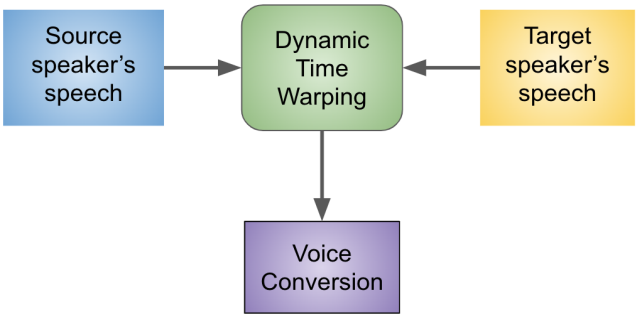

Fig. 2.5: A schematic of a traditional Voice Conversion framework

.

the speakers (especially in the health care domain for the privacy of the patients), personalization of synthetic voices, and movie dubbings, etc., There are three types of VC namely, parallel VC [108, 109], non-parallel VC [110], cross-lingual VC [111, 112, 113]. In a parallel VC, the speaker transformation is carried out for the same set of sentences being delivered by the source speaker and the target speaker. In non-parallel VC, the source and the target deliver a different set of sentences. Cross-lingual VC is a special type of non-parallel VC where the source and target deliver the speech samples in two different languages. Traditional VC techniques consist of the following steps: alignment of the source and target speech samples [114], spectral conversion, F0 transformation, and resynthesis of speech. Figure 2.5 displays the schematic of a traditional VC framework.

Initial VC research consisted of codebook-based spectrum conversion between the source and the target speakers using various vector quantization techniques [115, 116]. This is achieved in two steps, firstly the codebook for the source and the target speakers is prepared, secondly, a mapping of these codebooks (mapping of source features to the corresponding target features) is performed followed by the speech generation. The studies display that a decent conversion quality and speaker individuality can be achieved even with limited training data. In addition to quantized mapping of spectral features, authors in [116] also employ the speaker adaptation while performing the fuzzy mapping of codebooks derived from multiple speakers. The evaluation results presented the effectiveness of the technique in cross-gender conversion. Especially, in the male-to-female conversion (the converted voice in male-to-female conversion was recognized as the female voice most frequently). Later on, statistical models such as Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) were employed for speaker conversion [117, 118, 119]. [117] leverage the maximum likelihood estimation in achieving high-quality frame-based spectral conversion. Further, the over-smoothing of the converted speech is addressed by utilizing the global variance derived from the target's utterances. Con-

sequently, [118] describes the voice conversion system submitted to the Voice Conversion Challenge (VCC 2016). The authors deal with the over-smoothing (which leads to poor-quality speech) and parameterization errors through waveform modification enabled by a differential filtering of the spectral features. The evaluation results report higher speaker conversion accuracy and the generated speech's naturalness. Even though GMMs and HMMs can produce speaker individuality (a similar voice to that of the target speaker) in the converted speech, the conversion quality is affected due to their frame-based conversion. In order to address this problem, the authors in [119], proposed the probabilistic feature mapping that enables sequence-wise identity (speaker) conversion. The proposed method can effectively handle the inconsistencies that arise due to the use of dynamic features during the speaker conversion. The evaluation results display a better quality conversion than the (then) prevalent VC methods. Furthermore, [120] presents a survey on various VC techniques (then prevalent techniques) starting from the quantized map-ping to the probabilistic models such as GMMs and HMMs. The main limitation of the prevalent techniques as mentioned in this survey was the poor quality conversion (muffled voices) due to the frame-based mapping. On the other hand, [98] presents the speaker transformation using Artificial Neural Networks. The authors present six different experiments carried out with various NN architectures. Finally, the 4-layer architecture (25L, 50N, 50N, 25L) presents good conversion quality (based on the objective metric, Mel Cepstral Distortion, MCD). Comparing the objective, subjective evaluations performed for the converted voices with ANNs and GMM-Maximum Likelihood Parameter Generation (GMM with MLPG) displays a better speaker conversion through ANNs. Though NNs can outperform GMMs and HMMs in quality of conversion, due to the frame-based modeling they still suffer from restricted quality and loss of temporal information in the speech. Therefore, sequential networks have been investigated in [121] along with their counterparts, DNNs. The authors employ a deep bidirectional LSTM for a cross-gender conversion (male-to-female conversion). The study reports a significant improvement in the naturalness of the converted speech. Recently, Generative Adversarial Networks (GAN) have been explored for speaker conversion [122]. A GAN consists of 2 networks, a generator, and a discriminator. The generator is a neural network (it could be an RNN or CNN or DNN). It takes input speech frames and generates the possibly transformed output frames (produces new data or realistic predictions). The discriminator is also a NN that tries to classify if the generated frames belong to the target speaker. These are called the *adversarial* networks as they compete with each other. The generator network produces as realistic predictions as possible to fool the discriminator, while the discriminator aims at correctly distinguishing between the real and fake predictions. The model training and optimization are enabled by comparing the expected targets and the predicted targets. There have also been variants of GANs that were introduced such as STAR-GAN [123] and Cycle GAN [124]. GANs were efficient in addressing the over-smoothing problem encountered previously in VC voices. Therefore, different variants of GANs are widely used in Voice Conversion experiments for emotion conversion [125], limited data conversion [126], non-parallel conversion

[123], and many more. In this, thesis, the Star-GAN model has been explored for VC experiments presented in chapter 6.

2.4.2 Vocoders

Waveform synthesis (or speech signal reconstruction) in both TTS and VC is handled by the source-filter vocoder models such as STRAIGHT [127], WORLD [128]. STRAIGHT stands for Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum. STRAIGHT is used for both speech feature extraction and also high-quality speech reconstruction. WORLD is another source-filter model that enables speech analysis, manipulation, and reconstruction. In this thesis, we utilize the WORLD vocoder for the speech reconstruction of VC voices presented in chapter 6. The speech features extracted using the STRAIGHT or WORLD vocoders are spectrograms (spectral envelope), aperiodicities, and fundamental frequency.

Additionally, neural vocoders have gained much attention in the recent past [5]. Wavenet [5] is an efficient generative model that employs probabilistic autoregression for the generation of high-quality human-like sounding speech. Wavenet uses dilated causal convolutions and generates speech on a sample basis [129]. Even though wavenet uses the convolutional layers for training, the speech generation is slow due to its sequential processing. The network can also learn the speaker-specific characteristics by conditioning the model on the speaker labels (speaker IDs) when provided with a multi-speaker database. Apart from its sequential processing of speech samples (processes sample-by-sample), the wavenet needs to be trained using additional features like linguistic features, fundamental frequency, and durations for speech generation. Another commonly used speech signal reconstruction technique is the griffin-lim algorithm [130]. In this approach, the signal generation is carried out by phase reconstruction from the spectrograms without requiring the previous knowledge of the target signal to be synthesized. In this thesis, we utilize wavenet in chapter 7 for the TTS experiments and the griffin-lim reconstruction for the TTS experiments presented in chapters 3 and 4.

2.4.3 Text-to-Speech

Text-to-Speech synthesis (TTS) as the name suggests is a mechanism that generates speech when provided an input text. Typical TTS systems consist of the following components, a) text processing module, b) grapheme-to-phoneme conversion, c) acoustic modeling, duration modeling, and d) waveform generation. Text processing involves text normalization, tokenization, parts-of-Speech tagging, punctuations, and letter-to-sound rules. Festival [131], the speech synthesis toolkit is mostly preferred as the front-end for the text processing tasks in a conventional TTS setup. In this

thesis, Festival generated voices have been evaluated for their social perceptions (in chapter 4).

The evolution of TTS research over the years can be presented as a) rule-based synthesis [132], b) formant synthesizers [133], c) articulatory synthesis [134], d) concatenative synthesis [135, 136], e) HMM-based synthesis [137], f) DNNs [97], g) end-to-end TTS [4, 6, 23]. In rule-based speech synthesis, as the name suggests the speech generation is carried out by the rules designed by the experts. Formant synthesizer and articulatory synthesis are also rule-based synthesis techniques defined by the formants, formant frequencies, the vocal tract shape, and the speech articulators like tongue, lips, and jaws.

In concatenative speech synthesis, speech generation is achieved by concatenating smaller units of pre-recorded speech segments. Since the speech generation is carried out from the original speech segments, the perceived speech sounds natural. These small chunks of speech could be phones or diphones or syllables. Unit Selection Synthesis (USS) is a concatenative speech generation approach that utilizes a large data inventory for concatenative speech synthesis. The choice of appropriate units from a large inventory is carried out using a Viterbi search algorithm [138]. Further, speech generation is enabled through signal processing techniques such as Overlap Addition over the target units [139].

The speech synthesis carried out based on HMMs and NNs is called the statistical parametric speech synthesis (SPSS) as it utilizes the parametric representation of speech. The advantage of parametric models over concatenative synthesis is that they enable the modeling of acoustic parameters during the synthesis procedure. The traditional HMM-based speech synthesis techniques employ decision trees for the mapping of linguistic information onto acoustic space [137, 140, 141]. HMMs and DNNs as observed in VC literature, do not perform well for speech synthesis due to their frame-based modeling. This was handled to some extent by the use of dynamic features in a DNN-based speech synthesis [97]. However, eventually, sequence-to-sequence-based speech generation has been found to render better quality speech over the other parametric models [142].

Current TTS research employs end-to-end models such as Tacotron, and Tacotron 2 [4, 6]. The tacotron network can generate speech when provided with the characters as the input. It was introduced in 2017 to replace multiple blocks in the TTS framework (text processing block, acoustic modeling, duration modeling, speech generation) that were prevalent. Thus, the temporal modeling of speech sentences was enabled through the use of sequence-to-sequence networks. The network leverages the convolutional layers and the bi-directional GRUs and the attention-based decoding for the generation of raw spectrograms. Figure 2.6 displays the flowchart of a traditional tacotron. The speech signal reconstruction is carried out by the Griffin-Lim algorithm [130]. Later, a modification to the Tacotron network, was proposed which was eventually called, Tacotron2. This network eliminates the complex structure of Tacotron by the use of vanilla LSTM layers. Further, the waveform generation in Tacotron2 unlike in Tacotron is carried out by the Wavenet vocoder. In this thesis, the speech samples generated using the tacotron model were utilized for the subjective evaluations presented in chapters 3, 4. Further, the TTS experiments in chapter

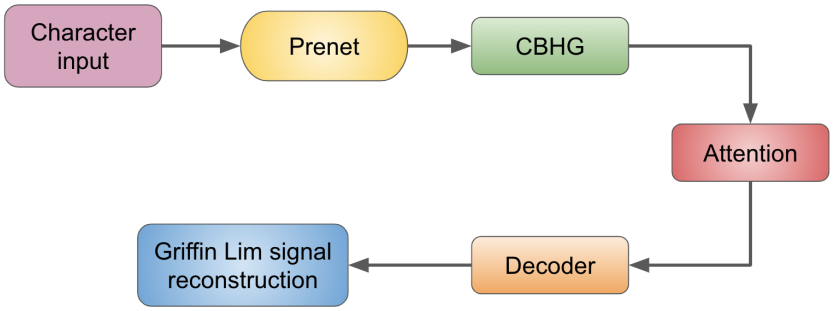

Fig. 2.6: Flowchart of the Tacotron.

.

7 were carried out on the tacotron model along with the wavenet vocoder for speech signal reconstruction.

On the other hand, several toolkits have been developed by various research institutes for speech synthesis across the world. Merlin is a speech synthesis toolkit developed at CSTR, (by the speech group at the University of Edinburgh) [143]. This toolkit performs a DNN-based speech generation which is trained using Theano [144]. Another DNN-based TTS open-source platform is IBM Watson. An online demo version is also available that supports 9 different languages[1]. The current demo version provides expressive neural voices. MARY TTS is an open-source TTS platform developed by the Language Technology Lab at the DFKI in association with the Phonetics department at Saarland University[2]. The platform supports USS and HMM-based speech generation. The Google TTS engine [3] supports more than 200 voices and more than 40 languages. Speech generation is available in three different voice types namely, standard, wavenet, and neural2. Amazon Polly [4] provides the TTS speech samples with two voice types, neural and standard. It supports 29 languages and around 61 voices. The API versions of the commercial TTS systems are also available for research purposes. One can generate and also download speech samples for one's research using these platforms. In this thesis, I have utilized the speech samples generated from Mary TTS, IBM Watson, Google, and Amazon Polly for the subjective tests presented in chapter 4.

[1] https://www.ibm.com/demos/live/tts-demo/self-service/home

[2] http://mary.dfki.de/

[3] https://cloud.google.com/text-to-speech

[4] https://eu-central-1.console.aws.amazon.com/polly/home/SynthesizeSpeech

2.5 Evaluation of TTS and VC

This section details the traditional evaluation setups prevalent in TTS and VC research followed by the details of the evaluation metrics included in this thesis. Further, details of in-lab and crowd-sourcing subjective tests are also provided.

Even though objective measures are also commonly used in TTS and VC research, in the scope of this thesis, only subjective evaluations of the generated voices are presented. The subjective perceptions are standard, and much more reliable than objective measures as they include human judgments [145]. Further, we are interested in understanding the social perceptions of synthetic voices. There have not been any prior works on the assessment of warmth, and competence in synthetic speech. Therefore, as an initial attempt, we explore the subjective perceptions of these characteristics from the generated speech. The well-known and widely used subjective assessment tests on the generated speech are a) intelligibility (how clearly one can comprehend the speech), b) naturalness, c) speech quality, and d) speaker similarity [146, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44]. Intelligibility is measured by calculating the number of words that are correctly identified from the generated speech. Speech quality, as the name suggests evaluates the quality of the perceived speech. Naturalness provides information on how close (natural) the listeners perceive the voice to be to that of human speech. Speaker similarity is calculated to interpret the similarity of the converted (in the case of VC) or generated (in a TTS setup) voice to that of the target/original speaker. In the current thesis, we discuss the evaluation of synthetic speech using naturalness, speech quality, and speaker similarity. All these metrics can be measured using absolute 5-point scales. ITU-T Recommendation P.85 [147] was developed in 1994 for the evaluation of synthetic voices on various scales such as pleasantness, listening effort, overall quality, etc., The quality labels as provided in the ITU-T Rec P.800 [148] are presented in table 2.1 along with the naturalness and speaker similarity. The subjective ratings collected for each utterance from different individuals (participants) are averaged and correspondingly Mean Opinion Scores (MOS) are derived for each test condition. These tests can also be called direct scaling tests as they define the questions (naturalness, quality, similarity) and also each point on the scale.

Table 2.1: The labels for each point on the 5-point absolute scale as used in the evaluation of speech quality, and naturalness.

Score	Speech quality	Naturalness	Speaker similarity
5	Excellent	Highly natural	Sounds very similar
4	Good	Sounds natural	Sounds similar
3	Fair	Somewhat natural	Sounds somewhat similar
2	Poor	Not natural	Does not sound similar
1	Bad	Completely unnatural	Not at all similar

Another commonly used subjective test for the generated speech is the AB preference test [149]. In this test, the participants are provided with two speech samples (A, B) and are free to choose any of those based on the questionnaire (naturalness or speech quality). Some experimenters also provide another option, "No preference" during this test. If the listeners are not provided with the third option then it's called the forced preference test (as they are compelled to choose between A or B). There is also another test called, the ABX preference test. An ABX test additionally consists of a reference speech sample. The participants listen to three speech samples, sample A generated from system 1, and sample B generated from system 2, sample X from the original target speaker. The participants should choose which among A or B is closer to that of the target speaker. In this thesis, along with speech quality, naturalness, and speaker similarity, an AB (with an option for No preference) preference test is carried out for the perception of warmth, and competence from the converted voices, ABX test for speaker similarity of the converted and the target speaker's speech (VC experiments in chapter 6). Additionally, there are also semantic differential scales that measure different perceptions of the participants of a speech sample over the bipolar scales (adjective-antonym pairs at the extremes) [150]. These scales as described previously, are widely used in behavioral research to evaluate various personalities or speaker characteristics. In this thesis, we utilize semantic differential scaling tests for the evaluation of warmth and competence from synthetic speech (in chapters 4, 6, 7).

2.5.1 In-lab and crowdsourcing-based subjective evaluation

The in-lab subjective evaluations are conducted in noise-free and acoustically damped room settings where the listening environment is in the control of the experimenters. In comparison, crowd-sourcing-based assessments are carried out on publicly available platforms like Amazon Mechanical Turk (AMT). In these studies, the listening environment of the participants is not under the control of the experimenter. The number of participants to take part in the study, nativity of the participants, age, gender, and the number of speech samples to be provided during the test (for both in-lab and crowd-sourcing studies) are determined by the experimenter. Even though an in-lab subjective test can create an ideal environment for listening tests, there are some drawbacks, a) not a realistic listening environment, b) costly when compared to crowd-sourcing-based studies, and c) difficulty in finding participants that belong to a specific group. Also, during the pandemic situation, performing subjective studies in an in-lab setting was practically not possible. Thus, crowd-sourcing research and platforms have gained much attention in the last couple of years. The ITU-T Rec. P.808 [151] presents the guidelines for conducting subjective evaluations through crowd-sourcing setups. This was further tested and shown to be reliable through the studies presented in [152].

According to the ITU-T Rec. P.808, there are some considerations one must not ignore while designing a crowd-sourcing-based evaluation. As per the ITU

recommendations, the test duration should last only a few minutes. Therefore, it is recommended to split the data into smaller chunks. The number of questions (speech stimuli), to be provided during the subjective tests via crowd-sourcing (for the crowd-worker to finish a task in a couple of minutes) should be limited to 5-15 [148, 151, 152]. As a result, each crowd-worker would take part in one or more tasks but not all the tasks (the raters would rate only a subset of the entire study). However, this would give rise to error variance because of the corpus effect in the collected subjective data. Therefore, the experimenter must design multiple (adequate) test conditions such that the raters in each pool would rate all the samples (one entire condition) in the given subset of data (5-15 stimuli). Also, one can encourage the crowd-workers to take part in multiple jobs by providing them with bonuses (extra rewards).

Based on the ITU recommendations, the crowd-sourcing assessments should consist of three jobs namely, a) qualification, b) training, and c) rating. Under the qualification job, the experimenter verifies the eligibility of the participant for the designed subjective test. It includes questions that would verify hearing impairments (if any) among the registered participants. Also, if the participants are comfortable they can also disclose their age and gender which would further aid the researcher in the analysis of the collected subjective data. Depending on the requirement of the experiment, the experimenter can then choose the group of participants for the study. Also, one can choose the participants based on their previous performances in the crowd-sourcing tasks (based on the number of approved tasks). The second job is to train/familiarize the participants with the evaluation setup. This job must include speech samples that are representative of both the best and worst quality speech samples. The participants are not given any information about the data distribution or the speech samples being provided in this job. This job would train the participants for the actual rating job. In turn, if the participants fail to perform the rating job within 24 hours of finishing the training job, they cannot perform the rating job and will be redirected to the training job. The rating job consists of some initial steps. Firstly, the participants are asked about the listening environment, system (headphones, speakers, etc.,), and the option to adjust the sound level before the actual test begins. The participants should be instructed to perform the task in a noise-free environment. This can be further ensured based on their responses to the quality comparison tests. If the frequency (number of occurrences) of correctly selected voice (voice with good quality speech should be selected in a comparison test when provided with two speech samples with a varied speech quality) is high then the experimenter can interpret that the participant was indeed in a noise-free environment. Also, the participants need to be informed to wear two-eared headphones. This condition can also be verified by providing a small maths question, where the numbers are played at random in the left and right (ear) headphones. Also, the participants should set the volume to a level that's most comfortable for them. They should be instructed to not change this volume during the test after this step. [153] presents a study on including a gold standard question (trapping question) during the subjective tests. This is an additional question included in the test along with all the other speech stimuli. The participants who pay attention during the test only can answer this correctly. The

researcher can thus filter out the participants based on their responses to this trapping question. Additionally, it is also recommended to include another trapping question where the speaker (within the speech stimuli presented during the test) asks to select an option from the given choices. The participants who play and listen to all the speech samples alone can be retained through this mechanism. Further, the collected subjective data is finally processed to verify if all the above-mentioned conditions are met. The participant information can be discarded if a) they cannot answer one or both the trapping questions correctly, b) listening environment conditions are not as specified, and c) the listening system was not set up properly. Also, the experimenter can remove the participants' data if found any outliers or abnormal patterns in the subjective responses.

In this thesis, the subjective evaluations were carried out through both in-lab (in chapter 3, 4, 6, 7) as well as crowd-sourcing-based experiments (in chapter 4). The evaluation of various perceptual dimensions of synthetic speech (except the speech quality, naturalness, and speaker similarity) is performed using the continuous 100-point scales available in TheFragebogen [154, 151, 152]. It is an open-source platform that supports the assessment of speech samples through various questionnaires related to Quality of Experience (QoE). The software is available in Javascript and the scales, and questionnaires can be modified for different experiments in HTML. The continuous 100-point scale displays the adjective-antonym pair at the extremes of the scale. This would enable the participants to choose any point on the scale, instead of restricting them to specific points as in absolute scales or Likert scales. The evaluation of various adjectives was carried out using continuous 100-point scales in chapter 4. The speech quality and naturalness assessments of the VC and TTS voices in chapters 4, 6, 7 are performed using 5-point Likert scales. The VC experiments in chapter 6 were evaluated using both the AB tests, and 5-point direct scaling tests for each warmth, and competence. Finally, chapter 7 presents the evaluation of the TTS for SSC on a 5-point scale.

2.6 Summary

This chapter presents the necessary background and relevant studies from the literature for the current work. All of this information is structured into four different parts. The first part provides insights into understanding various behaviors, personalities, characteristics, emotions, and the differences between each of these dimensions in humans and machines. The second part deals with the machine-learning approaches that could be utilized for the study of various aspects of SSC from the synthetic speech (specific to this thesis). The third part details the evolution of VC and TTS research. The experiments carried out in this thesis using TTS and VC setups are derived from the state-of-the models discussed in this part. The last part of this chapter presents information on various aspects of the subjective evaluation of synthetic speech.

Chapter 3
Choice of datasets and adjectives

This chapter presents the datasets used for different studies carried out in this thesis followed by the adjectives used in the subjective evaluation setup. These details are provided while discussing the challenges encountered in the initial subjective tests. Apart from the publicly available datasets, the others deployed in this thesis are hand-picked by myself from various sources (corresponding links are provided). I have alone done the collection and finalized the adjectives through different subjective evaluations.

3.1 Related work

This section discusses the datasets previously used for the study of social speaker characteristics in different experimental setups. The behavioral studies performed on humans (discussed in the previous chapters) consisted of a complete assessment of a person by a known acquaintance [56, 155]. Further, [156] provides warmth and competence judgments based on non-verbal cues. They report the association of the non-verbal cue, "smile" with an increased perception of warmth than competence. In [157] authors present the effect of articulatory cues during the pronunciation of the vowels, /i:/ and /u:/. The study claims that the speakers were perceived as much warmer and more competent when they utter the usernames with the vowel, /i:/ over the usernames with /u:/. This is due to the facial gestures created while pronouncing the vowel, /i:/. The speaker characteristics such as age, gender, and dialect [158] were previously studied from the phoneme sequences. Authors in [75, 76] collect semi-spontaneous conversations (ordering a pizza over a phone call) from 300 German speakers for the studies on speaker characterization. In their study, perceptual analysis was carried out in zero-acquaintance scenarios. [159] investigate the automatic prediction of personality traits from synthetic speech. The speech samples used were 5 sentences among which two were nonsensical words in the native language and the rest three were, 'Thank you', 'How are you', and 'I love you' (in the listener's native language).

Nevertheless, for this work, we were interested in understanding the social perceptions of the generated speech (TTS). The goal was to further include these additional dimensions in the evaluation of speech generation systems (TTS and VC). Accordingly, the choice of datasets for our studies was dependent on three criteria, a) utterances that would aid in the perception of warmth and competence from speech, b) study of SSC in a speech-alone scenario (focus on acoustic features alone), and c) perceptual studies on synthetic speech. Correspondingly, various datasets and adjectives were investigated. The details of these studies are provided further.

3.2 Challenges

This section provides the challenges encountered in the selection of datasets and the preparation of questionnaires (or scales or adjectives or speaker attributes) for subjective tests. These challenges are discussed while providing the details of the preliminary studies carried out for the social perceptions of synthetic voices. The preliminary studies were therefore carried out to examine the following, a) choice of datasets, and b) scales to be used in the subjective evaluation.

3.2.1 Choice of datasets

All the datasets used in this thesis are in English (speakers with a US accent). In the initial studies, the utterances that displayed care, compassion, and assurance in addressing the customer's/patient's problems were derived from different sources [1] [2]. The generated dataset was termed as WC dataset (WC = warmth, competence). Further, a preliminary study was designed with this dataset for the scales to be used in the evaluation (details provided in the section 3.2.2). The analysis of the subjective responses (actual test), has shown that the content (text) has a significant influence on the perception of warmth and competence (discussed in chapter 4). The primary aim of this thesis was to examine the perception of warmth and competence from ***synthetic speech*** and incorporate those using different ***acoustic feature*** modification techniques if necessary. Therefore, in order to narrow down the dependence of the ***speech perception*** on the content, the subsequent perceptual studies were carried out on neutral speech (Chapter 4, 5, 6). However, later on, perceptions of compassionate speech have also been studied and presented in chapter 5 (presents a small study on how duration affects the perception of warmth and competence in compassionate speech), chapter 7 (investigates whether the acoustic correlates of warmth and competence derived for neutral speech can be generalized for compassionate speech).

[1] https://www.fluentu.com/blog/business-english/how-to-talk-with-customer-in-english/

[2] https://www.verywellmind.com/what-to-say-when-someone-is-depressed-1067474

3.2.2 How to evaluate warmth and competence from synthetic speech?

A preliminary study was carried out for the perception of warmth and competence from synthetic speech. The study consisted of 15 participants (male = 7, female = 8). Their ages ranged between 21-32 (mean = 24.5, std = 2.3). The speech samples used for the subjective tests consisted of 2 male (bdl, rms) and 2 (slt, clb) female voices from the CMU arctic database [160]. These voices were generated using the traditional Tacotron model following a similar architecture as presented in [4]. The sentences produced from these four voices for the initial subjective tests were from the WC dataset. During the evaluation, the participants were asked to rate the speech samples on a scale of 5 for warmth and competence (1 = cold/incompetent, 5 = warm/competent). The behavioral studies presented in [15] state that the impressions of the intellectual dimension (competent) are made easily and also prior to the social dimension, warmth. However, in our study, the questions provided to the participants are randomized and the phenomena mentioned above cannot be observed based on the design of the evaluation setup. Nevertheless, we made an observation that might be in line with the above-mentioned study to some extent. The participants could easily rate the speech samples on the scale of competence. But, in order to evaluate the perceptual dimension, warmth, they required additional adjectives that would best describe the characteristic. Based on the interactions with the participants and the analysis of the subjective responses, it is understood that the evaluation of warmth would require a different set of questions during the subjective evaluation.

3.3 Datasets

This section details the datasets used for different studies conducted in the thesis. These studies can be divided into two parts, a) Perceptual studies (studies on social perceptions of synthetic voices, and their acoustic analysis), and b) Modeling of the synthetic speech (VC and TTS experiments). The datasets used for each of these studies are further provided in detail.

3.3.1 Datasets for perpetual studies

The perceptual studies presented in chapter 4 included two different datasets namely, a) WC dataset, and b) neutral speech. The WC dataset consists of 10 sentences. The text used to generate synthetic speech is provided in Appendix (A.1). The experiments on the neutral speech were carried out using two phonetically balanced datasets, namely the Harvard database[3] and CMU arctic database[4] [160]. These are publicly

[3] https://www.cs.columbia.edu/ hgs/audio/harvard.html

[4] http://festvox.org/cmu_arctic/

available datasets and are widely used in speech research [121, 108]. The perceptual analysis (on neutral speech) in chapter 4 and the acoustic analysis in chapter 5 were carried out on the Harvard database (number of sentences used = 32). The perceptual analysis of the generated speech has shown that the speaker characteristics remain constant across different speech segments used in the study. Therefore, by leveraging this advantage, in the latter experiments, CMU arctic database (same data type = neutral speech, number of sentences = 1132) was used for training different NN models. The NN-based classification of SSC (automatic prediction of SSC) using the acoustic correlates of warmth and competence presented in chapter 5 was performed on the arctic dataset.

3.3.2 Datasets for modeling of synthetic speech

This section presents the details of the datasets used for modeling of the synthetic speech using VC in chapter 6 and, training and testing the TTS models employed in chapter 7.

Chapter 6 presents the VC experiments carried out on the synthetic voices for altering the negatively perceived voices into positive ones. The modeling of the synthetic speech using VC techniques is carried out using the CMU arctic database.

Artificial speech generation through neural models requires abundant data for synthesizing natural-sounding speech [4, 6, 5, 161]. The end-to-end TTS models, Tacotron and Tacotron 2 were developed using 24.6 hours of female speaker's speech [4, 6]. The Deep Voice utilizes 20 hours of speech that comprises 13,079 utterances [161]. Additionally, the model was trained on a subset of Blizzard 2013 data [36]. Wavenet vocoder [5] was built with 44 hours of speech data derived from the VCTK corpus rendered by 109 speakers [162]. Further, in the TTS experiments (using wavenet as the vocoder), the model was trained on single-speaker databases, which were 24.6 hours of English speech, and 34.8 hours of Mandarin Chinese. Similarly, the Tacotron model (used in this thesis in chapter 7) was trained on 24 hours of speech data in the current work. The description of the dataset is provided below.

3.3.2.1 LJSpeech dataset

The LJSpeech dataset consists of approximately 24 hours of a single speaker's speech data rendered by a female speaker [163]. The speaker reads out passages from non-fiction books which constitute about 13,100 speech samples (each sample duration is approx. 10 seconds). This dataset is publicly available and is widely used in TTS research [164, 165, 166]. The tacotron model employed in chapter 7 was trained on this dataset.

3.3.2.2 Twitter data

The evaluation of the generated speech samples using the TTS models presented in chapter 7 was done with the Twitter dataset (choice of the dataset for the experimental setup is explained in detail in chapter 7). The sentences that expressed compassion, empathy, and generosity were hand-picked from various Twitter posts and threads. While doing so, the experimenter made sure that there were no sentences/words that would stimulate any negative impressions on the listeners. The sentences used in the evaluation are provided in Appendix (A.1).

3.4 Choice of adjectives

From the section 3.2.2, we have observed that the perception of warmth requires additional adjectives during the subjective evaluation. Also, from the literature on human behavior, it is evident that behaviors or characteristics, or personalities of a person are determined based on multiple dimensions (through the assessment of a person on various adjectives). Therefore, we employ a similar procedure for our line of work. The adjectives to be included in the perceptual studies were determined through a 2-step procedure. In the first step, an in-lab subjective evaluation was carried out with 15 participants. Their ages ranged between 21 to 40 (mean = 29.3, std = 5.3). In this study, the participants were asked to describe the speaker's voice using various adjectives. Later on, the participants were requested to provide the adjectives they felt as essential for health care professionals and customer service agents. In the second step, inspired by [69, 70, 75, 76], various adjectives employed in the studies of human perception were accumulated. Some of the adjectives collected from the first step correlated with the ones found in the literature. Finally, through these two steps, a list of 66 adjectives has been derived. The finalized adjectives list is further presented in Appendix (A.2). However, how many among these 66 can be perceived from the synthetic speech is not known from this study (since some of the adjectives were included a) that are specific to the desired application domains, and b) some from the studies on understanding human behavior and personality). Therefore, follow-up work on the current 2-step procedure is presented in chapter 4 to determine the list of adjectives that can be perceived from synthetic speech.

3.5 Summary

This chapter discusses the initial steps toward assessing the SSC from synthetic speech. The chapter begins with a discussion of the datasets used previously for various perceptual studies. Further, we provide some insights into the challenges encountered in the preliminary studies which aided in the design choices of our experimental setup. Later on, the datasets used for different experiments in this thesis

are presented. Finally, the chapter concludes with a description of the questionnaire preparation (adjectives) required for the subjective tests.

Chapter 4
Social perceptions of synthetic speech

This chapter details the studies performed on social perceptions of synthetic voices. Two different studies are presented, a) perceptual studies on a wide range of TTS systems, and b) two commercial TTS systems. A discussion on the experimental setup was done between me, Benjamin Weiss, Sebastian Möller, and Babak Naderi. The second half of the chapter (studies on two commercial systems) are from the work published in [167]. Therefore the content outlined would be closely related to that work.

