Culture and Ambience: Investigating the Role of Social Environments on Classification and Generation of Facial and Verbal Expressions

by

Emma Lynn Hughson

B.Sc., Simon Fraser University, 2020
B.Sc.(Hons), University of Victoria, 2017

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Computing Science
Faculty of Applied Sciences

SIMON FRASER UNIVERSITY
Spring 2022

Declaration of Committee

Name:	**Emma Lynn Hughson**
Degree:	**Master of Science**
Thesis title:	**Culture and Ambience: Investigating the Role of Social Environments on Classification and Generation of Facial and Verbal Expressions**
Committee:	**Chair:** Angel Chang Assistant Professor, Computing Science
	Angelica Lim Supervisor Assistant Professor, Computing Science
	Mo Chen Committee Member Professor, Computing Science
	Philippe Pasquier Examiner Professor, Interactive Arts and Technology

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

Social environments play a critical function in how humans express themselves in non-verbal communication with other humans and robots. For instance, culture may change how one expresses emotions on their face, while ambiance may change how they express themselves vocally. Little research has been done in affective computing to take into account such social environments when building affective systems. The current thesis aims to investigate these two sources of influence. In study 1, a support vector machine was leveraged to classify three negative emotions across three different cultures using a curated dataset compiled from YouTube videos. In addition, a one-way ANOVA was used to analyze the differences that exist between each culture in terms of the level of activation of underlying social signals. In study 2, we investigate modifications to a robot's speech, using various speech generation tools, to maximize acceptability in various social and acoustic contexts, starting with a use case for service robots in varying restaurants. An original dataset was collected over Zoom with participants conversing in scripted and unscripted tasks given seven different ambient sounds and images. Voice conversion, and altered Text-to-Speech that matched ambient specific data, were implemented for speech generation tasks. A subjective perception study showed that humans favour generated speech that matches the ambient environment, ultimately preferring more human-like voices. This work provides three important contributions to culture-specific emotion expression recognition, as well as ambient appropriate generated voices: (1) understand how different cultures express themselves through their facial expressions, (2) understand how humans adapt their voices to different ambiences, and (3) taking data-driven approaches to perform classification and generation tasks using context-sensitive machine learning methods and novel data collection protocols.

Keywords: affective computing, human-robot interaction, emotion expression recognition, social signals, voice adaptation, voice signal processing, voice generation, classification

Dedication

To my parents who have supported me throughout my studies and to my sister Sarah.

Acknowledgements

I would like to acknowledge the hard work of Paige Tuttosi, who assisted me in study 2 with data collection, literature review, results, and analysis. I would like to acknowledge Akihiro Matsufuji, who assisted in analysis of study 2. I would also like to acknowledge the work of Roya Javadi and Kevin San Gabriel who helped with data collection for study 1. Finally, I would like to state how appreciative I am for the advice, guidance, and knowledge that Dr. Angelica Lim has provided me over my studies. Without her I would not have realized my love for affective computing, and for that I am truly grateful.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

There are a variety of different social environments that guide the way humans communicate. Context, according to Merriam-Webster, is "the interrelated conditions in which something exists or occurs" [2]. A social environment is a subset of context that includes non-linguistic social contextual influences that impact the meaning of linguistic communication and facial expressions. For example, one study done by [28] found that the social environment of a social interaction, whether it be imagined or "real", was more indicative of facial displays of happiness than the emotion itself. In addition, the display of a given emotion is more intense and changes more frequently when one imagines social interactions within a social environment compared to when they imagine being alone [29]. Given this implicit influence on human communication, two unique social environments will be discussed in this thesis: (1) culture and (2) ambience.

Whether it be by vocal or facial signals, humans have the innate capability to communicate thoughts with high fidelity. This natural ability can be deliberate, but more-often it is done automatically. More specifically, gestures, speech features (e.g., pitch), facial expressions (e.g., facial landmarks), gaze, and much more, are all used simultaneously to communicate our thoughts and our feelings instinctively [79]. These social signals can be displayed differently from person to person for a given emotion. For example, one person may display disgust by wrinkling their noise at a terrible smell, while another may show no reaction at all. In addition, humans adapt these social signals to a variety of different environments [9, 75, 33]. By way of explanation, specific social signals, such as speech features and/or facial landmarks, may change depending on what environment you are placed in. How these social signals change also depend on how one has been taught to act in certain environments. Culture is one way knowledge is handed down and guides ones' emotional expression in different environments [19, 14]. Researchers call this concept "social learning", and it is driven by experiences and interaction within a individual's culture [71].

Knowledge, handed down through cultural experiences, plays a large role in how one expresses themselves and perceives the world around them. Psychological studies [20, 19], have established a strong case that different countries express emotions differently, due in part to how members of a given culture are exposed to, and taught, to display their own emotions. It is expected that different rules are applied in different ambiences; this may or may not depend on ones culture. Ambience is the "feeling" or "atmosphere" of a physical surrounding, such as a restaurant [1]. One may not express themselves the same way when placed in a loud, noisy upbeat restaurant versus a quiet, empty library. This also depends on how we are taught what are or are not appropriate social signals to display in certain ambient environments [65, 9, 82, 19, 71].

Given the variety of social signals used to communicate in different contexts, it is important that robots follow the same set of rules provided by cultural and ambient surroundings. This innate capability of seamlessly adapting to culture and ambiences that humans possess is something that is, currently, difficult for robots to reproduce. Not only is it crucial that robots perceive emotions accurately across multiple different contexts (e.g., culture), it is also important that robots generate realistic expressions that match that of the current environment (e.g., ambience). In this way, human-robot interaction will become more harmonious as it will allow for correct responses by the robot. In other words, robot's will be able to accurately perceive and display appropriate social signals cross-culturally in a variety of ambient contexts. As such, in the scope of this thesis, I focus on facial expression recognition and speech generation across different cultures and ambiences. The goals of the current thesis are to:

- Understand how different cultures express themselves through their facial expressions.
- Understand how humans adapt their voices to different ambiences.
- Taking data-driven approaches to perform classification and generation tasks using context-sensitive machine learning methods and novel data collection protocols.

1.2 Emotion Expression Recognition

Understanding how humans have the ability to move their face to communicate has seen a good deal of discussion in the psychological community [20, 19, 86, 81, 62, 65, 9, 82, 19, 71]. In affective computing, classifiers have been used to "recognize" basic emotions, with little insight into how specific facial movements allow one to express their thoughts and feelings, non-verbally [94, 76, 67, 103, 87]. Some forms of emotion expression recognition rely on a coded system that maps facial landmarks with facial movements. One such coding system is called Facial Action Coding System (FACS) consisting of Action Units (AUs) (Figure 2.2), ranging from 1 to 64, that is commonly used to map facial landmarks to emotions

[99]. For example, [58] used FACS to code the expression of pain for the use of frame-by-frame automatic detection in video. Using a similar idea other emotions, such as anger and happiness, can also be coded on a frame-by-frame basis. Commonly used machine learning models, such as convolutional neural networks (CNN), have been used to extract AUs, using popular image datasets containing human faces, such as AffectNet or FER+ [67, 103, 87]. Using these extracted AUs we can then start to classify an emotion.

Traditionally in psychology understanding how certain AUs are activated for a given emotion in order to categorize extracted AUs with an emotion label is normally done [58, 106, 86, 81, 62]. Another approach, commonly seen in affective computing would be to provide an emotion label first for image data based off how human annotators perceive them and then extract AUs [67, 103, 87]. Combining these two approaches opens the door to allow us to use human perception at the forefront of data collection by understanding how one expresses themselves in a given image and how one perceives an expression to provide an emotion label. In addition, most studies centered around AU activation and emotion expression take on a basic emotion approach, which establishes that all emotions are expressed similarly across cultures. Yet, the basic emotion approach is outdated and will be discussed further in chapter 2 [20, 19, 37]. This puts forward the issue of how appropriate it is to collect data without considering who is in the image. For example, is it preferable to collect data by understanding how one's cultural background could impact how they display emotions in a given image? Furthermore, is training emotion expression recognition systems by feeding data in without considering culture the best method?

1.3 Speech Generation

Gaining insight into how the human voice changes in different ambient settings has also seen little exploration in affective computing, in particular in the area of speech generation. Speech generation has seen steady advancements in recent years and involves a computer synthesizing a voice that sounds human-like [8]. There are several techniques used to conduct such speech synthesis, for example, voice conversion [88] or style transfer [18] to name a few. Many implementations use deep learning models and will be discussed further in Chapter 3.

One example of a readily available and easy to use speech generation system is text-to-speech (TTS), where a user inputs text and a computer synthesizes voice features to generate speech. TTS was first developed in 1939, with the invention of Voder [83], since then TTS has seen significant developments. Currently, many TTS systems use Speech Synthesis Markup Language (SSML) to manipulate how generated voices should sound. For example, to increase loudness of a voice a SSML "volume" element can be used to increase the loudness, or to make pitch higher a SSML "pitch" element can be used to raise the pitch. SSML is simple to use and TTS have also been developed to be expressive

[17, 55]. However, off-the-shelf TTS systems have a limited set of SSML parameters and, like other speech generation implementations, have not investigated ambience adaptation. As such, another issue raised, in addition to those mentioned in section 1.2, is should a data-driven approach be taken to investigate how speech generation should adapt to different ambiences? More specifically, should we understand how humans adapt their voices to a wide range of ambiences first to improve robot speech? In addition, are speech generation systems that adapt to different ambiences considered more socially acceptable versus ones that do not?

1.4 Thesis Outline

As we have seen, culture and ambience provide nuanced social environments for affective computing. Culture and ambience may also prove important when improving classification performance for emotion expression recognition and creating more realistic generated voices that adapt to ambient specific surroundings. As such, addressing culture and ambience as social environments (or context) for machine learning may improve everyday human experiences with affective computing technologies.

In chapter 2, I focus on emotion classification which incorporates a variety of different cultures into emotion expression recognition using a traditionally seen machine learning model. More specifically, chapter 2 investigates how culture influences negative emotion expression for North American (Canadian and American), Persian, and Filipino data in the wild using machine learning.

In chapter 3, I focus on speech generation, more specifically, the use of voice conversion and TTS pitch adaptation. Furthermore, a perception study of different generated voices to use human perception for evaluating speech generative models was conducted.

Chapter 2

Culture and Negative Emotion Expression Recognition

Culture has been identified to play a significant role in our upbringing, influencing the perception and expression of emotional experiences [65, 9, 82]. Culture embeds in humans an innate set of rules to help select the appropriate display of emotions to navigate social environments in one's culture [19]. In the area of affective computing, culture has had some exploration investigating emotion expression recognition technologies role in mental health cross-culturally [77, 78]. Nevertheless, there is still a large body of research regarding emotion expression recognition technologies that treats all cultures equally using a basic emotion, or common view, approach. This approach emphasizes a common underlying structure for each emotion expressed cross-culturally [9]. On the other hand, a social constructivist approach emphasizes individuality of social influence on humans emotion expression and perception [27]. Furthermore, researchers, such as [43], have illustrated that different cultures demonstrate emotions differently. Taking on a basic-emotion approach could impact the performance of emotion expression recognition technologies by treating training data that originates from different cultures, equally. Furthermore, current datasets primarily contain Caucasian and North American training samples or do not categorize by culture by incorporating culture tags during dataset generation [24]. Therefore, these available datasets potentially increase algorithm bias, hindering the inclusion of all cultures [85]. In this chapter, we aim to address the literary gap regarding culture and current emotion expression recognition technologies by:

1. Collecting a novel dataset of negative emotions across 3 different cultures.
2. Using a data-driven analysis of different AUs and their level of activation to investigate culture's influence on social signals associated with negative emotion expression.
3. Evaluating a traditional machine learning algorithm on the aforementioned dataset to understand if adding cultural categories to our data could increase model perfor-

mance, providing a more versatile and culturally-aware emotion expression recognition system.

2.0.1 Cultural Insights into Emotion Mapping

Historically, researchers have found an implicit ability in humans to detect emotions with higher accuracy when interacting with someone who is in their cultural group (in-group advantage)[22, 20]. For example, [22] conducted a meta-analysis that reviewed 87 articles investigating performance of cross-culture vs within-culture emotion expression recognition in humans. They found that although there seemed to be a universal biological core to each emotion, within-group accuracy was slightly higher. Culture can be broken down into individualistic and collectivist subsets which can be ranked using Hoftsede's Individualism (IDV) score by placing cultures on a scale between 1 (purely collectivist) to 100 (purely individualistic) [37]. Individualism characterizes individuals who are loosely tied to others, prioritizing themselves along with immediate family. On the other hand, collectivism characterizes individuals who are tightly linked with others, prioritizing the well-being of the group over themselves. Cultures who have low IDV scores have shown to suppress negative emotions to maintain harmony in social interaction, whereas cultures with high IDV scores idealize self-expression and open communication [25, 82, 37]. Furthermore, one study assessed social robots' cultural sensitivity by using AUs from East Asian culture, as the definition of universality of facial expressions was more tailored to Western cultures[15]. East Asian participants identified the culturally derived expressions with higher recognition accuracy compared to the the existing set of expressions that were placed on a robot's face. Therefore, indicating that there is also cultural discrepancy regarding how one expresses and perceives emotions. In addition, [43] found significant differences in the level of activation of AUs associated with a given emotion when mapping the intensity of each AU across Western Caucasian and East Asian cultures.

2.0.2 Negative Emotion Expression Recognition Performance

Most affective computing research regarding emotion expression recognition technologies focus on 7 distinct emotions (i.e., happiness, sadness, anger, disgust, contempt, fear, and surprise) [53, 84]. There is a tendency to investigate all seven emotions together instead of separating them into positive and negative groups. Although covering more ground by including all emotion categories, these studies can become too broad and overshadow underlying differences of negative emotions. Furthermore, negative emotion expression recognition in both technology and humans has been found to have the worst accuracy discrepancy when performance is compared across different cultures. For example, [15] reviewed 15 different studies and found that overall human participants from western cultures were able to recognize fear, disgust, and anger with high accuracy, whereas other cultures' recognition was much lower. [87] found that using the popular AffectNet dataset, resulted in accuracy be-

tween 45% and 54% for each of contempt, anger, and disgust. Happiness, on the otherhand, had achieved 77% accuracy. In addition, contempt is usually the worst performing negative emotion. When using the FER+ dataset, [87] found contempt could only reach 20% accuracy, whereas happiness reached 95%. [103] also found that not only did contempt perform the worst, with 23% accuracy, but also that it had fewer examples available in emotion datasets. This may be due to differences that occur in the display of contempt, which could be solved when analyzed cross-culturally. As such, although images themselves can be used for classification purposes (e.g., using a CNN), the current study aims to address how AUs vary amongst different cultures and how using AUs as attributes can be used to classify negative emotions in various culture groups.

2.0.3 Negative Emotions and Underlying Social Signals

For the current study, we focused on the emotions of contempt, anger and disgust (the so-called CAD Triad [80]) across three different cultures that range from highly individualistic to highly collectivist. As previously mentioned, negative emotions, in particular contempt, have been hard to predict cross-culturally in both humans and emotion expression recognition technologies [87, 103]. According to the CAD Triad model, each of the mentioned emotions is evoked when one of the three moral codes are violated. Contempt is elicited when community codes are violated, anger is elicited when individual rights are violated, and disgust is elicited when divinity codes are violated causing impurity against oneself, others or God [80]. In the context of the current study, negative emotions will be broken down into their social signals, or AUs. As briefly mentioned in chapter 1, AUs map facial landmarks that are activated when a given emotion is expressed [58, 106]. Disgust, described as a wrinkled nose and mouth, is associated with the activation of AU4 (brow lowerer), AU9 (nose wrinkler), AU10 (upper lip raiser), AU17 (chin raiser), and AU20 (lip stretcher) [86, 81]. Anger, described as protruding teeth and tightened eyes looking downward, is associated with the activation of AU4 (brow lowerer), AU5 (upper lid raiser), and AU27 (mouth stretch). Contempt, shown in Figure 2.1, involves a raised and tightened unilateral lip corner and is associated with the activation of AU4 (brow lowerer), AU7 (lid tightener), AU10 (upper lip raiser), AU25 (lips part), and AU26 (jaw drop) [86, 81, 62]. Using the associated AUs for each emotion we can map their descriptive qualities (e.g., activation of AU9 can be mapped to the wrinkled nose description of disgust). Yet these basic emotion expression templates do not explain the variance of expressions in the wild [9].