4.1 Related work

This work was inspired by [75, 76] in speaker characterization through peer-ratings on unknown speakers. The study was carried out on human speech in German. [76] presents a 34-item semantic differential scaling test for interpersonal speaker characterization from semi-spontaneous conversations. The first impressions of the listeners as observed by the authors in the study were warmth, attractiveness, confidence, compliance, and maturity. The motivation of the work was that it is not easy to provide personality judgments (on BIG FIVE traits, O, C, E, A, N) when only confronted with speech (zero-acquaintance scenarios) [75]. Therefore, the study includes multiple adjectives representing various behaviors and personality traits. According to [66] the sample adjectives that represent different personality traits are as follows: a) Openness to Experience = Artistic, Curious, Imaginative, Insightful, Original, Wide interests, b) Conscientiousness= Efficient, Organized, Planful, Reliable, Responsible, Thorough, c) Extraversion = Active, Assertive, Energetic, Outgoing, Talkative, d) Agreeableness = Appreciative, Kind, Generous, Forgiving, Sympathetic, Trusting, e) Neuroticism = Anxious, Self-pitying, Tense, Touchy, Unstable, Worrying. Further [168] presents the paralinguistic speaker challenge conducted for assessing personalities from speech using these adjectives. The experiments were performed on the SPC corpus. The study incorporates the BFI-10 (a subset of the BFI-44) questionnaire for personality assessments. Similar to studies presented in [75, 76],

authors in [86] assess the personality of the speaker in zero-acquaintance scenarios. They present the automatic prediction of personality from a single speaker's speech followed by a comparison of the human evaluations. The study shows that the trait, openness to experience cannot be understood from speech using the NEO-FFI questionnaire. [169] also investigates the relationships between personality traits and their perceptions from speech. The analysis shows that extroverted speakers speak louder without any pauses or stalling in between. However, all these studies were previously carried out on natural (human) speech.

[170] present the personality assessments of synthetic speech in the case of concatenative as well as parametric speech synthesizers. The study investigates the effect of the content being said, the naturalness of the generated speech, and the voice type (tensed voice, lax voice) on speech perceptions. The study displays a significant impact of the voice quality and the synthesizer used, on the personality judgments. In [171], authors investigate the perceptions of the brand personalities (sincerity, excitement, competence, sophistication, ruggedness) from the synthetic speech. The study presents the perceptions of these dimensions from German speech generated through Mary TTS. The attributes that were representing the brand dimension competence were reliability, intelligence, and success. These attributes were obtained through factor analysis of the personality dimensions that associate with specific brands in markets. This analysis was presented in [172]. In this work, the authors present the study on customers' behaviors and responses to specific brands analogous to the studies on human personality. [173] presents the study on the multi-facet nature of the fundamental dimensions. They propose to distinguish the social dimension into warmth and morality, and the intellectual dimension into assertiveness and competence. The study also claims that the perceptions of these dimensions are stable across different cultures and regions.

Overall, from these studies, we can conclude that a) the study of personality (or any human behaviors) requires additional dimensions/adjectives in the case of speech-alone scenarios (we have also observed this from our preliminary studies presented in chapter 3. Therefore, this holds true even in the case of synthetic voices.), and b) the first impressions made by humans are not limited to other humans or animals (living beings) but also other domains (non-living things) such as brands (also generated speech). Therefore, as a follow-work for these studies, in the current work, we focus on the assessment of speech perceptions from synthetic voices in English on multiple perceptual dimensions. Following the studies, [75, 76], various adjectives were derived from the behavioral and personality research for the same [69, 70]. Further, the first impressions (social dimensions) of synthetic speech were obtained using Exploratory Factor Analysis (EFA) on the collected subjective data. These studies aid in answering the following research question.

? Research question: ***"What social speaker characteristics do people perceive from synthetic speech?"***

4.2 Study A: A case study on wide-range TTS systems

This section provides the details of the preliminary research conducted with the choice of speech data and the adjectives prepared so far.

4.2.1 Experimental setup

4.2.1.1 TTS voices

Although the research on the adjectives to be included in the subjective setup was derived for the Neural TTS voices (Tacotron, as presented in chapter 3), in the current study, the subjective perceptions of a wide variety of TTS systems were analyzed. This was done to ensure that all the variations of synthetic speech perception could be included in the study. The TTS systems that are available in both industry and academia were researched for the same. Finally, the study consisted of five different TTS systems. Among them, the academic systems used in the study were, Festival (Clustergen) [131], Mary TTS [1] [174], Tacotron [4]. Correspondingly, the commercial ones were, Google (Wavenet-A, B, C, D, E, F, standard = B, C, D, E)[2], IBM Watson[3]. A total of 41 voices (male = 23, female = 18) were collected from these TTS systems. The list of the TTS systems and the number of voices collected from each system are provided in table 4.1.

Table 4.1: List of the TTS voices used in the study.

TTS system	Number of voices	Male	Female
Tacotron	5	3	2
Festival	5	3	2
Mary TTS	9	6	3
IBM Watson	6	2	4
Google	16	9	7

The academic systems, Tacotron and Festival were trained on the CMU arctic database [160]. The 3 male voices from Tacotron and Festival were of bdl, rms, jmk. The 2 female speakers are slt and clb. The Mary TTS voices were also from the arctic database. The TTS system consisted of three speech generation mechanisms namely, Unit Selection Synthesis (USS) [175], Hidden Semi Markov Models (HSMM) [141], and USS with prosody modification. The voices generated from these models were as follows: 2 male voices (bdl, rms) * 3 models (USS, HSMM, USS with prosody

[1] http://mary.dfki.de/

[2] https://cloud.google.com/text-to-speech/

[3] https://cloud.ibm.com/apidocs/text-to-speech

modification), 1 female (slt) * 3 models. The commercial TTS IBM Watson, had 3 models namely, enhanced DNN, transformable TTS, and expressive TTS. The voices generated from these models were as follows: 1 male * 2 models (transformable TTS, enhanced DNN), 1 female * 2 models (enhanced DNN, transformable TTS), and 1 female * 2 models (expressive TTS, transformable TTS). The Google voices used were from the Wavenet. There were 6 voices (3 male, 3 female). Additionally, the pitch variation and the speaking rate variation were also incorporated for 2 male and 2 female voices (male = B, D; female = C, E). Thus, the total number of voices derived from the Google TTS system was 16.

4.2.1.2 Questionnaire preparation

We have observed the list of adjectives that are derived from the 2-step procedure employed for questionnaire preparation in chapter 3. However, the perception of those adjectives from the synthetic speech is not known. Therefore, in the current study, the number of adjectives among the list of 66 adjectives that can be perceived from the synthetic speech is examined. The test was conducted with a new set of 15 participants. Their ages ranged between 25 to 48 (mean = 31.8, std = 5.87). The speech samples (2 male, 2 female) provided were the same as that in the study presented in chapter 3. A final list of questionnaires was compiled based on the frequency of the responses received (the adjectives that were selected as being perceived from TTS voices by more than 10 participants were selected). This list consisted of 34 adjectives and is provided in Table 4.2.

Table 4.2: Attributes for 34-Dimensional Semantic Scaling Test

Attributes			
Kind	Confident	Energetic	Outgoing
Distant	Talkative	Proactive	Tense
Empathetic	Calm	Introvert	Unsympathetic
Trusting	Worrying	Not irritated	Indecisive
Emotional	Secure	Old	Friendly
Relaxed	Reliable	Hearty	Arrogant
Assertive	Agreeable	Anxious	Pleasant
Responsible	Active	Cynical	
Enthusiastic	non-likable	Accessible	

Apart from the speech perceptions, we have also examined the text perceptions in this study. At the end of the listening test, the participants were provided with the sentences used in the survey and are asked to rate the sentences on a scale of 5 for different adjectives. Additionally, ten speech samples from the Harvard database (neutral speech) were included in the survey. The comparison of the subjective responses between the WC dataset (10 sentences) and the neutral speech (comparison

between speech as well as text separately) has shown that the text has a significant (2-sample t-tests between texts and speech samples indicated statistically significant differences, $p<.05$) influence on the perceptions of various speaker attributes.

4.2.1.3 Design of the survey

One of the challenges encountered while designing the actual survey was deciding the number of questions to be included in the test. The total questions with the current design choices would include, the number of TTS voices (41) * the number of sentences (10 from the WC dataset defined in chapter 3) * and the number of adjectives (34) = 13940 questions. The study was determined to be carried out in a lab environment. And the list of questions was rather high for such an evaluation. Even if we split the test into multiple parts each spanning for an hour it would require around 6-7 appointments for the participants to complete the study. In order to handle this situation, the number of sentences to be used in the study was reduced to 2. This is because the study aimed at the analysis of a wide range of TTS voices (so the number of voices could not be reduced). The sentences thus selected for the perceptual studies are provided below.

- Is there anything I can do to help?
- I am sure we can reach a solution.

Therefore, finally, the study consisted of TTS voices (41) * sentences (2) * adjectives (34) = 2788 questions. The evaluation setup was prepared using the publicly available framework, TheFragebogen [154]. The survey consisted of continuous 100-point scales prepared with the adjectives presented in table 4.2. The adjectives and their antonyms are defined at the extremes of the scale (semantic differential scaling test) [176]. The positive adjectives were at the extreme right and the negative ones were on the left. A short example of the survey is provided in figure 4.1.

For the current study, the nativity of the listeners, age, and gender were not considered as the deciding factors for their participation in the study. The number of participants to take part in the study was decided to be 25.

4.2.1.4 Subjective test

Among the listeners who signed up for the study, there were 15 male and 10 female participants. The age of the participants ranged from 20 to 56 (mean = 28.35, std=9.5). The education status of all the participants varied between high school and University degrees. The participants were provided with Sennheiser HD 449 headphones. The test was conducted in an acoustically damped room. On average, the time taken by all the participants to complete the test was 54 minutes. They were allowed to take a break during the test, every 10 minutes to avoid any fatigue. They could listen to the speech samples multiple times during the test. All the participants were compensated for participating in the test.

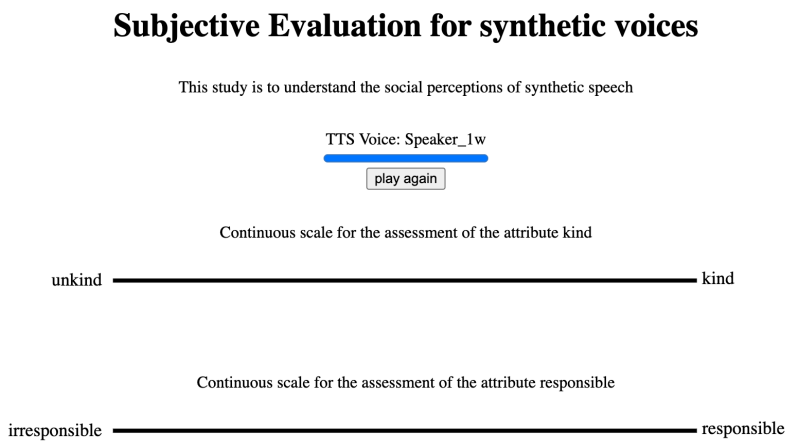

Fig. 4.1: Sample of the semantic differential scaling test used during the subjective evaluations
.

4.2.2 Analysis of subjective responses

4.2.2.1 Inter rater agreements

Intra-class correlation coefficient (ICC) is a reliable indicator and is highly recommended in deriving the inter-rater reliability scores from the subjective responses. The coefficient value provides the degree to which the raters concur with each other on a specific choice of ratings. Generally, the higher the coefficient value, the more closer and reliable the subjective responses (> 0.6 to a maximum of 1) [177, 178]. The ICC values were calculated for each of the adjectives used in the study. The coefficient values were computed for male and female speakers separately. The ICC range for female speakers was from 0.51 (for hearty) to 0.83 (for old) with an average of 0.72 (std = .08). The range for male speakers was from 0.35 (Secure) to 0.75 (Enthusiastic) with an average of 0.57 (std = 0.6).

4.2.2.2 Exploratory Factor Analysis

Exploratory Factor Analysis (EFA) as the name suggests, explores the complex data and defines the underlying structures [179]. EFA is a form of dimensionality reduction technique applied to a set of observed variables (factor loadings) in order to determine underlying latent variables (factors)[180]. Researchers choose EFA over other factor analysis approaches when the number of factors to be derived is

unknown. The variables loading under each factor and their values (factor loading values) would represent the amount of variation they have in that factor. If the value of a variable is the highest in a specific factor compared to others, then it most likely belongs to that factor. The number of factors to be derived is determined based on how well the data is structured in each experiment (factors = 3,4,5). Once the number of factors is determined, various rotations and a second-factor analysis are implemented to interpret the belongingness of a variable under a specific factor. There are different types of rotations such as Oblimin [181], Promax [182], Varimax [183]. The rotation of factors is essential for a better factor representation and to maximize the factor loading values. From the literature, we have observed that the social dimensions are orthogonal [55] and independent of each other [12]. In order to understand this phenomenon in the case of synthetic speech, in the current study, we have examined both varimax and oblimin rotations with a minimum residual factoring method (varimax performs an orthogonal rotation of the factors, oblimin is used when the factors are correlated). The factor analysis was carried out using the "factor_analyzer" package available in python.

Table 4.3: Factor loadings for female speakers

Attributes	Warmth	Competence	Extraversion
Hearty	0.79		
Distant	-0.78		
Pleasant	0.81		
Reliable	0.79		
Trusting	0.79		
Emotional	0.78		
Agreeable	0.78		
Energetic	0.75		
Unlikable	-0.83		
Sympathetic	0.82		
Enthusiastic	0.77		
Calm		0.78	
Tense		-0.72	
Relaxed		0.63	
Anxious		-0.6	
Introvert			-0.74
Outgoing			0.8
Talkative			0.67

A gender-dependent factor analysis was carried out on the collected subjective data. The number of factors that could best represent the data as observed from multiple analyses is three (the best representation of factors was obtained with varimax rotation). After the first analysis, the attributes were retained based on three criteria: i) when the main loading of an attribute was greater than 0.5, ii) the difference between the main loading and the cross-loading was at least 0.2, iii) the communality [184] (provides the variance information) values were higher than 0.4. The

Table 4.4: Factor loadings for male speakers. (*) indicates the attributes that are common for both male and female speakers under each of the derived factors.

Attributes	Warmth	Extraversion	Competence
Kind	0.8		
Hearty*	0.83		
Arrogant	-0.64		
Friendly	0.83		
Pleasant*	0.83		
Trusting*	0.77		
Agreeable*	0.75		
Empathetic	0.75		
Emotional*	0.71		
unlikable*	-0.79		
Sympathetic*	0.81		
Responsible	0.68		
Active		0.7	
Introvert*		-0.76	
Energetic		0.75	
Outgoing*		0.79	
Talkative*		0.75	
Proactive		0.66	
Calm*			0.73
Tense*			-0.66
Secure			0.61
Relaxed*			0.74
Anxious*			-0.68
Not irritated			0.57

communality values for all the attributes loaded under different factors were higher than 0.4. Hence, none of the attributes were removed based on the third criterion. Of over 34 attributes that are used in the subjective tests, 18 attributes for females and 24 attributes for males were retained after the factor analysis. The attributes retained in each of male and female voices are presented in tables 4.3, 4.4 respectively. A second-factor analysis was performed on the remaining attributes which explained the variance of 80% for female and 73% for male voices. In addition, we have also verified the goodness-of-fit ($p<.05$) using the chi-squared test to examine the three-factor model for the best representation of the subjective data. Further, The factors were defined/named based on the items/adjectives they represent. The derived factors were named warmth, competence (social speaker characteristics), and extraversion, a personality trait. In table, 4.3 we find the factors and the factor loadings of female speakers. The factor loading values range between -1 and 1. The value indicates the amount of influence the factor has on the variable/attribute. The negative sign indicates that the antonym of the adjective is the underlying speaker attribute. For example, the attribute, "distant" has a negative loading of 0.78. Therefore, the antonym of "distant" (friendly) is the underlying speaker attribute for the factor,

warmth. Corrspondingly, a similar analysis was performed on the male speakers and the results of the factor analysis are provided in table 4.4.

4.2.2.3 Discussion

- A reliability test was conducted to verify the internal consistency of the derived factors (warmth, competence, and extraversion). This was estimated using Cronbach's alphas [185]. Cronbach's alpha values range between 0 and 1. The higher the alpha value (> 0.5) the more reliable the factor analysis. The Cronbach's alphas calculated for each of the characteristics and the personality trait are provided in table 4.5 separately for male and female speakers. The internal consistency values of the subjective responses presented in the table are calculated irrespective of the gender of the participants.

Table 4.5: Cronbach's alphas

Factors	Male speaker	Female speaker
Warmth	(12 adjectives) 0.83	(11 adjectives) 0.76
Competence	(6 adjectives) 0.78	(4 adjectives) 0.87
Extraversion	(6 adjectives) 0.71	(3 adjectives) 0.74

- The effect of listeners' gender on the perception of various adjectives is interpreted using the 2-sample t-test carried out separately for female and male speakers (since I have employed gender-dependent analysis throughout the study). Statistically, significant differences are observed, when female listeners perceived the male speakers to be more relaxed with p=0.029, energetic with p = 0.01, and responsible with p = 0.012 ($p<0.05$). Correspondingly, the male listeners felt the female speakers are more agreeable with p =0.00 and reliable with p = 0.02.
- The current studies propose that the derived factors are orthogonal to each other (since we have finalized the varimax rotation for the factor analysis). This observation is in line with the theory proposed in [55] (social and intellectual dimensions are orthogonal to each other). Also, the order of the derived factors is different for male and female synthetic voices. Table 4.3 displays the factors derived for female voices to be, warmth, competence, and extraversion. While in the male voices, as shown in table 4.4 the order of the factors extraversion and competence are reversed (warmth, extraversion, and competence).
- ***Warmth***: The first identified factor or the first impressions of the participants as observed from the factor analysis were related to the social dimension, warmth [15, 186]. Therefore, the first factor was termed warmth. From the results of the factor analysis, we can observe the number of adjectives contributing to warmth in female and male synthetic voices to be slightly different. There are eleven and twelve attributes accounting for the characteristic warmth in female and male speech, respectively. Among them, seven attributes are commonly loaded in both genders (hearty, pleasant, trusting, agreeable, emotional, unlikable, sympathetic).

The attribute "energetic" was found under the trait warmth in female speakers. While, it was found responsible for extraversion in male speech. The studies on interpersonal relationships among humans [15, 186] displayed that the attributes, friendliness, and trustworthiness are related to the characteristic, warmth. Similar behavior is also observed in the current studies (friendly and trusting contribute to male warmth; trusting related to female warmth).

- ***Competence***: The second factor was termed competence because the speakers who are calm and relaxed under pressure can come across as powerful, confident, and competent individuals [187]. The targeted domains were health care and customer service. Being calm and composed is therefore essential for the conversational agents that are utilized in these application domains. There are four and six attributes representing competence in female and male speech, respectively. Among them, all four attributes identified in female speakers' speech are found in male speech too (calm, tensed, relaxed, anxious). Additionally, the male speakers had the attributes "secure", and "not-irritated" under competence.
- ***Extraversion***: The third factor derived from the subjective responses was extraversion. The goal of this study was to identify the first impressions of listeners from the given synthetic voices. However, among the adjectives provided to the participants (collected from behavioral and personality studies), there were also the ones that would describe the personality of a person (among the 66 adjectives). The participants indicated that it is also possible to perceive personality-related attributes from synthetic speech (can be observed from the initial subjective test presented in the questionnaire preparation (section 4.2.1.2) described in the experimental setup (section 4.2.1)). From the current factor analysis, we can therefore interpret that apart from the social and the intellectual dimensions, the listeners could also perceive the personality trait, extraversion from the synthetic speech (from both the genders). This supports the theory that extraversion can be easily predicted from speech [169] and further validates its relevance in synthetic speech. There are three and six attributes representing extraversion in female and male speech, respectively. Among them, all three attributes identified in female speech are also observed in male speech (introvert, outgoing, talkative). Further, there were three more attributes underlying extraversion in male speech, "energetic", "active", and "proactive".
- Other interesting observations were that the attributes "hearty", "sympathetic", and "non-likable" are commonly contributing to warmth in both natural [76] and synthetic voices. The attribute "pleasant" was related to the physical factor, "attractiveness" in natural speech [76, 75]. However, similar to research shown in [188], attractiveness is associated with "warmth" in this work. Subsequently, the attribute "secure" is observed to be related to the characteristic, competence in synthetic speech, but it was found to contribute to the perception of "confidence" in natural speech [76]. Further, we can see that the adjectives, "outgoing" and "talkative" were commonly found (in both genders) responsible for the perception of the personality trait, extraversion in natural as well as synthetic speech [168]. As outlined in the behavioral studies [15], the adjectives, "active" and "energetic" were also contributors to extraversion in male synthetic voices.

4.2.3 Perceptual analysis of TTS voices

This section presents the dissection of the experimental data in order to comprehend the performance of TTS voices (speech quality, naturalness, perception of warmth, competence, etc.,). In order to achieve the same, I present each TTS voice and its corresponding analysis, instead of providing the system-level (TTS system) performance. The details of the male and female voices and the TTS system they are derived from, are provided in table 4.6. The speakers/voices that displayed an equal amount of ratings for warm-cold (almost similar ratings for the bipolar adjectives) and competent-incompetent judgments were removed. Finally, there were 18 male and 18 female speakers (36 voices).

Initially, subjective evaluations were carried out for the perception of speech quality and naturalness. The study was carried out with 20 participants with their age ranges between 22-35 (mean = 24.4, std = 3.2, male = 13, female = 8). The participants were provided with a 5-point Likert scale where 1 = not at all natural/poor quality, 5 = very natural/Excellent. They could listen to the samples any number of times during the test. The speech samples were randomized for all the participants during the study.

Table 4.6: List of the TTS voices and the corresponding TTS systems used in the study. USS = Unit Selection Synthesis, UPM = USS with prosody modification, HSMM = Hidden Semi Markov Models, DNN = Deep Neural Networks, exp = Expressive DNN, fest = Festival, taco = Tacotron. Bp0, Dp0, Dp2, Cp0, Cp2, Ep0, Ep2 = pitch variations

TTS system	Male voice	Female voice
Mary TTS	bdl_hsmm	slt_hsmm
Mary TTS	bdl_Upm	slt_Upm
Mary TTS	bdl_USS	slt_USS
IBM Watson	Michael_transformable	Allison_dnn
IBM Watson	Michael_dnn	Allison_exp
IBM Watson	-	Lisa_transformable
IBM Watson	-	Lisa_dnn
Mary TTS	rms_hsmm	-
Mary TTS	rms_Upm	-
Google TTS	A_wavenet	C_wavenet
Google TTS	B_wavenet	E_wavenet
Google TTS	D_wavenet	F_wavenet
Festival	bdl_fest	clb_fest
Festival	jmk_fest	slt_fest
Festival	rms_fest	-
Tacotron	bdl_taco	clb_taco
Tacotron	jmk_taco	slt_taco
Google TTS	Bp0_standard	Cp0_standard
Google TTS	Dp0_standard	Ep0_standard
Google TTS	Dp2_standard	Cp2_standard
Google TTS	-	Ep2_standard

Figure 4.2 displays the MOS scores calculated for the speech quality and the naturalness of the male TTS voices used in the study. We can observe that the speech quality and naturalness scores for the parametric approaches (DNNs, Wavenet, Google standard voices) were higher compared to other systems. Even though the Tacotron model employs neural speech synthesis, the performance of the model (in terms of speech quality and naturalness) was the least compared to other NN-based TTS voices. This could be because of a) less training data (CMU arctic database), and b) speech signal reconstruction by griffin-lim reconstruction. I have observed that Google voices (wavenet and standard) had the highest MOS ratings for speech quality (4.3 for wavenet voice D, 4.1 for standard voice D) and naturalness (4.5 for wavenet voice D, 3.5 for standard D) among others. The voice generated from the Festival TTS system trained on CMU arctic database (jmk) had the lowest MOS ratings for speech quality (2.1) and naturalness (1.3). A similar observation was made from the subjective responses of female TTS voices. The performance of female TTS voices is presented in the figure, 4.3. The speech quality and naturalness ratings are the highest for Google's wavenet voice, F (4.3 and 4.5 respectively), and the standard voice, E (4.1 and 3.5 respectively). Correspondingly, the voice generated from the Mary TTS system trained on CMU arctic database (slt_USS) had the lowest MOS ratings for speech quality (1.9) and naturalness (1.3).

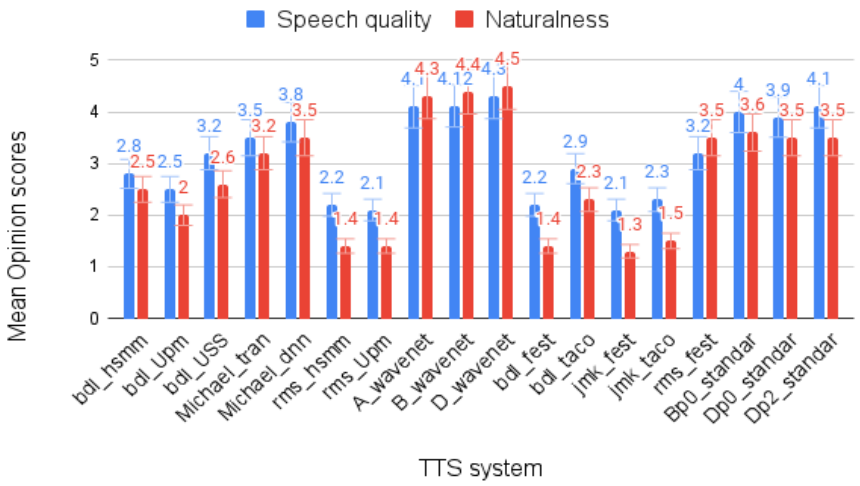

Fig. 4.2: The Mean Opinion Scores calculated for the male TTS voices for the scales, speech quality and naturalness
.

Further, the perception of warmth, competence, and extraversion for each of the male and female TTS voices is studied. In order to derive the perception of social speaker characteristic, warmth, the subjective ratings of all the adjectives (commonly

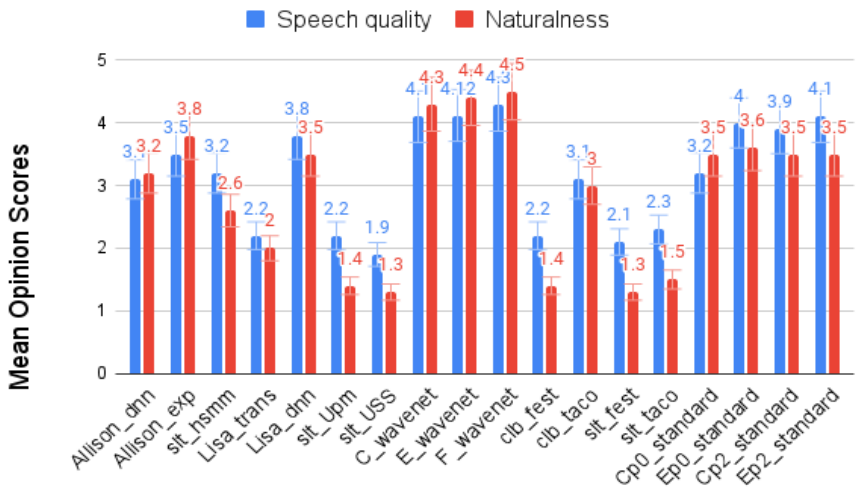

Fig. 4.3: The Mean Opinion Scores calculated for the female TTS voices for the scales, speech quality and naturalness

.

found in both male and female voices) loaded under the factor were averaged. Therefore, the warmth ratings for male and female TTS voices were derived by averaging the subjective ratings of the adjectives, hearty, pleasant, trusting, agreeable, unlikable, emotional, and sympathetic (for the negative adjectives, like unlikable, tensed their antonyms are considered). Similarly, the competence ratings are obtained from the subjective ratings of the adjectives, calm, tensed, relaxed, and anxious, the extraversion from the ratings of, introvert, outgoing, and talkative.

Figure 4.4 displays the perception of warmth from the male TTS voices along with the 95% confidence intervals. Correspondingly, the warmth ratings for female speech are presented in figure 4.5. The TTS voices which displayed a better speech quality and naturalness also received the highest ratings on warmth. We can observe that among the male voices, the Google Wavenet voice, D is perceived to be highly warm (with an averaged warmth rating, wavenet =77, Google standard = 76). The bdl voice generated from the Festival (clustergen) had the lowest perception of warmth (29.25) among the male voices. Among female TTS voices, there were two speakers with the highest perception of warmth. One was developed with IBM watson (Allison_expressive = 78) and the other with Google TTS (C_wavenet = 78). While clb voice generated from the Festival (clustergen) had the lowest perception of warmth (21) among others. We can also observe that warmth rating among highly warm voices (male voice D = 77, female voices, Allison, C = 78) are close to each other, while there is a significant difference in the perception of warmth from the voices, bdl (warmth rating = 29.25, male), and slt (warmth rating = 21, female) (voices with lowest ratings of warmth).

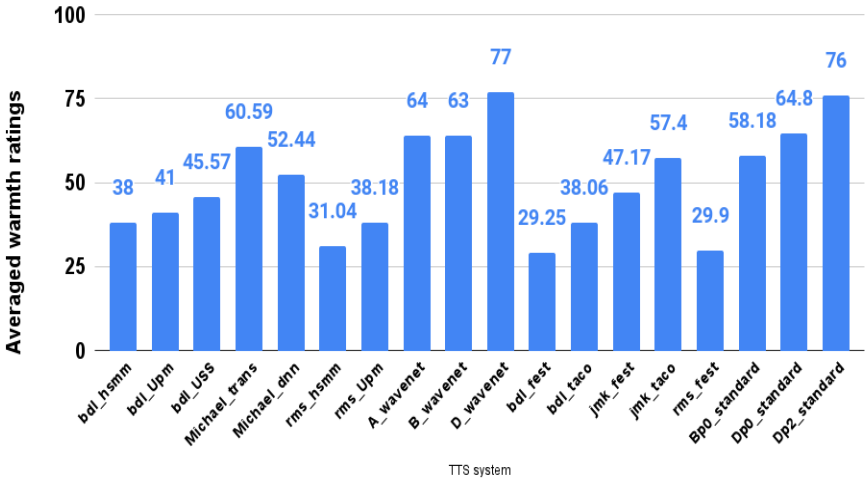

Fig. 4.4: Perception of warmth from male TTS voices. The subjective ratings of the speaker attributes loaded under the characteristic warmth were averaged for the analysis.
.

Figure 4.6 shows the perception of competence from the male TTS voices along with the 95% confidence intervals. The perception of competence has also been found to be affected by speech quality. Once again, Google Wavenet voice, D has the highest averaged competence ratings (Wavenet = 75, google standard = 74). While the bdl (Festival, clustergen) was again the voice with the lowest ratings for competence (34). Even among the female voices, we can find the similar pattern as in the case of warmth. Figure 4.7 depicts the competence ratings of female TTS voices. The Google Wavenet voice, C has the highest averaged competence ratings (81). The clb voice generated from the Festival (clustergen) had the lowest perception of competence (45) among the female voices. However, the perception of competence from clb is still higher than the bdl. We can observe that the warmth and the competence ratings aligned with each other (to some extent) in the current studies. Nevertheless, if these characteristics are independent of each other as proposed in [12] is not obvious from this study.

Figure 4.8 displays the perception of extraversion from the male TTS voices along with the 95% confidence intervals. Similar to the analysis seen in the perception of warmth and competence, Google voice D exhibited the highest ratings on extraversion. However, the voice with the highest rating was from Google standard model with a pitch variation (Google standard voice with the pitch set to 2 = 79). Unlike in the studies on warmth and competence, the voice with the lowest extraversion ratings was rms, generated from the Mary TTS with HMM-based synthesis. On the other hand, the female voice with the highest extraversion ratings was F, from Google's

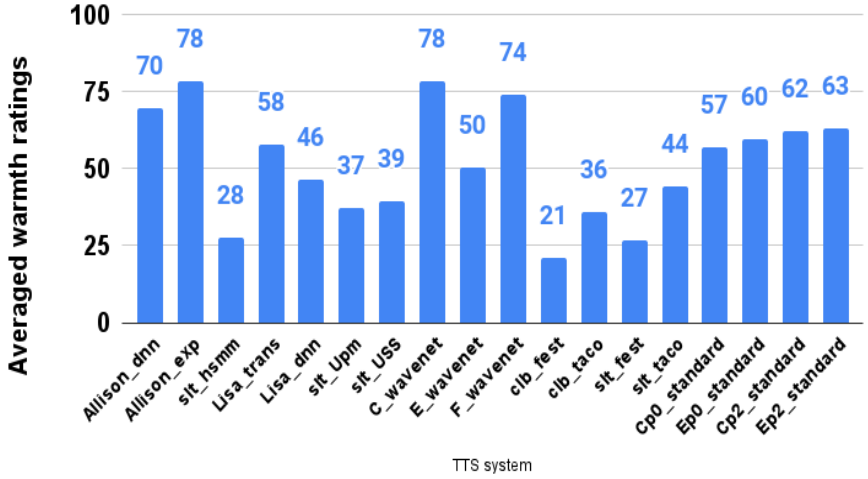

Fig. 4.5: Perception of warmth from female TTS voices. The subjective ratings of the speaker attributes loaded under the characteristic warmth were averaged for the analysis.
.

wavenet (74) which is different from the results seen in the studies on female voices for warmth and competence (voice, C had the highest warmth and competence ratings). The obvious difference between the voices, C and F is the pitch (F has a higher pitch when compared to C). The TTS voice with the lowest extraversion ratings as observed from the results was, clb generated using Festival. Figure 4.9 displays these perceptions of extraversion from the female TTS voices.

In contrast to the findings presented in [170], the perceptual analysis of the factors derived in the current study displays a similar response to that of the subjective data of speech quality and naturalness (speech quality and naturalness have an impact on the social perceptions). However, there were some primary differences in both the experimental setups a) the choice of the TTS voices used in the studies, b) the speech data (content and also the length of the utterances) used in the studies, and c) the type of evaluation employed in each of the studies. Apart from these observations, we have not performed any extensive analysis on the relationship between the perceptions of speech quality/naturalness and the SSC from the synthetic speech in our study.

4.3 Study B: Deriving the ground truth information

The observations made from study A are as follows: a) speech quality and naturalness of the TTS voices have a significant impact on the perception of the social speaker

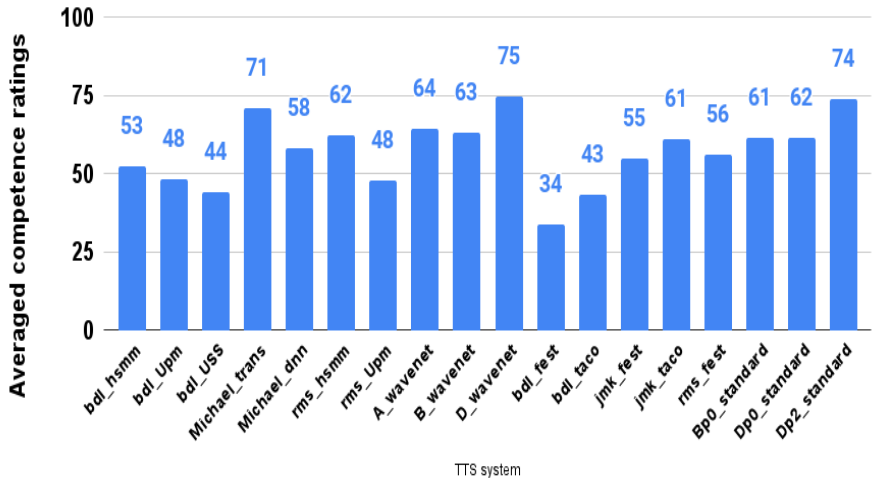

Fig. 4.6: Perception of competence from male TTS voices. The subjective ratings of the speaker attributes loaded under the characteristic competence were averaged for the analysis.
.

characteristics, warmth, and competence, b) the content has an impact on the social perceptions of the speakers (based on the feedback received from the participants obtained through the preliminary test conducted for questionnaire preparation (in section 4.2.1.2)), c) the subjective evaluation was carried out only with two sentences (the results might not be reliable or can be generalized), and d) too many adjectives used in the study (also contains adjectives with similar meaning). A subsequent study was designed considering these observations [167]. In order to handle the first task, the TTS systems that could deliver good-quality speech were researched further. Since, the commercial TTS system, Google has displayed the highest subjective response for both quality and naturalness, the idea was to examine publicly available commercial TTS systems which could produce good-quality speech. In order to achieve this, the TTS engine that supports neural voices (Amazon Polly) similar to the Google TTS engine (voice type = Wavenet) was chosen. Therefore, finally, two commercial TTS systems, Google Wavenet[4] and Amazon Polly[5] were used in the current study. Secondly, in order to avoid the effect of the content on speech perception, neutral speech was utilized in the study. The speech data used in this study was generated using the Harvard database [6]. Later on, a few modifications were made to the evaluation setup in order to a) include multiple sentences in the study,

[4] https://cloud.google.com/text-to-speech/

[5] https://aws.amazon.com/polly/

[6] https://www.cs.columbia.edu/ hgs/audio/harvard.html

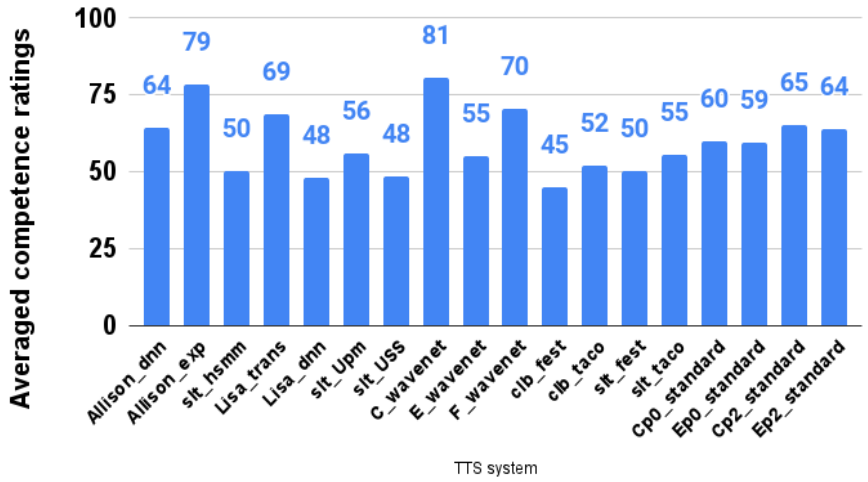

Fig. 4.7: Perception of competence from female TTS voices. The subjective ratings of the speaker attributes loaded under the characteristic competence were averaged for the analysis.
.

and b) recompiled the adjectives list. Firstly, multiple speech files were combined to form the speech segments of duration approx. 20sec (each speech segment had 7-8 sentences). Also, the sentences generated by each TTS voice were randomized (Those 8 sentences were in a different order for each voice). Secondly, the attributes that have a similar meanings have been removed. The finalized list consisted of 15 adjectives and the details are provided in table 4.7.

Another modification to the previous experimental setup is the platform chosen to conduct the subjective tests. The subjective tests were again carried out using TheFragebogen [154] but are handled using the Amazon Mechanical Turk (AMT) [189]. Also, the participants of the subjective tests were all native English speakers (US English). A detailed description of the study is presented in [167]. Table 4.8 and 4.9 display the derived factors and the speaker attributes loaded under each of the factors for female and male synthetic voices respectively. Similar to the study on the wide range of TTS voices, the current study has also provided three factors namely, warmth, competence, and extraversion.

4.3.1 Comparison of the studies

This section provides the similarities and differences found in both the studies, a) wide-range TTS systems, and b) two commercial TTS systems.

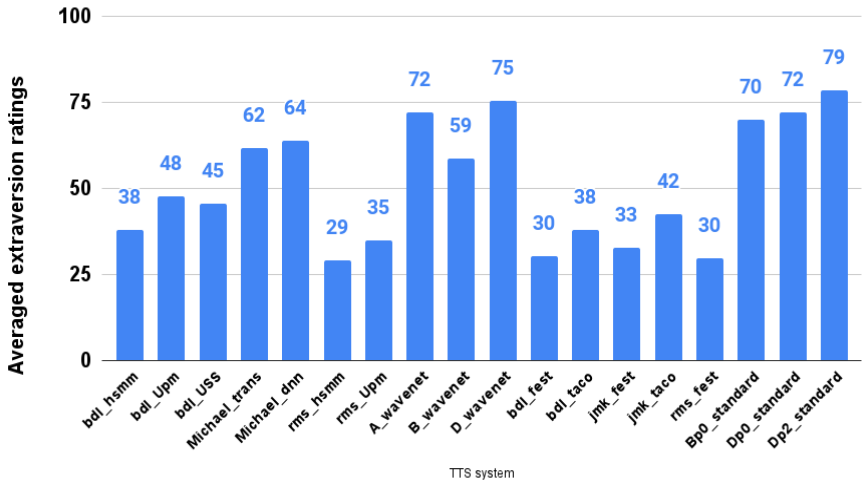

Fig. 4.8: Perception of extraversion from male TTS voices. The subjective ratings of the speaker attributes loaded under the factor (personality trait) extraversion were averaged for the analysis.
.

- The number of factors and the type of factors derived from both studies are the same. The derived factors were warmth, competence (social speaker characteristics), and extraversion (personality trait).

Table 4.7: Adjectives used in the study with Google Wavenet and Amazon Polly voices

Adjectives	
relaxed	not relaxed
confident	not confident
enthusiastic	unenthusiastic
energetic	not energetic
friendly	unfriendly
arrogant	not arrogant
pleasant	unpleasant
likable	unlikable
responsible	irresponsible
reliable	unreliable
accessible	inaccessible
sympathetic	not sympathetic
skilful	not skilful
kind	unkind
extrovert	introvert

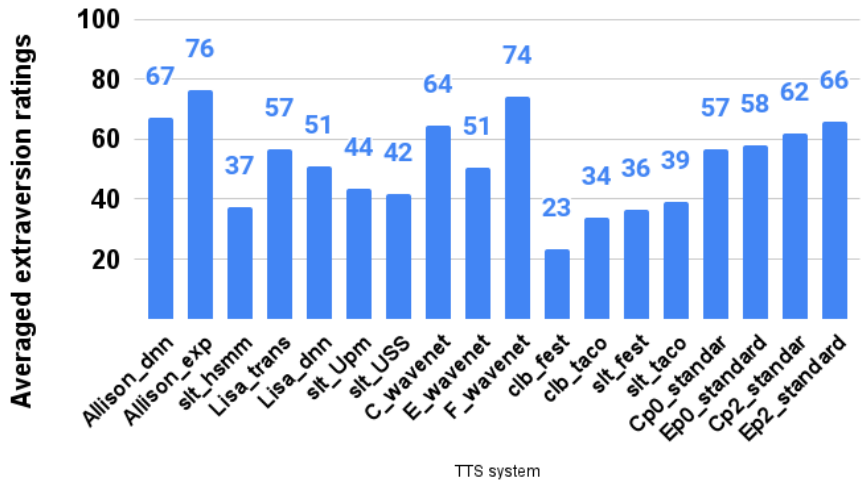

Fig. 4.9: Perception of extraversion from female TTS voices. The subjective ratings of the speaker attributes loaded under the factor/personality trait extraversion were averaged for the analysis.
.

Table 4.8: Factor loading for female speakers

Attributes	Warmth	Competence	Extraversion
friendly	0.72		
kind	0.82		
likeable	0.79		
pleasant	0.77		
sympathetic	0.78		
confident		0.79	
reliable		0.87	
responsible		0.89	
skillful		0.88	
energetic			0.89
enthusiastic			0.83
extrovert			0.71

- ***Warmth:*** The adjectives, friendly, kind, likable, pleasant, and sympathetic were observed to be contributing to warmth in both male and female synthetic voices (from the study on 2 commercial TTS systems). Among these, the adjectives, pleasant, likable, and sympathetic are commonly found to be contributing to warmth in synthetic speech (found in both studies).
- ***Competence:*** The factor loadings responsible for competence in synthetic voices are confident, reliable, responsible, and skillful. These results are in line with the dimensions proposed for the multi-facet study presented in [173]. Additionally, we

Table 4.9: Factor loading for Male speakers

Attributes	Warmth	Competence	Extraversion
accessible	0.63		
friendly	0.72		
kind	0.75		
likeable	0.71		
pleasant	0.67		
sympathetic	0.77		
confident		0.65	
reliable		0.81	
responsible		0.93	
skillful		0.84	
energetic			0.81
enthusiastic			0.81
extrovert			0.79

find that the adjectives that would best represent competence in the case of a) wide range TTS systems and b) 2 commercial TTS systems are completely different. The attributes derived from a wide range of TTS systems under competence were calm, anxious, tense, and relaxed. This phenomenon could be because of the datasets used in both the studies (wide-range TTS = WC dataset, 2 commercial TTS = neutral speech). Another reason could be that, with the inclusion of a wide variety of voices, there were certain vocal cues that contributed to varied perceptions of different speaker attributes in the first study (wide-range TTS). We further provide the acoustic analysis of a wide range of TTS voices (in chapter 5) and the voices from 2 commercial TTS systems (in chapter 7) to understand the acoustic correlates of SSC in each of these studies.

- ***Extraversion:*** The personality trait extraversion was found to be a combination of the adjectives, energetic, enthusiastic, and extrovert in the study with the commercial TTS systems. On the other hand, the adjectives contributing to the perception of extraversion as observed from both studies were, energetic and introvert (negative adjective).

4.3.2 Defining the ground truth voices

There has not been any previous work on synthetic speech's perception of warmth or competence (in speech-alone scenarios). Correspondingly, there are no standard databases for warm/cold or competent/incompetent synthetic voices. In our work, we are interested in studying these characteristics from synthetic speech and further analyzing the speech features, and later on, investigating the modifications of the synthetic speech for the positive perceptions of the voices. In this regard, we intend to define the standard voices for our studies in this thesis. The current study was designed while considering the shortcomings of the previous study on a wide variety

of TTS voices. Considering the above-mentioned line of work, in this thesis, I define the voices derived from these two commercial TTS systems as the reference voices or the "ground truth voices". Accordingly, throughout the thesis, these voices are hereafter referred to as ground truth TTS voices, and any information derived from these voices is called ground truth information.

4.4 Limitations

This section details the limitations of the work based on the conclusions drawn from the analysis.

- **Adjectives list:** The aim of this thesis is to understand the perceptions of warmth and competence and later on derive the acoustic correlates and further model the speech generation mechanism for positive perceptions of synthetic speech. However, through the preliminary studies, we have observed that perception of the desired speaker characteristics required the inclusion of multiple adjectives in the subjective evaluation. In other words, forming the first impressions of synthetic speech (in speech-alone scenarios) has thus required an analysis of the synthetic speech in various perceptual dimensions. In this regard, an analysis of a different set of speaker characteristics might require a similar analysis (collection of adjectives, subjective tests, factor analysis). This means that the results of the current study cannot be directly used for multiple application domains or SSC. Since the experimental results cannot be generalized for multiple application domains, we consider this to be a limitation of such an analysis of speaker characteristics from synthetic speech.

4.5 Summary

This chapter outlines the analysis of the subjective tests carried out to interpret social perceptions of synthetic voices. In order to address the research question posed at the beginning of this chapter, two studies were provided, a) perception of SSC from a wide variety of TTS voices, and b) two commercial TTS voices. From the behavioral studies [15, 54, 56, 60, 173], it is evident that the characteristics, warmth and competence are considered the ***"universal dimensions of social perception"***. These studies were performed on human behavior (known humans). In the current work, similar studies were performed in a speech-alone (unknown voices) scenario in the case of synthetic voices. Through these studies, it is apparent that the characteristics, warmth, and competence can also be perceived from synthetic voices (US English). In this work, we also define the ground truth TTS voices that are further used throughout this thesis for various studies. Apart from the social dimensions, the current study has also provided us with the personality trait, extraversion. Moreover, we have observed that females with high-pitched voices are interpreted as extroverts.

Chapter 5
Acoustic correlates

This chapter details the studies carried out for the prediction of vocal cues contributing to warmth and competence in synthetic voices. Feature extraction employed in the experiments was using the publicly available OpenSMILE toolkit [190]. The suggestions and discussion on the prediction of acoustic correlates were done between me, Benjamin Weiss, and Sebastian Möller. Further, automatic prediction of the social speaker characteristics is presented using the derived vocal cues of warmth and competence in male and female synthetic voices.

5.1 Related work

A considerable amount of research has been previously performed on the analysis of acoustic correlates of various behaviors, personalities, and characteristics of speech (natural as well as synthetic speech). These works can be broadly classified into two types, a) works that directly investigate the impact of F0 (and its dynamics), loudness, and speaking rate of various speech perceptions [78, 82], and b) works that derive the acoustic correlates of various characteristics or personalities. [78, 82] presents that the speaking rate, durations (longer speech, fewer pauses, shorter-phrase durations), and intonation are highly correlated with the charismatic speech of celebrities. High pitch (high standard deviation of pitch) not only contributes to charismatic speech but also contributes to being a "good speaker" (increases the perception of the attributes, expressive, powerful, involved, and trustworthy) [80]. In [191] authors present the acoustic feature prediction and feature importance in emotion estimation from speech. They derive 46 acoustic features using pitch and energy contours. Along with the pitch, speaking rate, and intensity as seen previously, this study provides the relevance of spectral features (mfccs) in understanding emotions from speech. The order of importance for these features in deriving various emotions according to their work is as follows, 1) mfcc, 2) energy, 3) pitch, and 4) duration. Correspondingly, the authors in [192] also perform the classification of emotions using acoustic features, pitch, spectral features, and energy. [193] employ the OpenSMILE features for

studying emotion recognition in the case of virtual agents [190]. OpenSMILE features can capture various speaker characteristics present in the speech signal. Hence, these are widely used in paralinguistic studies on speech signals [194]. Accordingly, [75] utilize the openSMILE features for understanding various social and physical factors in the speech in case of zero acquaintance scenarios. Their work proposes the dependence of the perceptual dimensions, confidence, apathy, and serenity on the fundamental frequency. Previously, voice likability and attractiveness have also been found to be correlated with F0 [195]. The study shows that lower F0 values result in more pleasant and attractive voices. Additionally, [196] investigates the effect of the feature, Harmonics-to-Noise ratio (HNR) as a reliable indicator for attractiveness in the perceived speech. HNR is generally associated with the voice quality in the speech signal (nature of voice, hoarseness). Through their work, they state that the regular vocal fold vibrations contributed to the attractiveness of the voice irrespective of gender. Also, the speaking rate was proposed to be a reliable indicator of the characteristic, competence [197]. While the speakers with lower speaking rates were considered less truthful, and less empathetic. Consequently, the male voices with high pitch were found to be less believable and less truthful [198].

5.2 Overview

This chapter presents the acoustic analysis carried out on the wide range of TTS systems employed previously in chapter 4. The goal is to identify the acoustic correlates of warmth and competence from various voices that involved different synthesis procedures. In this work, we employ the second approach discussed in the previous section to interpret the vocal cues of SSC. Figure 5.1 displays the overview of the work carried out in this chapter. We extract the OpenSMILE features for each of the speech samples collected over the wide range of TTS systems and further compute the feature relevance using various dimensionality reduction techniques. Finally, we also present the automatic prediction of SSC from the derived acoustic correlates. This chapter addresses the following research question.

? Research question: ***"What are the acoustic features that contribute to the social speaker characteristics?"***

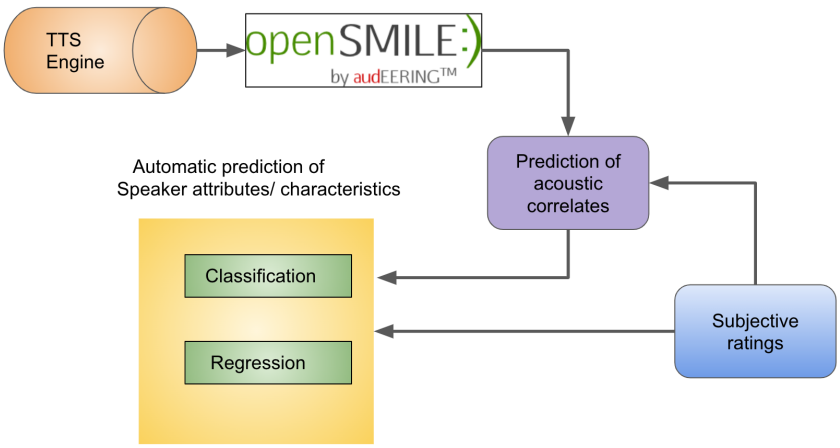

Fig. 5.1: Flowchart of the current workflow.

.

5.3 Preparation of the experimental setup

So far, we have examined a) various speaker attributes that can be perceived from synthetic speech b) the attributes that can contribute to each of warmth and competence in synthetic speech, and c) the synthetic voices with the highest and lowest subjective ratings of warmth and competence (from chapter 4). The follow-up work is towards understanding the acoustic features contributing to different speaker attributes and characteristics.

5.3.1 Input data: OpenSMILE features

[199] propose the parameter set that could capture the paralinguistic information (extra information other than the content, for example, age, gender, mood, mental states, etc.,) present in the speech. In the current study, we are interested in interpreting the acoustic features contributing to various speaker characteristics and social perceptions of synthetic speech. Therefore, inspired by the functionality of the OpenSMILE features, in the current work, the Geneva Minimalistic Acoustic Parameters Set (eGeMAPS) configuration has been explored [199]. The feature extraction using eGeMAPS provided an 88-dimensional feature vector for each speech file. These acoustic features can be categorized as Low-Level Descriptors (LLDs) and functionals: loudness, 4 mfcc, alpha ratio, slope (0–500 Hz, 0.5–1.5 k Hz), Hammerberg Index, Spectral Flux, F0 (semitones), formants F1, F2, F3 (frequency,

mean, amplitude), log. HNR, Jitter, shimmer, Harmonic difference H1–H2, H1–A3. A list of these acoustic features is further provided in the Appendix.

5.3.2 Output data: Subjective data for warmth and competence

Through the studies in chapter 4, the factors (warmth/competence) and the factor loading (adjectives) information is obtained. A quick look into this information (as it is relevant to the current experiments) is presented in this section. Table 5.1 presents the speaker attributes corresponding to warmth and competence in female and male TTS voices. In order to derive the output features for the acoustic feature prediction, the subjective ratings of the adjectives/factor loadings that are commonly contributing to each of warmth and competence in both genders are averaged. Thus, the output data for warmth is obtained from the average of 7 adjectives (Hearty, Pleasant, Trusting, Agreeable, Emotional, Unlikable, Sympathetic) and competence from 4 adjectives (calm, tensed, relaxed, anxious).

Table 5.1: The list of speaker attributes loaded under each of SSC for both genders. (*) indicates the attributes are commonly found in both male and female voices.

Male		Female	
Warmth	Competence	Warmth	Competence
Kind	Calm*	Hearty*	Calm*
Hearty*	Secure	Distant	Tensed*
Arrogant	Tensed*	Pleasant*	Relaxed*
Trusting*	Anxious*	Reliable	Anxious*
Friendly	Relaxed*	Trusting*	
Pleasant*	Not-irritated	Agreeable*	
Unlikable*		Emotional*	
Agreeable*		Unlikable*	
Empathetic		Sympathetic*	
Responsible		Enthusiastic	
Emotional*		Energetic	
Sympathetic*			

5.3.3 Prediction of the vocal cues

In this section, we can find various experiments carried out for the a) prediction of acoustic correlates, and b) automatic prediction of social speaker characteristics.

5.3.3.1 Data

The number of training examples available for each for male (18 voices* 2 sentences* 2 characteristics = 72) and female (18 voices* 2 sentences* 2 characteristics = 72) was less than the input dimensions (88 dimensional acoustic features). As a result, feature processing and modeling the data were challenging.

5.3.3.2 Feature processing: Removal of redundant features

The first step in the data processing was to remove the redundant features. The redundant features are those that either a) do not contribute to the perception of the desired attributes or b) the ones that are not any different from the other significant features. The first task was handled using Pearson's Correlation Coefficient (r). The correlation values were calculated between the normalised acoustic features and the averaged, transformed subjective ratings of warmth and competence. The acoustic features that exhibited a positive or negative correlation ($|r| > .8$) with the ratings were retained.

The second task in the feature processing was to remove the multi-collinearity. Principal Component Analysis (PCA) was employed to disentangle and remove the collinearity between the features. PCA linearly reduces the high dimensional data into principal components that are less correlated with each other. Dimensionality reduction is widely used in both statistics and machine learning research [200]. It refers to the projection of the high dimensional data onto a low dimensional space. This is done while keeping the necessary information intact. Apart from EFA (as seen in chapter 4), there are other techniques for dimensionality reduction such as: Principal component Analysis (PCA), Linear Discriminant Analysis (LDA), Backward Feature Elimination (linear regression technique), and many more. Besides, i) eliminating the collinear features, and ii) dimensionality reduction, PCA can also deal with the over-fitting caused by too many variables in the dataset. One can leverage all the advantages of PCA by carefully choosing the number of principal components. The first principal component holds the maximum amount of variability in the data, the second component holds the second highest, and so on. The details of these feature processing steps (a) removal of redundant features, b) removal of multi-collinearity) and the prediction of vocal cues of SSC from the remaining features are presented in the sections below.

5.3.3.3 Feature modeling

After the initial pre-processing of the input acoustic feature vector, the data distribution was examined to verify the models that would best fit the feature-label pairs. Since the dataset is limited, the models explored in the current studies include Decision trees, Support Vector Regressor, and linear regressor.

- ***Decision Trees*:** One of the approaches examined for the prediction of relevant acoustic features was decision-tree-based regression. The modeling of the data was handled using the decision tree regressor package available in the sklearn library. Due to the limited amount of training examples, a leave-one(speaker)-out-of-cross validation along with the mean squared error is employed in the current experiments. There were 18 male and 18 female speakers and the experiments were carried out separately for both genders. Therefore, for the cross-validation, the data was divided into 18 equal parts (male and female separately). Among these sets, 1 speaker set was held out at a time and the remaining 17 speakers were used in the training. This step is repeated for all the speakers. The details of the experiments (inputs fed to the model and the predictions) and the analysis of results are provided in sections 5.3.3.4 and 5.3.3.5 respectively.
- ***Support Vector Regressor*:** Secondly, SVR was investigated for the relevant feature prediction from the input feature vector. Due to the availability of limited data (88-dimensional input vector; fewer training examples per gender (female voices = 18, male voices = 18, number of sentences =2)), in the current studies, we choose Leave-One-Speaker-Out Cross Validation (LOSO-CV) with a linear kernel. The values of C (regularisation parameter) and epsilon (no penalty for the prediction loss in this limit) were set to 1 and 0.2 respectively. The input and output data were normalized using the standard scaler available in the scikit-learn. The regression was also carried out using the SVR package available in the sklearn library.
- ***Linear Regressor*:** The model linearly transforms the independent variables (acoustic features) into continuous dependent values (social speaker characteristics). Additionally, there are multiple regression techniques that can be employed for feature selection using linear regression in the case of multi-variate inputs. For instance, a) forward step-wise regression or forward selection, and c) backward step-wise regression or backward elimination. A step-wise regression iteratively selects the features contributing to the desired task. In a forward selection method, the predictions start with the model with no variables at the beginning and the addition of the most reliable features at each step. While in a backward elimination, the model predictions at first are made with all the available input features. Further, with each iteration, the least contributing variable is removed. This continues until the highly relevant features are retained. In this chapter, I have utilized both a) a traditional linear regressor without any step-wise regression and b) a regressor with a backward elimination technique. The implementation of the model was enabled through the use of the *LinearRegressor* package from sklearn. The model predictions were carried out using leave-one-speaker-out cross-validation.

5.3.3.4 Experimental setup

This section details various dimensionality reduction techniques employed in order to achieve feature importance.

- ***Experiment 1*:** In the first experiment, the acoustic feature processing involved dimensionality reduction using PCA. The feature normalization was carried out using the standard scaler available in the sklearn library. The principal components were calculated on the normalized features. The number of principal components employed in the experiments was 5 with an explained variance of 78%. These principal components were then fed (input dim = 5) to three models namely, Linear Regressor, Support Vector Regressor, and Decision Trees for the prediction of warmth and competence.
- ***Experiment 2*:** This experiment consists of two steps. Firstly, a Pearson correlation coefficient is computed between the acoustic features and the speaker characteristics. Secondly, the acoustic features that had the highest correlation (positive/negative) with the characteristics were normalized and principal components were derived. The derived principal components were later trained with the 3 models.
- ***Experiment 3*:** PCA not only does the dimensionality reduction but also finds a combination between the derived components. Hence, in this experiment, I calculate the Pearson correlation between the Principal components derived from experiment 1 and the input acoustic features. The acoustic features that displayed the highest correlations positive/negative with the principal components were retained. These acoustic features were further passed through a PCA and the derived principal components were fed to the three models. The diagrammatic representation of experiments 1, 2, and 3 is presented in 5.2.

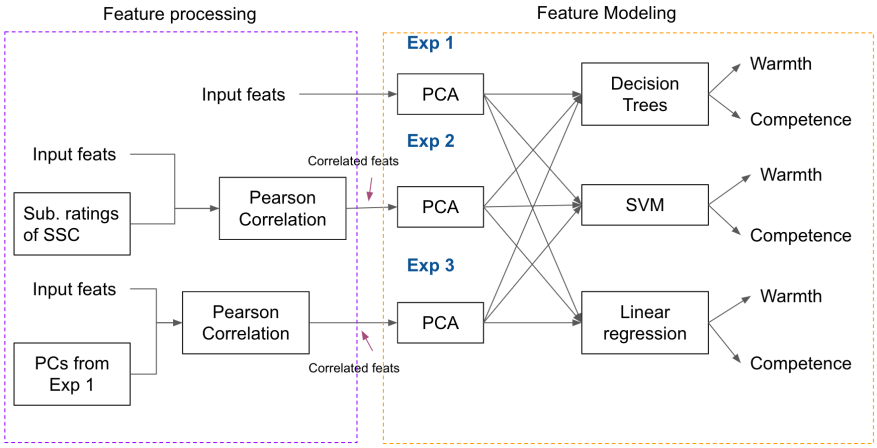

Fig. 5.2: The diagrammatic representation of experiments, 1, 2, and 3. In the figure, Exp = experiment, PCA = Principal Component Analysis, SVM = Support Vector Machines, Sub. Ratings = Subjective ratings, PCs = Principal Components

- ***Experiment 4*: Ablation study** This study examines the impact of the principal components on the performance of the models. Instead of feeding the principal components to the three models, the acoustic features derived in experiments 1,2,3 are directly fed to the three models. This study was conducted to verify the effect of the feature information (combination of different acoustic features) present in the principal components to determine different speaker characteristics (In other words, we examine the role of the PCA layer in the above-mentioned experiments). The schematic of the studies carried out under this experiment is further displayed in 5.3.

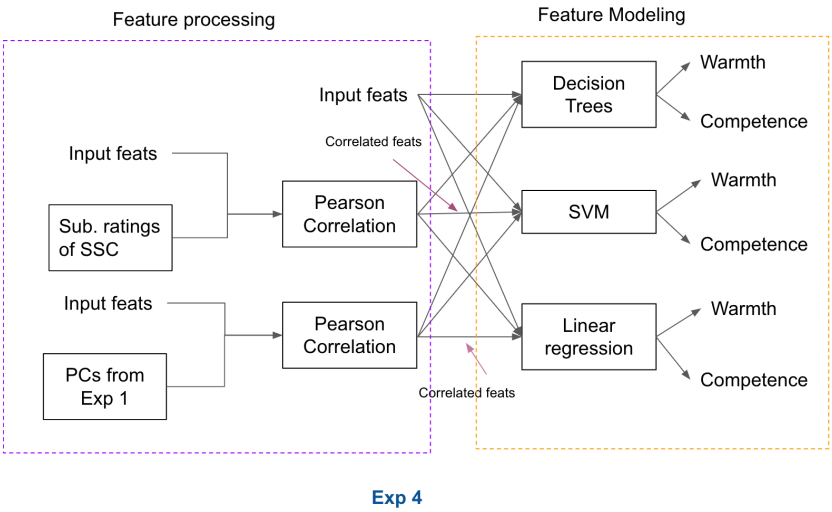

Fig. 5.3: The diagrammatic representation of experiment 4. In the figure, Exp = experiment, SVM = Support Vector Machines, Sub. Ratings = Subjective ratings, PCs = Principal Components

- ***Experiment 5*:** This experiment consists of the multi-variate linear regression applied to the input features using a backward elimination algorithm. The study was performed using the "statmodelsregression" available in the "linear model" package of the sklearn library. The significance level was set to 0.05 (SL=0.05). The p-values higher than this significance level were eliminated in each step. The derived acoustic features and the explained variance for each warmth and competence in the case of male and female speakers are provided in the section 5.3.3.5.

5.3.3.5 Results and Observations

For experiments 1-4, (except for the backward elimination technique), the Mean Squared Error (MSE) scores of the models were determined to be deciding factor for the feature selection. The lower the MSE, the higher the reliability of the model and the relevance of the corresponding acoustic features. Table 5.2 displays the experimental results of experiments 1,2,3 for warmth in synthetic speech.