Figure 2.1: Prototypical contempt expressions from North American, Persian, and Filipino cultures.

2.1 Dataset

2.1.1 Data Collection Protocol

Our dataset consists of 257 short video clips (between 1 to 11 seconds) collected from YouTube that depicted contempt, anger, and disgust (the CAD Triad) across Canadian and American (High Individualism Score) cultures, along with Persian (Medium Individualism Score) and Filipino (Low Individualism Score) cultures. The Canadian and American groups were categorized together and considered as one culture called North American culture. Video clips in each culture were collected and annotated by three data collectors, one from each culture. Six volunteers in addition to the three data collectors, two from each culture category, added additional annotation for the clips in their identified culture. A clip was labeled with a given emotion label where at least two of the three annotators agreed on the given label, therefore establishing ground truth for each clip. If no annotators agreed on a given label then that clip was held-out of the experiment as ground truth for that clip could not be established. These in-the-wild clips were taken from either professionally acted (e.g., movies or TV shows) or spontaneous (e.g., reality TV or "vlogs") scenarios. Filipino culture contained 74 videos displaying either anger (25), contempt (30), or disgust (19). Persian culture contained 75 videos displaying either anger (27), contempt (28), or disgust (20). Finally, North American culture contained 108 videos displaying either anger (48), contempt (39), or disgust (21). The entire dataset is over a total of 15 minutes and contains 27020 frames.

2.1.2 Dataset Features

After video collection, OpenFace [7] was used to extract social signals (i.e., activation levels of AUs) for each frame in a given video. Only 17 AU attributes were collected (i.e., AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU23, AU25, AU26, and AU45) describing relative values for each AU. Refer to Figure 2.2, which contains AUs

Action Unit (AU)	Description
AU1	Inner Brow Raiser
AU2	Outer Brow Raiser
AU4	Brow Lowerer
AU5	Upper Lid Raiser
AU6	Cheek Raiser
AU7	Lid Tightener
AU9	Nose Wrinkler
AU10	Upper Lip Raiser
AU12	Lip Corner Puller
AU14	Dimpler
AU15	Lip Corner Depressor
AU17	Chin Raiser
AU23	Lip Tightener
AU25	Lips Part
AU26	Jaw Drop
AU45	Blink

Figure 2.2: Collected AUs with their associated definitions.

and their associated definitions. Relative values measure the amount of activation for each AU ranging from 0 to 5 (e.g., high values indicate high activation). Confidence level and success level, which indicate how confident the software is that a given frame successfully exhibited a given set of AUs and the associated level of activation, were also collected. For each frame the originating filename, frame number, and culture were noted indicating origin of each data point in our dataset.

During preprocessing, extra features produced by OpenFace, such as head movement, head position, and gaze direction, were removed. Furthermore, we only chose the frames that had a high confidence level, above 80%, that a given AU was activated. Frames with success that was less than 1 were also removed. In addition, if a clip contained multiple subjects, data for the subjects not displaying the emotions were discarded. Since the available clips for each emotion in each culture had varying amounts, the training dataset for a given emotion category was balanced by randomly removing videos to make each emotion category have the same amount of available training data within a given culture. In addition, the smallest available dataset (i.e., disgust) in each culture is approximately 20 clips. Thus each culture and emotion category will be trained on a similar amount of clips in order to remove bias towards one culture or one emotion. For example, the North American dataset had clips removed randomly for emotion categories contempt and anger, to ensure that they contained the same amount as the smallest emotion category, disgust.

2.2 Experimental Methods

2.2.1 Choice of Model for Classification

Establishing a robust machine learning model to classify emotions is important as it can greatly impact performance outcomes. [94] used a cascade of both a binary and multi-class support vector machine (SVM) to classify emotions. They found that using an SVM for classification outperformed state-of-the-art algorithms, specifically when detecting contempt and disgust. Others have also found success using an SVM in comparison to other commonly used classification methods when recognizing different emotions in older adults and children [59, 69]. There has also been a rise in research that uses SVM in conjunction with deep learning models. For example, [68] used a CNN to extract features in images and then used a SVM to classify images based on the extracted features with 70% to 85% accuracy. Although results using a CNN with a SVM are good, [7] found even more promise using a Convolutional Experts Constrained Local Model (CE-CLM) for the toolkit OpenFace, beating state-of-the-art methods in regards to extracting AUs. Furthermore, given the current study is exploring the differences in the underlying expressions of negative emotions, the additional implementation of a deep learning pipeline could reduce interpretability, as feature embeddings are not as visualizable as activation levels of AUs. As such, OpenFace was used to extract the activation levels for the various AUs in addition to using a SVM for classification, given its previous success on negative emotion expression recognition.

A SVM uses binary classification to separate training vectors along with class labels into two categories using decision boundaries [38]. The training vectors are n-dimensional and in our case we have 16 AUs, therefore, n is 16. The category labels are emotion categories. The SVM then maps the input vectors onto a higher-dimensional feature space. The goal is to optimize the distance between the line that separates the two categories, called a hyperplane, and support vectors that indicate points that are closest to that line. Therefore, optimizing our decision boundary and increasing the separation of categories. The hyperplane is constructed using a kernal function. In the current study, the kernal function is linear and the cost $c = 1.8$, as it introduced the least amount of error during cross-validation. While the linear kernel had the best performance on this dataset, other work on detecting emotions in still images showed better accuracy using quadratic kernels [3]. The SVM and k-fold cross-validation was implemented using an open source python library called Scikit-learn [74].

2.2.2 Experiments

The current chapter has two hypotheses:

1. Cultures express the emotions of the CAD Triad differently, and

2. The CAD Triad cannot be predicted using machine learning models across three different cultures when considering all cultures equivalently.

Differing Expressions of CAD Triad

We first hypothesize that for a given CAD emotion, AU activation will be different for each culture, indicating that social signals used to display the CAD triad are not universal. To investigate this hypothesis, each AU related to an emotion category (i.e. contempt, anger, and disgust) will be analyzed using a one-way analysis of variance (ANOVA). The AUs associated with each emotion will be compared cross-culturally to identify if there is a significant difference in means of AU activation across each culture. Furthermore, a heat map, or activation map, will be created using all AUs for each culture to identify what AUs, with a activation level threshold of at least 2.5, are highly activated for each emotion. Differences in activation maps would further support the premise that AUs activated for each culture are different.

Within-Culture and Cross-Cultural Recognition of CAD Triad

Our second hypothesis is the the CAD Triad cannot be predicted cross-culturally and as such, model performance will be better when cultures are considered separately. We will test our second hypothesis with two experiments, examining: (1) the accuracy when performing within-culture training and testing, and (2) the accuracy when performing cross-cultural training and testing. Both experiments will include each of the three culture datasets: North American, Persian, and Filipino. Both experiments use a SVM as a training and prediction model. Since the dataset is relatively small, a 5-fold cross-validation process will be used. Before feeding data into the two models, around 25% of the clips will be randomly held out for testing to assess the models ability to generalize. Accuracy for each fold will also be calculated and averaged over all folds as an evaluation metric. Both experiment will use prediction performance (i.e., accuracy) to compare cultural categories.

The goal of the first experiment is to assess model's recognition performance of each culture independently, by training an SVM on one culture and testing on the same culture. The comparison between cultures will provide a baseline to assess how well the model can predict using our feature set, independently of other cultures. This experiment will also provide information on the similarity of CAD within a given culture. The second experiment will be to train an SVM on one of the three culture datasets and test on one of the remaining two cultures. The accuracy for each training and testing combination will be compared with one another. We expect that within-culture accuracy will be higher than cross-culture accuracy, indicating a in-group advantage.

ANOVA			Sum of Squares	df	Mean Square	F	Sig.
Contempt	AU4	Between Groups	1449.15	2	724.57	1586.52	<.001***
		Within Groups	3913.96	8570	0.46		
		Total	5363.11	8572			
	AU7	Between Groups	300.07	2	150.03	197.25	<.001***
		Within Groups	6518.55	8570	0.76		
		Total	6818.62	8572			
	AU10	Between Groups	180.20	2	90.10	155.74	<.001***
		Within Groups	4957.93	8570	0.58		
		Total	5138.13	8572			
	AU25	Between Groups	649.54	2	324.77	507.10	<.001***
		Within Groups	5488.59	8570	0.64		
		Total	6138.14	8572			
	AU26	Between Groups	380.52	2	190.26	468.70	<.001***
		Within Groups	3478.83	8570	0.41		
		Total	3859.35	8572			
Disgust	AU4	Between Groups	764.44	2	382.22	466.37	<.001***
		Within Groups	3843.74	4690	0.82		
		Total	4608.18	4692			
	AU9	Between Groups	451.93	2	225.97	601.62	<.001***
		Within Groups	1761.55	4690	0.38		
		Total	2213.48	4692			
	AU10	Between Groups	143.24	2	71.62	85.61	<.001***
		Within Groups	3923.68	4690	0.84		
		Total	4066.93	4692			
	AU17	Between Groups	503.58	2	251.79	521.83	<.001***
		Within Groups	2262.99	4690	0.48		
		Total	2766.57	4692			
Anger	AU4	Between Groups	1065.14	2	532.57	861.68	<.001***
		Within Groups	6140.43	9935	0.62		
		Total	7205.57	9937			
	AU5	Between Groups	717.56	2	358.78	453.49	<.001***
		Within Groups	7859.99	9935	0.79		
		Total	8577.54	9937			

Significance levels are denoted as follows *** α = 0.001, ** α = 0.001, * α = 0.05, . α = 0.1

Table 2.1: Descriptive statistics for the one-way ANOVA for each emotion across all cultures.

2.3 Results

Statistical Analysis. Fleiss' Kappa Inter-rater agreement was 0.83 for the North American dataset, 0.40 for the Filipino dataset, and 0.59 for Persian dataset. An one-way ANOVA (Table 2.1; implemented using Python's SciPy library[102]) was conducted using on each emotion category across each culture category to compare the effect of culture on each emotions' associated level of AU activation. For contempt, analysis of variance demonstrated that cultures effect on AU4 ($F(2, 8570) = 1586.52$, $p < .001$), AU7 ($F(2, 8570) = 197.25$, $p < .001$), AU10 ($F(2, 8570) = 155.74$, $p < .001$), AU25 ($F(2, 8570) = 507.10$, $p < .001$), and AU26 ($F(2, 8570) = 468.70$, $p < .001$), were statistically significant. A Post-hoc Tukey Test (Table 2.2) indicated that each cultural group differed significantly in activation for each AU, $p < .05$, except AU10. AU10 did not significantly differ on North American and Filipino datasets ($p = .49$; Figure 2.3). For assessing cultures effect on anger, a one-way ANOVA was

Post-hoc		Group 1	Group 2	Mean Difference	Standard Error	Sig.	Lower	Upper
Contempt	AU4	North America	Persian	0.88	0.016	<.001***	0.84	0.92
		North America	Philippines	0.14	0.020	<.001***	0.09	0.18
		Persian	Philippines	-0.74	0.020	<.001***	-0.79	-0.70
	AU7	North America	Persian	-0.28	0.021	<.001***	-0.33	-0.23
		North America	Philippines	-0.49	0.026	<.001***	-0.54	-0.43
		Persian	Philippines	-0.20	0.025	<.001***	-0.26	-0.14
	AU10	North America	Persian	0.30	0.018	<.001***	0.26	0.35
		North America	Philippines	0.03	0.022	.49	-0.03	0.08
		Persian	Philippines	-0.28	0.022	<.001***	-0.33	-0.23
	AU25	North America	Persian	-0.61	0.019	<.001***	-0.66	-0.57
		North America	Philippines	-0.43	0.023	<.001***	-0.48	-0.38
		Persian	Philippines	0.19	0.023	<.001***	0.13	0.24
	AU26	North America	Persian	-0.47	0.015	<.001***	-0.51	-0.44
		North America	Philippines	-0.23	0.019	<.001***	-0.27	-0.18
		Persian	Philippines	0.25	0.019	<.001***	0.21	0.29
Disgust	AU4	North America	Persian	0.89	0.031	<.001***	0.82	0.96
		North America	Philippines	0.04	0.033	.49	-0.04	0.11
		Persian	Philippines	-0.85	0.035	<.001***	-0.94	-0.77
	AU9	North America	Persian	-0.61	0.021	<.001***	-0.66	-0.56
		North America	Philippines	-0.64	0.022	<.001***	-0.69	-0.59
		Persian	Philippines	-0.02	0.024	.57	-0.08	0.03
	AU10	North America	Persian	0.20	0.032	<.001***	0.12	0.27
		North America	Philippines	-0.27	0.033	<.001***	-0.35	-0.19
		Persian	Philippines	-0.47	0.036	<.001***	-0.55	-0.38
	AU17	North America	Persian	0.69	0.024	<.001***	0.63	0.74
		North America	Philippines	-0.07	0.025	.02	-0.13	-0.01
		Persian	Philippines	-0.75	0.027	<.001***	-0.82	-0.69
Anger	AU4	North America	Persian	0.84	0.021	<.001***	0.79	0.89
		North America	Philippines	0.22	0.019	<.001***	0.18	0.27
		Persian	Philippines	-0.62	0.019	<.001***	-0.67	-0.57
	AU5	North America	Persian	-0.54	0.024	<.001***	-0.59	-0.48
		North America	Philippines	-0.63	0.022	<.001***	-0.68	-0.58
		Persian	Philippines	-0.09	0.022	<.001***	-0.14	-0.04

Significance levels are denoted as follows *** α = 0.001, ** α = 0.001, * α = 0.05, . α = 0.1

Table 2.2: Descriptive statistics for the Post-hoc Tukey Test for each emotion across all cultures.

applied to AU4 and AU5 independently. An analysis of variance on AU4 and AU5 resulted in significant variation amongst culture, $F(2, 9935) = 861.68$, $p < .001$; $F(2, 9935) = 453.49$, $p < .001$. Post-hoc Tukey test also indicated that each cultural group differed significantly, $p < .001$. Finally, to assess disgust, a one-way ANOVA was applied to AU4, AU9, AU10 and AU17 independently. An analysis of variance on AU4 ($F(2, 4690) = 466.37$, $p < .001$), AU9 ($F(2, 4690) = 601.62$, $p < .001$), AU10 ($F(2, 4690) = 85.61$, $p < .001$), and AU17 ($F(2, 4690) = 521.83$, $p < .001$) resulted in statistically significant variation amongst cultures. A Post-hoc Tukey test also confirmed the statistical significance for each AU and culture, $p < .05$. Except, AU9, $p = .57$, and AU4, $p = .49$, did not have a statistically significant result between the Persian dataset and the Filipino dataset (Figure 2.4) and the North American and the Filipino dataset (Figure 2.5), respectively. It is important to note that some AUs were not available for collection or could not be collected reliably using OpenFace, such as AU20 and AU27 which are associated with disgust and anger, respectively.

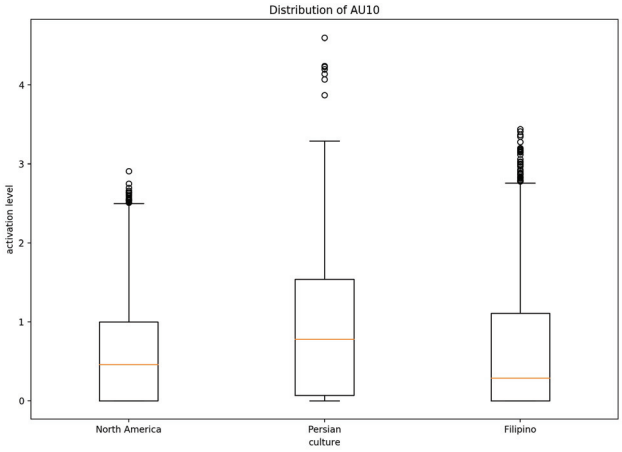

Figure 2.3: Distribution of AU10 cross-culturally for contempt.

Activation Map: Culture-specific social signals of CAD triad. In addition to conducting a one-way ANOVA, an activation map (shown in Figure 2.6; implemented using R) was used to illustrate the activation of all AUs in the datasets for each culture. Each AU is mapped to one or more 2-dimensional facial points as proposed by [30]. A threshold of at least 2.5 selects AUs that are activated. To simulate the facial movement captured by AUs, we added random variance in the direction of movement and a small amount of noise to produce smoother density plots. The activation map indicates which AUs are more likely to be activated for a given culture and emotion as compared to other AUs within that same culture and emotion.

2.3.1 Data Analysis

Contempt. North American contempt shows AU25 (lips part), AU2 (outer brow raiser), AU45 (blink), AU7 (lid tightener) and AU12 (lip corner puller) to be highly activated. Persian contempt shows the AUs normally found to be highly activated with contempt, AU4 (brow lowerer), AU10 (upper lip raiser), AU1 (inner brow raiser), AU15 (lip corner depressor) and AU25 (lips part). Filipino contempt has both AU1 (inner brow raiser), AU2 (outer brow raiser), and AU25 (lips part) highly activated. Figure 2.7 shows an example of the underlying differences in each culture for contempt given the mentioned activated AUs.

Anger. North American anger is reflected by AU5 (Upper Lid Raiser), AU1 (Inner Brow Raiser), AU25 (lips part) and AU10 (upper lip raiser). Persian anger has AU4 (brow lowerer), AU10 (upper lip raiser), AU1 (inner brow raiser) and AU25 (lips part) highly activated. In terms of Anger, Filipino anger shows high activation of AU1 (inner brow

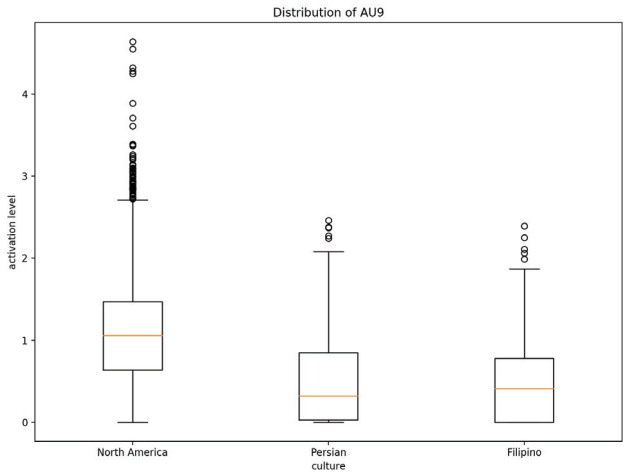

Figure 2.4: Distribution of AU9 cross-culturally for disgust.

raiser), AU25 (lips part), AU4 (brow lowerer), AU10 (upper lip raiser) and AU26 (jaw drop).

Disgust. North American disgust has AU7 (lid tightener), AU25 (lips part), AU9 (nose wrinkler), and AU10 (upper lip raiser) highly activated. Persian disgust has high activation of AU4 (brow lowerer), AU10 (upper lip raiser), AU1 (inner brow raiser), AU15 (lip corner depressor) and AU17 (chin raiser). Finally, Filipino disgust has high activation of AU4 (brow lowerer), AU10 (upper lip raiser), AU1 (inner brow raiser), AU26 (jaw drop), AU25 (lips part), AU7 (lid tightener).

2.3.2 Experimental Results

Within-culture testing accuracy. Figure 2.8 shows the confusion matrix for within-culture experiments (the diagonal). Testing accuracy (i.e., percentage of correct emotion labels) for the North American dataset was the highest with 66% accuracy. The next highest testing accuracy was the Filipino dataset with 39%, followed by the Persian dataset, with 37%. This decreasing trend in recognition accuracy is similar to the agreement scores.

Within-culture confusions. Both disgust and contempt for the within-culture experiment on the North American dataset (Figure 2.9) had the highest accuracy with 72% and 69%, respectively. Anger performed the worst with 45% accuracy and was incorrectly classified as disgust 34% of the time. Using the Persian dataset (Figure 2.10), anger performed best with 43%. Anger was also confused 38% of the time with contempt. Disgust was commonly confused with anger, with 41% of the disgust clips being incorrectly labelled as anger and only 26% of disgust clips being correctly classified. On the other hand, contempt was split between either being correctly classified 39% of the time or misclassified, 36% of

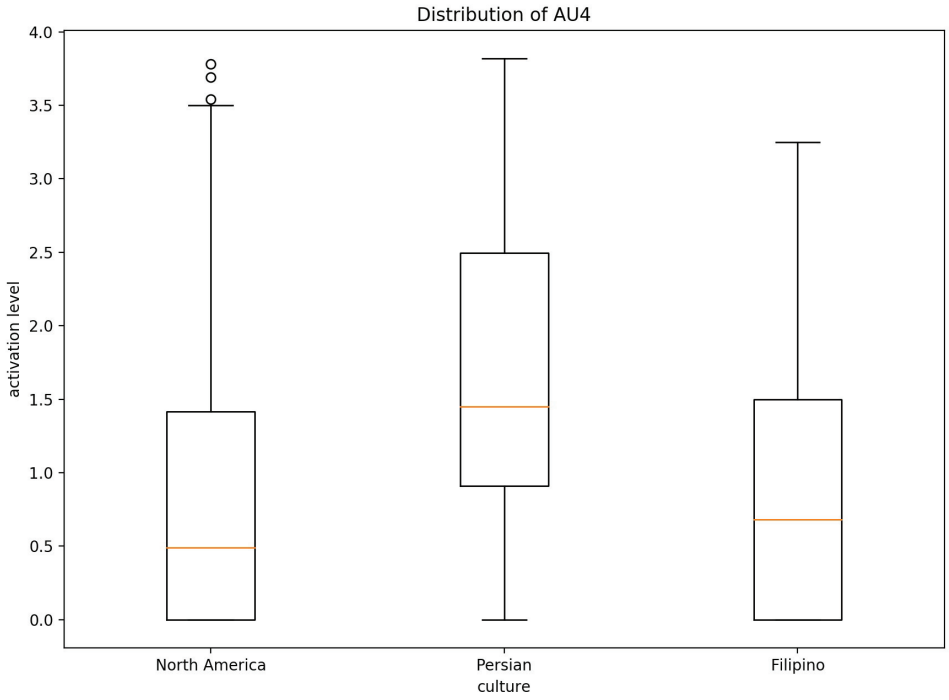

Figure 2.5: Distribution of AU4 cross-culturally for disgust.

the time, with anger. Finally, the within-culture Filipino dataset (Figure 2.11) performance was highest for anger with 55% accuracy. However, contempt and disgust were frequently mistaken for anger with 51% of contempt and 72% of disgust being mistaken as anger for the Filipino dataset.

Cross-culture confusions. The North American dataset achieved the best test results in the cross-culture experiment (Figure 2.8) with training on Persian with approximately 47% accuracy and training on Filipino with approximately 46% accuracy. The Persian dataset for the cross-culture experiment performed similarly to the within-culture experiment with 30% accuracy when trained on the Filipino dataset and 40% accuracy when trained on the North American dataset. When trained on the Persian dataset, the Filipino dataset resulted in a small increase of 40% accuracy, while 36% accuracy was present when trained on the North American dataset for the cross-culture experiment.

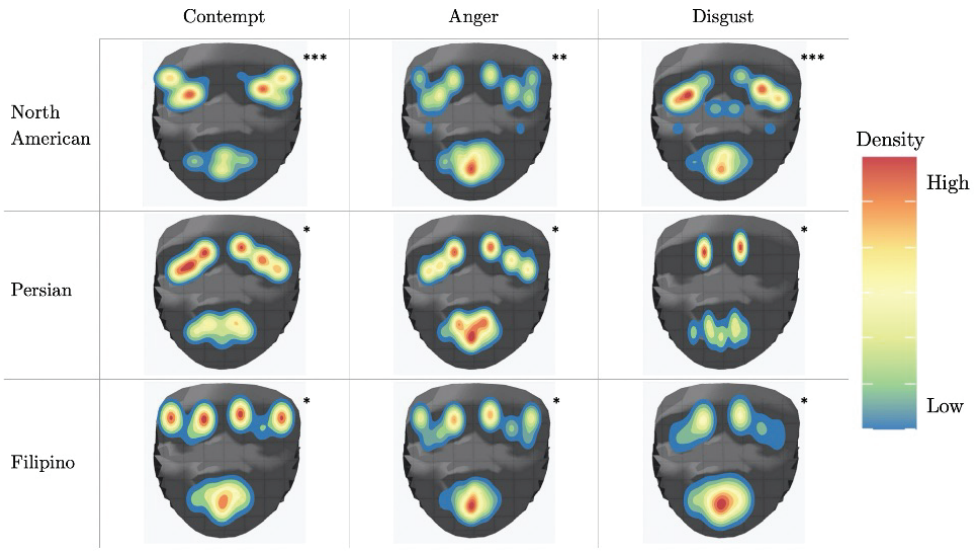

Figure 2.6: AU activation map that shows the localization of activation using a threshold of 2.5. *Movie/TV shows as primary source ; ** mix of Movie/TV shows, Reality TV and Vlogs as primary source; ***Reality TV and Vlogs as primary source.

Figure 2.7: Examples showing the differences in underlying social signals of contempt.

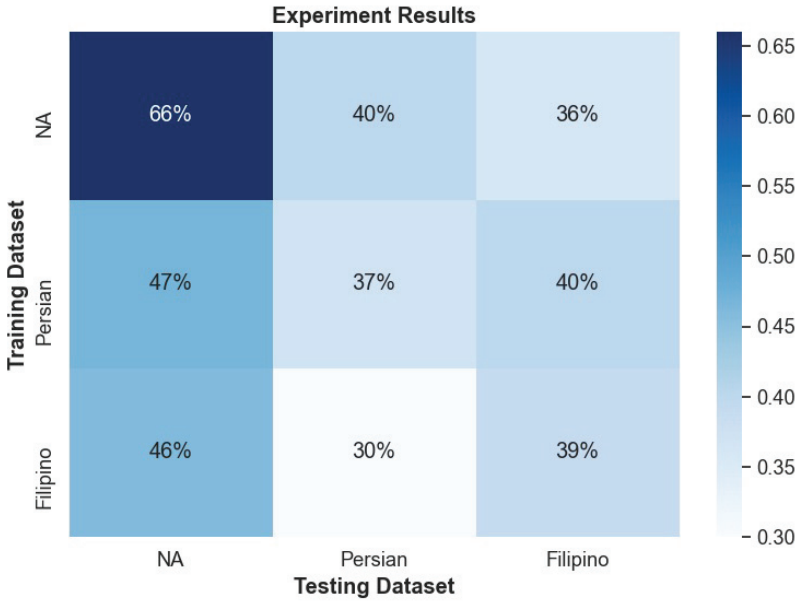

Figure 2.8: Confusion matrix illustrating the results of all within-culture experiments.

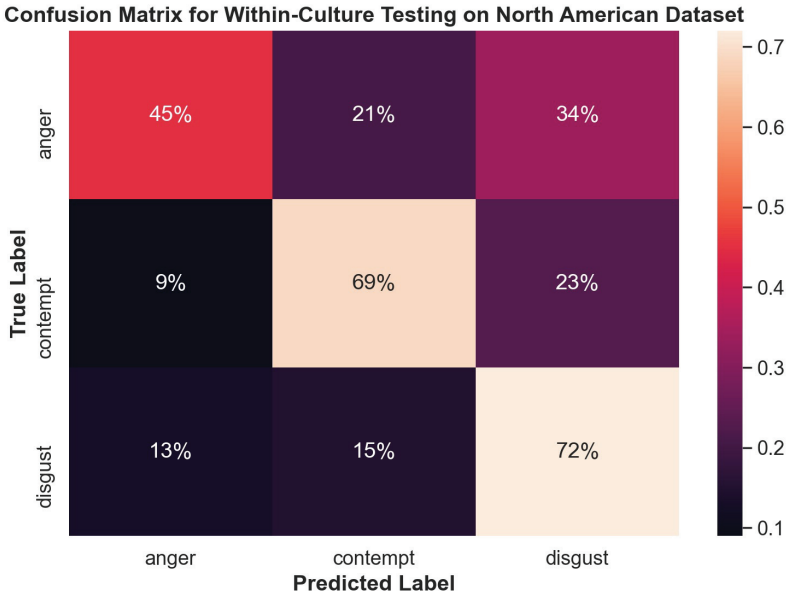

Figure 2.9: Confusion matrix illustrating the results of the within-culture experiment on the North American dataset.

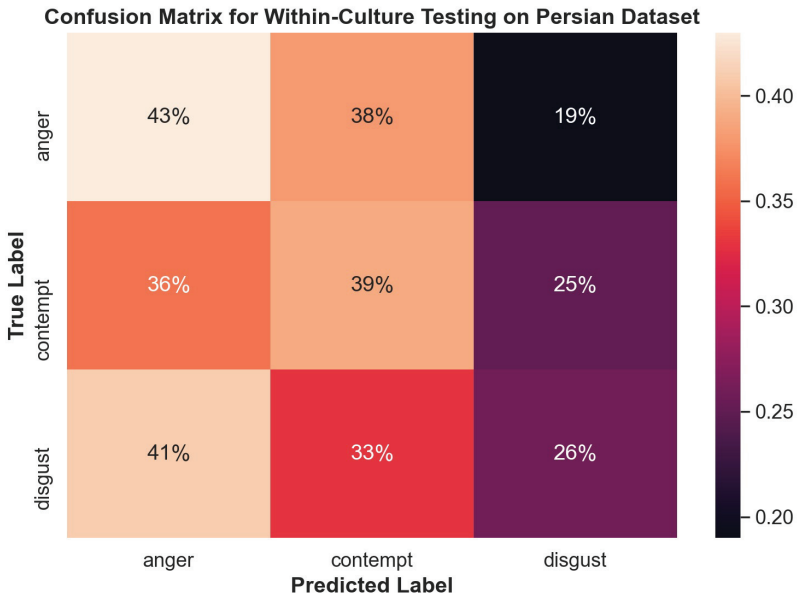

Figure 2.10: Confusion matrix illustrating the results of the within-culture experiment on the Persian dataset.