Table 5.2: Results of regression techniques implemented for the perception of warmth in synthetic speech. IFs= number of input features fed to the models, PCs = Principal Components, CFs = acoustic features correlated with the speaker characteristics, CPs = acoustic features correlated with the principal components, DTree = Decision Tree, LR = Linear Regression, SVR = Support Vector Regressor, MSE = mean squared error

Model	Male		Female	
	IFs	MSE	IFs	MSE
DTree	PCs (5)	0.69	PCs (5)	0.63
SVR	PCs (5)	0.34	PCs (5)	0.59
LR	PCs (5)	0.26	PCs (5)	0.63
DTree	CFs+PCs (5)	0.63	**CFs+PCs (5)**	**0.52**
SVR	CFs+PCs (5)	0.32	**CFs+PCs (5)**	**0.37**
LR	CFs+PCs (5)	0.36	**CFs+PCs (5)**	**0.41**
DTree	**CPs+PCs (5)**	**0.55**	CPs+PCs (5)	0.63
SVR	**CPs+PCs (5)**	**0.32**	CPs+PCs (5)	0.62
LR	**CPs+PCs (5)**	**0.24**	CPs+PCs (5)	0.64

The first block in the table 5.2 consists of the details of the first experiment. The principal components obtained from experiment 1, were the input features provided to each of the three models. In this experiment, the 88-dimensional input vector is reduced to 5 principal components. Further, the prediction of the SSC was carried out with the derived principal components. The performance of the systems for each gender is provided separately in terms of MSE scores. The second block in the table represents the details of experiment 2. The correlations between the 88-dimensional acoustic features and the speaker characteristics were calculated. This has resulted in a reduction of dimensions from 88 to 51 features in female and 54 in male speakers. These dimensions were further reduced to 5 using PCA. Further, the derived principal components were modeled by Decision trees, SVR, and linear regression. The third block in the table displays the details of experiment 3. The correlation coefficient calculated between acoustic features and the principal components (derived in experiment 1) provided 79 and 76 acoustic features for female and male speakers respectively. A second dimensionality reduction was performed using PCA on the derived features. Later on, the principal components were modeled

for the prediction of SSC. From the results, we can observe that (presented in bold in table 5.2) the acoustic features that correlated with the speaker characteristics (experiment 2) contribute to variations in the SSC in the case of female speakers (by considering MSE as a deciding factor). Similarly, the acoustic features that correlated with the principal components were found to affect the speaker characteristics in male synthetic voices (experiment 3).

Table 5.3 represents the results of the ablation studies. Here instead of using the principal components, the acoustic features either directly or after calculating the correlation with the speaker characteristics and principal components (from experiment 1) are alone examined. The first block shows the performance of the models when they are directly provided with the input feature set (88-dimensional vector). The second block in the table represents the details of experiment 2. The input vector consists of the acoustic features that correlated with the speaker characteristics (female = 51 features, male = 54 features). These were directly passed through the three models by eliminating the PCA layer seen in the previous experiments. The third block in the table displays the details of experiment 3. The acoustic features correlated with the principal components were fed to the models by eliminating the second stage of dimensionality reduction. From the table, we observe that the acoustic features correlated with the speaker characteristics are acoustic correlates (experiment 2 after eliminating the PCA layer) in the case of both genders (shown in bold in table 5.3). Further, we can also observe that removing the PCA layer before training these models has contributed to the model performance significantly. The list of retained acoustic features (51 for female speakers, 54 for male speakers) are provided in table 5.4 for female speakers (warmth) and table 5.5 for male speakers (warmth).

Table 5.3: Results of ablation studies performed for the perception of warmth in synthetic speech. AFs= the acoustic features fed to the models, CFs = acoustic features correlated with the speaker characteristics, CPs = acoustic features correlated with the principal components, DTree = Decision Tree, LR = Linear Regression, SVR = Support Vector Regressor, MSE = mean squared error

Model	Male		Female	
	AFs	MSE	AFs	MSE
DTree	88	0.45	CPs (79)	0.49
SVR	88	0.33	CPs (79)	0.36
LR	88	0.59	CPs (79)	0.69
DTree	**CFs(54)**	**0.29**	**CFs (51)**	**0.34**
SVR	**CFs(54)**	**0.25**	**CFs (51)**	**0.31**
LR	**CFs(54)**	**0.47**	**CFs (51)**	**0.59**
DTree	CPs (76)	0.34	CPs (79)	0.41
SVR	CPs (76)	0.29	CPs (79)	0.35
LR	CPs (76)	0.56	CPs (79)	0.71

Therefore, from the above experiments, we can assume that the acoustic features retained through experiment 4 (features correlated with subjective responses of warmth) are the vocal cues of warmth in synthetic voices. However, tables 5.4, 5.5 provide a very long list of acoustic features, and the dependence of each acoustic feature on individual speaker characteristics is not obvious. Correspondingly, a subsequent study was carried out for the relevant acoustic feature prediction using the backward elimination-based linear regression (experiment 5).

The results of the linear regression are provided in the tables 5.6, 5.7, 5.8, 5.9. Table 5.6 displays the list of acoustic features retained after step-wise regression for the female warmth. The coefficient values represent the amount of change in the perception of a particular speaker characteristic for a unit change in the corresponding acoustic feature. For instance, one unit change in the F0 mean falling slope of female speakers would increase the perception of warmth by the corresponding coefficient value (0.4718). Correspondingly, the negative sign in the coefficient value suggests that the acoustic feature would negatively affect the perception of the specific characteristics in the generated speech. Accordingly, the acoustic features contributing to female competence are provided in table 5.7. The acoustic features accountable for male warmth and competence are presented in tables 5.8, and 5.9 respectively. The explained variance (R squared) values for each of those experiments are also provided in the respective tables (table caption).

Table 5.4: Derived acoustic correlates of warmth in the female speech as obtained from ablation studies.

Voicing specific LLDs, functionals	**Number of features**
F0	7 (pitch, intonation contour,(mean falling slope), dynamics, percentiles)
Jitter	1
Shimmer	2
Harmonic difference	3
Harmonics (F1,2,3)	12 (dynamics)
Voiced segment	1
HNR	2
Spectral LLDs, , functionals	**Number of features**
mfcc [1-4]	8 (dynamics)
Hammerberg Index	2
Alpha ratio	2 (dynamics)
slope (UV, V)	3 (dynamics)
spectral flux	2
Energy specific LLDs, functionals	**Number of features**
loudness	6 (dynamics, percentiles,loudness peaks per sec, equivalent sound db)

Table 5.5: Derived male acoustic features through the ablation study.

Voicing specific LLDs, functionals	**Number of features**
F0	7 (pitch, intonation contour,(mean falling slope), dynamics, percentiles)
Jitter	2
Shimmer	1
Harmonic difference	3
Harmonics (F1,2,3)	11 (dynamics)
Voiced segment	1 (voiced segment per sec, mean)
Spectral LLDs, functionals	**Number of features**
mfcc [1-4]	8 (dynamics)
Spectral energy	2 (dynamics)
Hammerberg Index	2 (dynamics)
slope (UV, V)	5 (dynamics)
spectral flux	4 (dynamics)
Energy specific LLDs, functionals	**Number of features**
loudness	8 (rising and falling slopes, dynamics, percentiles, loudness peaks per sec, equivalent sound db)

5.3.3.6 Automatic prediction of warmth and competence

Now that the acoustic correlates of SSC for female and male TTS voices are known, we performed automatic prediction of warmth and competence using the derived acoustic features. These experiments were carried out using both classification and regression algorithms.

- **Regression**: In this experiment, we predict the values corresponding to each of warmth and competence using a linear regressor and a support vector regressor. The results of this study are presented in table 5.10. The first block provides the results of the regression models applied to the derived acoustic features (through linear regression obtained from table 5.6 for female warmth and from table 5.8 for male warmth) and the subjective ratings of the characteristic, warmth. As mentioned previously, the warmth ratings were obtained by combining all the adjectives that were commonly found in both genders (7 adjectives commonly found under the factor warmth). Similarly, the second block in table 5.10 provides the details of regression experiments carried out on the characteristic, competence. The acoustic features used for this experiment were derived from table 5.7 for female competence and table 5.9 for male competence. MSE was used as the metric to evaluate the performance of the models. As the number of speech samples was limited (18 male voices and 18 female voices each uttering two sentences), LOSO-CV (leave-one-speaker-out cross-validation) was devised. We can observe that there is a significant improvement in the systems' performance when trained on acoustic correlates of warmth and competence as opposed to high dimensional inputs (when compared to the MSE scores in the tables 5.2 and

5.3). The number of acoustic features used in each of these experiments for male, female, warmth, and competence is detailed in the table 5.10.

- **Classification**: Further, we classify the synthetic voices into warm/cold and competent/incompetent. For this, we chose the voices that had highest and the lowest ratings of warmth/competence. From the subjective ratings of a wide range of TTS systems performed in the previous chapter (4), we can find that Google's voice, D (male) has the highest ratings for warmth and competence among the male voices (from the figure, 4.4, 4.6 in chapter 3). The female voices, Google's voice C and F exhibit highest subjective ratings for warmth and competence among others (from Figure, 4.5, 4.7 in chapter 4). Correspondingly, the voices generated through Festival (male = bdl, rms, female = slt, clb) had lower subjective ratings for warmth and competence. Therefore, we utilize these 8 voices (4 male, 4 female) voices for the classification task. 1132 sentences were generated from each of these voices using the CMU arctic database. Further, the gender-dependent classification of SSC was carried out by the models, SVM, and Neural Networks. The input provided to the models for warmth was 5 acoustic features from table 5.9 for male, table 5.6 for the female voices. Correspondingly, four acoustic features for male competence (from table 5.9) and five features for female competence (from table 5.7). The NN model used in this experiment was a four-layered Neural Network with the architecture [(input nodes)R,12R,8R,1S] (input nodes = 4 for male competence and 5 for the other experiments, R = relu, S= Sigmoid). The model training was carried out using an Adam optimizer [201] with a batch size of 16. The performance of the models was high in the case of both the characteristics and the genders. I assume that the data size also significantly impacts this performance. The results with 10-fold cross-validation are provided in table 5.11.

5.3.4 Observations

In the current study, we have examined the acoustic correlates of a wide range of TTS voices. In [202], the acoustic analysis of the ground truth TTS voices is presented. The study is motivated by the current workflow and examines the acoustic correlates of warmth and competence. However, the perception of warmth in [202] was examined using the perceptual analysis of the adjectives, friendliness, and likability collected from the two commercial TTS systems. Further, the analysis of competence was carried out using the subjective responses for the adjective, skilfulness from the ground truth TTS voices (obtained from chapter 4). Since the analysis was on TTS voices for the acoustic correlates of SSC, we intend to compare the results of the study presented in [202] with the current study. Also, we can observe that the adjective, likable was the common contributor to warmth in both the studies. Further, we also compare our results with the acoustic correlates of various emotions and speaker characteristics as previously observed in the literature.

- **Female warmth:** From table 5.6, we can observe that the acoustic features accountable for female warmth (from a wide range of TTS systems) are fundamental frequency, (F0 mean falling slope), F2 dynamics (standard deviation), Hammerberg Index, loudness, and unvoiced segment length. The fundamental frequency has also been previously found to contribute to the perception of different characteristics and emotions in natural speech [191, 195]. [195] discuss the correlations between voice likability, attractiveness, and fundamental frequency. As in the current studies, "likable" is one of the adjectives contributing to warmth in synthetic speech, our results are consistent with the studies presented in [195]. The acoustic correlates of SSC as derived from the two commercial systems (ground truth voices) are spectral flux, F1 mean, and F2 mean. Similarly, the authors in [203] show the dependence of the perception of the characteristic, warmth on F0 and its formats, F1 and F2 in German female speech (human speech). Hammerberg index provides information on the vocal quality based on the articulatory effort involved in producing speech. The value is obtained by computing the energy difference between the bands, 0-2kHz and 2-5kHz. Authors in [203] discuss the influence of the Hammerberg Index in the case of male attractiveness (natural german speech). In addition to the Hammerberg index, loudness has also been found to be an indicator of female warmth in synthetic voices (our studies on wide range TTS). [203] display the relevance of loudness in the perception of confidence in female German speech. Also studies in [78, 82] display the dependence of charismatic speech on speech intonation. [203] present the unvoiced segment lengths as the reliable indicator of the perception of attractiveness in male speech (human). In our studies, it was found to affect the perception of female warmth of synthetic voices.
- **Female competence:** The acoustic correlates of female competence are presented in table 5.7. The derived acoustic features accountable for female competence (from a wide range of TTS systems) are F0 dynamics (standard deviation), mfcc dynamics (mean), F1 amplitude, F3 dynamics (mean), and spectral flux. The acoustic features derived from ground truth TTS voices are spectral flux, voiced slope, mfcc. Apart from the previously observed features in female warmth, female competence is also dependent on spectral features such as mfcc, spectral flux, and slope. The mfccs were previously found to contribute to the perception of maturity in female speech from the studies in [203]. Additionally, mfccs were regarded as the first among other acoustic features to be contributing to various emotions in speech [191]. The spectral flux was found to commonly contribute to female warmth and competence in both studies (wide range TTS and 2 commercial TTS).
- **Male warmth:** Table 5.8 displays the derived acoustic correlates of male warmth in synthetic voices. The relevant acoustic features as presented in the table are loudness, mfcc3 dynamics (mean), HNR, F3 bandwidth, and spectral flux. The acoustic features derived from two commercial systems are the F1 mean, spectral slope, and loudness. F3 bandwidth was previously displayed to be an indicator of the characteristic, maturity in male German voices [203]. The spectral flux has been found to be contributing to all three, female warmth, female competence,

and male warmth in synthetic voices. HNR denotes the ratio of the energy of the harmonic sound (periodic component) to the energy of the noise (non-periodic components) in the speech signal. The higher the HNR, the lesser the noise in the signal, and vice versa. As previously seen, it relates to the roughness or hoarseness of the voice which would affect the perception of attractiveness in speech [196]. From our studies, we can observe that the HNR value is directly proportional to the perception of warmth in male synthetic speech. The feature spectral slope was commonly found to be contributing to male warmth in studies on natural as well as synthetic speech [203].

- **Male competence:** The acoustic correlates of competence in male synthetic voices are detailed in table 5.9. The derived acoustic features accountable for male competence (from a wide range of TTS systems) are loudness, mfcc4 dynamics (mean), F1 dynamics (mean), and Hammerberg Index. The acoustic features derived from ground truth TTS voices are F0 mean, voiced segment length. As discussed before, the Hammarberg index was found to contribute to male attractiveness in natural speech [203]. In the current studies, the value seems to positively affect the perception of competence in male synthetic voices. We have already observed that the lower pitch would affect the perception of trustworthiness in male speech [198]. Correspondingly, the studies on two commercial studies presented in [202] display negative correlations of the feature F0 semitone in the perception of competence from synthetic speech.
- **Additional observations on acoustic feature relevance:** Other than the observations made from the derived OpenSMILE features, we have also found the dependence of other acoustic features on different speech perceptions. From the different sets of studies carried out, we found the significance of speech pauses and speaking rate in the perception of warmth and competence from synthetic speech. While the insertion of pauses in neutral speech has not provided any notable differences in the perception of warmth, the participants could associate the voices with high speaking rates and no speech pauses to be highly competent over others. These observations are on par with the studies presented in [197]. Further, a similar survey was also conducted on compassionate speech (both WC dataset and Twitter sentences). An example sentence (derived from Twitter data) that displayed varied perceptions of warmth is presented below.

> "Don't put time on it. Relax! Maybe nap and get back to it when you get up"

In the above sentence, a speech pause inserted in a TTS speech sample (with lower F0) between the words, "Relax" and "Maybe nap" has been observed to be highly warm compared to the voices that did not include this pause. However, a systematic description of where to insert such pauses was not understood in the current studies. Overall, we can conclude that the insertion of speech pauses and lower F0 would positively affect the perceptions of warmth in synthetic speech.

On the other hand, high speaking rates contribute to the perception of competence in the generated speech.

5.4 Limitations

This section provides the limitations of the current studies.

- **Limited data:** Even though its captivating to understand the acoustic features contributing to warmth and competence in synthetic speech, the availability of the labeled data was limited (obtained from our previous studies in chapter 4). This has further affected the acoustic feature prediction for the relevant vocal cues of SSC. The input dimensions were higher than that of the number of training examples. Thus, detecting the multi-collinearity and the feature selection were challenging. The multi-collinearity was identified by the use of Pearson correlation between the features and adjectives. The feature selection was enabled by the recursive feature elimination approach. The availability of abundant data would have aided in effective feature modeling and predictions.
- **Acoustic features:** In this chapter, we derive the OpenSMILE features and discuss various dimensionality reduction techniques to predict the acoustic correlates of SSC. Investigation of different sets of acoustic features for the task and a fusion of different feature sets could have also been interesting. Therefore, one of our future works would focus on examining varied feature representations for the prediction of vocal cues of SSC.

5.5 Summary

In this chapter, we have examined the acoustic correlates of SSC in synthetic speech. The studies were carried out on the wide range of TTS systems introduced previously in chapter 4. Further, the comparison of the results was done with the studies previously performed on natural speech as well as the experiments carried out on two commercial TTS systems in [202]. The acoustic feature relevance was derived by employing various dimensionality reduction techniques on the derived 88-dimensional OpenSMILE features. Finally, the features derived from the backward elimination approach in linear regression were determined to be the acoustic correlates of SSC. The approach has also provided information on the impact of each feature on the perception of warmth and competence from synthetic speech. The study shows that the f0 and formant frequencies along with spectral flux contribute to the warmth in female speech (common features in wide-range TTS as well as ground truth TTS). While, loudness, mfccs, formants (F1 mean, F3 bandwidth), slope, and spectral flux contribute to male warmth (both wide-range and ground truth TTS). Correspondingly, contributors of competence as observed from both wide-range and ground

truth TTS for males are loudness, mfccs, F0 semitone, F1 mean, Hammarberg Index, and voiced segment length. For females, the contributors are F0, its formats (F1 mean, F3 mean), voiced slope, flux, and mfccs. Additionally, we have also observed that insertion of speech pauses associated with lower F0 would contribute to higher warmth and a high speaking rate leads to the perception of competence in synthetic speech.

Table 5.6: Acoustic features contributing to female warmth and their corresponding coefficients. The explained variance for female warmth (R squared) = 98%

Acoustic features	Coefficients
F0 meanFallingSlope	0.4718
F2 stddevNorm	0.6379
hammarbergIndex stddevNorm	-0.5254
loudnessPeaksPerSec	0.4224
StddevUnvoicedSegmentLength	0.1285

Table 5.7: Acoustic features contributing to female competence and their corresponding coefficients. The explained variance for female competence (R squared) = 93.3%

Acoustic features	Coefficients
F0 stddevNorm	1.1833
mfcc4 mean	-0.7423
F1amplitudeLogRelF0 stddevNorm	0.8561
F3frequency mean	-0.3834
spectralFlux amean	0.1285

Table 5.8: Acoustic features contributing to male warmth and their corresponding coefficients. The explained variance for male warmth (R squared) = 96.7%

Acoustic features	Coefficients
loudness mean	-2.7000
mfcc3 amean	0.7069
HNRdBACF stddevNorm	0.4708
F3bandwidth amean	-0.2639
spectralFlux amean	0.3586

Table 5.9: Acoustic features contributing to male competence and their corresponding coefficients. The explained variance for male competence (R squared) = 93%

Acoustic features	Coefficients
loudness stddevNorm	-0.4453
mfcc4 mean	0.5610
F1frequency mean	0.2816
HammarbergIndex stddevNorm	0.3903

Table 5.10: Results of regression techniques. AFs= number of acoustic features fed to the model, At/Ch. = attributes/characteristic, (W) warmth, (C) Competence, LR = Linear Regression, SVR = Support Vector Regressor, MSE = mean squared error

Model	Male			Female		
	AFs	At/Ch	MSE	AFs	At/Ch	MSE
LR	5	1 (W)	0.08	5	1 (W)	0.11
SVR	5	1 (W)	0.35	5	1 (W)	0.08
LR	4	1 (C)	0.11	5	1 (C)	0.22
SVR	4	1 (C)	0.10	5	1 (C)	0.33

Table 5.11: Classification of low/high warm/competent voices. AFs= number of acoustic features fed to the model, Ch=characteristic ((W) warmth in block 1 and (C) competence in block 2), LC = Linear Classifier, SVM = Support Vector Machine, Acc = Accuracy

Model	Male			Female		
	AFs	Ch	Acc	AFs	Ch	Acc
LC	5	1 (W)	99.1	5	1 (W)	97.5
SVM	5	1 (W)	99.3	5	1 (W)	100
LC	4	1 (C)	98.4	5	1 (C)	98.8
SVM	4	1 (C)	99.3	5	1 (C)	100

Chapter 6
Modeling using Voice Conversion

This chapter presents the modeling of synthetic speech for the positive perceptions (warm and competent) of negatively perceived (cold or incompetent) voices. Voice Conversion experiments carried out for both intra and inter-gender transformations are presented. The Star-GAN model utilized for the VC experiments detailed in this chapter is adapted from [123]. A discussion on the subjective evaluations of the converted speech was done between me and Sebastian Möller.

6.1 Related work

Besides their efficiency in generating images, GANs [122] have also displayed their excellence in modeling speech data [124, 123]. Different variants of GANs have been investigated for various Voice Conversion experiments and have proved to outperform the Variational Auto-encoders (VAEs) [204]. A Variational Autoencoder is a modified version (a probabilistic model) of a basic autoencoder. The model consists of an encoder and a decoder. The encoder takes the input speech data (let's say, x) and projects it onto a latent space. The sampled information from this latent space is fed to the decoder and the decoder generates the samples that are as close as possible to that of the x. The model optimization aims at reducing the distance between the true posterior and the predicted variational posterior and the loss is estimated using the Kullback-Leibler divergence (KL loss). Eventually, Conditional VAEs were proposed which would consist of an additional label (an auxiliary attribute) provided to accommodate different speech attributes (for example, speaker identity) [205]. This additional dimension (the auxiliary variable or speaker identity) in the encoder-decoder models could further facilitate the speaker conversion based on the speaker ID (or the relevant speech attribute) provided to the decoder during the conversion. Also, due to their architecture, these networks do not require parallel data from the source and the target speakers. However, the VAEs suffer from poor conversion quality (generation of over-smoothed speech). GANs were reported to handle this

drawback of VAEs in speaker conversion and have been found to produce better conversion quality than VAEs. [124] employs a cycle-GAN for unpaired speech samples of source and target speakers. Cycle-GAN is a variant of GAN which relies on the consistency of the mappings between the pair of the unaligned speaker's data. The network consists of two generators G, F, where G tries to map the input features (I) to a target's voice (O), $G: I \rightarrow O$ for which the inverse mapping by F is represented as $F: O \rightarrow I$. The generator network in the cycle-GAN, therefore, ensures to produce the predictions are as realistic as possible through this cycle consistency network. Also, there is a discriminator D network that tries to differentiate between the real and fake predictions of the target. The model optimization is therefore handled while minimizing a) the adversarial loss due to the predictions made by the generator (G tries to produce fake predictions and fool the discriminator (i.e., to maximize the discriminator loss), D), and b) the adversarial loss calculated for the discriminator network in classifying the real/fake targets (tries to maximize the generation loss). In addition, the cycle-GAN also consists of cycle consistency loss computed from the forward and the reverse mappings functions, G, F. Therefore, the training loss is a combination of adversarial losses and cycle-consistency loss. Even though the cycle-GANs were efficient compared to previously used VC approaches, they had a major shortcoming in their applications for different speech domains (the use of auxiliary variables) [124]. With the increase in the number of speech attributes, the number of parameters to be learned would also increase while the number of training examples is still the same. Therefore the model is limited to the one-to-one mapping of speakers. Follow-up work by the same research group was proposed using Star-GANs for many-to-many voice conversion with non-parallel data from the source and the target speakers [123]. The VC experiments presented in this chapter are carried out following the work provided in [123].

6.1.1 Description of the Star-GAN model

The Star-GAN-based VC presented in [123] consists of multiple advantages over the previously proposed Cycle-GAN VC [124]. Apart from the aforementioned advantages (does not require parallel utterances from source and target speakers, and can carry out many-to-many VC), the Star-GAN model proposed in [123] also leverages a) the transcription-free VC (text and speech alignment errors can be avoided), b) no need for time alignment of the source and the target utterances (time alignment techniques such as dynamic time warping and the errors due to these warping of speech signals can be avoided), c) one generator network can learn multiple tasks from the provided speech samples (such as speaker identity, speaker characteristics, language information, style information, etc.,), and d) can generate good quality speech with limited data (speech data spanning less than an hour or several minutes).

The generator (G) is an encoder-decoder type network with an auxiliary attribute, s. This auxiliary attribute can hold multiple pieces of information present in the

speech signal (as mentioned before, the speaker's identity or speaker characteristics or language information). This information can be stored in the form of one-hot vectors and multiple tasks can be concatenated while training the network (for instance, speaker id + speaker characteristics + language id). In the current study, the auxiliary variable, s only consists of the speaker identity of the target speaker. The generator performs mapping of the input features from the feature space (X) to that of the target's voice specified in the auxiliary variable or the target attribute s (in one-hot representation), $Y^{'} = G(X,s)$ where Y' is the generated feature space.

The generator network should produce speech data that is as close as possible to that of the target's voice. Further, if the generated speech is close to that of the target speaker, is verified by a real/fake discriminator network through its predicted probabilities $D(Y,s)$. Besides, the generator and the discriminator, the experimental setup additionally consists of a domain classifier (C) to examine the class probabilities of the generated speech (class probability is represented as $p_{cl}(s/Y)$). Since the current study involves only speaker identity as the auxiliary attribute, the class probabilities are computed only for the belongingness of the given speech features to a specific class/speaker. The loss functions for each of these networks are provided further. The adversarial loss for the discriminator is provided in the equation 6.1, where $\mathbb{E}$ is the expectation, $Y \sim p(Y|s)$ denotes the feature sequences in the acoustic feature space, Y or of real speech with the attribute, s, and $X \sim p(X)$ represents it with an arbitrary attribute. The generator training is carried out using the adversarial loss defined as $\mathcal{L}_{advG}$ and is computed as shown in equation 6.2. Further, the classification losses computed from the domain classifier ($\mathcal{L}_{clC}$) and the generator network ($\mathcal{L}_{clG}$) are provided in the equation 6.3 and 6.4. The loss values ($\mathcal{L}_{clC}$) indicate the error in the predictions made by the classifier network in classifying the acoustic feature sequences from Y, $Y \sim p(Y|s)$. Therefore, the training of the classification aims at minimizing this error with respect to the domain classifier (C). Correspondingly, ($\mathcal{L}_{clG}$) represents the error in the predictions of G while verifying their belongingness to the speaker label as described in s. Therefore, ($\mathcal{L}_{clG}$) is to be minimized with respect to the generator network.

$$\begin{aligned}\mathcal{L}_{avdD} = &-\mathbb{E}_{s\sim p(s),Y\sim p(Y|s)}[log(D(Y,s))] \\ &-\mathbb{E}_{X\sim p(X),s\sim p(s)}[log(1-D(G(X,s),s))]\end{aligned} \tag{6.1}$$

$$\mathcal{L}_{avdG} = -\mathbb{E}_{X\sim p(X),s\sim p(s)}[log(D(G(X,s),s))] \tag{6.2}$$

$$\mathcal{L}_{clC} = -\mathbb{E}_{s\sim p(s),Y\sim p(Y|s)}[log(p_{cl}(s|Y))] \tag{6.3}$$

$$\mathcal{L}_{clG} = -\mathbb{E}_{X\sim p(X),s\sim p(s)}[log(p_{cl}(s|G(X,s)))] \tag{6.4}$$

All the networks used in the Star-GAN model (Generator, Discriminator, domain classifier) are designed using CNNs. Therefore, the model leverages sequential processing and learns the acoustic feature sequences for speaker conversion instead of frame-to-frame mapping. Among these, the generator is an encoder-decoder model where only the decoder is fed with the auxiliary attribute while training the model. Unlike CVAEs and cycle-GAN VC, for the current star-GAN VC setup, there is no

need to feed the generator with the auxiliary attribute (or the target attribute) at the test time. In this chapter, we utilize the star-GAN model proposed in [123] to examine the positive perceptions of the converted speech.

6.2 Overview

This chapter addresses the following research question.

? Research question: ***"What modifications of synthetic speech enable their positive perceptions?"***

The focus of this chapter is to alter the negatively perceived synthetic voices into positive ones. In order to achieve the same, we employ voice conversion in the current studies. The contributions of this chapter are detailed below.

- **Why VC?:** Spectral conversion has been found to impact the perceptions of various emotions in speech [125, 206]. [125] provides a speaker-independent emotion conversion using the Variational Autoencoding Wasserstein Generative Adversarial Network (VAW-GAN). The study investigates two separate VAW-GAN pipelines for each of spectral conversion and prosody conversion. The conversion results display an improved emotion conversion performance when conditioning the model on the Continuous Wavelet Transform (CWT) based F0. Similarly, [206] utilizes a highway network to model the pitch, energy, and spectral information of the source to that of the target's emotion. The subjective evaluations of the converted speech were performed for the classification of perceived emotions. Also, the authors investigate this phenomenon in the case of human speech as well as wavenet-generated speech. Inspired by the previous works on emotion conversion, we examine VC for the transformation of SSC in synthetic speech through this work.
- **How to evaluate the generated speech?:** In this work, we are interested in the conversion of negatively perceived voices into positive ones. Therefore, the evaluation of the converted samples must display a degree of variation in the perception of SSC from that of the original voices. In order to do so, firstly we need to define the highly warm/competent, and highly cold/incompetent voices. We show the details of deriving the ground truth warm/cold, competent/incompetent voices in the section 6.3. Secondly, our the converted speech was validated using the previously defined ground truth warm/cold, competent/incompetent (from section 6.3). Correspondingly, our evaluation setup included the AB preference test between the converted voice and the negatively perceived (highly cold/incompetent) voices. We thus try to interpret how better/positive is the converted speech from

that of the original voice (negative voice). Alternatively, we also evaluate the converted speech on a 5-point scale for the social perceptions of the converted speech. Finally, we provide a comparison of AB preference tests and the 5-point scaling tests.

6.3 Experimental setup

This section provides the experimental details of VC studies performed using the Star-GAN model. In order to investigate the perception of SSC in converted speech, both intra-gender and inter-gender experiments were carried out. A traditional VC setup requires a source speaker and a target speaker. The source and the target speakers for each of the experiments presented in this section are derived using the subjective ratings obtained for the ground truth TTS voices proposed in chapter 4.

6.3.1 Choice of speakers and adjectives

From the studies in chapter 4, it is evident that the speech quality and the naturalness of the generated voice have a significant impact on the perception of SSC. Additionally, performing VC over the synthetic speech would further alter the quality of the converted speech. Therefore, in order to retain the speech quality, the experiments were carried out on ground truth TTS voices alone.

6.3.1.1 Deriving ground truth adjectives for warmth and competence

The subjective evaluations provided in chapter 4 utilize a long list of adjectives. As we have already seen previously, this would increase the number of questions provided to the participants during the subjective tests. Therefore, considering the relevance of the adjectives and the results obtained from the previous studies, in this section, I derive two adjectives (from the long list provided in chapter 4) for each of warmth and competence. These adjectives are derived by considering their factor loading values. The factor loading value provides information on how well the adjective fits under the particular factor. The higher the factor loading value, the better it fits and represents the corresponding factor.

Table 6.1 presents the adjectives and the corresponding factor loadings for each of male and female voices under the factor/characteristic, warmth [167] (from chapter 4). The adjectives commonly found in both male and female voices are kind, sympathetic, likable, and pleasant. Among these, the adjectives kind (male = 0.82, female = 0.75), and sympathetic (male = 0.78, female = 0.77) display the highest factor loadings for the characteristic, warmth. Here, even though, the adjective likable displayed a slightly higher factor loading (0.79) than the adjective sympathetic

Table 6.1: Results of the ground truth experiments from chapter 4 [167]. The adjectives and the corresponding factor loadings of the characteristic/factor, warmth. The adjectives in bold display the highest factor loadings among others under the factor, warmth.

Female		Male	
Adjectives	Factor loadings	Adjectives	Factor loadings
Kind	**0.75**	**Kind**	**0.82**
Sympathetic	**0.77**	**Sympathetic**	**0.78**
Likable	0.71	Likable	0.79
Pleasant	0.67	Pleasant	0.77
Accessible	0.63	Friendly	0.72

(0.75), we chose to use the adjective, sympathetic. This is in order to be in line with that of the female voices and define common ground truth adjectives irrespective of gender. Therefore, hereafter the adjectives, kind and sympathetic are determined to be the metrics in the perception of warmth from synthetic speech.

Table 6.2: Results of the ground truth experiments from chapter 4 [167]. The adjectives and the corresponding factor loadings of the factor, competence. The adjectives in bold display the highest factor loadings among others under the factor, competence.

Female		Male	
Adjectives	Factor loadings	Adjectives	Factor loadings
Responsible	**0.93**	**Responsible**	**0.89**
Skilful	**0.84**	**Skillful**	**0.88**
Reliable	0.81	Reliable	0.87
Confident	0.65	Confident	0.79

Similarly, table 6.2 shows the factor loadings of competence for the ground truth voices. Among the four adjectives commonly loaded under the factor, competence, the adjectives responsible (male = 0.89, female = 0.93) and skillful (male = 0.84, female = 0.88) display the highest factor loadings. Hence, from now on, the perception of competence from synthetic voices is evaluated through the scales, responsible and skillful.

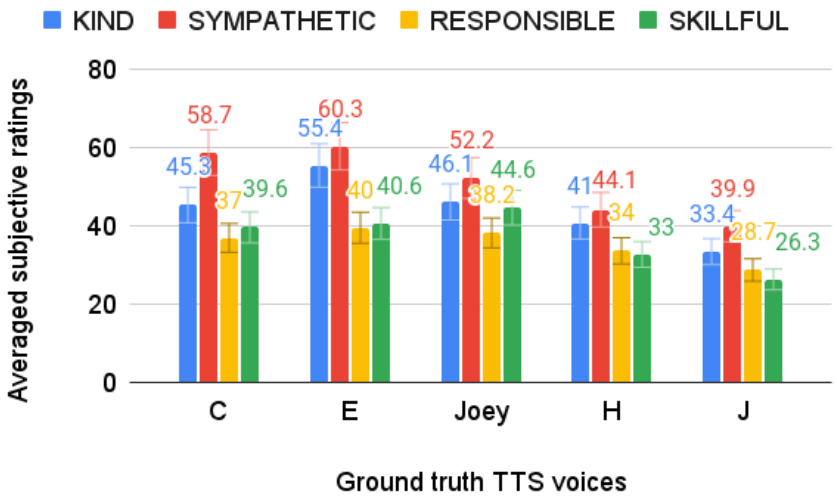

Fig. 6.1: Averaged subjective ratings for ground truth TTS voices. C, E, H = Google wavenet female voices, J = Google wavenet male voice, Joey = Amazon Polly male voice.

6.3.1.2 Deriving ground TTS truth voices for the characteristics, warm/cold and competent/incompetent

The ground truth voices for warm/cold and competent/incompetent are derived from the averaged subjective ratings computed over the ground truth adjectives (warmth= kind+sympathetic, competence = responsible+skilful) for both genders separately. The ground truth TTS voices that displayed the highest averaged ratings among others were considered the ground truth for highly warm/competent voices. Similarly, the voices with the least averaged ratings among others were treated as ground truth for cold/incompetent voices. The details of the male and female ground truth voices for each of warmth and competence are provided below.

- ***Highly warm and highly competent TTS voices***: Figure 6.1 presents the averaged subjective ratings collected for the adjectives, kind, sympathetic, responsible and skillful (from chapter 4). The figure shows the averaged subjective ratings of three female (C, E, H) and two male (Joey, J) voices that displayed the highest and the lowest ratings among other voices (5 voices out of 20 (Google =10+ Amazon Polly=10) were used in the current study). The TTS voice, E (Google wavenet female) displays the highest averaged ratings for warmth (kind = 55.4, sympathetic=60.3), and competence (responsible = 40, skillful = 40.6). Therefore, E is considered the ground truth for highly warm and highly competent female TTS voices. Correspondingly, the male voice, Joey (male) exhibited the highest warmth

(kind = 46.1, sympathetic = 52.2) and competence (responsible = 38.2, skillful = 44.6) among other male voices. Thus, the voice Joey was regarded as the ground truth for a highly warm and highly competent male TTS voice. Additionally, the TTS voice C (female, kind = 45.3, sympathetic = 58.7, responsible = 37, skillful = 39.6), was also considered in the current studies as it displayed the highest warmth and competence ratings in experiments with a wide range of TTS voices (chapter 4).

- ***Highyly cold and highly incompetent TTS voices***: Of all the female voices, H displayed the lowest averaged ratings for warmth (kind (41) +sympathetic (44.1)) and competence (responsible (34)+skillful (33)). Accordingly, the voice H was regarded as the ground truth female voice for cold and incompetent voices. Among male voices, J displayed the lowest ratings for kind (33.4), sympathetic (39.9), responsible (28.7), and skillful (26.3). Therefore, the TTS voice J was used as representative of the cold and incompetent TTS male voice. The details of these voices are further summarised in the table 6.3.

Table 6.3: Details of ground truth warm/cold/competent/incompetent voices

Female		Male	
TTS voice	Characteristic	TTS voice	Characteristic
E	Highly warm/competent	Joey	Highly warm/competent
C	Highly warm/competent	J	Highly cold/incompetent
H	Highly cold/incompetent	-	-

6.3.2 Data preparation for VC setup

The dataset used for the VC experiments was the neutral speech, CMU arctic database [160]. The goal is to achieve positive perceptions of the generated speech. Therefore, the voices with the lowest warmth and competence ratings were to be transformed into highly warm and highly competent voices. In a traditional VC setup (without any feature-specific modeling), the converted voice sounds similar to the target speaker but retains the speaking style of the source speaker. This property of the conventional VC setup is leveraged for the perception of SSC in the generated speech samples. In other words, instead of modifying the SSC of negatively perceived voices, in our study, we focus on altering the voices of the highly warm/competent speakers to sound like the target speakers (negatively perceived speakers) while retaining the characteristics of positively perceived voices. The hypothesis is that the conversion should render a voice that sounds as close as possible to that of the target (cold/competent) speaker, but should contain the characteristics (SSC) of the highly warm/competent voices (source speaker's characteristics). That being the case, in all of the VC experiments presented in this section, the ground truth voices for high

warmth and competence (E, C, Joey = highly warm and competent voices) were considered as the source speakers. Accordingly, the TTS voices, H and J (highly cold and incompetent voices) were the target speakers for the different intra and inter-gender experiments.

Even though the conversion was carried out only between the source and target speakers, the training of the Star-GAN model was accomplished with all the ground truth voices (except for the baseline. Baseline was trained only on 2 speakers, C and H). This was done to a) improve the quality of the converted speech, and b) train the VC model with the voices that displayed varied perceptions of social speaker characteristics. The assumption was that including other TTS voices (voices with different degrees of warm/competent ratings) in the study, apart from the source and target voices alone (voices at opposite extremes, highly warm - highly cold; highly competent - highly incompetent) would aid in bridging the gap between the social perceptions of the converted voices. Thus, along with the source and target speakers, the speech samples were also generated (for CMU arctic database) for the remaining ground truth voices using Google and Amazon Polly TTS systems.

Further, all the speech samples were converted to '.wav' format and are sampled at 22.5kHz using the SOX command. The acoustic features were derived using the WORLD vocoder [207]. Therefore, the feature vector consisted of the spectral envelope (36-dimensional Mel Cepstral Coefficients), logarithmic fundamental frequency (log F0), and aperiodicities (ap).

6.3.3 Experimental details

The experimental setup consists of seven different experiments, a) baseline, b) 3 intra-gender conversions, and c) 3 inter-gender conversions. The model used for all the experiments is the same. The differentiation between the baseline and the other conversions is only in terms of the number of speakers used in the training phase. The baseline model was developed with two voices, C (source, female) and H (target, female). The intra-gender conversion consisted of three experiments, a) E (highly warm/competent female) as the source and H (highly cold/incompetent female) as the target, b) C (highly warm/competent female) as the source speaker and H (highly cold/incompetent female) as the target speaker, and c) Joey (highly warm/competent male) as the source and J (highly cold/incompetent male) as the target. Similarly, three experiments were carried out for inter-gender conversion, a) E (highly warm/competent female) as the source speaker and J (highly cold/incompetent male) as the target, b) C (highly warm/competent female) as the source and J (highly cold/incompetent male) as the target, and c) Joey (highly warm/competent male) as the source and H (highly cold/incompetent female) as the target. The experiments were carried out on an NVIDIA 1080 Titan GPU 12GB. Each experiment mentioned in this section took about 12 hours to complete the conversion process and synthesize speech. The data split was as follows, training data = 80%, validation data = 10%, and the test data = 10%. Except for the baseline model, all the other conversions con-

sisted of 20 voices in the training phase. The generated speech samples are available at [1].

6.4 Subjective evaluation

The subjective evaluation of the converted speech was threefold, a) speech quality, naturalness, and speaker similarity, b) AB preference test (with an option for No preference) for warmth and competence, and c) a 5-point scale (direct scaling test) for warmth and competence. The number of participants took part in the subjective tests were 28 (male = 18, female =10, age range = 23 to 33, mean =26.3, std =1.4). The participants were university students and are compensated for their participation.

6.4.1 Speech quality, naturalness, and speaker similarity

The first step in the evaluation consists of two parts, a) collection of the subjective ratings on an absolute 5-point scale for i) speech quality, ii) naturalness, and b) an ABX preference test for speaker similarity. The 5-point scale description for the evaluation of quality and naturalness is as follows, 1 = poor quality/not at all natural, 5 = high quality/ very natural. The speaker similarity test was carried out between the baseline model and the intra-gender conversion (C_H). Since the source and the targets are the same in these experiments, we compare the baseline with this system's output (intra-gender conversion C_H) throughout our studies. In the ABX preference test (speaker similarity test), the participants were provided with the original target speech sample (X) (in our study, its H) and the converted samples from the baseline model, and the intra-gender conversion model (C_H). They would listen to all three speech samples and choose the voice (baseline converted voice or the voice from inter-gender conversion (in ABX test, either A or B)) that is close to that of the target's voice (H). The participants can also select "No preference" if they cannot decide between A or B. The speech samples provided in all three tests (quality, naturalness, and similarity) were randomized to avoid any bias in the ratings. The number of speech samples provided in this evaluation setup from each conversion was 15 (CMU arctic database). Therefore, the number of responses collected was 15 (conversions) * 3 (speech quality, naturalness, speaker similarity) = 45.

6.4.2 AB preference test for warmth and competence

Similarly, to validate the social perceptions of the converted speech, the subjective evaluation was carried out for the voices obtained from 7 VC systems (baseline [C

[1] https://saisirishar.github.io/VCsamplesforSSC.github.io/

as the source, H as target], 3 inter-gender conversions, and 3 intra-gender conversions). Accordingly, an AB preference test was designed. In order to perceive the characteristic, warmth, the converted samples were evaluated on the metrics namely, kindness and sympathy. Similarly, the characteristic, competence was evaluated on the scales, responsible and skillful. The AB tests were carried out separately for the target speakers, H (female), and J (male). In total, 12 comparisons (6 for male, 6 for female) were made for the perception of each metric in the voices, H and J. Among these 12 comparisons, 6 were between the target speaker and the converted voice from each of the intra-gender and inter-gender experiments (AB test on female voice, 1) H Vs E_H (E as the source, H as target), 2) H Vs C_H, 3) H Vs Joey_H; AB test on male voice, 4) J Vs E_J, 5) J Vs C_J, 6) J Vs Joey_J). The other six were comparisons within the converted speech samples (AB test on female voice, 1) E_H, Vs C_H, 2) C_H Vs Joey_H, 3) Joey_H Vs E_H, AB test on male voice, 4) E_J Vs C_J, 5) C_J Vs Joey_J, 6) E_J Vs Joey_J). The voices A and B (in AB preference test) for each comparison are different (provided in figures 6.4, 6.5, 6.6, 6.7, 6.8, 6.10, 6.9, 6.11). The participants were free to listen to the speech samples any number of times during the study. They were also allowed to take breaks in between to avoid any fatigue.

6.4.3 5-point direct scaling test

Apart from the preference tests, we have also carried out the direct scaling test in the current study with the target speakers (H, J) and the converted voices. As opposed to the AB preference test, in the current evaluation, the participants were provided with scales that consist of adjective-antonym pairs. The metrics used for the evaluation of warmth and competence are the same as seen in the previous AB preference test (warmth = kind, sympathetic; competence = responsible, skillful). The converted speech samples were rated on each of those metrics with the bipolar adjectives defined at the extremes of the scale. The evaluation of warmth and competence using the ground truth voices was previously performed in [167]. However, the study was carried out on averaged speech samples (1 speech segment = 8 speech samples), and on more variety of scales (warmth = friendly, kindness, likable, pleasant, sympathetic; competence = confidence, reliable, responsible, skillful). Further, the current evaluation is carried out on arctic speech samples (similar data type, neutral speech).

6.5 Observations

- ***Speech quality:*** Figure 6.2 presents the results of the subjective ratings collected for speech quality and naturalness. One of the observations made through the subjective tests was that the TTS speaker, Joey has a breathy voice. Therefore, the conversion carried out using Joey as the source rendered a poor quality conversion

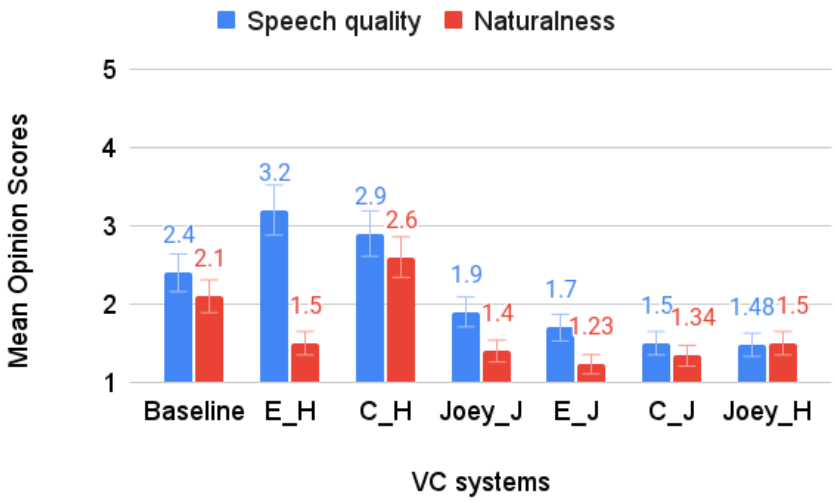

Fig. 6.2: Mean Opinion Scores calculated for subjective ratings of speech quality, and naturalness along with their 95% confidence intervals. Baseline = model trained with only 2 speakers (C, H), intra-gender experiments = E_H (female-to-female), C_H (female-to-female), Joey_J (male-to-male), Inter-gender experiments = E_J (female-to-male), C_J (female-to-male), Joey_H (male-to-female).

over the other source speakers. The inter-gender conversion between Joey (source) and H (target) has the lowest conversion quality (1.48) of all the systems (inter-gender conversion systems). Similarly, the intra-gender conversion between Joey (source) and J (target) has the lowest speech quality among the intra-gender system performances (1.9). On the other hand, the intra-gender conversion between the female voices, E (source) and H (target) displayed the highest speech quality (3.2) over the other conversions. Since, the amount of data used for VC experiments other than the baseline system, was much higher, the quality of the converted speech was better than the baseline conversion (baseline = 2.4, C_H = 2.9). We can also observe that the speech quality in the case of inter-gender conversions is bad when compared to intra-gender conversions.

- ***Naturalness:*** The subjective ratings of the metric, naturalness seemed to have been influenced by the intelligibility of the perceived speech. The intelligibility of the converted speech was good when the voice, C was used as the source speaker compared to others. Therefore, from figure 6.2 we can observe that the naturalness ratings were higher when C was the source speaker (C_H = 2.6). Accordingly, the inter-gender conversion between E (source) and J (target) had the least MOS ratings for naturalness (1.23) (less intelligible speech). Among the intra-gender experiments, the conversion between Joey (source) and J (target)

(male-to-male) displayed the lowest ratings of naturalness (1.4). Among inter-gender conversions, the conversion between Joey (source) and H (target) showed the highest naturalness (1.5).

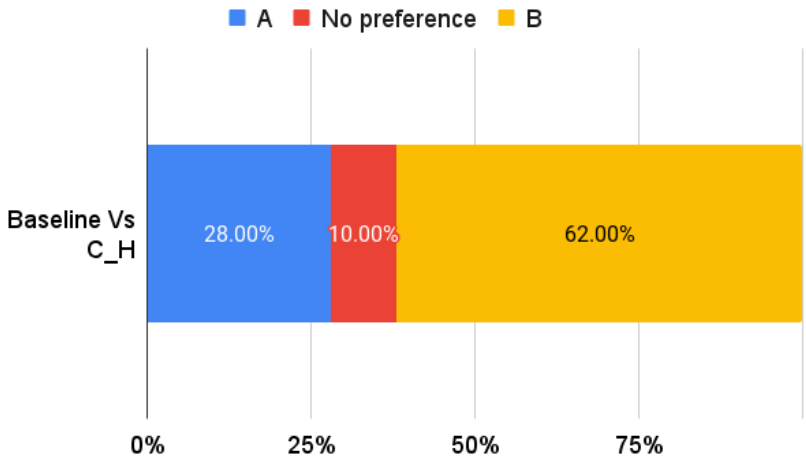

Fig. 6.3: Results of ABX preference tests carried out for the metric, speaker similarity for each of baseline (C_H) and the intra-gender conversion (C_H).)

- ***Speaker similarity:*** Figure 6.3 displays the results of the ABX preference test collected to interpret the speaker similarity of the converted speech to that of the target's voice. We present the comparison between the baseline model (C_H) and the intra-gender conversion (C_H) for better interpretability of the speaker conversion. We can find the improved speaker similarity when the conversion involved the training with additional data (baseline involves data only from the desired source and the target voices; The system performance has enhanced with the inclusion of additional data). The participants could identify the voice generated from the intra-gender conversion (C_H) to be close to that of the original target's voice (H).

6.5.1 AB preference test for warmth

- ***Perception of warmth and competence in the female voice, H:*** The Pearson correlation coefficient (r=.62; $p < 0.05$) (person correlation values calculated on the subjective responses collected for the perceptual studies presented previously in chapter 4) displayed similarities between the ratings of the adjectives, kind and sympathetic. Accordingly, the perception of warmth was directly correlated

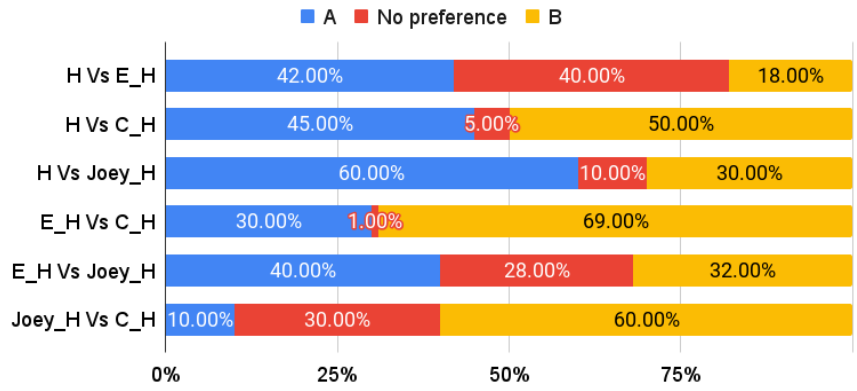

Fig. 6.4: Results of AB preference tests carried out for the metric, kindness for the female voice, H.

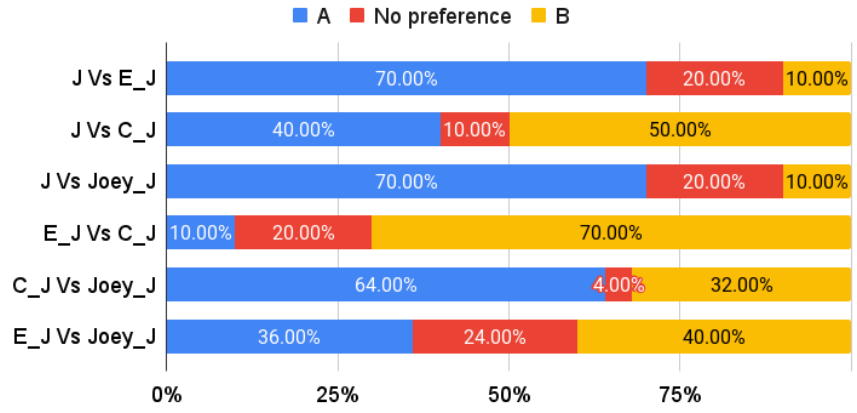

Fig. 6.5: Results of AB preference tests carried out for the metric, kindness for the male speaker, J.

with these subjective responses. Figure 6.4 and figure 6.6 present the results of the AB preference test carried out for the adjectives, kind and sympathetic respectively. Similarly figure 6.8, 6.9 present the results of subjective responses for the metrics, responsible and skilful respectively. There are some observations made from the speech samples of different speakers which have further influenced the perception of SSC from these voices. The female voice, H displayed a higher speaking rate, and high pitch than the female voice, C. Therefore, due to the high speaking rate, the voice was perceived to be skillful (when the quality

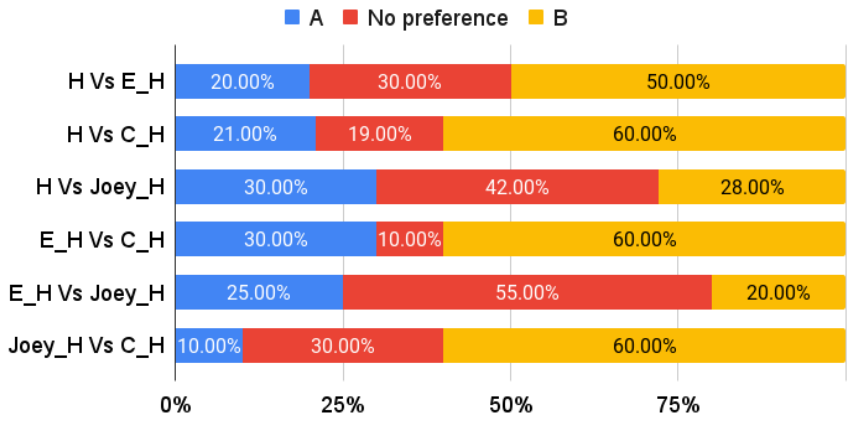

Fig. 6.6: Results of AB preference tests carried out for the metric, sympathetic for the female speaker, H.

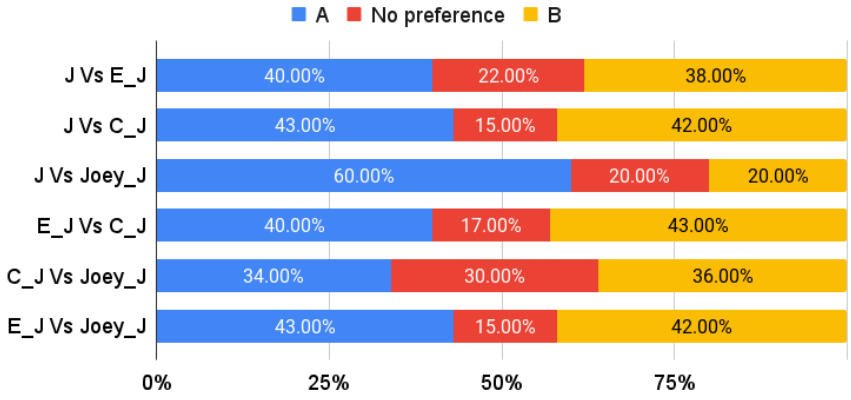

Fig. 6.7: Results of AB preference tests carried out for the metric, sympathetic for the male speaker, J.

of the converted speech was good and the speech was intelligible). Similarly, the pitch has also contributed to an increased perception of warmth. However, this occurred only when voice, C was used as the source in the conversion. As the speaking rate of C was less compared to H, the conversion resulted in a reduced speaking rate and original pitch of H. Therefore, reduced speaking rate and increased pitch in the intra-gender conversion (C_H) have the highest perception of warmth (on the scales, kind, sympathetic). The voice with high

speaking rates though considered skillful (also from studies presented in chapter 5; high speaking rate=highly competent) was not found to be responsible. This was also corroborated by the Pearson correlation (r= -.07; $p<0.01$) calculated between the subjective responses of these adjectives. Further, we found that the subjective responses for the adjective, responsible were found to be correlated with the responses of warmth ratings (kind ($r = .706$; $p<0.05$)) and sympathetic ($r = .75$; $p<0.05$)). The voices with moderate speaking rates were considered warm (kind, sympathetic) and responsible. While the voices with high speaking rates were considered competent (skillful) but less warm and have a lower perception of responsibility. All the subjective results were found to be statistically significant ($p<0.05$).

- ***Perception of warmth and competence in the male voice, J:*** The perception of male warmth through AB tests is displayed in the figures 6.5 (Kind), 6.7 (Sympathetic). The competence ratings are presented in the figures 6.10 (responsible) and 6.11 (skillful). As opposed to studies on female voices, the male voices did not exhibit a high speaking rate. However, we have observed that a moderate speaking rate and lowered pitch have improved the positive perceptions of the male-converted voices. Similar to that of the observations made in female voices, the conversion with C as the source had the highest perception of the dimension, kind. On the remaining dimensions (sympathetic, responsible, and skillful) we can observe equal preference for the target voice, and converted voice (in comparisons between the target and the converted speech) and also the comparison between the conversions. The explanations for this phenomenon are not very obvious from the feedback from the listeners or through listening to the speech samples.

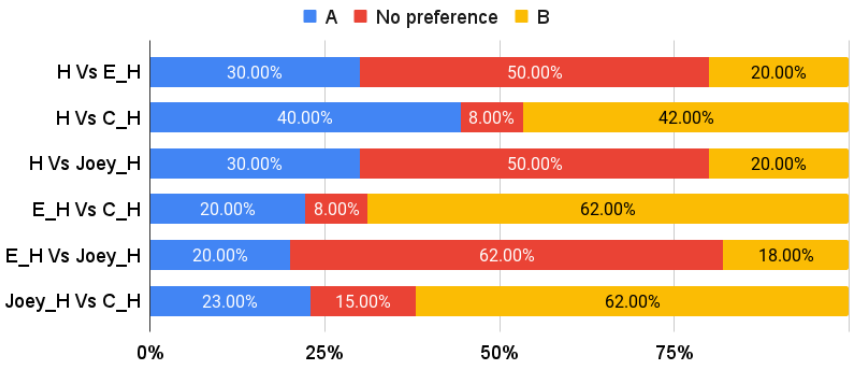

Fig. 6.8: Results of AB preference tests carried out for the metric, responsible for the female voice, H.

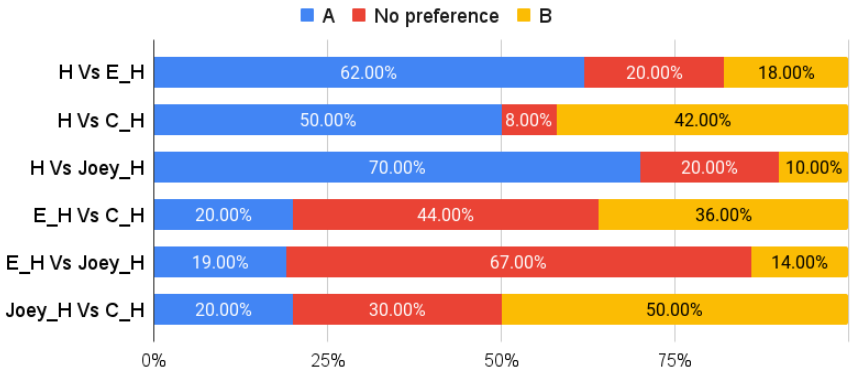

Fig. 6.9: Results of AB preference tests carried out for the metric, skillful for the female voice, H.

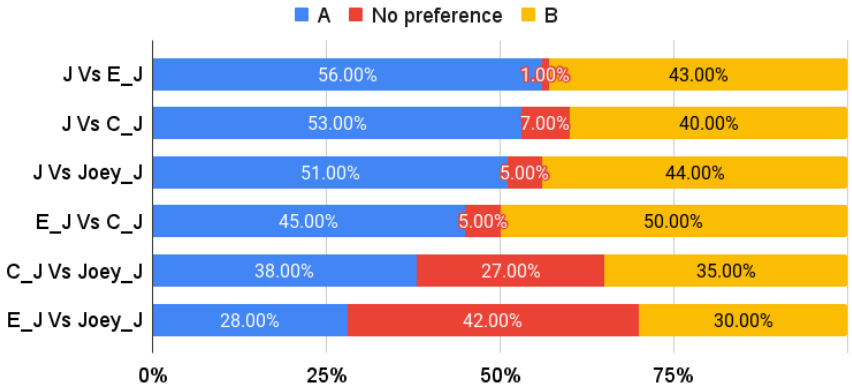

Fig. 6.10: Results of AB preference tests carried out for the metric, responsible for the male voice, J.

6.5.2 Direct scaling test

Figure 6.12 displays the results of the 5-point scale-based evaluation of warmth in a) the negatively perceived TTS voices (ground truth cold voices = H, J), and b) the converted voices (intra and inter-gender conversions). In this figure, we presented the mean opinion scores collected for the target voices/cold ground truth voices (H, J) to verify and compare their perceptions before and after the VC experiments. We can observe that in the intra-gender conversion between C (source) and H (target),

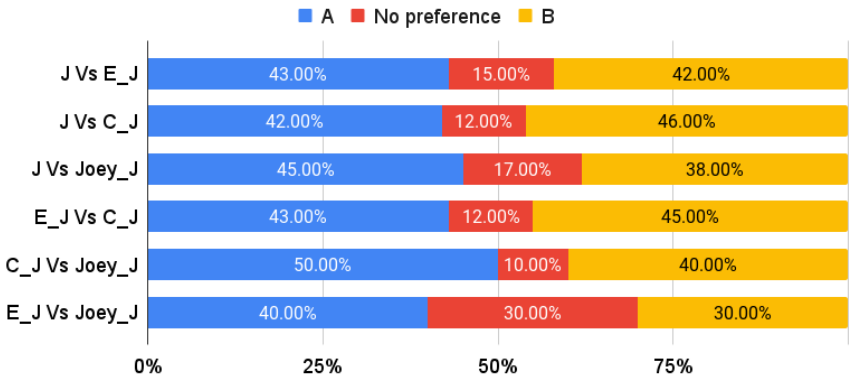

Fig. 6.11: Results of AB preference tests carried out for the metric, skillful for the male voice, J.

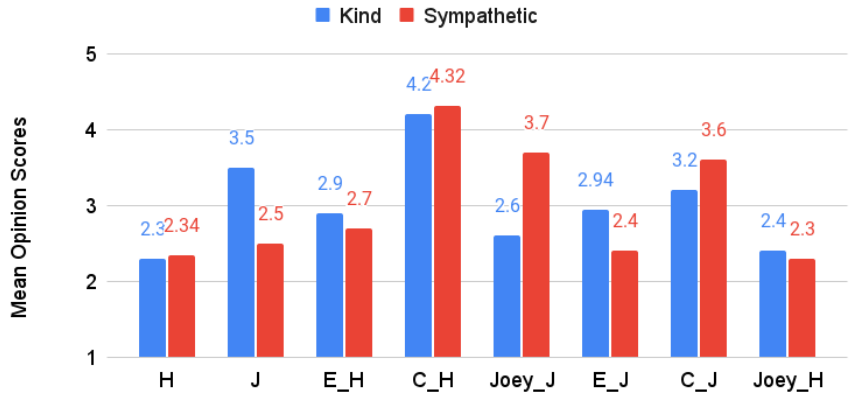

Fig. 6.12: Results of the evaluation carried out for warmth on the 5-point continuous scales, kindness, and sympathy.

the converted voice was perceived as more positive (kind = 4.2, sympathetic = 4.32) than that of the original (target) voice (kind = 2.3, sympathetic = 2.34). Accordingly, the male voice, J was perceived to be more sympathetic (sympathetic = 3.6) after the conversion (C as the source and J as the target) than the original (target) voice (sympathetic = 2.5). The other inter-gender conversions (E_J, Joey_H) did not fetch positive perceptions of the converted voice. In the comparison between the target voice, H, and the inter-gender conversion, Joey_H displays almost similar perceptions of the converted and the original target's voice. However, the comparison between

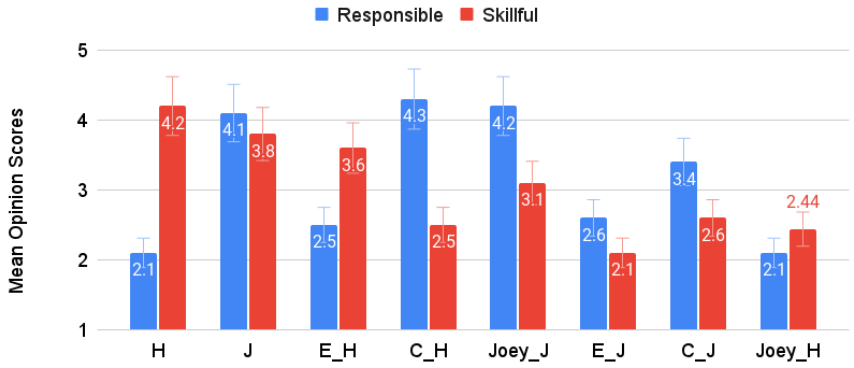

Fig. 6.13: Results of the evaluation performed for the characteristic, competence on the 5-point continuous scales, responsible, skillful.

the target voice, J (kind = 3.5, sympathetic, 2.5), and the inter-gender conversion, E_J (kind = 2.94, sympathetic, 2.4) present negative perceptions of the converted voice.