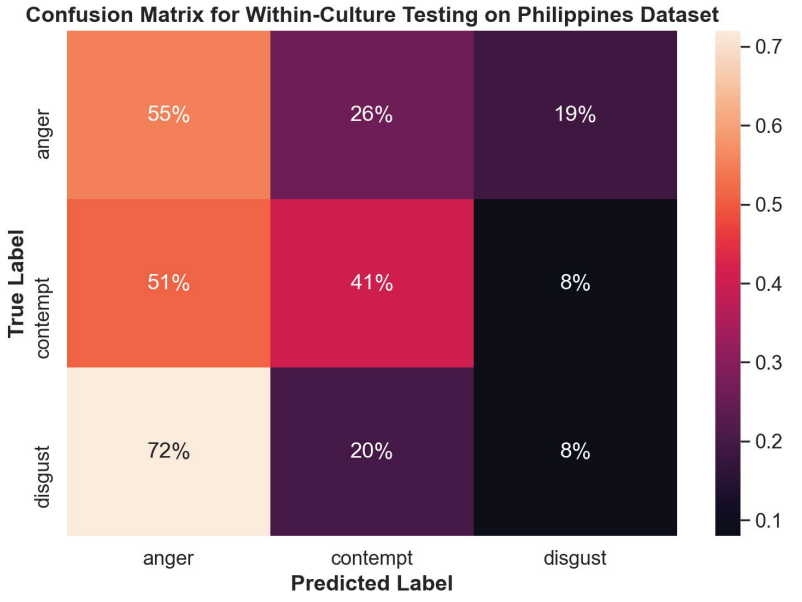

Figure 2.11: Confusion matrix illustrating the results of the within-culture experiment on the Filipino dataset.

Chapter 3

Ambient Adaptive Speech Generation

Humans have the innate ability to adapt their voice to different contexts and social situations. Although we consider linguistic vocal phenomenon to be our primary means of communication amongst humans, a significant portion of our communication results from non linguistic vocal features that can completely alter the meaning of phrases [60]. For example, someone may say "it's over." at the end of a vacation with a sense of sadness, or they may say "It's over!" with a sense of enthusiasm after completing an examination they have been studying for. Without the ability to produce and understand these contextual non-linguistic vocal features we would have a difficult time navigating everyday life. Voice is important and it solidifies trust in not only human-to-human interaction but also human-to-robot interaction [23]. Given the features associated with our speech are important in communicating in different contexts, it is important for robot's to be able to communicate in a similar fashion.

Providing a robot with the opportunity to adapt their voice to different contexts and social situations has not yet been thoroughly explored. Robots are used every day in vastly different contexts, therefore, the need for a robot to successfully adapt itself into both the ambience and the social environment proves important when integrating robots into humans' everyday lives [98, 42, 34]. State-of-the-art social robots, such as Pepper, Nao and iCub, utilize expressive voices custom built by companies, such as Acapela [1]. These voices are exceedingly expensive and resource intensive to generate, and as such, are often outside the means of individuals and small scale companies developing interactive robots. In these cases the developers often rely on widely available text-to-speech (TTS) services that only allow minor adjustments and voice selections.

Our solution is to use a data-driven approach to generate a robot's voice. For example, using a corpus of human voices that are collected in different ambiences, we use human

[1] https://www.acapela-group.com/

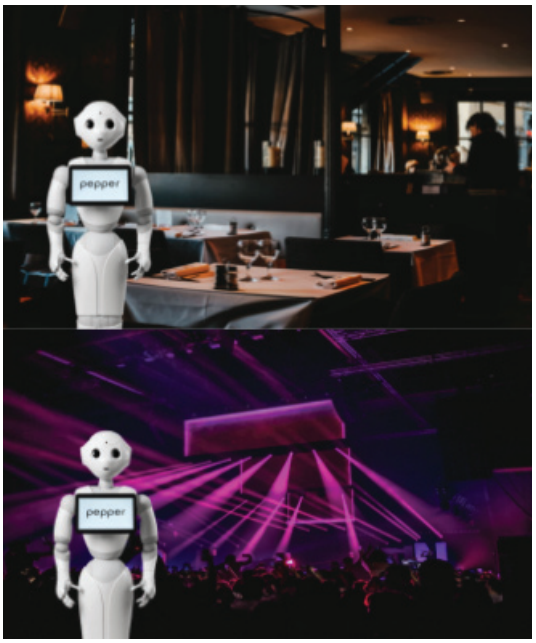

Figure 3.1: A robot in a fine dining restaurant vs. a night club should adapt its voice to the ambience.

speech features to modify those of a robot's voice. However, the issue with this method is that it is difficult to collect clean recordings of realistic data in noisy and crowded environments. The current study plans to circumvent these challenges by utilizing a readily available video-conferencing platform. Placing participants into a simulated environment that is in the comfort of their own home opens the door to easily collect naturalistic data more efficiently. As such, in this chapter we hope to bridge the following gaps in the literature:

1. Implementing a novel protocol for collecting realistic contextual audio voice data
2. Investigating human voice adaptation to determine relevant features that can improve robot voices in different ambient environments
3. Testing human perception to better understand how humans perceive robot voices, in particular: (a) Comparing baseline TTS, adaptive TTS, voice conversion, and human voices, (b) How humans perceive voice conversion as adaptive to the environment, and (c) How humans perceive pitch in TTS against a common social environment

3.0.1 Human Contextual Vocal Modification

Most often human vocal modifications are for the purposes of creating "deliberately clear speech," when the listener is experiencing reduced comprehension [10]. These modifications may occur due to auditory hindrance, as is the case when a speaker is distant [33] or when

in a noisy environment [33]. For example, one of the most well researched and understood vocal phenomena is the Lombard effect [56]. The Lombard effect is an involuntary increase in vocal effort, often due to the presence of background noise in order to gain clarity in noisy environments [33]. Modifications may also be targeted to the listener, such as in speech directed towards infants and children [11], hearing impaired [50], and machines [63]. In some cases speech is not modified for clarity, but rather to communicate a specific intent, such as politeness [12]. For example, modifications are made to get an infant's attention not to increase clarity [16]. Altogether, vocal modifications are produced with or without conscious effort to elicit a specific auditory feature as a result of achieving the aforementioned goals. Although it is well understood that humans produce these vocal phenomena in response to social environments, the reproduction of these effects towards ambience in generative speech is relatively new and sparsely studied.

3.0.2 Current State-of-the-Art Methods for Speech Generation

Text-to-Speech. As previously mentioned, TTS has become an inexpensive and efficient means to create realistic voices for the purpose of simulating robotic behaviors [17, 101, 72]. Currently, state-of-the-art speech generation tools for TTS use signal processing algorithms in order to reconstruct speech (e.g., vocoders). For example WaveNet, uses speech samples to synthesize wave forms for speech reconstruction [88]. Tacotron, another popular method of speech generation, uses spectrogram synthesis to reconstruct speech. Moreover, Tacotron-GST's synthesized speech has the capability to express basic emotions [90, 55, 91, 5]. Companies like Google[2], Amazon[3], and Microsoft[4] all have their own variations of these vocoders.

Nevertheless, TTS has multiple shortcomings. Firstly, for many years the primary concern for TTS was intelligibility; this resulted in voices being produced by the state-of-the-art that can be mistaken for a human voice. TTS is also traditionally considered flat and inexpressive. Although, TTS has become more expressive in recent years [55], features such as rate-of-speech, are still manipulated linearly, extending vowels, words and pauses evenly throughout an utterance [61]. Furthermore, TTS still does not have the ability to adapt to both physical and social environments. This is primarily due to TTS being a rule-based system and as such, as mentioned in chapter 1, is constrained by SSML. Although the available features have broadened and include loudness, pitch, and rate-of-speech [5], it is not clear whether this system of speech generation is sufficient for a robot to flexibly and automatically adapt its voice to different ambiences.

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

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

[4]https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/

[5]https://cloud.google.com/text-to-speech/docs/ssml

Voice Conversion. Voice conversion is a method whereby a source speaker's speech waveforms are adjusted to that of a target speaker [18], as such allowing for the modification of source speaker's auditory features to match that of a target speaker [108]. The traditional method of voice conversion takes on a one-to-one approach, which involves using a single target speaker and a single source speaker. This method has provided promising results, using Deep Mind's WaveNet vocoder in generating a personalized voice with independent speakers [88]. In addition, the use of parallel data has also been the traditional norm, where utterances and linguistic information must be identical between source and target speaker. It, however, has become increasingly more popular to use non-parallel speech, which has seen the establishment of useful benchmarks and protocols from Voice Conversion Challenge (VCC) 2016 [88, 46]. Non-parallel speech allows for us to keep intact the underlying linguistic information (e.g., words) of the source utterances without restricting the target data to be the same duration and contain the same utterances as the source [36].

A popular method for voice conversion is to make use of statistical methods, such as the commonly employed Gaussian Mixture Models (GMM) for parallel voice conversion tasks [18]. With the recent advancements of deep learning models, specifically Generative Adversarial Networks (GAN), non-parallel voice conversion has seen significant improvements. In addition, there has been heightened attention to supplementing GANs with auto-encoders, specifically variational auto-encoders (VAE) [88, 109, 88, 36]. Supplementing GANs with VAEs has provided improvements in not only non-parallel voice conversion, but cross-lingual voice conversion tasks [89]. For example, allowing the voice of an native-english source speaker to have the same auditory speech features of a native-german target speaker. Given the success of parametric statistical models it is also no surprise that utilizing vector quantizations, such as Vector-Quantized Variational Autoencoders (VQ-VAE), has improved voice conversion output. VQ-VAE's have seen a rise in popularity for speech generation, specifically voice conversion tasks, and have comparable results to state-of-the-art speech generation tools, such as TacoTron [88, 107, 104]. Due to its success, voice conversion may prove versatile enough to produce Generative speech that is adaptive to a surrounding ambience.

3.0.3 Voices in Robotics

A recent survey of robotics research found that the vast majority choose robot voices by convenience rather than considering contextual information to adjust auditory features [64]. This approach to voice generation can be problematic as it has been continually agreed upon that first impressions with a robot will determine the course of the user experience [97, 51, 100, 44]. Furthermore, several studies have shown that humans show preferences for voice depending on context and task [98, 61, 61, 73, 39]. In [98], participants rated the appropriateness of robotic speech given different contexts including schools, restaurants, homes and hospitals. They found that even given the same physical appearance, participants

selected varying voices depending on context and concluded that a robot voice created for a specific context is likely not generalizable.

Some studies have suggested the incorporation of context based methods such as sociophonetic inspired design [93] and acoustic-prosodic adaption to match user pitch [57]. Further research has also made an attempt to produce the Lombard effect, with research relying on incremental adaptation of loudness given the distance and target user [26] or the adjustment of volume based on environmental noise levels [32]. Although these methods have shown promising results, few are readily available for general use. Even curated voices of state-of-the-art robots are not necessarily perceived as appropriate in all contexts [98, 64]. As such, it is important to develop robotic voices that are context-specific, yet readily available and generalizable.

3.1 Dataset

Due to the trying times of the global pandemic we have created a pioneering method of virtual data collection using readily available tools that will allow researchers to collect data with no physical human interaction. Our dataset contains speech utterances and extracted vocal features from 12 participants.

3.1.1 Data Collection Protocol

During the pandemic, the ability to interact one-on-one in a public area became difficult, and was prohibited by governmental restrictions in several countries across the world. As such, it has become particularly difficult for researchers to conduct field studies involving human participants. One of the novelties of the current study was how we overcame this issue by devising a protocol mimicking a naturalistic environment over Zoom[6].

Zoom is a teleconferencing service which allows individuals to communicate from anywhere in the world online using both face and audio. Using Zoom, we paired two participants and had them listen to ambient sounds while conversing with one another in the roles of a waiter and a restaurant-goer. There were a total of 6 ambient sounds and 1 additional baseline measure that included no sound. The baseline condition was placed between a randomly chosen pair of ambience conditions. This baseline condition was used to reduce carry-over from the previous condition and obtain speakers baseline levels. In addition to sound, participants were asked to change their Zoom background to an image that was pre-selected to match the given ambience (see Figure 3.2). Between each ambient condition there was a 1 minute period to update participants' Zoom backgrounds and prepare for the next condition serving as a washout to reduce carry-over effect from the previous condition. Each ambience condition was further broken down into 2 subsets: (1) scripted and

[6]www.zoom.us

(2) unscripted; the assigned scripted roles were maintained for the unscripted condition. The restaurant ambiences included: fine dining, café, lively restaurant, quiet bar, noisy bar, and night club. The baseline condition was a bakery with no sound or image. The ambient sounds can be listened to here[7].

Scripted condition. Participants first read a brief summary of their character at a specific restaurant. For example, in the fine dining condition the restaurant-goer was on a date, while at the noisy bar the restaurant-goer was with a group of friends to watch the Stanley Cup finals. Once each participant read the summary for their character, they then read from a script that was slightly tailored for the given ambience; food and drink choices matched what is usually offered at that given restaurant. Consistency amongst scripts allowed for comparison of speech features across each condition. The differences between scripts created a more realistic environment and reduced redundancy to maintain participant attention.

Unscripted condition. Participants were then told to remain in character and proceed with the initial scenario description of a waiter taking a customer's order, but this time there was no script to read. This allowed participants to be more authentic as they were free to adapt other features to the ambience, such as their choice of words.

3.1.2 Dataset Demographics

The dataset consisted of 8 females and 4 males. There were a total of 1545 female utterances, here defined as a single sentence, and 796 male utterances. We used only the female speakers for the current study. Altogether there were 2341 clips ranging from 1 to 7 seconds. All recruited participants were undergraduate students at Simon Fraser University who had either a customer service background, experience in improv, or experience in theater.

3.1.3 Dataset Features

Clips were first converted to a monophonic channel and sampling frequency was set to 24000. We collected 10 features, which can be broken up into (1) loudness features, (2) spectral features, including pitch, and (3) rate-of-speech features. Our toolbox for vocal feature extraction can be found here.

Loudness Features

Humans have shown that they increase and project their voice in loud environments in order to increase quality and sound clearer [56, 92, 75]. Increase of vocal intensity, often leading to Lombard speech, is commonly employed in noisy environments [13]. As such, we collected 3 loudness features: (a) mean intensity, (b) energy, (c) maximum intensity. Mean

[7] https://ehughson.github.io/ambiance.github.io/

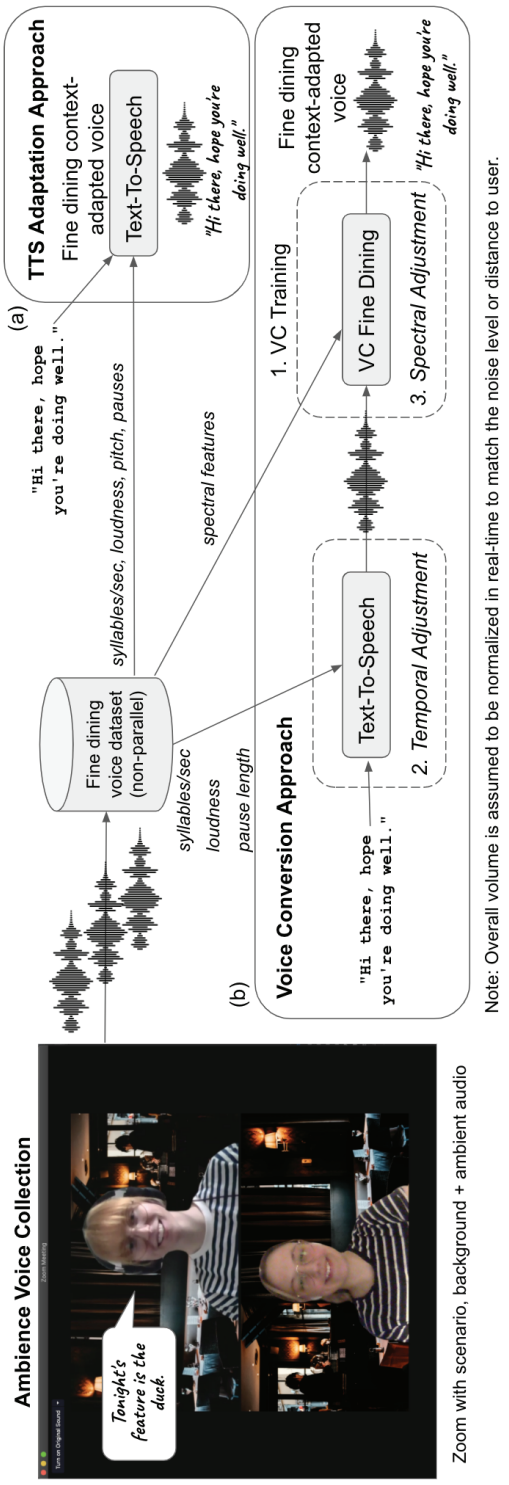

Figure 3.2: Left: Overview of data collection using Zoom: participants' backgrounds and shared audio match the current ambient condition. Right: Ambience voice adaptation approaches for a given ambience: (a) TTS Adaptation using temporal and pitch features, or (b) Voice conversion for spectral features.

intensity and energy features were calculated using the Praat [8] library via Parselmouth[9], a Python wrapper. Librosa [10] was used to extract the sound wave to calculate maximum intensity (power) following the formula provided in Section 1.3.3. of [66].