Figure 6.13 depicts the competence ratings collected over the 5-point scales, responsible and skillful. We observe that the target voice, H, was perceived to be highly skillful (H = 4.2), and the voice, J was perceived as highly responsible (4.1). The intra-gender conversion between C (source) and H (target) improved the perception of the adjective, responsible (H = 2.1, C_H = 4.3) in the voice, H. Also, from the figure, we can observe that the converted voice, C_H was the most positively perceived voice over the scale, responsible when compared to other voices. However, the perception of the adjective, skillful was negatively affected by the conversion (H = 4.2, C_H = 2.5). This could be because of the speaking rate of the source speaker. Similarly, the conversion between Joey (source) and J (target) improved the positive perceptions of the target speaker J over the scale, responsible (J = 4.1 Joey_J = 4.2). Nevertheless, the conversion has negatively affected the perception of the adjective, skillful (J = 3.8, Joey_J = 3.1). The remaining intra and inter-gender conversions have also contributed to negative perceptions of the converted voices.

6.5.3 Comparison between the AB preference test and the direct scaling test

This section summarises the comparison between the two subjective tests.

- **Negative perceptions:** Through the direct scaling test, we have discovered the negative perceptions of the converted speech which were not evident from the AB preference tests. The inter-gender VC experiment between the voices, E (as the source) and J (target) yielded negative perceptions (less warm) of the J when

compared to the original voice. Correspondingly, the perception of the adjective, skillful in H was negatively affected due to the VC experiment between C (source) and H (target). A similar observation has been made in intra-gender conversion between Joey (source) and J (target). The conversion resulted in an improved perception of the adjective, responsible, in the converted voice of J. However, this conversion has reduced the perception of the adjective, skillful, in the converted voice, J.

- **Orthogonal attributes:** Similar to the AB preference test, the 5-point direct scaling test has also displayed the opposite effect on the converted speech in the case of the adjectives, responsible and skillful (except for voice J in the AB preference test). Therefore the negative correlations between these two adjectives which were displayed from the AB preference tests were also corroborated by the subjective responses of the direct scaling test.

6.6 Discussion

In this chapter, we focus on altering the negatively perceived TTS voices into positive ones. For this, we have chosen VC and converted the voices of highly positive (warm/competent) voices into the voices of the speakers who are perceived to be less warm and less competent. Thus, we target only the *transfer of voices* and not the *transfer of characteristics*. Further, in the training of the Star-GAN model (other than the baseline model), we have included all the ground truth TTS voices and not just the highly positive and negative ones. We have hypothesized that the inclusion of voices that ranged from positive to negative on the scales of warmth and competence could render a unique representation of speaker characteristics that could be leveraged in the conversion process. However, the advantages or disadvantages of such inclusion were not studied extensively in the current work. Nevertheless, from the subjective responses, we can observe that there is definitely an improvement in the positive perceptions of converted speech when trained on many speakers (comparison between the baseline and intra-gender conversion, C_H).

6.7 Limitations

This section highlights the points that would restrict the reliability of the interpretations made from the studies.

- **VC Model:** As mentioned before (in chapter 1), the aim of this thesis is not to build new frameworks or propose new approaches for VC/TTS, but to investigate the possibility of achieving the positive perceptions of the generated speech using existing models. Thus, in the current study, we utilize a single VC model for all our experiments. Also, we rely on spectral conversion alone and do not implement any feature-specific modifications. Through this study, we only show that it is possible

to achieve positive perceptions of synthetic voices through VC techniques and further present the inclusion of various speaker attributes (kind, sympathetic, responsible, skillful) in the evaluation of VC voices.

6.8 Summary

In this chapter, we have investigated the transformation of negatively perceived synthetic voices into positive ones. A Star-GAN model was employed for both intra-gender and inter-gender conversions. The conversions were carried out between the ground truth warm/competent voices (as source) and ground truth cold/incompetent voices (as target). The converted speech samples were further evaluated using both AB preference tests as well as the direct scaling test. The evaluation of warmth and competence was performed using the ground truth adjectives (warmth = kind, sympathetic; competence = responsible, skillful). The evaluation results show that it is indeed possible to alter the negative perceived synthetic voices for their positive perceptions and vice versa. Some of the voice characteristics such as breathiness and high speaking rate seem to affect the perception of the generated speech. Breathiness in the male voice, Joey has negatively affected the speech quality of the VC voices that had Joey as the source speaker. Further, speaking rate has positively contributed to the perception of competence in synthetic speech.

Chapter 7
Modeling using TTS

This chapter presents the studies carried out on the modification of the synthesis procedure for positive perceptions of the generated speech (female speech). The studies were implemented using a conventional end-to-end TTS, Tacotron [4]. The suggestions on the subjective evaluation setup were provided by Sebastian Möller. The studies presented in this chapter (except the prediction of ground truth vocal cues of warmth and competence in synthetic voices) are similar to the work presented in [208]. Therefore, the content presented here is closely related to the experimental setup presented in the paper.

7.1 Introduction

In the previous chapter, we observed the manipulation of synthetic speech using VC techniques for the perception of SSC. In this chapter, we can examine the modification of a traditional synthesis procedure for the perception of SSC. The current end-to-end TTS mechanisms [4, 6] enable various alterations to the existing frameworks. Hence, modeling acoustic features (especially prosody) for expressive speech synthesis has been of high interest in the recent past [209, 210, 211, 23]. Prosody includes the extra-linguistic information (intonation, stress, style, etc.,) present in the speech signal. Modeling the prosody would therefore convey the person's moods, intentions, and mental states. [209] presents the prosody transfer in a traditional TTS framework by conditioning the model on the latent representations of the input acoustic features. Finally, they employ an AXY discrimination test for the subjective assessments of prosody transfer. A is the reference speech sample and X, and Y are two different prosody-modified speech signals. The participants could rate which among X or Y was perceived to be close to that of A on a 7-point scale. In [210] authors address the prosody transfer in case of an unseen speaker's speech. This was attained by combining the phonemic information with that of the corresponding prosody variations while training the end-to-end TTS framework. In this work, the

authors employ a Variation Autoencoder (VAE) for the latent representations of the prosody which resulted in a stable prosody transfer. [211] introduce the style modeling in a TTS framework using the unsupervised style tokens. In order to achieve expressive speech synthesis, the authors employ an additional module called the style attention network. This module captures and stores the weights necessary for the generation of various styles of speech signals. However, this model captures only the local variations in the F0 contour. Follow-up work was presented in [23] that models the speed along with the speaking style of synthetic speech. The model uses both the local and the Global style tokens through the use of a reference encoder. Additionally, the decoder was also conditioned using the style embedding layer which further improved the style control and transfer. [212] investigate the vector quantization to disentangle the prosody features such as speaking velocity, style, and pitch. The framework consists of Tacotron2 for text-to-spectrogram conversion, a WaveRNN [27] for speech signal reconstruction, and an auxiliary encoder that handles the Vector Quantization (VQ). The prosody transfer with the disentangled vectors prepared using the auxiliary encoder seems to have performed better than the model in [23]. Inspired by the line of the work on prosody modeling by conditioning the TTS framework on a variety of acoustic feature spaces, in the current work, we employ conditioning of a traditional TTS on vocal cues of warmth and competence.

7.2 Overview

This chapter addresses the following research question.

? Research question: ***"What modifications of synthesis procedure contribute to positive perceptions of generated voices?"***

The above research question is further divided into the following sub-tasks.

- ***Step 1:*** What are the acoustic features that need to be conditioned during the synthesis procedure?

  Throughout the literature, various research groups have investigated the modeling of prosodic features for expressive speech synthesis. Nevertheless, modeling warmth and competence in a TTS setup has not been done before. Therefore, the first step towards synthesizing these social aspects would be to identify the vocal cues of warmth and competence. In chapter 6, we find the ground truth adjectives derived for each of warmth (kind, sympathetic) and competence (responsible, skillful). An acoustic analysis similar to the studies presented in chapter 5 is car-

ried out using the subjective responses of these adjectives (obtained from chapter 4) in the current studies for the prediction of vocal cues of SSC. Thus, the acoustic features to be modeled for the perceptions of SSC is determined through this step.

- ***Step 2:*** How should one condition these acoustic correlates in an end-to-end TTS setup?
  This sub-question is two-fold: a) dataset, and b) modeling techniques. Most of the earlier works on expressive speech generation have included datasets with a variety of speaking styles. However, in the current work, our primary goal was to investigate the acoustic correlates of SSC without the effect of the content on the speech. Accordingly, the studies presented in the previous chapters were car-ried out on neutral speech (studies on ground truth TTS voices). Nevertheless, the ground truth voices (2 commercial TTS systems; Google and Amazon Polly) have also been trained on a wide variety of datasets (their internal datasets). Hence, we hypothesize that acoustic feature conditioning of an end-to-end TTS framework on a different dataset (LJspeech, audiobook dataset) with the acoustic correlates of SSC derived from neutral speech would still be relevant. Secondly, the condi-tioning of the TTS setup (tacotron) was enabled through quantized vocal cues of SSC.

- ***Step 3:*** How to evaluate the generated speech for the positive perceptions of synthetic voices?
  In chapter 4, we have observed the evaluation of TTS voices on a direct scaling test (semantic differential scales). In the current work, we follow a similar approach for the evaluation of the generated speech. The evaluation of warmth and com-petence was carried out using a 5-point scale with the adjective-antonym pairs at the extremes. The questionnaire provided during the evaluation was as follows:
  a) kindness, sympathetic (to measure warmth), b) responsible, and skillful (for the perception of competence) (ground truth adjectives obtained from chapter 6). From the evaluations conducted on VC experiments, we learned that the partici-pants found the sentences to be rather short (CMU arctic was presented in chapter 6) to provide any judgments such as kindness/sympathetic/responsible/skillful. This feedback from the listeners was also found to be in line with the studies in the literature (the utterances presented previously in various expressive speech generation research spanned around 20 seconds [23]). Further, we were inter-ested in understanding the generalisability of the vocal cues of SSC to different data types. Therefore, considering the above two points, the sentences provided during the evaluation phase consisted of Twitter sentences (long sentences and compassionate speech).

7.3 Experimental setup

7.3.1 Which acoustic features should be conditioned?: Ground truth vocal cues

This section discusses the vocal cues of warmth and competence derived from ground truth voices (Google and Amazon Polly). These acoustic correlates are further termed the ground truth vocal cues of SSC for synthetic speech. In chapter 5, the acoustic correlates of warmth and competence in the case of only the wide-range TTS were presented (acoustic feature prediction experiments). In the current work, a similar approach was employed to derive the vocal cues of ground truth TTS voices. The dataset chosen for training the TTS model is of a female speaker. Hence, for this study, the acoustic correlates of only female warmth and competence were derived using the backward elimination approach in linear regression. The input to the linear regression model was the 88-dimensional acoustic feature sequence obtained from the OpenSMILE toolkit. The output fed to the model was the subjective ratings of SSC as collected in chapter 4 for the ground truth TTS voices. The adjectives used for calculating the warmth and competence ratings are obtained from ground truth adjectives defined previously in chapter 5 (warmth = kind, sympathetic; competence = responsible, skillful). The ground truth vocal cues derived for SSC in female TTS voices are presented in Table 7.1. From the acoustic analysis, we can observe that the acoustic features contributing to female warmth are spectral flux, F1 mean and F2 mean. Correspondingly, the vocal cues of competence in female TTS voices are slope and spectral flux. These results seem to be in line with the studies presented in [202]. This phenomenon is due to the choice of adjectives used in the study [202]. The adjectives friendliness and likable were highly correlated with the ground truth adjectives, kindness (friendly $r=.86$, $p<0.09$, likable $r=.91$, $p<0.01$) and sympathetic (friendly $r=.907$, $p<0.04$, likable $r=.908$, $p<0.05$). Correspondingly, the correlation coefficient for the adjectives skillful and responsible was $r=.89$, $p<0.01$. We can observe that the formant frequencies (F1 mean, F2 mean), spectral features (flux, slope), and as seen previously (in chapter 5) durations (speaking rate and speech pauses) displayed dependence on the perception of SSC. Therefore, the modeling of these features in the case of expressive speech (LJSpeech) than in a neutral speech (neutral speech) would be more challenging and interesting at the same time. Hence, the choice of data (LJspeech and not neutral speech) seems to be consistent with the desired line of work. Further, the conditioning of these vocal cues in a TTS setup is performed using Quantisation technique.

Table 7.1: The ground truth vocal cues of warmth and competence derived for the ground truth TTS voices. The acoustic correlates of warmth are derived using the subjective ratings of the adjectives kind and sympathetic. Similarly, the vocal cues of competence are derived using the ground truth adjectives, responsible and skillful.

Female	
Warmth	Competence
Spectral flux	Slope
F1 mean	Spectral Flux
F2 mean	-

7.3.2 Data preparation for feature conditioning (Quantisation of acoustic features)

Quantisation is a well-known form of compression technique used in data transmission. It involves dividing a larger set of vectors into multiple smaller groups. Quantisation techniques have been widely used in TTS and VC for learning the latent representations through various neural models for more convincing quality, and expressivity in the generated speech. In our experiments, we chose to quantize the acoustic features into 3 different clusters which we refer to as, classes (class 0 = cold/incompetent, class 1 = Neutral, class 2 = warm/competent). The assumption is that this would enable the label-based training of SSC in an end-to-end TTS setup. Figure 7.1 displays the flowchart of the experimental setup.

By now, the vocal cues to be controlled during the speech generation mechanism are known. The goal is to modify the acoustic correlates of SSC in the synthesis procedure for positive perceptions of generated speech. For this, each of the spectral flux, F1 mean, F2 mean and slope derived for the LJSpeech were quantized into three different classes. This segregation was performed based on their respective acoustic feature distributions. The acoustic feature values for all the speech samples in the LJspeech were calculated using OpenSMILE. Further details of the class distribution for each of the vocal cues representative of SSC in synthetic speech are provided below.

7.3.2.1 Experiment 1: Spectral flux

The spectral flux provides information on the variations in the spectrum within the speech signal (difference between frames of a speech signal) and also between different speech signals (spectral differences between two separate speech signals). The spectral flux values derived from the OpenSMILE for LJSpeech ranged between 0.157-0.706. Feature quantization on the spectral flux values resulted in three classes. The class division and the number of examples obtained per class are provided in table 7.2.

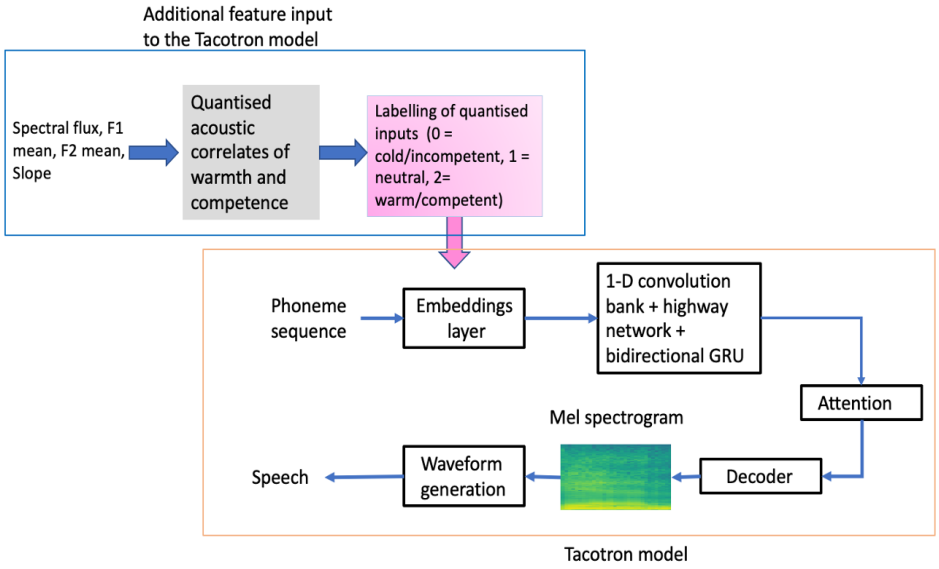

Fig. 7.1: A flowchart with the workflow of current TTS experiments.

Table 7.2: Feature quantization for spectral flux values of LJspeech

Class	Range of spectral flux	Number of speech samples
0 (less warm/cold)	0.15 - 0.30	3193
1 (Neutral)	0.31-0.44	5193
2 (warm)	0.45 and above	4714

7.3.2.2 Experiment 2: F1 mean

Formants provide the amount of acoustic energy contained in the frequency component. In humans, the first formant, F1, is related to the height of the tongue while generating speech. If the height of the tongue is elevated during the generation of speech, then the content has a lower F1 and vice versa. From our acoustic analysis, we have observed that the F1 mean contributes to female warmth in synthetic speech. Therefore, the F1 mean values obtained using the OpenSMILE toolkit are quantized in the current experiment for our studies on TTS framework. The range of F1 mean values calculated over the LJSpeech database and the number of speech samples used for training per class are provided in table 7.3.

Table 7.3: Feature quantization for F1 mean values of LJspeech

Class	Range of F1 mean	Number of speech samples
0 (less warm/cold)	400-515	4701
1 (Neutral)	516-540	3598
2 (warm)	541 and above	4801

7.3.2.3 Experiment 3: F2 mean

The formant F2 is directly proportional to the movement (forward) of the tongue. F2 mean has also been considered one of the contributing factors of warmth in female synthetic speech. The range of the F2 mean derived for LJspeech and the class distribution is provided in table 7.4.

Table 7.4: Feature quantization for F2 mean values of LJspeech

Class	Range of F2 mean	Number of speech samples
0 (less warm/cold)	1280 - 1550	3410
1 (Neutral)	1551-1600	4431
2 (warm)	1601 and above	5259

7.3.2.4 Experiment 4: linear combination of spectral flux + F1 mean + F2 mean

Apart from investigating the effect of individual features on the positive perceptions of the generated speech, we were inquisitive about knowing the combined effect of these features. However, how to combine these features was obscure. Therefore, as an initial attempt, we chose the linear combinations of the individual acoustic correlates of warmth in this experiment. A linear combination is a linear equation combining all the relevant vocal cues, with the individual coefficient values summed to 1. The linear combinations of various ML models have been leveraged in different fields previously for effective feature modeling [213, 214]. In [213], authors explore the effectiveness of an ensemble of ranking algorithms in recommender systems. Their research posits better recommendations by the proposed ensemble model over the traditional stochastic optimization techniques. [214] presents the study on the automatic selection of an ML model from a linear combination of models over manually examining individual models for function approximation. The authors propose a better choice of models through the use of this automatic selection that would best address the research problem. Similarly, in our studies, we leverage the linear combination of the acoustic correlates of warmth for the positive perceptions of the generated speech. The class distribution over this linear combination of features is provided in the table 7.5. The expression employed for the combination of the

features is also displayed in equation 7.1.

$$Linear\_comb\_warmth = 0.33 * Spectral\ flux + 0.33 * F1\ mean + 0.33 * F2\ mean \tag{7.1}$$

Table 7.5: Feature quantization for the linear combination of flux, F1 mean, and F2 mean

Class	Combination of spectral flux+F1 mean+ F2 mean	Number of speech samples
0 (less warm/cold)	569 - 690	5073
1 (Neutral)	691-715	4458
2 (warm)	715 and above	3569

7.3.2.5 Experiment 5: Slope

In this experiment, we quantize one of the acoustic correlates of competence as found from ground truth TTS voices. The spectral slope provides information on the voice quality of a speaker in the speech signal. The voice quality includes information such as husky voice, creaky voice, etc., The class distribution over the quantized slope values of LJspeech is provided in the table 7.6.

Table 7.6: Feature quantization for slope values of LJspeech

Class	Slope	Number of speech samples
0 (less warm/incompetent)	0.07 - 0.11	3193
1 (Neutral)	0.112-0.116	5193
2 (warm/competent)	0.1161 and above	4714

7.3.2.6 Experiment 6: Linear combination of spectral flux + slope

This experiment provides the linear combination of the ground truth vocal cues of competence in synthetic speech. From table 7.1, we find the features contributing to competence in female TTS voices are spectral flux and slope. In experiment 1, we have already computed the spectral flux values of LJspeech. Therefore, in this experiment, we combine the quantized spectral flux obtained from experiment 1 and slope values obtained from experiment 4. The class distribution over the linear combination of features is provided in the table 7.7. The equation employed for the

linear combination is provided in equation 7.2.

$$Linear\_comb\_competence = 0.5 * Spectral\ flux + 0.5 * Slope. \quad (7.2)$$

Table 7.7: Feature quantization for linear combination of spectral flux and slope

Class	Combination of spectral flux and slope	Number of speech samples
0 (incompetent)	0.13 - 0.22	4882
1 (Neutral)	0.23-0.26	4365
2 (competent)	0.27 and above	3853

7.3.3 Model details

A traditional Tacotron model would take the character inputs and provide the raw spectrograms [4]. Nevertheless, in the current framework, instead of character sequence, the model is fed with the phoneme sequences extracted using the EHMM labeling available in the festvox [131]. Two losses were calculated from the acoustic modeling namely, L1 divergence loss between a) predicted and original mels, and b) predicted and original linear spectrograms. The loss values of the padded frames were not masked which aided in the prediction of the sentence endings. The tacotron model was trained on the LJSpeech dataset. The acoustic feature extraction consisted of the speech frames spanning 50ms with a frameshift of 12.5 msec using hamming windows. The features obtained were a 1025-dimensional linear spectrogram and an 80-dimensional mel spectrogram. The speech signal reconstruction from mel spectrograms employed the Wavenet vocoder [5]. Similar architecture to the one provided in the [5] was employed in the current studies. The speech samples were power normalized to a range between -1 and 1. The individual speech samples were encoded using the 16-bit μ law quantization. The acoustic frames were upsampled using linear interpolation instead of transparent convolutions. This upsampling has enabled the correspondence between the acoustic frames and the time resolution of the speech samples. Training of the models took around 10-16 hours (Tacotron = 10 hours, Wavenet = 16 hours) on an NVIDIA 1080 Titan GPU 12GB.

7.3.3.1 TTS experiments for warmth and competence

The study of warmth consisted of four experiments, a) spectral flux, b) F1 mean, c) F2 mean, d) linear combination of flux, F1 mean and F2 mean. In these experiments, the model training is performed by the conditioning of the TTS framework on each of

the mentioned acoustic features, derived from the LJspeech. All the acoustic features were conditioned independently except for the linear combination experiments. The conditioning of these acoustic features is enabled by an additional embedding layer introduced at the input. Further, the phoneme sequences are concatenated with the acoustic feature information present in this additional embedding layer. Those additional embedding layers are created using the quantized acoustic features. Similarly, for the perception of competence, the acoustic feature conditioning was implemented using the slope and spectral flux values of the LJSpeech database. Further, a linear combination of spectral flux and slope was also investigated using the same model configuration. A baseline model was developed without providing any class information (no additional embedding layer) in the input.

7.4 Subjective evaluations

The subjective tests on the generated speech consisted of two stages: a) evaluation of speech quality and naturalness from the baseline model, b) perception of SSC from baseline, feature conditioned studies (4 experiments on warmth, and 3 experiments for competence). 25 (male = 13, female = 12) university students were recruited for the subjective tests. Their ages ranged between 24-43 (avg = 35.4, std = 1.32).

The evaluation of speech quality and naturalness of the generated speech was carried out on a 5-point Likert scale (1=poor quality/not at all natural, 5 = very good quality/natural). In this test, the participants rated 10 speech samples synthesized from the baseline model. The duration of the speech samples ranged from 7 to 15 seconds (approximately avg =12sec). All the speech samples presented in the test were randomized for each participant. Further, the listeners could play the speech samples multiple times during the test. The subjective evaluations for the perception of SSC in the generated speech are performed using 5-point scales with bipolar adjectives at the extremes. The number of questions provided for each of warmth and competence is as follows, 10 (speech samples) * 2 adjectives (kind, sympathetic for warmth) * 4 (number of experiments carried out for warmth = 3 + baseline) = 80 questions for warmth; 10 (speech samples) * 2 adjectives (responsible, skillful for competence) * 3 (number of experiments carried out for competence =2 + baseline) = 60 questions. The perception of warmth was obtained from the averaged subjective ratings collected for the scales, kind and sympathetic. Similarly, the competence ratings are the averaged subjective scores of the scales, responsible and skillful. The generated speech samples are available at [1].

[1] https://saisirishar.github.io/TTSforSSCspeechsamples/

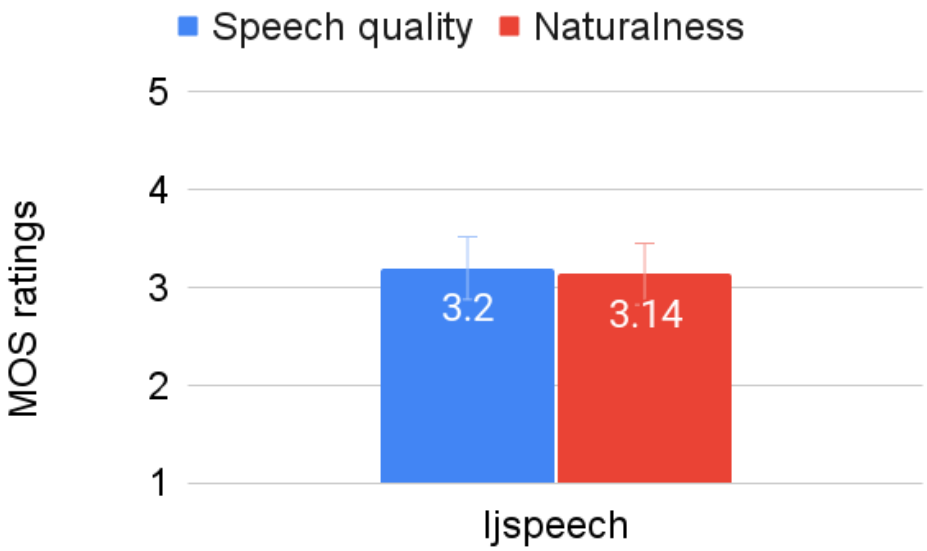

Fig. 7.2: Results of the subjective responses collected for the speech quality and the naturalness of the generated speech (for the baseline model).

7.5 Observations

Figure 7.2 presents the Mean Opinion Scores calculated for the subjective responses collected corresponding to the speech quality and the naturalness of the generated speech. The plots also display the 95% confidence intervals for each of the metrics evaluated on the baseline. The subjective responses displayed acceptable speech generation quality, and naturalness (along with good intelligibility). Therefore, later on, feature conditioning experiments were carried out using the current Tacotron model.

Figure 7.3 displays the results of the subjective evaluation carried out for each of the experiments on warmth. Since, the warmth ratings were computed by averaging the subjective responses of the adjectives, kind and sympathetic, the correlation between these two adjectives as observed from the subjective responses is calculated. The correlation between kind and sympathetic ratings was r =.93 ($p<0.01$). Correspondingly, the correlation between the scales, responsible and skillful was found to be, r=.85 ($p<0.06$). As opposed to the observations made in the VC experiments provided in the previous chapter, in the current experimental setup, the ratings of the adjectives skillful and responsible were correlated with each other (this could be because of the nature of the voice used in the current studies). From the subjective responses (MOS for warmth shown in 7.3) it is evident that when

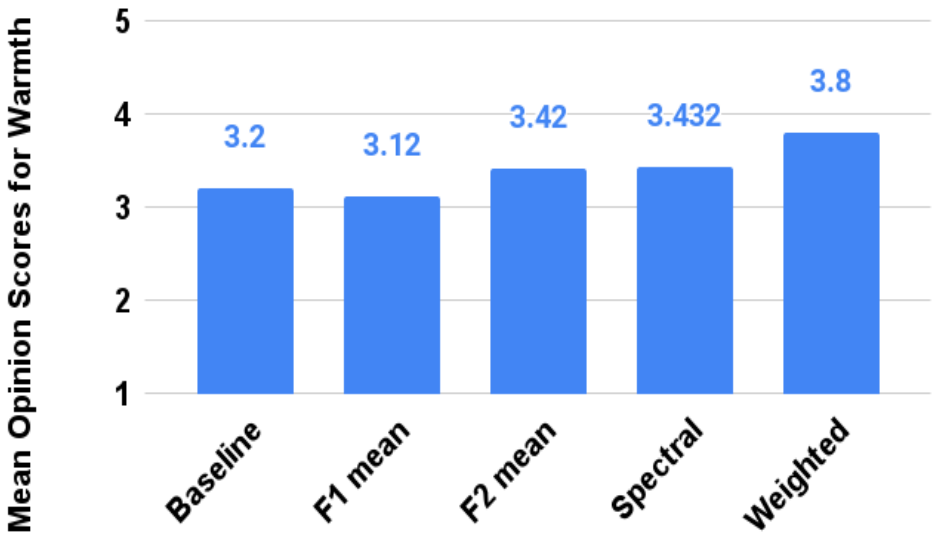

Fig. 7.3: Results of the subjective evaluation carried out for the perception of warmth in the generated speech (speech samples with class label 2 = warm).

compared to conditioning the model on individual acoustic features, conditioning it on a linear combination of these acoustic features provided higher perceptions of warmth from the generated speech. A similar observation was made even in the case of competence ratings (shown in figure 7.4). We have also observed that even though the TTS setup utilized in the current studies did not include any explicit duration modeling, the generated speech displayed speech pauses which have also contributed to the perception of warmth and competence (we drew this conclusion based on the sentence-wise subjective analysis). An example sentence generated and the observations made from each of the experiments are detailed below.

Following is the sentence provided during the subjective evaluation,

Suggestions for improvements means a person believes in your core idea and thinks their comments will help your work.

In the following sentences, the highlighted parts represent the stressed (emphasized) words in the generated speech (as perceived and indicated by the listeners). The speech pauses inserted in the generated speech in each of the experiments and the corresponding effect on speech perception are also discussed. For this analysis, the participants were provided with additional questions during the evaluation. They

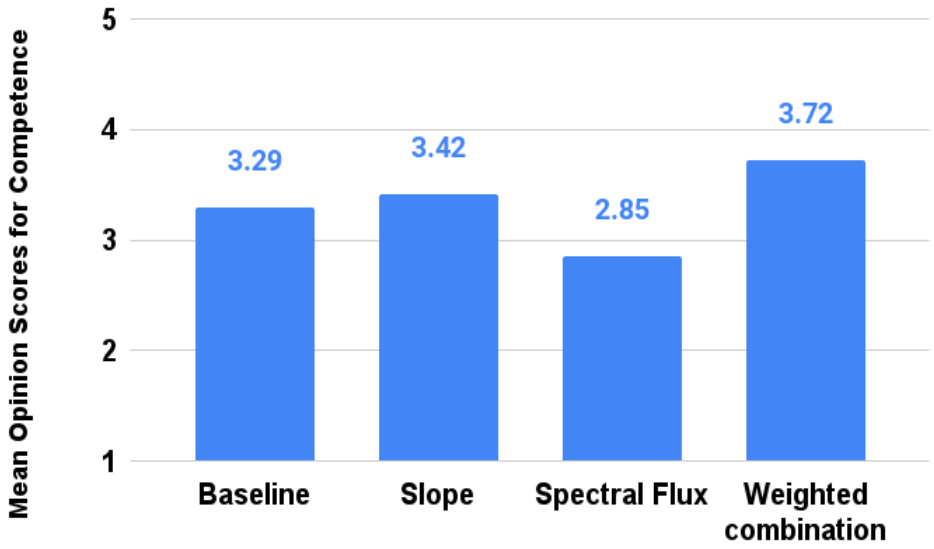

Fig. 7.4: Subjective responses collected for the perception of competence in the generated speech (speech samples with class label 2 = competent.

were provided with text blocks to indicate the words that they found were emphasized. Similarly, they were asked to provide pairs of words that consisted of a speech pause in between.

Baseline: *Suggestions for improvements* ***means*** *a person believes in your core* ***idea [pause]*** *and thinks their comments will help your work.*

The generated speech sample consisted of pitch variations throughout the sentence and the word, "means" was emphasized in the first part of the sentence. Further, a speech pause was inserted between the words, "idea", and "and". Also, before the insertion of the speech pause, the word, "idea" was emphasized along with a pitch variation.

F1 mean: *Suggestions for improvements means a* ***person*** *believes in your core* ***idea [pause]*** *and thinks their comments will help your work.*

The generated speech sample when the model was conditioned on F1 mean had a neutral voice in the first part of the sentence. Similar to the baseline, an emphasis and speech pause was inserted at the word, "idea". However, due to the neutral voice, the perception of warmth was less when compared to that of the baseline model.

Feedback collected from the participants at the end of the test suggested that the samples generated with this experiment were intimidating/dominating. This has also resulted in a lower perception of warmth from the samples generated when the model was conditioned on F1 mean.

F2 mean: Suggestions *for improvements means a* ***person*** *believes* ***in*** *your core* ***idea [pause]*** *and thinks their comments will help your* ***work****.*

In the above example, even though there were no speech pauses inserted in the first part of the sentence, due to the pitch variations throughout, the sentences generated in this experiment were considered warm compared to the previous experiment. Additionally, in the above example, we could perceive the emphasis on the words, “suggestions”, “person”, “in”, “idea”, and “work”.

Spectral Flux: *Suggestions for improvements* ***means [pause]*** *a person believes* ***[small pause]*** *in your core idea* ***[pause]*** *and thinks their* ***comments will help your work.***

In this experiment, we could perceive multiple speech pauses in the generated speech. This has further contributed to positive perceptions of the voice as the listeners felt the person was “patient”. However, on the other hand, there were no pitch variations (the voice was close to neutral speech). Therefore, the perception of warmth was only slightly higher than in the previous experiments (speech pauses contributed to warmth but neutral speech limited the positive perceptions). On the other hand, the pauses have negatively influenced the perception of competence from the generated speech.

linear combination for warmth: Suggestions *for improvements means a person believes in your core idea* ***[pause]*** *and* ***thinks*** *their* ***comments*** *will* ***help*** *your* ***work****.*

This experiment combines all the properties of the previous experiments, as it is a linear combination of F1 mean, F2 mean, and spectral flux. Therefore, the generated speech had pitch variations (as in F2 mean), there was only one speech pause (similar to the F1 mean, F2 mean), and emphasis on the words, “comments”, “help”, and “work” (as in spectral flux experiment) and the speech pause between ”idea”, and ”and” as in all the previous experiments. A combination of all these properties has definitely contributed to higher perceptions of warmth in the generated speech.

Slope: *Suggestions for improvements means a person believes in your core* ***idea [pause]*** *and* ***thinks*** *their comments will* ***help*** *your work.*

The generated speech samples did not have multiple speech pauses (as seen in spectral flux experiments). Further emphasis was on the words, “idea”, “thinks” and “help”. Also, there were pitch variations throughout the sentence. Moreover, the

perception of competence has improved compared to the baseline model.

linear combination for competence: *Suggestions for improvements means* ***[pause]*** *a person believes in your core idea* ***[pause]*** *and thinks their* ***comments*** *will help your* ***work.***

The combination of spectral flux and the slope has resulted in combined properties of conditioning the TTS model on these individual features. The speech pauses were found next to the words, "means" (as in spectral flux) and "idea" (as in all previous experiments). Further, there were pitch variations in the generated sample (as seen in the experiments on the slope). Overall, reduced number of speech pauses (improvement over spectral flux experiment) and improved pitch variations (from the experiments on slope) have led to increased perception of competence from the generated speech samples.

Overall, from the subjective analysis, we can declare that the positive perceptions of the generated speech through the conditioning of the TTS model on the ground truth vocal cues of SSC are possible.

7.6 Outlook on the dataset and the adjectives used in the study

The wide-range TTS systems (discussed in chapter 4) were trained on different datasets. For example, the academic systems were trained on CMU arctic (neutral speech), while the commercial systems such as Google's voices were trained on internal datasets and the evaluations were carried out on the WC dataset. The WC dataset was designed to display the characteristics such as care, compassion, and assurance (described in chapter 3). Also, the 2 commercial systems (Google and Amazon Polly) were trained on the internal dataset and evaluated on neutral speech (Harvard sentences). Therefore, in the current study, the use of LJspeech (audiobook data, prosodically rich sentences) seems in line with the previous studies. Further, we can observe that the adjectives, pleasant, likable, and sympathetic were common among the adjectives contributing to female warmth from both the studies (wide-range TTS and 2 commercial TTS). However, the adjectives contributing to competence were different. From the ground truth adjectives derived in chapter 6, we can find the adjective, sympathetic to be commonly found in both the studies irrespective of the TTS systems, and datasets used. Subsequently, in the current study, we also present the correlation values between all the ground truth adjectives. From the correlation values (presented in the section 7.3 (Experimental setup) before conducting the TTS experiments, and also in the section 7.5 (Observations), evaluations of the generated speech) we can presume that all the adjectives used in the study are correlated with each other. Thus, from a) the correlation values between the adjectives, and b) the adjectives commonly found in both datasets, we can state that the adjectives used in the current study are relevant to both datasets (neutral speech and compassionate speech).

7.7 Limitations

Limitations of the study are as follows, a) the acoustic correlates of SSC derived for neutral speech were used in this study, b) experiments on a single-speaker database, c) only one experiment on the TTS framework.

- **Acoustic correlates of neutral speech**: The acoustic correlates of SSC used in the current studies are derived using the subjective ratings of the adjectives, kind, sympathetic, responsible, and skillful collected from the studies presented in [167]. The experimental setup in [167] consists of the speech segments prepared by combining 8 speech samples into one file (each speech segment spanning approx. 20 seconds of speech). This implies that the subjective ratings were collected for the averaged information present in the speech segments (as the ratings were not collected for individual speech samples). Thus, even though the current studies provide positive perceptions of the generated speech, the solidarity of the derived acoustic correlates might be little.
- **Single-speaker database**: Another limitation of the study is that we utilize only one female speaker in all the experiments presented. The choice of the dataset for the current studies was highly inspired by the previous works on style generation using a TTS framework. Thus, we opted prosodically rich dataset (audiobook dataset = LJspeech) over a neutral speech (neutral speech). We propose to address the generalisability of acoustic features (vocal cues of SSC derived from neutral speech) to data types other than neutral speech but due to the use of a single-speaker dataset, the results cannot be generalized with the current setup.
- **Only one TTS experiment**: Even though there has been a lot of research on expressive speech generation in the TTS community, the features described in the current studies have not been explicitly modeled before (to the best of our knowledge). Our aim was to investigate the modifications of the derived acoustic correlates of SSC in a TTS setup. Correspondingly, this was our initial attempt toward incorporating positive perceptions of synthetic speech. Thus, we do not present any comparisons of our study with other works on expressive speech generation. In our future works, we intend to design an experimental setup considering all the limitations and further compare the results with the other state-of-the-art models.

7.8 Summary and future works

In this chapter, we address the research question that focuses on the modifications of the synthesis procedure for positive perceptions of synthetic speech. A traditional TTS framework has been employed for feature conditioning. This feature conditioning was enabled by quantized vocal cues of warmth and competence derived from the female ground truth TTS voices. The acoustic features found to be con-tributing to warmth and competence are spectral flux, F1 mean, F2 mean; slope,

and spectral flux respectively. The subjective evaluations of the generated speech show that the linear combinations of the relevant acoustic features contribute to the positive perceptions of the synthetic voices over the individual feature-based conditioning of the TTS. Further, the acoustic feature and spectral flux have been found to contribute to negative perceptions of synthetic speech on the intellectual dimension (competence). Now that, we have observed that the combinations of relevant vocal cues contribute to positive perceptions of generated speech, in our future works, we intend to investigate different combinations of these acoustic features for socially acceptable synthetic voices. Also, from the studies on the wide range of TTS systems (acoustic analysis on wide-range TTS systems provided in chapter 5), we can observe that F0 contributes to female warmth and competence. However, in the current study, we rely only on the ground truth vocal cues for the perceptions of SSC. Hence, we are interested in investigating the modeling of F0 and other vocal cues derived from previous studies (chapter 5) in female speech. A comparison of social perceptions of the generated speech when the TTS model is conditioned on vocal cues of SSC derived from ground truth TTS voices and the ones derived from the wide range of TTS voices would be one interesting follow-up work. Correspondingly, a similar analysis of male TTS voices could also provide interesting findings and future research directions.

Chapter 8
Summary, challenges and future work

8.1 Summary of contributions

The goal of this thesis is two-fold, a) understanding the social perceptions of synthetic voices, and b) manifesting the positive perceptions of synthetic voices through the acquired knowledge. We achieved these tasks through the following steps,

- **Step 1:** Interpreting the perception of warmth and competence (SSC) in different synthetic voices.
- **Step 2:** Prediction of the vocal cues contributing to multiple speaker attributes or SSC.
- **Step 3:** Manipulations over the synthetic voices using VC (spectral conversions) for positive perceptions of negatively perceived voices.
- **Step 4:** Feature conditioning of an end-to-end TTS framework through quantized vocal cues of SSC.

Steps 1 and 2 correspond to the first part of the goal (understanding the social perceptions of synthetic speech), and the latter two address the last part of the goal (manifesting positive perceptions). I dedicate the chapters, 4, 5, 6, 7 respectively for each of the steps mentioned above. Chapter 1 provides the motivation for choosing the social speaker characteristics, warmth, and competence in the current studies. Our target domains are health care and customer service. Even though there are many more characteristics that the agents in these domains should possess, in the scope of this thesis, we address only warmth and competence.

8.1.1 Addressing the objectives and research questions

Chapter 1 presents the objectives and the research questions we intend to address through this thesis. This section provides a summary of the work presented corresponding to each of the objectives and research questions along with the takeaways from each of them.

Objective 1: Postulate the significance of investigating the social perceptions of synthetic speech.

The conversational agents designed for different application domains still lack humanness in many ways (such as social speaker characteristics). The development of personal assistants/conversational agents that can outperform humans in this aspect would require the study of their current social perceptions.

The above objective is addressed in chapters 1 and 2. With the development of computing abilities, human-like natural-sounding speech has been achieved. However, these synthetic voices still lack humanness in many ways ("how" it is being said). Hence, a lot of work has been performed focusing on the expressivity of the generated speech in the recent past. Through this thesis, we propose to also consider the perceptions of different speaker attributes/characteristics from the synthesized speech via subjective evaluations. Chapter 2 provides the literature which shows that the interpretation of various speaker characteristics from speech (human and synthetic) is possible. Further, we employ similar techniques to understand warmth and competence from the synthetic speech in this thesis.

Research question 1: What social speaker characteristics do people perceive in synthetic speech?

Similar to the first impressions made in human-to-human interactions, human-machine interactions can also render interesting observations. The ***universal dimensions of social perception*** which were prevalent from behavioral studies have also been identified from the machine-generated voices. In addition to the social speaker characteristics, the studies show that personality traits can also be perceived from machine-generated voices.

Chapter 4 deals with the social perceptions of synthetic voices. Here we find two studies, perceptual analysis of a) a wide range of TTS systems, and b) two commercial TTS systems. The initial experiment consists of 36 synthetic voices evaluated on a 34-item semantic differential scaling test on two utterances. While this study includes a variety of synthetic voices, they have been evaluated on only

two utterances each spanning less than 5 seconds. Also, the speech quality and the naturalness of the generated speech influenced the social perceptions of the voices. Hence, a subsequent study was designed to overcome the shortcomings of this study. Two commercial TTS systems (Goole and Amazon Polly) were chosen for the study. The speech segments used in the subjective test were prepared by combining around 8 individual speech samples (4 speech segments= 4*8 =32 sentences). Further, we define these voices to be the ground truth (reference) TTS voices and use them throughout this thesis. The subjective analysis of both studies has provided us with the factors, warmth, competence (social speaker characteristics), and additionally the personality trait, extraversion.

Research question 2: Which acoustic features of synthetic speech affect the subjective perceptions of social speaker characteristics?

The acoustic correlates of female warmth are spectral flux, F1 mean, and F2 mean. While the male warmth is dependent on the vocal cues, F1 mean, loudness, and unvoiced slope (in the range of 500-1500). Correspondingly, female competence is perceived through the acoustic features, voiced slope, spectral flux, and mfccs. Similarly, male competence is influenced by F0 and voiced segment lengths.

Chapter 5 presents the studies on deriving the acoustic correlates of SSC from a wide variety of synthetic voices. The acoustic feature extraction was carried out using the eGeMAPS configuration available in the OpenSMILE toolkit. The feature extraction was inspired by the previous studies on examining various speaker characteristics and paralinguistic information from speech using this toolkit. The acoustic correlates of female warmth were F0 falling slope, F2 (standard deviation), Hammerberg Index, loudness, unvoiced segment length, spectral flux, F1 mean and F2 mean (overall observations from studies on both wide-range as well as ground truth TTS voices). The features, F0 standard deviation, mfcc4 mean, F1 amplitude, F3 mean, spectral flux, and slope were found to be responsible for female competence. On the other hand, loudness mean, mfcc3 mean, HNR, F3 bandwidth, spectral flux, F1 mean, slope, and F1 mean seem to affect the male warmth in synthetic speech. Correspondingly, the contributors to male competence as observed from the studies are, loudness, mfcc4 mean, F1 mean, Hammerberg Index, slope, and spectral flux. We can also observe a similar behavior even in the case of natural speech from studies presented in the literature. Further, the prediction of the SSC from the derived acoustic correlates was performed using both classification and regression techniques. The results of the automatic prediction support the relevance of the derived vocal cues in the perception of SSC.

Objective 2: Transform the negatively perceived synthetic voices to positive ones.

Voice Conversion was devised for the transfer of negatively perceived voices into positive ones. The transformation was enabled through the use of the ground truth TTS voices.

Chapter 6 presents the VC experiments carried out on the TTS voices for the transformation of negatively perceived voices into positive ones. The converted speech was evaluated using different subjective evaluation metrics, speech quality, naturalness, speaker similarity, warmth, and competence. The metrics, speech quality, and naturalness are evaluated on 5-point Likert scales. Due to the breathy voice, the conversion that involved Joey had lower quality compared to others. The speaker similarity was evaluated using the ABX preference test between the baseline (C_H, model trained on 2 speakers, C (positive female), H (negative female)) and the C_H (model trained on many speakers). The similarity scores were observed to be higher in the case of the multi-speaker model than in the baseline model. Later on, the perceptions of SSC from the converted speech were evaluated using two tests, a) AB preference test, and b) direct scaling test (5-point scales with the adjective-antonym pairs). The scales used for the evaluation of warmth are kind and sympathetic; adjectives for competence are, responsible and skillful. The subjective responses display that speaking rate has positively affected the perceptions of the adjective, skillful. The responses for the adjectives, kind, sympathetic, and responsible are correlated with each other. On the other hand, the adjective skillful was found to be orthogonal (this observation might be because of the speaking rate) to the perception of the adjectives, kind, sympathetic, and responsible. Additionally, the direct scaling test has shown negative perceptions (conversion between E (positive female) and J (negative male) had negative perceptions of warmth; conversion between C (positive female) and H (negative female) negatively affected the perception of skillfulness in H (negative female); similarly reduced perception of skillfulness was also seen in the conversion between Joey (positive male) and J (negative male)) of the converted speech which was not obvious from the analysis of the AB preference test.

Research question 3: Which alterations of the synthesis procedure lead to positive perceptions of speakers?

The acoustic correlates of SSC were quantized and are embedded along with the input phoneme sequences in an end-to-end TTS framework. This feature quantization provided the labels corresponding to the positive, negative, and neutral perceptions of the synthetic voices as derived from the ground truth TTS voices. The generation of positively perceived synthetic voices was therefore controlled using the label information present in the additional embedding layer in the input.

Chapter 7 details the experiments carried out on a traditional end-to-end TTS framework for the positive perceptions of synthetic voices. The research question is divided into three parts namely, a) figuring out the acoustic correlates of the SSC from ground truth TTS voices, b) the dataset and the modeling technique to be employed for the task, and c) evaluation of the generated speech. Even though previously (in chapter 5) we have derived the vocal cues of synthetic speech, the study was conducted on a wide range of TTS voices. Also, due to the limitations of the study (a) only 2 utterances were used in the study, b) the acoustic feature extraction was done from a pool of TTS voices with a varied speech quality), in the current work, we intend to investigate the acoustic correlates of ground truth TTS voices and conditioning the TTS model on the derived acoustic features. The derived acoustic correlates were spectral flux, F1 mean, F2 mean (for warmth); slope, and spectral flux (for competence). The experiments were carried out on a single-speaker database (female, LJspeech). In the current work, we condition the model only on the derived ground truth acoustic correlates of SSC. The conditioning of the TTS model was carried out using quantized acoustic features. Along with the individual features, a linear combination of these features was also examined for the positive perceptions of the synthetic voices. The evaluation results show that the convex combination of the acoustic correlates contributes to positive perceptions of the generated voices in the case of both warmth and competence. Additionally, we found that the spectral flux has negatively affected the perception of competence in the generated speech.

8.2 Challenges

In this section, I discuss the challenges we have encountered in the course of this work followed by possible future works.

Throughout this thesis, we conduct a wide range of subjective tests for the analysis of SSC from synthetic speech. The challenges in conducting such evaluations were three-fold, a) how many questions (adjectives, speech samples, number of TTS systems, number of male and female voices) should one include in the evaluation?, b) which datasets should we use? c) collecting a variety of subjective responses (native speakers, speech and NLP experts).

- **Questions**: We were interested in understanding the social perceptions of a variety of synthetic systems and voices. However, with the increase in the number of TTS systems, the number of voices (male, female) has also increased. Further, evaluating the perceptions of warmth and competence was found difficult without including the additional adjectives. As a result, the number of questions to be included in the test setup has increased enormously. One approach to address this would be to divide the task into multiple sub-tasks (division per gender or TTS system type) and conduct multiple subjective tests. But, the aim of this thesis was also to investigate methods for incorporation of SSC in the generated speech (apart from understanding social perceptions of synthetic voices through

subjective tests). Therefore, the speech samples were combined and evaluated for the SSC. Further, the acoustic feature prediction and modeling of the synthetic speech were performed on this averaged information (we assume that the acoustic and the subjective information would have been averaged because of combining the speech samples; this part is also discussed in the limitations of chapter 7).

- **Dataset**: Earlier works on understanding charisma from speech [83], warmth, and competence judgments from the interactions with virtual agents [53], speaker characterization from speech-alone scenarios [76, 75], personality judgments from speech [86], consisted of rather long speech utterances (or long conversations). However, in our work, we focus on understanding acoustic correlates of SSC from speech alone (perceptions without being affected by the content). Hence, we chose neutral speech for our studies. Nevertheless, the commercial TTS systems (Google's Wavenet, Amazon Polly) used in the studies were previously trained on internal databases and the details of these databases (data type, length of the speech utterances) are not known. As the variability of data was evident, later on, we chose to stick to the nativity of the speakers (US natives) for our studies. Since we were keen on modeling the acoustic correlates of SSC in a TTS setup for the positive perceptions of the generated speech, the TTS experiments in chapter 7 were performed on LJspeech (approx. 24 hours of data) and were evaluated on compassionate speech following the works on expressive speech synthesis [23].
- **Listeners**: Finding the participants (native speakers, speech experts) for the subjective tests was another challenge. Since, the beginning of the COVID-19 pandemic, the use of crowd-sourcing studies has received much interest. This has facilitated the availability of native US speakers through AMT. However, obtaining speech experts (signal processing experts, TTS experts, NLP researchers) would have provided a different perspective on the studies (we carried out some informal listening tests with the experts). Therefore, through this thesis, we would propose to include different speaker attributes in the evaluation of TTS and VC systems on the platforms like Blizzard Challenge and VC challenges. This would provide the subjective responses (and additional insights on how to include such attributes in subjective evaluations) of researchers from all over the world. I hypothesize that the inclusion of area expertise (speech or NLP experts) and language expertise (we already included the native speakers in the current studies) would provide a much more interesting analysis of the studies.

8.3 Future Work

This section provides some suggestions for future works while highlighting the limitations of this thesis. The most important limitation of this work is the use of subjective evaluation setups. This can be discussed in two different parts, a) questionnaire, b) evaluation scales

- **Questionnaire:** Our goal through this thesis was to investigate the perceptions of SSC from synthetic speech. Through the initial subjective tests, we realized that the evaluation of these characteristics would require us to provide multiple adjectives. Thus, our evaluation setup was extensive with the inclusion of multiple adjectives and synthetic voices. However, for future researchers, we propose to investigate the perceptions of one or two speaker attributes/adjectives at a time from synthetic voices rather than investigating the speaker characteristics (a combination of different adjectives). This would reduce the number of adjectives to be provided during the test. Further, this arrangement would facilitate the inclusion of multiple speech samples. As a result, one can also handle the drawback of modeling limited data as seen in this work. Finally, a series of such works (when all the adjectives can be grouped into one cluster) can be combined and utilized for the analysis of different speaker characteristics or personalities, or emotions from synthetic voices. In particular, the set of adjectives representing various speaker characteristics or personalities would be different (may or may not be). Thus, carrying out such an extensive evaluation setup (a long list of adjectives for each characteristic or personality) constantly would be cumbersome and is also not pragmatic. Thus, if interested in a similar line of work, we propose to study the perceptions of one or two adjectives from the synthetic voices rather than a long list of them.
- **Evaluation setup:** We employ different evaluation scales to understand the social perceptions of synthetic voices throughout this thesis. Especially, in chapter 6, we have used AB preference tests for VC voices. Since we were investigating SSC from the converted voices, we have employed 4 scales, kind, sympathetic, responsible, and skillful in these studies. However, this has increased the number of comparisons carried out in the study. Generally, such an extensive setup of comparisons is not preferred and also not practical. Therefore, as mentioned before, an analysis of one or two adjectives per study would be ideal. In addition, as discussed before, the inclusion of additional perceptual dimensions in the evaluations of TTS and VC systems through platforms like Blizzard and VC challenges would provide adequate subjective data. This would further aid in developing objective metrics for the evaluation of different speaker attributes/adjectives. The introduction of objective metrics for social perceptions of synthetic voices can also reduce the other challenges in conducting subjective evaluations such as recruiting reliable participants, and investment of time and resources to conduct the listening tests. Such objective metrics can be designed in multiple ways. Here, we provide two approaches for each TTS and VC setup.
  A typical end-to-end tacotron model utilizes the L1 loss function for model optimization. In addition to it, one can include a classification loss obtained based on the classification of the generated voice into warm/cold, competence/incompetent. Further, the model training can be enabled through the cost function obtained from the L1 loss and the classification loss (social perceptions). For this setup, the class information (class labels) should be obtained prior to the TTS training or at least for the baseline voices. Therefore preliminary work for this experimental setup would be a) determine the scales/adjectives to be used in the evaluation, b) sub-

jective evaluation for class information (or labels), c) train an additional network or integrate the class information into the TTS setup. For instance, in this thesis, we derive the ground truth adjectives for warmth as kind, and sympathetic; and for competence, the adjectives are responsible and skillful.
A similar approach can also be included in the VC setup. Apart from feeding only the speaker identity as an additional input, the speaker characteristics can also be fed (use of labeled data) to the VC setup. Further, an additional classification loss for social perceptions of the generated voice can be computed for the converted speech (since we have used Star-GAN in our experimental setup and it allows such a modification). However, we discuss these future works while considering only the experimental setups employed in this thesis. On the other hand, other future directions could be a) a different choice of source and target speakers (in our study, the source was the highly warm/competent voice, the target was the highly cold/incompetent voice), and b) modeling of acoustic correlates of SSC in a VC setup (we rely on spectral conversion alone).
Nevertheless, such an inclusion of objective valuation for SSC into the training mechanism of a TTS or VC setup would require abundant labels to train the classifiers (as discussed in the case of TTS, this requires some preliminary work). Therefore, in this respect, through this thesis, we would like to request the TTS and VC community to not only include additional dimensions in the evaluation of their systems, but also to open-source the subjective responses. This would enable the availability of a large database of labeled information for different speaker characteristics or emotions or personalities of synthetic voices (similar to the data that is currently available for speech quality and naturalness of synthetic voices).

8.3.1 Other related works

In this thesis, we have only examined the social perceptions of synthetic voices in the case of English speakers' speech (in the US accent). Also, apart from the listening tests conducted in the preliminary studies, all the remaining listening tests presented in this thesis included only US native speakers (Overall, US speakers' speech was evaluated by native US listeners). However, the speech perceptions can vary with language, accent, speaking style, nativity, age, and gender of the speakers. For instance, [215] discusses the differences in the perception of speech among native English and Spanish speakers. The study examines such differences in the perception of stop consonants, /ba/, /pa/. It consists of two tasks, a) identification task, and b) discrimination task. The study reports significant differences in the speech perceptions (identification and discrimination task) by the Spanish and English speakers. Similarly, [216] explores the speech perceptions and perceptual mapping of the language input (native) in infants. This mapping has further been found to contribute to the language-specific knowledge (phonetic units, stress patterns, prosodic cues, etc.,) in those infants before one year. In addition to age, authors in [217] have also investigated the effect of gender on the speech perceptions (phonetics)

of Persian speakers in the case of L2 language (English). The study claims that there were no differences as observed in their studies between the speech perceptions of children and adults. On the other hand, the study does not show any impact of gender of the person in perceiving an L2 language (perception of English speech by native Persians). [218] discuss the influence of the speaker's age, dialect, and gender of the listener in interpreting different speaker traits from speech. The traits analyzed in this study are prestige, confidence, credibility, and pleasantness. The study was carried out with a total of 48 participants with an equal number of New Zealanders and Utahns. The study reveals the correlations between the dialect of the speaker and the gender of the listener. They posit that the female listeners perceived higher confidence and pleasantness from the speech of the New Zealanders. The study also shows that the nativity of the listeners does not influence (negatively) the perception of the traits such as confidence, prestige, and pleasantness. Utahns provided the highest ratings to New Zealanders over native speakers on these perceptual dimensions. A similar observation was reported in [219]. The study investigates the speech perceptions (English) of New Zealanders, Americans, and Australians by around 400 students from these countries. The subjective responses display that American speakers were unanimously preferred by all the listeners irrespective of their nativity. The questionnaire of the study included 22 traits. The list was a combination of personality traits (13 dimensions), voice quality dimensions (5 dimensions), and the status index (4 dimensions). The factors derived from the subjective responses of these speakers were clustered and labeled as power, solidarity, competence, status, and voice quality. However, contrary to these studies, [220] put forward the negative effects on speech perceptions by native listeners in the case of non-native speakers. The survey involved perceptual analysis of Greek Australian English and the native (standard) Australian English. The targeted perceptual dimensions were solidarity and the status. The perceptual studies involved listening to three different passages by each of the native and non-native speakers (goal oriented, social interactions in public, and interactions in home (friendly and intimate)). The listeners of both the groups (Greek-Australian listeners and Anglo-Australian listeners) have reported lowest ratings on the status dimension for the Greek-Australian speakers.

Yet, we do not address all these elements in this thesis. However, a parallel analysis of these components in the future works could provide interesting findings.

8.4 Conclusion remarks

Through this thesis, we provide a foundation for the investigation of the social speaker characteristics, warmth, and competence from synthesis speech (speech alone). Although the amount of data was limited and the system building or evaluation methods are not highly efficient, we provide the groundwork and key insights for the generation of socially acceptable voices. The contributions of this thesis are four-fold. Firstly, our work proposes to include additional perceptual dimensions in the evaluation of synthetic voices (TTS and VC). Secondly, we derive different

mechanisms for determining the acoustic feature relevance in social perceptions of synthetic speech. Then we present a rudimentary procedure for the transformation of negatively perceived voices into positive ones. Finally, we show the modeling of a TTS framework focusing on acoustic features other than the commonly used pitch and durations. This thesis has thus highlighted and addressed the parts that were not very prevalent in the TTS or VC research before. Hopefully, our work would encourage the researchers and further pave a way for new future directions in these research fields.

Appendix A
Appendix

A.1 Datasets

A.1.1 WC Dataset

- Is there anything I can do to help?
- Do you need someone to talk with?
- There is hope.
- Have you told your doctor how you are feeling?
- It is okay to feel this way.
- I am sure we can reach a solution
- I can remedy the situation.
- I understand that you feel upset.
- Feel free to call us anytime if you have questions.
- If you need any further assistance, I would be happy to help.

A.1.2 Twitter dataset

- Don't put time on it. Relax! Maybe nap and get back to it when you get up.
- Suggestions for improvements means a person believes in your core idea and thinks their comments will help your work.
- Don't be disheartened, that's normal. It's part of the process.
- Maybe the lesson here is that it's very hard to have a totally relaxed interpersonal relationship!
- I have been exactly here. I am always ready if you ever need support. I'm here for you!
- Feedback isn't punishment. Your perspective shows you understand its purpose.
- I know I'm finding that self-care is more challenging than ever right now, but also more important than ever.

- When I'm out running and have a big hill to climb, I'm teaching myself to lean into the discomfort. It's helping me cope with things I can't change in real life also.
- Keep strong and find someone to help support you through. Take care of yourself; keep your friends close - I'm sure there are many out there.
- I love this. I started telling people "don't be sorry" when they apologize for things. most often what's necessary is gratitude rather than remorse!!!

A.2 Adjectives

Table A.1: List of adjectives derived from the 2-step procedure described in chapter 3.

Speaker attributes or adjectives			
Kind	Confident	Energetic	Outgoing
Distant	Talkative	Proactive	Tense
Empathetic	Calm	Introvert	Unsympathetic
Trusting	Worrying	Not irritated	Indecisive
Emotional	Secure	Forgiving	Friendly
Relaxed	Reliable	Hearty	Arrogant
Assertive	Agreeable	Anxious	Pleasant
Responsible	Active	Cynical	Organised
Enthusiastic	Unlikable	Accessible	Efficient
Affectionate	Generous	Trusting	Intelligent
Insightful	Active	Dominant	Curious
Thorough	Attractive	Benevolent	Decisive
Bored	Indifferent	Competent	Distant
Appreciative	Planful	Cynical	Submissive
Extrovert	Anxious	Self-pitying	Touchy
Stable	Compassionate	Imaginative	Dependable
Cynical	Approachable		

A.3 OpenSMILE features used in the study

- **Frequency specific parameters**

  - **F0 semitone at 27.5Hz** - mean, standard deviation (std dev), percentiles (20,50,80), percentile range (0-2), mean rising slope, std dev rising slope,

mean falling slope, std dev falling slope

- **Jitter** - mean, std dev
- **Shimmer** - mean std dev
- **Formant F1** - mean, std dev, bandwidth mean, bandwidth std dev, amplitude mean, amplitude std dev
- **Formant F2** - mean, std dev, bandwidth mean, bandwidth std dev, amplitude mean, amplitude std dev
- **Formant F3** - mean, std dev, bandwidth mean, bandwidth std dev, amplitude mean, amplitude std dev

- **Energy/Amplitude specific parameters**
  - **Loudness** - mean, std dev, percentile (20, 50, 80), percentile range (0-2), mean rising slope, std dev rising slope, mean falling slope, std dev falling slope
  - **HNR** - mean, std dev
  - **Harmonic difference** - H1-H2 mean, H1 - H2 std dev, H1 -A3 mean, H1 - A3 std dev
  - **Voiced segment length** - mean, std dev

- **Spectral parameters**
  - **Spectral flux** - mean, std dev
  - **mfcc1** - mean, std dev
  - **mfcc2** - mean, std dev
  - **mfcc3** - mean, std dev
  - **mfcc4** - mean, std dev
  - **Alpha Ratio** - mean, std dev
  - **Hammarberg Index** - mean, std dev

- **Slope 0-500Hz, 500-1500Hz** - mean, std dev

References

1. Scott Reed, Konrad Żolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas, "A Generalist Agent," *arXiv*, 2022.
2. Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver, "Mastering Atari, Go, chess and shogi by planning with a learned model," *Nature*, 2020.
3. David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George Driessche, Thore Graepel, and Demis Hassabis, "Mastering the game of Go without human knowledge," *Nature*, 2017.
4. Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous, "Tacotron: Towards End-to-End Speech Synthesis," *in Proc. Interspeech*, 2017.
5. Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," in *CoRR, vol.abs/1609.03499*, 2016.
6. Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," *arXiv*, 2017.
7. Yaniv, Leviathan and Yossi, Matias, "Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone," in *Google AI Blog*, 2018.
8. Khyathi Chandu, Sai Rallabandi, Sunayana Sitaram, and Alan Black, "Speech synthesis for mixed-language navigation instructions," in *Proc. Interspeech*, 2017.
9. Jasna Petrović and Mladjan Jovanovic, "Conversational Agents for Learning Foreign Languages a Survey," in *Proc. International Scientific Conference*, 2020.
10. Kinga Lachowicz-Tabaczek, "Perceived competence and warmth influence respect, liking and trust in work relations," *Polish Psychological Bulletin*, 2016.
11. Duane Francis Alwin, Michael Braun, and Jacqueline Scott, "The Separation of Work and the Family: Attitudes Towards Women's Labour-Force Participation in Germany, Great Britain, and the United States," *European Sociological Review*, 1992.