Spectral Features

We expect that humans adapt spectral features in their voice to both social and ambience specific contexts. For example, average pitch tends to increase with vocal effort [92], often as a result of a loud environment or as a sign of politeness[12].

To investigate these expectations we collected 5 spectral features: (a) median pitch, (b) pitch range, (c) shimmer, (d) jitter, and (e) spectral slope. Parselmouth was used to extract (a)-(d). Median pitch and pitch range (the difference between the minimum and maximum pitch in a given segment) are calculated in Hz after removing silences and obtaining pitch values from voiced utterances. Local shimmer and local jitter, variations in the fundamental frequency, are perceived as vocal fry and hoarseness. Spectral slope gives an indication of the slope of the harmonic spectra. For instance a -12dB slope is typical for breathy voice, and a -3dB slope is typical of richer vocal tones [40]. Spectral slope was calculated with the use of Parselmouth and Librosa given the formula provided in section 3.3.6 of [52].

Rate-of-Speech Features

Research has pointed to the decrease in speaking rate when intelligibility becomes increasingly important [47], for example, in loud environments. We also posit that formal speech may be slower and clearer than informal speech. Subsequently, we collected 3 rate-of-speech features: (a) voiced (silences removed) syllables per second (b) overall (silences are not removed) syllables per second, and (c) pause rate. We employed syllables per second by taking the ratio of number of syllables over the duration(s) for both the voiced and overall utterances. We defined pause rate as the number of pauses, where a pause is defined as a silence of at least 50 ms between words, over the duration of a entire utterance (both voiced and overall components). Including the syllables per second for overall utterances along with pause rate may provide information regarding length of pauses and how pauses may impact the length of a utterance [61].

[8] http://www.praat.org/

[9] https://parselmouth.readthedocs.io/en/stable/

[10] https://librosa.org/doc/main/index.html

3.2 Experimental Methods

3.2.1 Voice Conversion Implementation

CRANK is a voice conversion software that implements several variations of a VQ-VAE to have a source speakers' speech match that of a target speaker [46]. For the current project, we used CRANK's best performing model from [46], which is the Cycle VQ-VAE GAN with Short Term Fourier Transform (STFT) loss with a speaker adversarial network and discriminator. The VQ-VAE implementation was a hierarchical implementation, similar to that of DeepMind's WaveNet architecture. The Cycle VQ-VAE GAN is a least squares GAN (LSGAN) with a cyclic VQ-VAE. 400,000 training epochs were performed and the adversarial training for the discriminator was initiated at 200,000 epochs. In addition, training settings provided in "mlfb_vqvae_24000.yml" and "default.yml" contained in CRANK's template folder were used. Input features provided for CRANK are MLFB, pitch, aperiodicity and spectrum input [46]. CRANK also uses training speakers for evaluation and development, called a speaker-closed condition. A pretrained vocoder for speech reconstruction was trained on the LJ speech dataset [41], which contains entirely female speakers, and is implemented using the Parallel WaveGAN vocoder repository [105].

3.2.2 Pipeline

Our experimental pipeline can be found in Figure 3.2. We tested several data-driven approaches, including (1) TTS Adaptation based on speech rate, loudness, pause length and pitch, and (2) Voice Conversion, first setting the TTS speech rate, loudness, and pause length, then adjusting the voice's spectral components using voice conversion.

TTS Adaptation Approach

Our TTS samples used for the perception study contained 6 TTS voices selected from a set of 9 samples. We generated these samples using Google TTS [11], which allows for the overall manipulation of (a) loudness, (b) rate-of-speech, and (c) pitch. The first sample was a baseline TTS, which has no alterations of the stated SSML elements (TTS-bl). The next sample was a set of 6 TTS voices that were generated by setting the SSML elements to the average features of all female speakers for each ambience (TTS-avg). Finally, two TTS samples (TTS-low and TTS-high) were generated with matching loudness and rate-of-speech elements of TTS-avg but differing levels of the pitch. TTS-low had the pitch element set to one specific speakers' pitch, a female undergraduate student from our dataset (714). TTS-low's pitch sat between the pitch of TTS-bl and TTS-avg. TTS-high's pitch element was set to $pitch(\text{TTS-avg}) + (pitch(\text{TTS-avg}) - pitch(\text{TTS-low}))$. We narrowed in on pitch

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

alteration as the literature has focused more on loudness and rate-of-speech in different contexts (e.g., Lombard speech). The utterance used for all perception experiments was, "Hi there, I hope you are doing well".

Voice Conversion Approach

Our voice conversion approach involved two main steps: temporal adjustment, followed by spectral adjustment. The initial procedures involved separating the human speakers by speaker ID then further separating each speaker's clips into each of the 6 ambience conditions (speaker-ambience sample). Speaker-ambience training batches contained between 16 to 41 utterances. This was in line with the dataset setup used to train CRANK's cycle VQ-VAE GAN [46]. This allows speakers to be independent of one another and to demonstrate a given ambience's effect on voice conversion at an individual level. Here, we use TTS as the source speaker, and all female speaker-ambience samples in our dataset as the target speakers.

Temporal Adjustment. Initially, TTS-avg was selected as the source speaker. However, given the perceptually significant differences in voice characteristics (e.g. natural pitch) between subjects, it was deemed more appropriate to manipulate the source speaker TTS to match that of a individual speaker. As such, we used the features of speaker 714 to generate our source speaker training samples (TTS-714). The ambience specific pre-processing of TTS-714 was integral to the project as voice conversion software primarily uses human speech utterances as source speakers, whereas we are using a generated robot voice. CRANK's Cycle VQ-VAE GAN [46] primarily adjusts spectral features, therefore, non-spectral features (i.e., loudness, rate-of-speech and pause rate) were added before the voice conversion process. This resulted in 211 temporally adjusted TTS-714 clips containing speech utterances of the scripted portions for both waiter and restaurant-goer roles for each ambience condition.

Spectral Adjustment. We used TTS-714 clips, alongside all female speaker voice clips to train CRANK's Cycle VQ-VAE GAN model for each ambience. Because we used a non-parallel model we were able to use unscripted data, which is not present in the TTS samples, as a target. This means that utterances can be generated that have never been heard by the trained model before, further adding socially and contextually appropriate data to our speaker-ambience training batches. Finally, the waiter utterances for our source TTS were held out of training to be used as evaluation samples. We then used 6 generated audio samples, 1 for each ambience, for the perception study. Generated samples where the target was speaker 714 was used for the perception study.

Post-processing. Voice conversion using CRANK resulted in samples that were slowed down significantly compared to the original source and target speakers' rate-of-speech. This appeared to be due to the presence of TTS voices, as the slowed down effect was not as extreme in human-to-human voice conversion samples. As such, the audio samples generated

from our voice conversion approach were sped up using Audacity's tempo change function, which maintains spectral envelope and pitch, to match the rate-of-speech of the original human speaker, 714. In addition, the TTS samples and voice conversion samples were set to -10.0 dBFS as to be normalized against the background sound to compare the voice quality and rate-of-speech only.

3.2.3 Perception Study

The perception study leveraged Mechanical Turk and Survey Monkey with 25 Canadian participants who were fluent English speakers, with 100 Human Intelligence Tasks (HITs) completed with a 98% acceptance rate. There were 4 research questions explored:

1. How do generated voices compare to human voices within ambiences? (3a of research goals)
2. How are voice conversion generated voices perceived in context? (3b of research goals)
3. How are voice conversion generated voices perceived when paired with the incorrect ambience? (3b of research goals)
4. How does a data-driven pitch manipulation for TTS impact human perception? (3c of research goals)

Listeners were first asked to use headphones and calibrate their audio. Next for each of the above research questions participants were told the provided audio sample was the voice of Pepper the robot, who was about to take their order at one of the 6 given ambience locations. They were given the same scenario description provided during dataset collection. For example, "You are a customer dining at the city's fanciest restaurant. The atmosphere is warm, the music is slow and romantic, and the lights are dimmed. You have waited months to take your date out to this particular restaurant. In hopes to impress your date you wish to get the duck, the restaurants staple item. Pepper, the robot, is going to take your order."

After listening to Peppers' voice over the background sound, participants were asked to respond to 7 statements using a 7-point likert-scale ranging from 1 (strong disagree) to 7 (strongly agree). The following were the statements provided: (1) Pepper's voice sounds socially appropriate for the scene (Figure 3.1), (2) Pepper's voice sounds robotic, (3) Pepper is aware of the surrounding ambience, (4) Pepper makes me feel comfortable, (5) Pepper makes me feel like I am in the given ambience location, (6) Pepper is too loud, and (7) Pepper is too quiet.

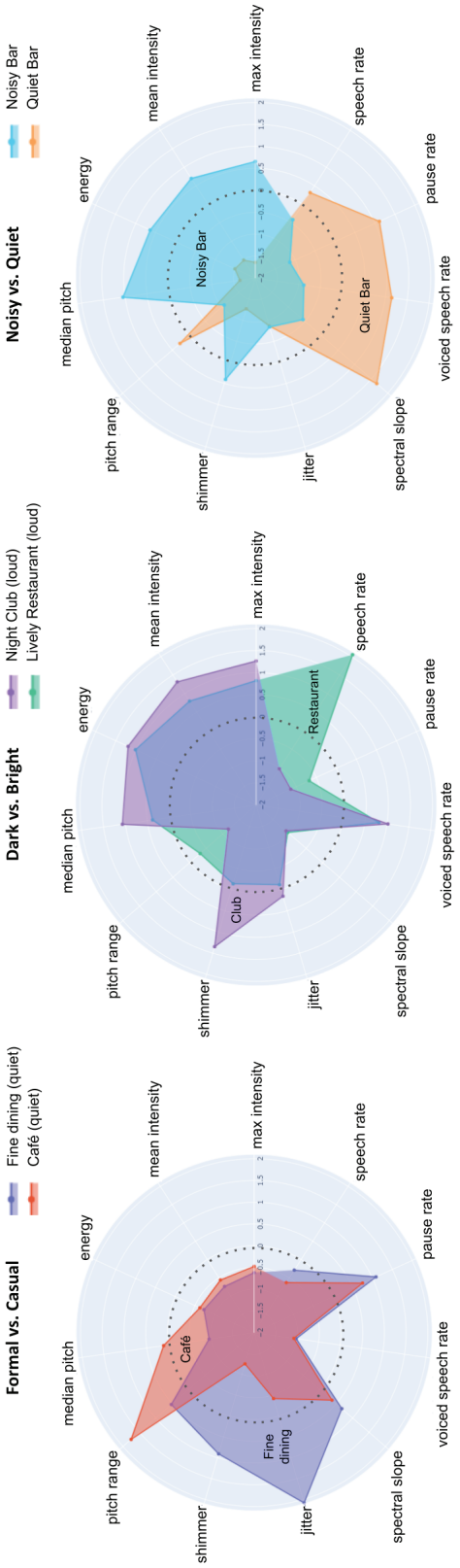

Figure 3.3: Radar graphs showing the differences amongst collected features across all female speakers in our dataset. The dotted circle represents the baseline ambience for all features.

rANOVA	DFn*	DFd*	F	P-value	sig.	Effect Size
Energy	5	35	9.97	< .001	***	.14
Mean Intensity	5	35	15.78	< .001	***	.31
Max Intensity	5	35	12.47	< .001	***	.28
Median Pitch	5	35	2.37	.06	.	.19
Pitch Range	5	35	.96	.46		.08
Shimmer	5	35	1.36	.26		.05
Jitter	5	35	1.01	.42		.06
Spectral Slope	5	35	9.43	< .001	***	.16
Speech Rate (voiced)	5	35	1.61	.18		.06
Speech Rate (overall)	5	35	1.93	.12		.09
Pause Rate	5	35	4.93	< .001	***	.23

*DFn is Degrees of Freedom of numerator and DFd is Degrees of Freedom of denominator to calculate F-statistic.
Significance levels are denoted as follows *** $\alpha = 0.001$, ** $\alpha = 0.001$, * $\alpha = 0.05$, . $\alpha = 0.1$

Table 3.1: Summary of statistical results of all features using a rANOVA.

3.3 Results

3.3.1 Trends in Collected Data

A subset of the overall dataset is displayed in Figure 3.3. These radar plots illustrate various features present in different ambient social environments in our underlying dataset. The dotted circle indicates average baseline features. They are based on the polarity of the two chosen ambient conditions. For example, the first radar plot shows the differences between two ambiences at a similar noise level but with differing social environments (i.e., fine dining (formal) and café (casual)). Loudness related features, in this condition, appear to be similar and occur at low levels. Indicating that these two ambiences appear to be on the more quieter side, which is unsurprising as both these ambiences are quiet at -34 dBFS and -42 dBFS, respectively. Spectral features appear notably different. Pitch range and median pitch are increased for the café ambience, while jitter and shimmer are increased for the fine dining ambience. Therefore, suggesting that in the casual setting of a café, a wider pitch range and higher median pitch may be preferred. Although spectral features, such as shimmer and pitch range, increased in the fine dining ambience, median pitch did not. This is surprising as a increase in median pitch was expected from a formal context not a casual one [12].

For the second radar graph, two similarly loud ambiences were compared. Participants experienced a bright, lively restaurant ambience with family-style polka music and trumpet

Pairwise T-tests	Group 1	Group 2	n1	n2	F	df	p-value	Adj. p*	Adj p sig.
energy	Fine Dining	Lively Res.	8	8	-3.98	7	.005	.08	.
	Lively Res.	Quiet Bar	8	8	4.57	7	.003	.04	*
	Quiet Bar	Noisy Bar	8	8	-5.48	7	< 0.001	.01	*
Mean Intensity	Fine Dining	Lively Res.	8	8	-4.21	7	.004	.06	.
	Cafe	Lively Res.	8	8	-4.87	7	.002	.03	*
	Cafe	Night Club	8	8	-4.47	7	.003	.04	*
	Lively Res.	Quiet Bar	8	8	5.90	7	< .001	.01	**
	Quiet Bar	Noisy Bar	8	8	-6.00	7	.001	.01	**
	Quiet Bar	Night Club	8	8	-5.20	7	.001	.02	*
Max Intensity	Cafe	Lively Res.	8	8	-3.87	7	.006	.09	.
	Lively Res.	Quiet Bar	8	8	5.42	7	.001	.02	*
	Quiet Bar	Noisy Bar	8	8	-4.56	7	.003	.04	*
	Quiet Bar	Night Club	8	8	-4.47	7	.003	.04	*
Pause Rate**	Lively Res.	Quiet Bar	8	8	-3.76	7	.007	.12	
Spectral Slope	Fine Dining	Lively Res.	8	8	4.18	7	.004	.06	.
	Quiet Bar	Night Club	8	8	5.65	7	.001	.01	*

* Bonferroni adjustments are used

** Pause rate is included as it showed significance in rANOVA, yet, it does not have significance considering p-value adjustment

Significance levels are denoted as follows *** $\alpha = 0.001$, ** $\alpha = 0.001$, * $\alpha = 0.05$, . $\alpha = 0.1$

Table 3.2: Summary of statistical results of significant features using a Post-Hoc Tukey Test.

(140 BPM; -15 dBFS) or a dark night club with electronic music (125 BPM; -8 dBFS). As expected with Lombard speech, both ambiences had a increase in median pitch and lower than baseline pitch range, however, the pitch range for the lively restaurant was slightly higher. It is possible that the joyful music in the lively restaurant with a large pitch range may have induced synchrony in vocal pitch patterns compared to the monotonous electronic beat. Shimmer, another feature representative of Lombard speech through hoarseness, appears to be more pronounced in the night club than the lively restaurant, perhaps another effect of joyful music. Lastly, rate-of-speech appeared notably different between conditions. There was an increase in overall speech rate and pause rate for the lively restaurant. The voiced speech rate, however, appeared similar for both night club and lively restaurant, indicating that pauses may play a role in differentiating the two conditions in-terms of rate-of-speech.

The last radar plot shows the differences between the noisy bar and quiet bar with a background ambiences of -21 dBFS and -31 dBFS, respectively. Features associated with Lombard speech are present in the noisy ambience, including an increase in energy, shimmer, median pitch and intensity with a decreased pitch range. Whereas the quiet bar had a increase in spectral slope, pitch range and voiced and overall speech rate.

Statistical Tests

The data for this study was collected as a repeated measures experiment. Six treatments and a baseline were present; each of the ambiences were applied to each of the study participants. Each pair of study participants were independent, however, within the pair of waiter and restaurant-goer we do not have independence as synchrony and mimicking is expected to occur. There was no randomization on the order of ambience, that is, the ambiences were applied in the same order for each experiment. As such, it is important to note that, in the future, it would be beneficial to increase the number of participants and complete a full Latin Square Design to better understand carry over effect. We completed a repeated measures ANOVA (rANOVA) for each of the extracted voice features. Due to the small sample size of participants and lack of randomization it is difficult to draw formal conclusions. Nevertheless, we suggest features, shown in Table 3.1, that may prove useful and warrant further investigation. Post-hoc results, using a pairwise t-test of the features that found significance in the rANOVA are shown in Table 3.2. In addition, the following will also showcase results that were significant at a threshold of $\alpha = 0.1$. Although this threshold is not considered community standard we want to highlight trends in the data and possible room for investigation for the future which could have been impacted by small dataset size.

Energy ($F(5,35) = 9.97$, $p < .001$), spectral slope ($F(5,35) = 9.43$, $p < .001$),Pause rate ($F(5, 35)=4.93$, $p < .01$), max ($F(5,35) = 12.47$, $p < .001$) and mean ($F(5,35) = 15.78$, $p < .001$) intensity were all significant. Median pitch ($F(5, 35) = 2.37$, $p = 0.06$) was also

ANOVA (Formal Dining)			Sum of Squares	df	Mean Square	F	p-value
RQ1	Appropriateness	Between Groups	96.51	3	32.17	12.63	<.001***
		Within Groups	242.00	95	2.55		
		Total	338.51	98	3.45		
	Comfort	Between Groups	81.60	3	27.20	14.47	<.001***
		Within Groups	180.40	96	1.88		
		Total	262.00	99	2.65		
	Awareness	Between Groups	32.11	3	10.70	4.25	.01**
		Within Groups	241.68	96	2.52		
		Total	273.79	99	2.77		
	Ambience Feeling	Between Groups	53.55	3	17.85	7.29	<.001***
		Within Groups	235.20	96	2.45		
		Total	288.75	99	2.92		

Figure 3.4: Summary of one-way ANOVA for RQ1 of the Fine Dining Ambience

significant at a threshold of $\alpha = .1$. Pitch range ($F(5, 35) = .96$, $p = 0.46$), shimmer ($F(5, 35) = 1.36$, $p = .26$), jitter ($F(5, 35) = 1.01$, $p = .42$), and voiced rate-of-speech ($F(5, 35)=1.61$, $p = .18$) and unvoiced rate-of-speech ($F(5, 35) = 1.93$, $p = .12$) were not significant.

For energy, a pairwise t-test for the pairs quiet bar and noisy bar,$p < .001$, and fine dining and lively restaurant ,$p < .05$, remained statistically significant when adjusted. Lively restaurant and quiet bar restaurant was significant, $p < .1$, at a threshold of $\alpha = .1$ after adjustment. However, when not adjusted, a pairwise t-test for lively restaurant and quiet bar restaurant ,$p < .01$, was significant.

A pairwise t-test for spectral slope showed two ambience pairings (fine dining and lively restaurant; quiet bar and night club) as significant, $p < .1$, after adjustment. A pairwise t-test for pause rate was significant for lively restaurant and quiet bar ,$p < .01$, however, this significance disappeared when the p-value was adjusted ,$p = .12$.

Both mean and max intensity showed the most statistical significance across ambiences post-hoc. For mean intensity, the pairs lively restaurant and quiet bar, $p < .01$, and quiet bar and noisy bar, $p < .01$, were the only ones statistically significant after adjustment. The remaining pairs, noted in Table 3.2, were significant at a threshold of $\alpha = .1$. All pairs for max intensity were statistically significant, $p < .1$, after adjustment at a threshold of $\alpha = .1$.

ANOVA (Night Club)			Sum of Squares	df	Mean Square	F	p-value
RQ1	Appropriateness	Between Groups	113.96	3	37.99	20.85	<.001***
		Within Groups	174.88	96	1.82		
		Total	288.84	99	2.92		
	Comfort	Between Groups	85.79	3	28.60	21.86	<.001***
		Within Groups	125.60	96	1.31		
		Total	211.39	99	2.14		
	Awareness	Between Groups	87.34	3	29.11	15.48	<.001***
		Within Groups	178.67	95	1.88		
		Total	266	98	2.71		
	Ambience Feeling	Between Groups	8.18	3	2.73	1.43	.24
		Within Groups	181.24	95	1.91		
		Total	189.41	98	1.93		

Figure 3.5: Summary of one-way ANOVA for RQ1 of the Night Club Ambience

3.3.2 Perception Study Results

RQ1: How do generated voices compare to human voices within ambiences?

Both Tables 3.3 and 3.4 contain summary information of a one-way ANOVA comparing the average perceptions scores of TTS-avg, TTS-bl, our voice conversion approach and human voice. The mentioned voices samples were rated by participants for the fine dining and night club ambiences, which were chosen due to their polarity in formality and loudness.

For the fine dining ambience, there was statistical significance for statement 1 (appropriateness; $F(3, 95)=12.63$, $p < .001$), statement 2 (comfort; $F(3, 96) = 14.47$, $p < .001$), statement 3 (awareness; $F(3, 96) = 4.25$, $p < .01$), and statement 4 (ambience feeling; $F(3, 96) = 7.29$, $p < .001$). For appropriateness and comfort, a post-hoc tukey test showed that the Human voice was significantly different to voice conversion, $p < .001$, and TTS-bl, $p < .001$. Comparing TTS-avg to the human voice, there was more significance in comfort, $p < .001$, compared to appropriateness, $p < .01$. For awareness, the human voice differed significantly with only TTS-avg, $p < .05$, and voice conversion, $p < .01$. Finally, for ambience feeling, the human voice was the most significantly different from the voice conversion, $p < .001$, followed by TTS-avg, $p < .01$, and TTS-bl $p < .01$.

The night club ambience was statistically significant for all of appropriateness ($F(3, 96)=20.85$, $p < .001$), comfort ($F(3, 96)=21.86$, $p < .001$), and awareness ($F(3, 95)=15.48$, $p < .001$). There was no significance for ambience feeling, ($F(3, 95)=1.43$, $p = 0.239$). For appropriateness, comfort, and awareness, using a post-hoc tukey test, the human voice was significantly different from TTS-avg, $p < .001$, and TTS-avg significantly differed from voice

Post-hoc (Formal Dining)		Group 1	Group 2	Mean Difference	Standard Error	p-value	Lower	Upper
RQ1	Appropriateness	Human	TTS-avg	-1.64	0.451	.003**	-2.82	-0.46
		Human	VC	-2.72	0.451	<.001***	-3.90	-1.54
		Human	TTS-bl	-1.84	0.456	<.001***	-3.03	-0.65
		TTS-avg	VC	-1.08	0.451	.09	-2.26	0.10
		TTS-avg	TTS-bl	-0.20	0.456	.97	-1.39	0.99
		VC	TTS-bl	0.88	0.456	.22	-0.31	2.07
	Comfort	Human	TTS-avg	-1.52	0.388	<.001***	-2.53	-0.51
		Human	VC	-2.48	0.388	<.001***	-3.49	-1.47
		Human	TTS-bl	-1.76	0.388	<.001***	-2.77	-0.75
		TTS-avg	VC	-0.96	0.388	.07	-1.97	0.05
		TTS-avg	TTS-bl	-0.24	0.388	.93	-1.25	0.77
		VC	TTS-bl	0.72	0.388	.25	-0.29	1.73
	Awareness	Human	TTS-avg	-1.24	0.449	.03*	-2.41	-0.07
		Human	VC	-1.48	0.449	.01**	-2.65	-0.31
		Human	TTS-bl	-1.08	0.449	.08	-2.25	0.09
		TTS-avg	VC	-0.24	0.449	.95	-1.41	0.93
		TTS-avg	TTS-bl	0.16	0.449	.98	-1.01	1.33
		VC	TTS-bl	0.40	0.449	.81	-0.77	1.57
	Ambience Feeling	Human	TTS-avg	-1.44	0.443	.01**	-2.60	-0.28
		Human	VC	-1.92	0.443	<.001***	-3.08	-0.76
		Human	TTS-bl	-1.56	0.443	.004**	-2.72	-0.40
		TTS-avg	VC	-0.48	0.443	.70	-1.64	0.68
		TTS-avg	TTS-bl	-0.12	0.443	.99	-1.28	1.04
		VC	TTS-bl	0.36	0.443	.85	-0.80	1.52

Table 3.3: Summary of Post-Hoc Tukey Test for RQ1 of the Fine Dining Ambience

Post-hoc (Night Club)		Group 1	Group 2	Mean Difference	Standard Error	p-value.	Lower	Upper
RQ1	Appropriateness	Human	TTS-avg	2.44	0.382	<.001***	1.44	3.44
		Human	VC	-0.32	0.382	.84	-1.32	0.68
		Human	TTS-bl	0.68	0.382	.29	-0.32	1.68
		TTS-avg	VC	-2.76	0.382	<.001***	-3.76	-1.76
		TTS-avg	TTS-bl	-1.76	0.382	<.001***	-2.76	-0.76
		VC	TTS-bl	1.00	0.382	.05*	0.00	2.00
	Comfort	Human	TTS-avg	2.16	0.324	<.001***	1.31	3.01
		Human	VC	-0.20	0.324	.93	-1.05	0.65
		Human	TTS-bl	0.56	0.324	.31	-0.29	1.41
		TTS-avg	VC	-2.36	0.324	<.001***	-3.21	-1.51
		TTS-avg	TTS-bl	-1.60	0.324	<.001***	-2.45	-0.75
		VC	TTS-bl	0.76	0.324	.09	-0.09	1.61
	Awareness	Human	TTS-avg	1.70	0.392	<.001***	0.67	2.72
		Human	VC	-0.87	0.392	.13	-1.89	0.16
		Human	TTS-bl	0.66	0.392	.35	-0.37	1.68
		TTS-avg	VC	-2.56	0.388	<.001***	-3.57	-1.55
		TTS-avg	TTS-bl	-1.04	0.388	.04*	-2.05	-0.03
		VC	TTS-bl	1.52	0.388	.001***	0.51	2.53
	Ambience Feeling	Human	TTS-avg	0.48	0.391	.61	-0.54	1.50
		Human	VC	-0.00	0.391	1.0	-1.02	1.02
		Human	TTS-bl	-0.33	0.395	.84	-1.36	0.70
		TTS-avg	VC	-0.48	0.391	.61	-1.50	0.54
		TTS-avg	TTS-bl	-0.81	0.395	.18	-1.84	0.22
		VC	TTS-bl	-0.33	0.395	.84	-1.36	0.70

Table 3.4: Summary of Post-Hoc Tukey Test for RQ1 of the Night Club Ambience

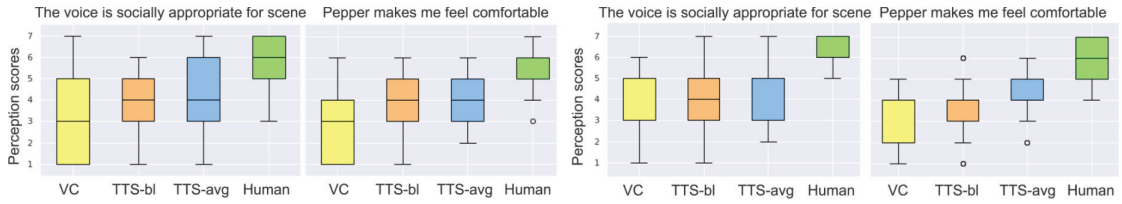

Figure 3.6: RQ1: Comparison of the perceptual rates of four voice types for fine dining (left two) and night club (right two)

conversion, $p < .001$, for all three statements. TTS-avg significantly differed from TTS-bl, $p < .001$, for appropriateness and comfort, while voice conversion significantly differed from TTS-bl, $p < .001$, for awareness. In addition, TTS-bl was slightly less significantly different from voice conversion, $p < .05$, for appropriateness, and was slightly less significantly different from TTS-bl, $p < .05$, for awareness. Overall, the voice conversion voice was rated the lowest for appropriateness, awareness, and comfort, followed by the TTS-bl, which was ranked as sounding the most robotic in both ambiences. Additionally, all generated voices were ranked noticeably lower than the human voice (see Figure 3.6). This indicates that a human voice, confirmed with the aforementioned one-way ANOVA and post-hoc tukey test results, is more comforting and socially appropriate for the ambiences.

RQ2: How are voice conversion generated voices perceived in context?

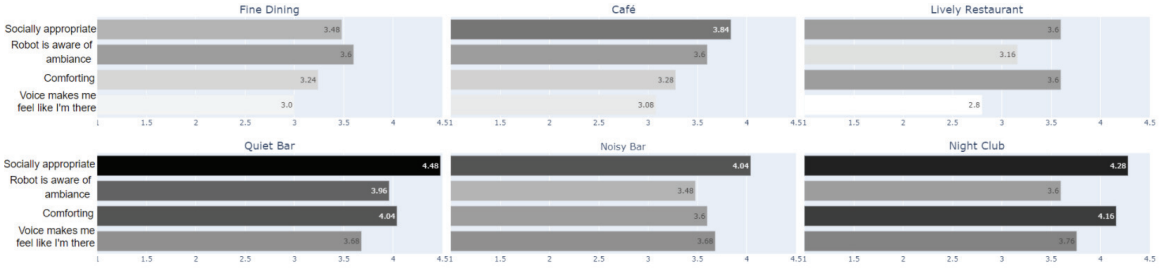

Figure 3.7: RQ2: Perception study comparing adapted voice conversion voices against each of the ambience conditions

Our 6 voice conversion samples for each of the 6 ambiences were assessed (see Figure 3.7). It is important to note, that we did not compare voice conversion samples together as that was not the goal of RQ2. The quiet bar, noisy bar and night club were rated the highest for appropriateness, awareness, comfort and ambience feeling. This indicates the Pepper with voice conversion was socially and contextually appropriate and comforting in conditions that required a Lombard effect. The fine dining condition was deemed to be the least socially appropriate, least comfortable and most robotic. The lively restaurant condition was rated the lowest for awareness and ambience feeling.