12. Mark P. Zanna and David Lewis Hamilton, "Attribute dimensions and patterns of trait inferences," *Psychonomic Science*, 1972.
13. Carroll E. Izard, "Emotion theory and research: highlights, unanswered questions, and emerging issues," *Annual review of psychology*, 2009.
14. Tobias Brosch, Klaus R Scherer, Didier Maurice Grandjean, and David Sander, "The impact of emotion on perception, attention, memory, and decision-making," *Swiss medical weekly*, 2013.
15. Susan T Fiske, Amy JC Cuddy, and Peter Glick, "Universal dimensions of social cognition: Warmth and competence," *Trends in cognitive sciences*, 2007.
16. Andrea E Abele, Thomas Rupprecht, and Bogdan Wojciszke, "The influence of success and failure experiences on agency," *European Journal of Social Psychology*, 2008.
17. Andrea Abele, Mirjam Uchronski, Caterina Suitner, and Bogdan Wojciszke, "Towards an operationalization of the fundamental dimensions of agency and communion: Trait content ratings in five countries considering valence and frequency of word occurrence," *European Journal of Social Psychology*, 2008.
18. Bogdan Wojciszke, Andrea E. Abele, and Wieslaw Baryla, "Two dimensions of interpersonal attitudes: Liking depends on communion, respect depends on agency," *European Journal of Social Psychology*, 2009.
19. Angeliki Kerasidou, Kristine Bærøe, Zackary Berger, and Amy Brown, "The need for empathetic healthcare systems," *Journal of Medical Ethics*, 2020.
20. Gordon Kraft-Todd, Diego Reinero, John Kelley, Andrea Heberlein, Lee Baer, and Helen Riess, "Empathic nonverbal behavior increases ratings of both warmth and competence in a medical context," *PLOS ONE*, 2017.
21. Judith Hall, Debra Roter, Danielle Blanch-Hartigan, and Richard Frankel, "Nonverbal Sensitivity in Medical Students: Implications for Clinical Interactions," *Journal of general internal medicine*, 2009.
22. Pascal Güntürkün, Till Haumann, and Sven Mikolon, "Disentangling the Differential Roles of Warmth and Competence Judgments in Customer-Service Provider Relationships," *Journal of Service Research*, 2020.
23. Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," *arXiv preprint:1803.09017*, 2018.
24. Daisy Stanton, Yuxuan Wang, and R. Skerry-Ryan, "Predicting Expressive Speaking Style from Text in End-To-End Speech Synthesis," *IEEE Spoken Language Technology Workshop (SLT)*, 2018.
25. Ya-Jie Zhang, Shifeng Pan, L. He, and Zhenhua Ling, "Learning Latent Representations for Style Control and Transfer in End-to-end Speech Synthesis," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
26. Ryan Prenger, Rafael Valle, and Bryan Catanzaro, "WaveGlow: A Flow-based Generative Network for Speech Synthesis," *arXiv*, 2018.
27. Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu, "Efficient Neural Audio Synthesis," *arXiv*, 2018.
28. Keiichi Tokuda Alan W. Black, "The Blizzard Challenge – 2005: Evaluating Corpus-Based Speech Synthesis on Common Datasets," *Blizzard Challenge Workshop*, 2005.
29. Mark E. Fraser Cereproc Edinburgh Matthew P. Aylett, Christopher J. Pidcock, "The Cerevoice Blizzard Entry 2006: A prototype Database Unit Selection Engine," *Blizzard Challenge Workshop*, 2006.
30. Mark Fraser and Simon King, "The Blizzard Challenge 2007," *Blizzard Challenge Workshop*, 2007.
31. Robert A. J. Clark Catherine Mayo Vasilis Karaiskos, Simon King, "The Blizzard Challenge 2008," *Blizzard Challenge Workshop*, 2008.
32. Simon King and Vasilis Karaiskos, "The Blizzard Challenge 2009," *Blizzard Challenge Workshop*, 2009.

33. Simon King and Vasilis Karaiskos, "The Blizzard Challenge 2010," *Blizzard Challenge Workshop*, 2010.
34. Simon King and Vasilis Karaiskos, "The Blizzard Challenge 2011," *Blizzard Challenge Workshop*, 2011.
35. Simon King and Vasilis Karaiskos, "The Blizzard Challenge 2012," *Blizzard Challenge Workshop*, 2012.
36. Simon King and Vasilis Karaiskos, "The Blizzard Challenge 2013," *Blizzard Challenge Workshop*, 2013.
37. Santosh Kesiraju Hema A Murthy Swaran Lata T. Nagarajan Mahadeva Prasanna Hemath Patil Anil Kumar Sao Simon King Alan W Black Kishore Prahallad, Anandaswarup Vadapalli and Keiichi Tokuda, "The Blizzard Challenge 2014," *Blizzard Challenge Workshop*, 2014.
38. Sai Krishna Rallabandi Santosh Kesiraju Hema Murthy T. Nagarajan Bira Singh Sajani T K Sreenivasa Rao Suryakanth V Gangashetty Simon King Keiichi Tokuda Alan W Black Kishore Prahallad, Anandaswarup Vadapalli, "The Blizzard Challenge 2015," *Blizzard Challenge Workshop*, 2015.
39. Simon King and Vasilis Karaiskos, "The Blizzard Challenge 2016," *Blizzard Challenge Workshop*, 2016.
40. Wei Guo Simon King, Lovisa Wihlborg, "The Blizzard Challenge 2017," *Blizzard Challenge Workshop*, 2017.
41. Amy Martin Lovisa Wihlborg Simon King, Jane Crumlish, "The Blizzard Challenge 2018," *Blizzard Challenge Workshop*, 2018.
42. Zhihang Xie Zhizheng Wu and Simon King, "The Blizzard Challenge 2019," *Blizzard Challenge Workshop*, 2019.
43. Simon King Xiao Zhou, Zhen-Hua Ling, "The Blizzard Challenge 2020," *Blizzard Challenge Workshop*, 2020.
44. Simon King Zhen-Hua Ling, Xiao Zhou, "The Blizzard Challenge 2021," *Blizzard Challenge Workshop*, 2021.
45. Avashna Govender and Simon King, "Using Pupillometry to Measure the Cognitive Load of Synthetic Speech," in *Proc. Interspeech*, 2018.
46. Deja Kamil, Sanchez Ariadna, Roth Julian, and Cotescu Marius, "Automatic Evaluation of Speaker Similarity," *arXiv*, 2022.
47. Gabriel Mittag and Sebastian Möller, "Deep Learning Based Assessment of Synthetic Speech Naturalness," in *Proc. Interspeech*, 2020.
48. Wen-Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi, "The VoiceMOS Challenge 2022," *arXiv*, 2022.
49. Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang, "MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion," in *Proc. Interspeech*, 2019.
50. Cheng-Hung Hu, Yu-Huai Peng, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang, "SVSNet: An End-to-End Speaker Voice Similarity Assessment Model," *IEEE Signal Processing Letters*, 2022.
51. Alice Baird, Stina Jørgensen, Emilia Parada-Cabaleiro, Simone Hantke, Nicholas Cummins, and Björn Schuller, "Perception of Paralinguistic Traits in Synthesized Voices," in *AM '17: Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences*, 2017.
52. Alice Baird, Emilia Parada-Cabaleiro, Simone Hantke, Felix Burkhardt, Nicholas Cummins, and Björn Schuller, "The Perception and Analysis of the Likeability and Human Likeness of Synthesized Speech," in *Proc. Interspeech*, 2018.
53. Kirsten Bergmann, Friederike Eyssel, and Stefan Kopp, "A Second Chance to Make a First Impression? How Appearance and Nonverbal Behavior Affect Perceived Warmth and Competence of Virtual Agents over Time," in *International Conference on Intelligent Virtual Agents*, 2012.
54. Solomon E Asch, "Forming impressions of personality," *The Journal of Abnormal and Social Psychology*, 1946.

55. Seymour Rosenberg, Cynthia Nelson, and Pathe S. Vivekananthan, "A multidimensional approach to the structure of personality impressions," *Journal of personality and social psychology*, 1968.
56. Robert Freed Bales, "A set of categories for the analysis of small group interaction," *American Sociological Review*, 1950.
57. Robert Bales, "Social Interaction Systems: Theory and Measurement: Book review," *Group Dynamics: Theory, Research, and Practice*, 2000.
58. Charles E. Osgood, William H. May, and Murray S. Miron, "Cross-Cultural Universals of Affective Meaning," in *University of Illinois Press*, 1975.
59. Robert R McCrae and Paul T Costa, "The structure of interpersonal traits: Wiggins's circumplex and the five-factor model," *Journal of personality and social psychology*, 1989.
60. Amy J.C. Cuddy, Susan T. Fiske, and Peter Glick, "Warmth and Competence as Universal Dimensions of Social Perception: The Stereotype Content Model and the BIAS Map," in *Advances in Experimental Social Psychology*, 2008.
61. Andrea Scarantino and Ronald de Sousa, "Emotion," in *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, 2021.
62. Silvan S Tomkins, *Affect imagery consciousness: the complete edition: two volumes*, Springer publishing company, 2008.
63. Simeng Gu, Fushun Wang, Nitesh P Patel, James A Bourgeois, and Jason H Huang, "A model for basic emotions using observations of behavior in Drosophila," *Frontiers in psychology*, 2019.
64. Olga Stavrova and Daniel Ehlebracht, "Cynical Beliefs About Human Nature and Income: Longitudinal and Cross-Cultural Analyses," *Journal of Personality and Social Psychology*, 2015.
65. Daniel Ehlebracht and Olga Stavrova, "Running head: CYNICAL GENIUS ILLUSION 1 The Cynical Genius Illusion: Exploring and Debunking Lay Beliefs about Cynicism and Competence," *Personality and Social Psychology Bulletin*, 2018.
66. Alessandro Vinciarelli and Gelareh Mohammadi, "A Survey of Personality Computing," *IEEE Transactions on Affective Computing*, 2014.
67. Robert R McCrae and Paul T Costa Jr, "The five-factor theory of personality," *The Five-Factor Model of Personality: Theoretical Perspectives*, 1999.
68. Reza Fazli Salehi, Ivonne Torres, Rozbeh Madadi, and Miguel Zúñiga, "The impact of interpersonal traits (extraversion and agreeableness) on consumers' self-brand connection and communal-brand connection with anthropomorphized brands," *Journal of Brand Management*, 2022.
69. Jerry S Wiggins, Paul Trapnell, and Norman Phillips, "Psychometric and Geometric Characteristics of the Revised Interpersonal Adjective Scales (IAS-R)," *Multivariate Behavioral Research*, 1988.
70. RR McCrae and OP John, "An Introduction to the Five-Factor Model and its Applications," *Journal of Personality*, 1992.
71. Robert M. Krauss, Robin Freyberg, and Ezequiel Morsella, "Inferring speakers physical attributes from their voices," *Journal of Experimental Social Psychology*, 2002.
72. Clifford Nass and Kwan Min Lee, "Does Computer-Generated Speech Manifest Personality? An Experimental Test of Similarity-Attraction," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2000.
73. Oliver P John and Kentle Robert L Donahue, Eileen M, "Big five inventory," *Journal of Personality and Social Psychology*, 1991.
74. "Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German," *Journal of Research in Personality*, 2007.
75. Laura Fernández and Gallardo and Weiss, Benjamin, "Towards Speaker Characterization: Identifying and Predicting Dimensions of Person Attribution," *in Proc. Interspeech*, 2017.
76. Laura Fernández Gallardo and Benjamin Weiss, "The Nautilus Speaker Characterization Corpus: Speech Recordings and Labels of Speaker Characteristics and Voice Descriptions," in *Proc. Eleventh International Conference on Language Resources and Evaluation (LREC)*, 2018.

77. Felix Burkhardt, Richard Huber, and Anton Batliner, *Application of Speaker Classification in Human Machine Dialog Systems*, 2007.
78. Andrew Rosenberg and Julia Hirschberg, "Acoustic/prosodic and lexical correlates of charismatic speech," in *Proc. Interspeech*, 2005.
79. Stephen M. Smith and David R. Shaffer, "Speed of Speech and Persuasion: Evidence for Multiple Effects," *Personality and Social Psychology Bulletin*, 1995.
80. Eva Strangert and Joakim Gustafson, "What makes a good speaker? Subject ratings, acoustic measurements and perceptual evaluations," in *Proc. Interspeech*, 2008.
81. Benjamin Weiss and Felix Burkhardt, "Voice attributes affecting likability perception," 2010.
82. Oliver Niebuhr, Jana Voße, and Alexander Brem, "What makes a charismatic speaker? A computer-based acoustic-prosodic analysis of Steve Jobs tone of voice," *Computers in Human Behavior*, 2016.
83. Oliver Niebuhr and Jan Michalsky, "Computer-generated speaker charisma and its effects on human actions in a car-navigation system experiment - or how steve jobs' tone of voice can take you anywhere," in *Computational Science and Its Applications – ICCSA 2019 - 19th International Conference, 2019, Proceedings*, 2019.
84. J A Bachorowski and Michael J. Owren, "Vocal Expression of Emotion: Acoustic Properties of Speech Are Associated With Emotional Intensity and Context," *Psychological Science*, 1995.
85. Petri Laukka, Patrik Juslin, and Roberto Bresin, "A dimensional approach to vocal expression of emotion," *Cognition and Emotion*, 2005.
86. Tim Polzehl, Sebastian Möller, and Florian Metze, "Automatically Assessing Personality from Speech," in *IEEE Fourth International Conference on Semantic Computing*, 2010.
87. Gelareh Mohammadi and Alessandro Vinciarelli, "Automatic personality perception: Prediction of trait attribution based on prosodic features extended abstract," in *2015 International conference on affective computing and intelligent interaction (ACII)*, 2015.
88. Anna Oleszkiewicz, Katarzyna Pisanski, Kinga Lachowicz-Tabaczek, and Agnieszka Sorokowska, "Voice-based assessments of trustworthiness, competence, and warmth in blind and sighted adults," *Psychonomic bulletin & review*, 2017.
89. Maimon Oded Rokach, Lior, "Decision trees," *Data Mining and Knowledge Discovery Handbook*, 2005.
90. Timo Baumann, "Decision tree usage for incremental parametric speech synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.
91. R.E. Donovan, "Segment pre-selection in decision-tree based speech synthesis systems," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1988.
92. Wern-Jun Wang, W. Nick Campbell, Naoto Iwahashi, and Yoshinori Sagisaka, "Tree-based unit selection for English speech synthesis," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1993.
93. Nello Cristianini and Elisa Ricci, "Support vector machines," *Encyclopedia of Algorithms*, 2008.
94. Armin Shmilovici, *Support Vector Machines*, Springer, 2005.
95. Kamil Aida-zade, Anar Xocayev, and Samir Rustamov, "Speech recognition using Support Vector Machines," in *IEEE 10th International Conference on Application of Information and Communication Technologies (AICT)*, 2016.
96. Zhizheng Wu, Oliver Watts, and Simon King, "Merlin: An Open Source Neural Network Speech Synthesis System," in *Proc. Speech Synthesis Workshop (SSW)*, 2016.
97. Heiga Ze, Andrew Senior, and Mike Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.
98. Srinivas Desai, E. Veera Raghavendra, B. Yegnanarayana, Alan W. Black, and Kishore Prahallad, "Voice conversion using Artificial Neural Networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.

99. David E. Rumelhart and James L. McClelland, "Learning Internal Representations by Error Propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, 1987.
100. Sepp Hochreiter, "The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 1998.
101. Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-term Memory," *Neural computation*, 1997.
102. Alex Graves and Jürgen Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, 2005, IJCNN.
103. Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *arXiv*, 2014.
104. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural Computation*, 1989.
105. Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller, "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," in *Proc. Interspeech*, 2021.
106. Jani Nurminen, Hanna Silén, Victor Popa, Elina Helander, and Moncef Gabbouj, *Voice Conversion*, 2012.
107. Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, 1998.
108. Srinivas Desai, E. Veera Raghavendra, B. Yegnanarayana, Alan W. Black, and Kishore Prahallad, "Voice conversion using Artificial Neural Networks," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009.
109. Mingyang Zhang, Berrak Sisman, Li Zhao, and Haizhou Li, "DeepConversion: Voice conversion with limited parallel training data," *Speech Communication*, 2020.
110. Athanasios Mouchtaris, Jan Van Der Spiegel, and Paul Mueller, "Non-parallel training for voice conversion by maximum likelihood constrained adaptation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2004.
111. Yi Zhou, Xiaohai Tian, Haihua Xu, Rohan Das, and Haizhou Li, "Cross-lingual voice conversion with bilingual phonetic posteriorgram and average modeling," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
112. Sai Sirisha Rallabandi and Suryakanth V Gangashetty, "An Approach to Cross-Lingual Voice Conversion," in *International Joint Conference on Neural Networks (IJCNN)*, 2019.
113. Zhenchuan Yang, Weibin Zhang, Yufei Liu, and Xiaofen Xing, "Cross-Lingual Voice Conversion with Disentangled Universal Linguistic Representations," in *Proc. Interspeech*, 2021.
114. Zhiwei Shuang, Raimo Bakis, Slava Shechtman, Dan Chazan, and Yong Qin, "Frequency warping based on mapping formant parameters," in *9th International Conference on Spoken Language Processing, Interspeech - ICSLP*, 2006.
115. M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1988.
116. K. Shikano, S. Nakamura, and M. Abe, "Speaker adaptation and voice conversion by codebook mapping," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 1991.
117. Tomoki Toda, Alan W. Black, and Keiichi Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, 2007.
118. Kazuhiro Kobayashi, Shinnosuke Takamichi, Satoshi Nakamura, and Tomoki Toda, "The NU-NAIST Voice Conversion System for the Voice Conversion Challenge 2016," in *Proc. Interspeech*, 2016.
119. Heiga Zen, Yoshihiko Nankaku, and Keiichi Tokuda, "Probabilistic feature mapping based on trajectory HMMs," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

120. Yannis Stylianou, "Voice Transformation: A survey," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.
121. Lifa Sun, Shiyin Kang, Kun Li, and Helen Meng, "Voice Conversion Using Deep Bidirectional Long Short-Term Memory based Recurrent Neural Networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.
122. Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Networks," *arXiv*, 2014.
123. Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo, "StarGAN-VC: non-parallel many-to-many Voice Conversion Using Star Generative Adversarial Networks," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018.
124. Takuhiro Kaneko and Hirokazu Kameoka, "Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks," *arXiv*, 2017.
125. Kun Zhou, Berrak Sisman, Mingyang Zhang, and Haizhou Li, "Converting Anyone's Emotion: Towards Speaker-Independent Emotional Voice Conversion," *arXiv*, 2020.
126. Kun Zhou, Berrak Sisman, and Haizhou Li, "Limited Data Emotional Voice Conversion Leveraging Text-to-Speech: Two-stage Sequence-to-Sequence Training," in *Proc. Interspeech*, 2021.
127. Hideki Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustical Science and Technology*, 2006.
128. Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, 2016.
129. Matthias Holschneider, Richard Kronland-Martinet, J. Morlet, and Ph Tchamitchian, "A Real-Time Algorithm for Signal Analysis with the Help of the Wavelet Transform," *Wavelets, Time-Frequency Methods and Phase Space*, 1989.
130. D. Griffin and Jae Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1984.
131. Paul Taylor, Alan W. Black, and Richard Caley, "The Architecture of the Festival Speech Synthesis System," in *Proc. Speech Synthesis Workshop (SSW)*, 1998.
132. R. Carlson and B. Granstrom, "A text-to-speech system based entirely on rules," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1976.
133. Dennis H. Klatt, "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America*, 1980.
134. Celia Scully, "Articulatory Synthesis," in *Speech Production and Speech Modelling*, 1990.
135. Yoshinori Sagisaka, Nobuyoshi Kaiki, Naoto Iwahashi, and Katsuhiko Mimura, "ATR μ-talk speech synthesis system," *2nd International Conference on Spoken Language Processing (ICSLP)*, 1992.
136. Robert E. Donovan and Ellen Eide, "The IBM trainable speech synthesis system," *5th International Conference on Spoken Language Processing (ICSLP)*, 1998.
137. Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," *6th European Conference on Speech Communication and Technology (Eurospeech)*, 1999.
138. Claude Sammut and Geoffrey I. Webb, "Viterbi algorithm," *Encyclopedia of Machine Learning*, 2010.
139. Alan W. Black, Paul Taylor, Richard Caley, Rob Clark, Korin Richmond, Simon King, Volker Strom, and Heiga Zen, "The Festival Speech Synthesis System, Version 1.4.2," in *Unpublished document available via http://www.cstr.ed.ac.uk/projects/festival.html*, 2001.
140. Heiga Zen, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura, "Hidden semi-Markov model based speech synthesis," in *Proc. Interspeech*, 2004.
141. Heiga Zen, Takashi Nose, Junichi Yamagishi, Shinji Sako, Takashi Masuko, Alan W. Black, and Keiichi Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. Speech Synthesis Workshop (SSW)*, 2007.

142. Yuchen Fan, Yao Qian, Feng-Long Xie, and Frank K Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Fifteenth annual conference of the international speech communication association*, 2014.
143. Zhizheng Wu, Oliver Watts, and Simon King, "Merlin: An Open Source Neural Network Speech Synthesis System," in *Proc. Speech Synthesis Workshop (SSW)*, 2016.
144. Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv*, 2016.
145. Daniel J. Hirst, Albert Rilliard, and Véronique Aubergé, "Comparison of subjective evaluation and an objective evaluation metric for prosody in text-to-speech synthesis," in *Proc. Speech Synthesis Workshop (SSW)*, 1998.
146. Mark Beutnagel, Alistair Conkie, Juergen Schroeter, Yannis Stylianou, and Ann Syrdal, "The ATT Next-Gen TTS System," *The Journal of the Acoustical Society of America*, 2000.
147. ITU-T Recommendation P.85, *A method for subjective performance assessment of the quality of speech voice output devices*, International Telecommunication Union, 1994.
148. ITU-T Recommendation P.800, *Methods for subjective determination of transmission quality*, International Telecommunication Union, 1996.
149. Maxine Eskenazi, Gina-Anne Levow, Helen Meng, Gabriel Parent, and David Suendermann, *Crowdsourced Assessment of Speech Synthesis*, 2013.
150. Hideyuki Takahashi, Midori Ban, and Minoru Asada, "Semantic Differential Scale Method Can Reveal Multi-Dimensional Aspects of Mind Perception," *Frontiers in Psychology*, 2016.
151. ITU-T Recommendation P.808, *Subjective evaluation of speech quality with a crowdsourcing approach*, International Telecommunication Union, 2018.
152. Babak Naderi and Ross Cutler, "An Open source Implementation of ITU-T Recommendation P.808 with Validation," in *Proc. Interspeech*, 2020.
153. Babak Naderi, Tim Polzehl, Ina Wechsung, Friedemann Köster, and Sebastian Möller, "Effect of trapping questions on the reliability of speech quality judgments in a crowdsourcing paradigm," in *Proc. Interspeech*, 2015.
154. Dennis Guse, Henrique Orefice, Gabriel Reimers, and Oliver Hohlfeld, "TheFragebogen: A Web Browser-based Questionnaire Framework for Scientific Research," in *Proc. QoMEX*, 2019.
155. Robert Freed Bales, "Social interaction systems: Theory and measurement," in *Routledge*, 2017.
156. Beatrice Biancardi, Angelo Cafaro, and Catherine Pelachaud, "Analyzing First Impressions of Warmth and Competence from Observable Nonverbal Cues in Expert-Novice Interactions," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017.
157. Margarida V. Garrido and Sandra Godinho, "When vowels make us smile: the influence of articulatory feedback in judgments of warmth and competence," *Cognition and Emotion*, 2021.
158. Tanja Schultz, "Speaker Characteristics," in *Speaker Classification I: Fundamentals, Features, and Methods*, 2007.
159. Alice Baird, Stina Hasse Jørgensen, Emilia Parada-Cabaleiro, Simone Hantke, Nicholas Cummins, and Björn Schuller, "Perception of paralinguistic traits in synthesized voices," in *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences*, 2017.
160. John Kominek and Alan W Black, "The CMU ARCTIC speech databases," in *Proc. Speech Synthesis Workshop (SSW)*, 2004.
161. Sercan Ö. Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, Shubho Sengupta, and Mohammad Shoeybi, "Deep voice: Real-time neural text-to-speech," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.
162. Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit," 2017, https://datashare.ed.ac.uk/handle/10283/3443.
163. Keith Ito and Linda Johnson, "The LJ Speech Dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

164. Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon, "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search," *arXiv*, 2020.
165. Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov, "Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech," *arXiv*, 2021.
166. Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, Frank Soong, Tao Qin, Sheng Zhao, and Tie-Yan Liu, "NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality," *arXiv*, 2022.
167. Sai Sirisha Rallabandi, Abhinav Bharadwaj, Babak Naderi, and Sebastian Möller, "Perception of Social Speaker Characteristics in Synthetic Speech," in *Proc. Interspeech*, 2021.
168. Björn W. Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, R. J. J. H. van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, Gelareh Mohammadi, and Benjamin Weiss, "The INTERSPEECH 2012 Speaker Trait Challenge," in *Proc. Interspeech*, 2012.
169. K. R. Scherer and U. Scherer, "Speech Behavior and Personality," *Speech Evaluation in Psychiatry*, 1981.
170. Matthew P. Aylett, Alessandro Vinciarelli, and Mirjam Wester, "Speech Synthesis for the Generation of Artificial Personality," *IEEE Transactions on Affective Computing*, 2020.
171. Jürgen Trouvain, Sarah Schmidt, Marc Schröder, Michael Schmitz, and William J. Barry, "Modelling personality features by changing prosody in synthetic speech," in *Proc. Speech Prosody*, 2006.
172. Jennifer L. Aaker, "Dimensions of Brand Personality," *Journal of Marketing Research*, 1997.
173. A. Abele, N. Hauke, K. Peters, E. Louvet, A. Szymkow, and Yanping Duan, "Facets of the Fundamental Content Dimensions: Agency with Competence and Assertiveness—Communion with Warmth and Morality," *Frontiers in Psychology*, 2016.
174. Marc Schröder, Marcela Charfuelan, Sathish Pammi, and Ingmar Steiner, "Open Source Voice Creation Toolkit for the MARY TTS Platform," in *Proc. Interspeech*. 2011, ISCA.
175. A.J. Hunt and A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, 1996.
176. Glen T. Evans, "Use of the semantic differential technique to study attitudes during classroom lessons," *Interchange*, 1970.
177. Domenic V Cicchetti, "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology," *Psychological assessment*, 1994.
178. Terry K Koo and Mae Y Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of chiropractic medicine*, 2016.
179. Dennis. Child, "The essentials of factor analysis / Dennis Child," in *The essentials of factor analysis*. 2006, Continuum.
180. W. Holmes finch, "Exploratory Factor Analysis," in *Handbook of Quantitative Methods for Educational Research*, 2013.
181. Urbano Lorenzo-Seva, "The weighted oblimin rotation," *Psychometrika*, 2000.
182. Edward E Cureton and Stanley A Mulaik, "The weighted varimax rotation and the promax rotation," *Psychometrika*, 1975.
183. Frank Rösler and Dietrich Manzey, "Principal components and varimax-rotated components in event-related potential research: Some remarks on their interpretation," *Biological Psychology*, 1981.
184. Robert Choate Tryon, "Communality of a variable: Formulation by cluster analysis," *Psychometrika*, 1957.
185. Richard A. Bernardi, "Validating Research Results when Cronbach'S Alpha is Below .70: A Methodological Procedure," *Educational and Psychological Measurement*, 1994.
186. Susan T. Fiske, "Stereotype Content: Warmth and Competence Endure," *Current Directions in Psychological Science*, 2018.
187. Marius Herberg and Glenn-Egil Torgersen, "Resilience Competence Face Framework for the Unforeseen: Relations, Emotions and Cognition. A Qualitative Study," *Frontiers in Psychology*, 2021.

188. Katherine Valentine, Norman Li, Andrea L. Meltzer, and Ming-Hong Tsai, "Mate Preferences for Warmth-Trustworthiness Predict Romantic Attraction in the Early Stages of Mate Selection and Satisfaction in Ongoing Relationships," *Personality and Social Psychology Bulletin*, 2019.
189. Kevin Crowston, "Amazon Mechanical Turk: A Research Tool for Organizations and Information Systems Scholars," in *Shaping the Future of ICT Research. Methods and Approaches*, 2012.
190. Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller, "Recent Developments in OpenSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proceedings of the 21st ACM International Conference on Multimedia*, 2013.
191. Dongrui Wu, Thomas Parsons, and Shrikanth Narayanan, "Acoustic feature analysis in speech emotion primitives estimation," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.
192. B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004.
193. A. Tickle, S. Raghu, and M. Elshaw, "Emotional recognition from the speech signal for a virtual education agent," *Journal of Physics Conference Series*, 2013.
194. Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, Gelareh Mohammadi, and Benjamin Weiss, "A Survey on perceived speaker traits: Personality, likability, pathology, and the first challenge," *Computer Speech Language*, 2015.
195. Molly Babel, Grant McGuire, and Joseph King, "Towards a more nuanced view of vocal attractiveness," *PloS one*, 2014.
196. Eiji Yumoto, Wilbur J Gould, and Thomas Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *The journal of the Acoustical Society of America*, 1982.
197. Bruce Smith, Bruce Brown, William Strong, and Alvin Rencher, "Effects of Speech Rate on Personality Perception," *Language and speech*, 1975.
198. Lynn A. Streeter, Robert M. Krauss, Valerie Geller, Carrie F. Olson, and W Apple, "Pitch changes during attempted deception," *Journal of personality and social psychology*, 1977.
199. Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al., "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE transactions on affective computing*, 2015.
200. Michail Vlachos, *Dimensionality Reduction*, Springer, 2010.
201. Diederik Kingma and Jimmy Ba, "Adam: A Method for Stochastic Optimization," *International Conference on Learning Representations*, 2014.
202. Sai Sirisha Rallabandi, Babak Naderi, and Sebastian Möller, "Identifying the vocal cues of likeability, friendliness and skilfulness in synthetic speech," in *Proc. Speech Synthesis Workshop (SSW)*, 2021.
203. Laura Fernández Gallardo and Benjamin Weiss, "Perceived Interpersonal Speaker Attributes and their Acoustic Features," in *Phonetik und Phonologie im deutschprachigen Raum*, 2017.
204. Diederik P Kingma and Max Welling, "Auto-Encoding Variational Bayes," *arXiv*, 2013.
205. Tetsuya Hashimoto, Hidetsugu Uchida, Daisuke Saito, and Nobuaki Minematsu, "Parallel-Data-Free Many-to-Many Voice Conversion Based on DNN Integrated with Eigenspace Using a Non-Parallel Speech Corpus," in *Proc. Interspeech*, 2017.
206. Ravi Shankar, Jacob Sager, and Archana Venkataraman, "A Multi-Speaker Emotion Morphing Model Using Highway Networks and Maximum Likelihood Objective," in *Proc. Interspeech*, 2019.
207. Masanori MORISE, Fumiya YOKOMORI, and Kenji OZAWA, "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE Transactions on Information and Systems*, 2016.
208. Sai Sirisha Rallabandi and Sebastian Möller, "On incorporating social speaker characteristics in synthetic speech," *arXiv*, 2022.

209. RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, and Rif A. Saurous, "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron," *arXiv*, 2018.
210. Viacheslav Klimkov, Srikanth Ronanki, Jonas Rohnke, and Thomas Drugman, "Fine-grained robust prosody transfer for single-speaker neural text-to-speech," *arXiv*, 2019.
211. Yuxuan Wang, RJ Skerry-Ryan, Ying Xiao, Daisy Stanton, Joel Shor, Eric Battenberg, Rob Clark, and Rif A. Saurous, "Uncovering Latent Style Factors for Expressive Speech Synthesis," *arXiv*, 2017.
212. Yutian Wang, Yuankun Xie, Kun Zhao, Hui Wang, and Qin Zhang, "Unsupervised Quantized Prosody Representation for Controllable Speech Synthesis," *arXiv*, 2022.
213. Erzsébet Frigó and Levente Kocsis, "Online convex combination of ranking models," *User Modeling and User-Adapted Interaction*, 2021.
214. Martina Echtenbruck, Thomas Bartz-Beielstein, and Michael Emmerich, "Building Ensembles of Surrogates by Optimal Convex Combination," in *Proc. of Bioinspired Optimization Methods and their Applications*, 2016.
215. T. Christina Zhao and Patricia K. Kuhl, "Linguistic effect on speech perception observed at the brainstem," *Proceedings of the National Academy of Sciences*, 2018.
216. Patricia K Kuhl, "Effects of language experience on speech perception," *Journal of the Acoustical Society of America*, 1998.
217. Zohreh Kassaian, "Age and gender effect in phonetic perception and production," *Journal of Language Teaching and Research*, 2011.
218. Ikeno Ayako and John Hansen, "The Effect of Listener Accent Background on Accent Perception and Comprehension," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007.
219. Donn Bayard, Ann Weatherall, Cynthia Gallois, and Jeffery Pittam, "Pax Americana? Accent attitudinal evaluations in New Zealand, Australia and America," *Journal of Sociolinguistics*, 2001.
220. Victor J. Callan, Cynthia Gallois, and Paula A. Forbes, "Evaluative Reactions to Accented English: Ethnicity, Sex Role, and Context," *Journal of Cross-Cultural Psychology*, 1983.