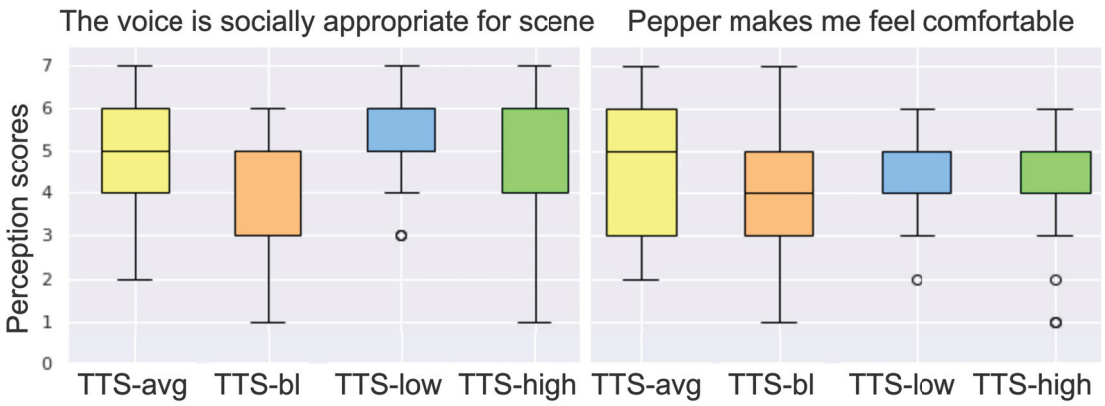

Figure 3.8: RQ 4: A comparison of perceptual rate in four voice types

RQ3: How are voice conversion generated voices perceived when paired with the incorrect ambience?

Three components were tested: (1) voice conversion sample for café overlaid on background sound for fine dining, (2) voice conversion sample for fine dining overlaid on background sound for café, and (3) voice conversion sample for night club overlaid on background sound for fine dining. Component (1) resulted in a boost for statements on appropriateness, awareness, comfort and ambience feeling compared to being overlaid with their respective matching ambiences in RQ2. Component (2) also resulted in improvements compared to their respective correct pairings. However, (2) was rated higher than (1). This indicates that the fine dining voice may have been more suited for the café. It is important to note that the fine dining condition was always first and so it may have taken time for participants to adjust to the experiment protocol. Appropriateness, awareness, comfort and ambience feeling were rated the lowest for (3). This result suggests that a voice suited for a loud social environment does not suit a quiet formal restaurant. In addition, ratings for (3) were similar to ratings for the fine dining condition in RQ2, where the voice was paired with the correct ambience.

Using a Welch's t-test, component (1) compared to the fine dining condition in RQ2 resulted in non-significance for appropriateness ($t(0.72) = 47.84$, $p = .48$), awareness ($t(1.48) = 47.92$, $p = .15$), comfort ($t(1.74) = 47.89$, $p = .09$), and ambience feeling ($t(1.08) = 47.74$, $p = .28$); component (2) compared to the café condition in RQ2 resulted in non-significance for appropriateness ($t(1.26) = 45.60$, $p = .22$), awareness ($t(1.09) = 47.95$, $p = .28$), comfort ($t(1.75) = 47.99$, $p = .09$), ambience feeling ($t(1.29) = 47.47$, $p = .20$); component (3) compared to the fine dining condition in RQ2 resulted in non-significance for appropriateness ($t(0.47) = 48.00$, $p = .64$), awareness ($t(-0.58) = 47.82$, $p = .56$), comfort ($t(0.85) = 48.00$, $p = .40$), and ambience feeling ($t(0.00) = 47.29$, $p = 1.0$).

RQ4: How does perception of a data-driven pitch manipulation for TTS compare to one randomly selected?

Finally, TTS-bl, TTS-avg, TTS-low, and TTS-high were overlaid on the café background sound and compared. The TTS-bl was rated the lowest for appropriateness, awareness, comfort and ambience feeling. TTS-bl was deemed to be the most robotic sounding, which could contribute to why it had low ratings in other categories. The results (as shown in Figure 3.8) indicate that the human pitched TTS (TTS-low) was deemed more socially and contextually appropriate, as well as comforting, rating highest on appropriateness, comfort and ambience feeling, and rated as least robotic sounding. TTS-high condition was rated second highest for awareness, followed by TTS-low. Altogether, using a data-driven method to alter pitch demonstrates that humans prefer pitch that matched one specific human's pitch and when that pitch matched the current ambient environment. For RQ4, there was no statistical significance for appropriateness ($F(3, 96) = 2.01$, $p = .12$), comfort ($F(3, 96) = 1.50$, $p = .22$), awareness ($F(3, 96) = 0.30$, $p = .83$), and ambience feeling ($F(3, 96) = 1.86$, $p = .14$), using a one-way ANOVA.

Chapter 4

Summary and Future Work

From humans' emotion expression cross-culturally to adaptation to different ambiences, a spectrum of different social environments was explored. The following two sections (sections 4.1 and 4.2) will go over the implications presented by both studies. Major findings are highlighted and a discussion of limitations and future studies will be presented in sections 4.4 and 4.5.

4.1 Summary of Culture and Negative Emotion Expression Recognition

Fundamentally, study 1 aimed to understand data that is used to train emotion expression recognition models as culture is currently being treated equally in state-of-the-art emotion expression recognition systems. Underlying differences in our cross-cultural dataset, which focused on negative emotion expressions, are identified. In recent years, it has become crucially important to recognize the impact culture has on developing rules to express oneself [9]. Therefore, to increase emotion expression recognition performance, one must consider culture. Our first hypothesis in this study was supported using a one-way ANOVA to illustrate that each culture in our dataset had significantly different means of activation for AUs related to CAD. As we saw from Figure 2.6, the level of activation for AUs associated with a given emotion significantly differed for each culture. In particular, AUs related to the upper half of the face, such as AU4 (brow lowerer), significantly differed. Several AUs mentioned for each emotion significantly differed cross-culturally, while AUs prototypical to contempt and disgust were similar cross-culturally. AU9 (nose wrinkler) did not significantly differ between Persian and Filipino, and AU10 (upper lip raiser) did not significantly differ between North American and Filipino datasets. AU10 demonstrates a sneer, which is a social signal commonly attributed to contempt [62], whereas AU9 demonstrates a wrinkling of the nose, which occurs in displays of disgust [86, 81]. In addition, AU4 (brow lowerer), which is common among most negative emotions, did not significantly differ for disgust between the Filipino and North American datasets [81].

The second hypothesis of study 1, which was that machine learning models perform poorly if all cultures are considered equal, was only partially supported. The results of the SVM showed that the North American culture category performed the best on training and testing in the within-culture experiment compared to the cross-culture experiment. However, the collectivist cultures had the best recognition accuracy when testing on the North American dataset. Therefore, the cross-culture experiment performed better than the within-culture experiment for the collectivist datasets. As previously mentioned, several studies have hinted at the suppression of negative emotions in collectivist cultures [22, 20, 21]. Thus the results could indicate that individualistic cultures allow emotion expression recognition systems to be more sensitive to emotion expressions as these cultures encourage an increase in activation of AUs. If collectivist cultures suppress their emotions this could indicate that the North American dataset provided stronger training and testing boundaries, giving a better testing dataset for the collectivist cultures.

In addition, contempt and disgust were commonly confused with anger for the Persian dataset. One study by [21] found that for both contempt and anger, activation of facial muscles occurred similarly cross-culturally, whereas, for disgust, activation of facial muscles differed. The differences noted by [21] could explain why there was confusion with anger, as some similarities exist in AU activation across all three negative emotions, or that features were missing that were not present in the Persian dataset, such as hand gestures.

In the Filipino culture dataset, anger was commonly confused with contempt, indicating similarities between the two emotions in terms of AUs. In addition, disgust was frequently confused with anger, also suggesting that AUs expressed are similar. Once again, however, this confusion could indicate features missing from the Filipino dataset that could be crucial for emotion expression recognition. The study by [21] also attributed that certain cultures depicted weaker expressions of disgust, leading to confusion with cultures that showed stronger depictions of disgust. As such, weaker displays of disgust could be why collectivist cultures' performance in both experiments was low compared to the within-culture experiment of the North American dataset.

Most interesting were the results obtained from the overall within-culture experiment. As culture became more collectivist, performance accuracy decreased, resulting in poorer recognition. Social signals related to the CAD Triad may slowly overlap and become similar. As mentioned, collectivist cultures tend to encourage the suppression of negative emotions to maintain harmony. Stemming from environmental factors, collectivist cultures, such as Filipino and other East Asian and Southeast Asian cultures, encourage low arousal (or activation) of social signals [54, 82]. Therefore, enhancing the idea that collectivist cultures do not display emotions the same as individualistic cultures. Consequently, the CAD Triad may either have low levels of activation or the AUs associated with the CAD Triad become similar due to suppression of social signals.

Together, the experiments show that the level of AU activation is different for each culture in our dataset. Even though the same AUs might have been present in all cultures, this difference in AU activation supports the idea that the underlying components of negative emotional expressions are not universal. Therefore, to train emotion expression recognition systems we must consider culture to remove bias in the dataset and potentially boost performance.

Explanatory power of selected facial action units. The confusion matrices in Figures 2.8, 2.9, 2.10, and 2.11 illustrate the limitations of the SVM model with the proposed feature set, specifically for both the Filipino and Persian datasets. As previously mentioned, most disgust samples for the Filipino dataset were confused with anger. Upon qualitative review, AU4 (brow lowerer) was a frequently occurring AU between Filipino anger and disgust. In addition, Filipino disgust also included an aversive gaze away (AU 51 or AU 52) which was not present in the study. As such, adding gaze-related action units may improve the distinction between Filipino disgust and other emotions. Furthermore, a distinctive feature of prototypical and North American disgust is AU9 (nose wrinkler), but this was not present in Filipino disgust. A likely explanation was that the North American clips contained more reality TV depictions of physical disgust (e.g. reactions to aversive food), whereas Persian and Filipino clips had more moral disgust towards a person. This points to the importance of distinguishing subtypes of disgust when performing in-the-wild data collection.

Secondly, Persian anger was often mistaken for contempt in the SVM experiment, and this confusion, is supported by anger and contempt's similar AU activations as shown in Figure 2.6. Qualitative interviews with a Persian annotator suggested that hand gesturing is an important social signal during these social displays, indicating that the speaker is referring to another person. Therefore, adding body poses or hand gestures is expected to improve performance.

The relationship between kappa values and classification. As previously mentioned, the North American dataset had the highest agreement amongst annotators with a Fleiss Kappa value of 0.83. Both Persian and Filipino datasets had a much lower agreement, with Filipino having the lowest at 0.40 and Persian with 0.59. Furthermore, in terms of within culture classification accuracy, North American (66%) had the highest, followed by Filipino (39%) and then Persian (37%). Although not linear, the less agreement in the dataset led to lower accuracy in classification. One reason for this low accuracy might be the suppression of social signals in the collectivist groups. As previously mentioned by [25], [82] and [37], collectivist cultures suppress their negative emotions more to maintain harmony. Therefore, collectivist cultures' expression of negative emotions can look vastly different from the stereotypical North American display. This is also demonstrated in Figure 2.6, which depicts the suppression of commonly seen AUs in collectivist cultures. As such, suppression of AUs might also explain the lower agreement amongst annotators for

the collectivist cultures. If collectivist cultures suppress commonly associated AUs, then it becomes increasingly difficult to decipher what differences do exist between emotions in these cultures

In summary, facial expressions, commonly considered under a common view (or basic emotion approach), have been the focus of many studies. The aforementioned approach results in relatively high accuracy for North American samples. However, our results suggest more investigation into non-facial features, which could improve recognition of CAD emotions in non-Western cultures. In addition, rather than considering social signals as suppressed for collectivist cultures, it could be that *facial features are not enough*; important distinguishing information is communicated through other channels, such as body pose and gesture as discussed here, voice [96] and context [45], and so on.

4.2 Summary of Ambient Adaptive Speech Generation

Study 2 provided a novel protocol to collect data online to gain insight into the human voice adaptation to different ambient environments. Human perception of generated voices using a data-driven approach was also explored. In addition, significant and notable features from eight female speakers were provided. Potential features that require further exploration are rate-of-speech, energy, max intensity, mean intensity, pause rate, spectral slope, and mean pitch. More nuanced features, such as jitter and shimmer, did not show significance statistically but do indicate a trend in the formal vs. casual ambiences in the first radar plot (Figure 3.3). Therefore, these two features could be further explored and may prove significant with a larger dataset. Furthermore, differences in pitch range and median pitch for the casual vs formal ambience suggested that more casual speech had an increase in spectral features, similar to Lombard speech [56, 75]. This is surprising as it was expected that in a formal setting, where politeness is at the forefront of social interactions, the median pitch would increase [12].

Some results in the second radar plot (Figure 3.3) indicated a level of synchrony to the background ambience. This synchrony could suggest the development of rapport between two participants as they want to adapt correctly and coincide in the same ambience [70]. Another avenue to investigate is the level of white noise, which may be perceived as higher in the night club, thus impacting the night club's effect on spectral features, such as pitch range. Moreover, the difference in rate-of-speech in the lively restaurant compared to the night club may be due to the pause rate, which was higher in the lively restaurant. More specifically, more pauses that were longer than 50 ms were present in the overall utterance of the lively restaurant. An increase in pauses could indicate that more breaks in-between words or phrases occurred. As such, pauses could play an important role in ambience adaptation, further suggesting that we need to develop TTS voices that can slow down by extending

pauses in a meaningful way, instead of slowing down the length of the overall utterance linearly [61].

For the third radar plot, the noisy bar appeared to show a Lombard effect, while the quiet bar showed an increase in rate-of-speech features. These results suggest that speakers provided more vocal effort to communicate in the noisy bar. In addition, for the quiet bar, intelligibility was not at the forefront of communication compared to the noisy bar, where communication was impaired with the louder background [40]. In other words, a more breathier voice was present in the quieter ambience, indicating that voice quality was not as important in the quiet bar as it was in the noisy bar[95, 31, 48]. Furthermore, due to the increase in rate-of-speech, spectral slope, and pitch range, the quiet bar may have induced an increased liveliness for this ambience [35]. Therefore, clusters of features should be explored, as this could indicate what specific features together result in an appropriate voice for a given ambience. For example, in the quiet bar ambience the increase in mentioned features could suggest these features together may be more appropriate for such an ambience, compared to a noisy bar, which had an increase in a different subset of features. It is important to note that shimmer, jitter, spectral slope, and energy, which combined with pitch and pitch range, impact overall voice quality [95, 31, 48, 56]. Using a rANOVA and a pair-wise t-test, we found that features, such as spectral slope and energy, do have a significant difference amongst ambiences. From these results, we can see that features related to voice quality assist in human voice adaptation for the ambiences. However, SSML for TTS does not have these features and as such, given the significance, should be added to the roster of modifications as they do assist in adaptation.

RQ1: how do generated voices compare to human voices within ambiences? One main takeaway from the results of RQ1 is that humans prefer a human voice that matches the social and ambient context. The TTS voice samples had significantly lower ratings than the human voice on appropriateness, comfort, awareness, and ambience feeling in study 2. This preference has arisen in other studies, suggesting that when interacting socially with a robot, a human voice is preferred [49]. On the other hand, others have suggested that humans favour robot companions that sound less human and more robotic, however, our study did not show support for this [98]. The voice conversion samples received the lowest ratings, but the night club ambience had less variability and higher ratings than the fine dining ambience in both appropriateness and comfort (Figure 3.6). Indicating that the ambience appeared to influence how we perceived a given voice. It is clear that since a human voice outperformed TTS there is still room to improve for current state-of-the-art generated voices. Although there was a preference for human voices, we saw that participants favoured TTS which is data-driven and altered to match the underlying ambient environment for the average of female speakers.

RQ2: How are voice conversion generated voices perceived in context? As voice conversion is a flexible and adaptive solution for speech generation, it shows promise,

as voice conversion ratings were noticeably different between quiet and loud ambiences as described by perceivers in RQ2. Specifically, it appeared that the ambiences that encourage Lombard speech (i.e., lively restaurant, noisy bar, and night club) had the highest ratings for the voice conversion [56, 75]. Qualitatively, some of our voice conversion samples had a Lombard effect, as it sounded like someone raising their voice and providing more vocal effort to be heard. Although the quiet bar was deemed to be a quieter ambience, it had a social environment that was similar to the noisy bar. Potentially due to the similarity in social environment, the quiet bar was still rated highly amongst the other voice conversion samples. The low rating for voice conversion samples present in RQ1 is most likely due to the low-quality samples generated by CRANK, which may indicate that more audio samples are required for each speaker-ambience training batch. It also might have to do with perceptual clarity experienced by the perceivers. In other words, since the CRANK output was not as clear as the TTS, it was rated poorly. Furthermore, it could be a result of the lack of significance amongst spectral features noted in the statistical tests in section 3.4.1. Thus, CRANK would not have been able to pick up on the differences in spectral features amongst ambiences because there are no meaningful differences.

RQ3: How are voice conversion generated voices perceived when paired with the incorrect ambience? For RQ3, the voice conversion for the fine dining seemed to be more appropriate for the café ambience. This signifies that a slower rate-of-speech is more appropriate for a casual setting. This goes against our hypothesis that a fine dining ambience would require a slower speech rate to induce politeness compared to a more casual setting where politeness is normally not as important [12]. Ratings for component (3) indicated that a voice with speech features required for Lombard speech would not suit a fine dining ambience [56, 75]. This highlights that a voice suited for a loud ambience is deemed inappropriate for a fine dining restaurant.

RQ4: How does a data-driven pitch manipulation for TTS impact human perception? Although human perceivers prefer the human voice, in RQ1, they also favour a TTS that has pitch altered. As shown in both RQ1 and RQ4, when the pitch is altered to match an ambience it is deemed more favourable than a vanilla TTS (TTS-bl), which has no alterations. More specifically, TTS-low, which used the pitch of a single person (speaker 714) for a given context in RQ4, was consistently rated high for appropriateness, comfort, and ambience feeling. The higher ratings indicate that perceivers prefer a TTS voice that is altered to match a single person. These results also add support that an adaptive TTS, that uses data-driven methods, is preferable for ambience-specific generated speech.

It is important to note, that our voice conversion approach was trained with approximately 16 to 41 utterances for each speaker-ambience training batch, compared to the 80 utterances used in [46]. Therefore, since it was able to generate perceivable voices on such a small dataset, it is expected that the quality will improve as the dataset increases. This also provides promise that voices built for robots do not require as much data. As such, instead

of recording hours (or days) worth of data of a single speaker in different ambiences from companies that custom build expressive voices, it may prove useful to just use one hour of data of a single speaker. However, this relates specifically to voice conversion and may not carry over into other speech generation techniques, like TacoTron [90, 55, 91, 5].

4.3 Limitations

There were six main limitations from both studies.

Data Collection Methods. The first such limitation is that, while the data from study 1 was collected as "in-the-wild" outside of a laboratory using internet sources, analysis revealed that the data contained both professionally acted material (e.g. scripted TV) and spontaneous material (e.g. reality TV). A majority of the clips from the North American dataset contained in-the-wild facial expressions collected from North American YouTube accounts which were "vlog" style content, reality television shows (e.g., Dance Moms), and talk shows (e.g., The Late Late show with James Corden). The Persian and Filipino datasets primarily contained data from YouTube accounts sharing clips from popular TV shows and movies from that area of the world. Furthermore, according to our Persian confederate, reality TV shows are uncommon in their culture. The result could point to the importance of distinguishing between in-the-wild acted vs. in-the-wild spontaneous videos, which currently is not considered during data collection in affective computing datasets. Consequently, we plan on diversifying the collectivist datasets with more realistic in-the-wild depictions in the future. Overall, study 1, attempted to move past solely acted depictions of Western expressions of contempt, anger and disgust, and this was achieved in this dataset. For study 2, on the other hand, the data was lab controlled, even though it was collected over Zoom outside of a physical laboratory setting. We tried to create a more authentic experience with the unscripted portion but it followed the scripted portion of the protocol giving way for influence in response and behaviour. As such, the data collected in this protocol might deviate from how one truly adapts to different ambiences in real life. Future studies would need to attempt to physically go to ambience locations, with the goal of not only collecting spontaneous in-the-wild data, but also improving the way one collects audio as to get the same quality as the zoom recorded clips.

Dataset Size. In addition to the type of data collected, the datasets in general were relatively small in both study 1 and 2. For study 1, each culture category had between 74 to 105 video clips, this amount was further reduced for each culture when considering each emotion separately. Although the data distribution for each emotion was similar for all cultures, more data could have increased the accuracy and f1-scores for each experiment. This is especially so for the Filipino dataset which not only performed the worst in both experiments but also had the fewest amount of overall video clips. In addition to the discrepancy of culture specific dataset size, the overall accuracy was low. The current study

only had 257 available video clips, which is relatively small. Other studies have used more advanced classification techniques to tackle the issue of small datasets. For example, [59] used XGBoost to classify 1323 video clips of to classify 6 different emotions. They found that XGBoost outperformed other classification methods, including an SVM. Future studies will explore the use of XGBoost in addition to increasing dataset size in order to improve overall classification performance. For study 2, there were only 8 female speakers with approximately 16 to 41 utterances per training batch (ambience-speaker pairing). CRANK was originally trained on 85 utterances for each speaker, which is double the amount used for study 2 [46]. Therefore, an increase in the amount of utterance per training batch would prove useful and may improve overall quality of the voice conversion samples. Furthermore, only 25 participants responded to the perception study on Mechanical Turk. As such, it was hard to gauge how significant the differences in responses were for each of the conditions in the survey. Future studies will require more participants to complete the perception study. This will allow for a clear indication of the overall trends occurring in human perception of generated voices with their respective ambience.

Annotation. Another limitation is that image frames were used as input data for study 1, whereas videos were annotated. We cannot be certain that each individual image from a given video is depicting the labeled emotion. For example, a video labeled "contempt" could have one or more frames that do not display the emotion "contempt". The results are involve relative findings under these constraints. In the future, we will run a pretrained state-of-the-art AU detector on the dataset and see what it results in, for a more absolute baseline. Future research can study the dynamic of negative emotions in culture by using a sequence of frames as input. It is also important to note that although the inter rater agreement was low for the Filipino dataset, other studies on large affective image datasets has garnered similar agreement [67].

Automatic Feature Extraction. We also had limited access to available AUs from OpenFace. There were several AUs that could not be collected that were important when describing prototypical emotions. For anger AU27 (Mouth Stretcher) was missing and AU20 (Lip Stretcher) was missing for disgust. Furthermore, contextual information could not be extracted from OpenFace. This contextual information includes hand gestures and audio information. Upon visual inspection of the Persian dataset, hand gestures were commonly present. As such, future studies will look into extracting these contextual features and exploring other modes of detecting more AUs.

Carryover Effects. An initial limitation in the speech generation study was the presence of a carryover effect. Carryover effects usually occurs when conditions are sequential, allowing for the effects of one condition to influence the effects of another. For example, the lively restaurants' upbeat and joyful music could have influenced the voice features of the subsequent condition, which was the quiet bar. While a 1 min break was introduced between ambient conditions, there may still have been some residual context from the pre-

vious ambience. This effect was further mitigated by randomly placing a baseline condition, which had no ambient background music or imagery. Nevertheless, because of the lack of speaker participants, the placement of the baseline condition was difficult to repeat. As such, without placing the baseline condition between any 2 given ambiences more than a handful of times it is difficult to say the carryover effect was truly mitigated.

In addition, in study 2 the fine dining restaurant was introduced first and speakers were initially adjusting to the experiment setup. As such, this could have impacted the vocal features for this initial condition. Fully randomizing the sequence of all ambience conditions would be a solution. Although randomization would be useful, this is difficult to ensure as it requires a significant amount of participants to partake in the data collection protocol. Another solution could be to shorten the number of ambience conditions to around 3, but this would require generalizing to the 6 ambiences. Moreover, scripts also varied slightly. As such, some features could have been further influenced through word choice. In order to limit independent variables, there was also only one phrase for the perceptual experiments, "Hi there, I hope you're doing well." This phrase has a positive valence to it, and it is possible the voice conversion samples did not match this valence.

Voice Conversion Tools. Finally, the tools used for the speech generation study may not have been the appropriate method for the current task. More specifically, CRANK only used spectral features for training. Other features, from the rANOVA results shown in Table 3.1, were deemed more appropriate for ambience adaptation, such as energy. Therefore, using a voice conversion implementation that specifically uses energy in conjunction with intensity and rate-of-speech in training, would be beneficial. Nonetheless, voice conversion tools that use features like energy, rate-of-speech or intensity are not currently accessible, or readily available. Thus, considering this gap, developing voice conversion tools that not only use spectral features, but also loudness, energy, and rate-of-speech would be greatly beneficial when creating voices that adapt to different ambiences.

4.4 Future Studies

When we, as humans, perceive emotion, we do not only pay attention to facial expression but at the body as a whole [6]. Hence, we cannot ignore the importance of body language such as head or hand movement. Adding data about body language in later research can lead us to testing more general hypotheses about role of culture in emotion. Moreover, it may help to build more accurate classifiers.

Another opportunity to further investigate the role of culture is to provide more detailed annotations, such as extracting more social signals from videos. As such future work may use a method other than OpenFace to extract more AUs, since OpenFace is not capable of extracting some AUs such as, AU27 (Mouth Stretcher), AU44 (Squinting), or AU24 (Lip Pressor) which we saw in our datasets for study 1.

Furthermore, potentially establishing culture categories in an already established dataset, such as AffectNet, could provide not only a larger dataset but an already established ground truth. Currently, we are establishing an effective method of data augmentation to utilize advanced deep learning tools, like a CNN, in order to boost recognition accuracy [4]. Finally, having members of one culture annotate the labels of another culture could increase the evidence that cultures perceive emotions differently and present further underlying differences in certain social signals.

In the future, we also hope to increase data collection in order to obtain every combination of ambient conditions. Future studies would also use the collected voice features in study 2 for further assessment. In addition, increasing speaker-ambience batch size and duration of conditions may provide higher quality output for our voice conversion approach. Other avenues for automatic speech generation that is adaptive to different ambience environments will also be explored.

Finally, we plan to incorporate culture into the speech generation process. Currently, data from Japan is being collected in order to compare to the English Canadian speaker samples collected for the current study. This would prove useful as differences in features that are adapted to a given ambience between Japan and Canada, may be due to culture. Therefore, providing more evidence that culture is a contextual feature that should be incorporated into machine learning that involves humans.

Bibliography

[1] Ambience. *Merriam-Webster*, 2022.

[2] Context. *Merriam-Webster*, 2022.

[3] Ibrahim A. Adeyanju, Elijah O. Omidiora, and Omobolaji F. Oyedokun. Performance evaluation of different support vector machine kernels for face emotion recognition. In *2015 SAI Intelligent Systems Conference (IntelliSys)*, pages 804–806, 2015.

[4] Saba Akhyani, Mehryar Abbasi Boroujeni, Mo Chen, and Angelica Lim. A sim2real approach to augment low-resource data for dynamic emotion expression recognition. In *Conference on Computer Vision and Pattern Recognition*, 2021.

[5] Jesus Alvarez, Holly Francois, Hosang Sung, Seungdo Choi, Jonghoon Jeong, Kihyun Choo, Kyoungbo Min, and Sangjun Park. Camnet: A controllable acoustic model for efficient, expressive, high-quality text-to-speech. *Appl. Acoust*, 186:108439, 2022.

[6] Hillel Aviezer, Yaacov Trope, and Alexander Todorov. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*, 338:1225–9, 11 2012.

[7] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66, 2018.

[8] Archana Balyan, Shyam Sunder Agrawal, and Amita Dev. Speech synthesis: A review. *International Journal of Engineering Research and Technology*, 2, 2013.

[9] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68, 2019.

[10] Ann R Bradlow. *Laboratory Phonology 7*, chapter Confluent talker- and listener-oriented forces in clear speech production, pages 241–274. Walter de Gruyter GmbH and Co. KG, 2008.

[11] Denis Burnham, Christine Kitamura, and Ute Vollmer-Conna. What's new, pussycat? on talking to babies and animals. *Science*, 296:1435, 2002.

[12] Jonathan A Caballero, Nikos Vergis, Xiaoming Jiang, and Marc D Pell. The sound of im/politeness. *Speech communication*, 102:39–53, 2018.

[13] Antonio Castellanos, José-Miguel Benedí, and Francisco Casacuberta. An analysis of general acoustic-phonetic features for spanish speech produced with the lombard effect. *Speech Commun.*, 20(1):23–35, 1996.

[14] Lei Chang, Miranda Mak, Tong Li, Bao Wu, Bin-Bin Chen, and Hui Jing Lu. Cultural adaptations to environmental variability: An evolutionary account of east–west differences. *Educational Psychological Review*, 23:99–129, March 2011.

[15] Chaona Chen, Laura Hensel, Yaocong Duan, Robin Ince, Oliver Garrod, Jonas Beskow, Rachael Jack, and Philippe Schyns. Equipping social robots with culturally-sensitive facial expressions of emotion using data-driven methods. In *2019 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019)*, pages 1–8. IEEE, 2019.

[16] Martin Cooke, Simon King, Maëva Garnier, and Vincent Aubanel. The listening talker: A review of human and algorithmic context-induced modifications of speech. *Comput. Speech Lang.*, 28(2):543–571, 2014.

[17] Scotty D. Craig and Noah L. Schroeder. Text-to-speech software and learning: Investigating the relevancy of the voice effect. *J. Educ. Comput. Res.*, 57(6):1534–1548, 2019.

[18] Zongyang Du, Berrak Sisman, Kun Zhou, and Haizhou Li. Expressive voice conversion: A joint framework for speaker identity and emotional style transfer, 2021. [Online] Available:arXiv:2107.03748.

[19] P. Ekman and W. V. Friesen. Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124–129, 1971.

[20] Paul. Ekman. Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation*, 17:207–282, 1971.

[21] H. A. Elfenbein, M. G. Beaupré, M. Levesque, and U Hess. Toward a dialect theory: cultural differences in the expression and recognition of posed facial expressions. *Emotion*, 7(1):131, 2007.

[22] Hillary Anger Elfenbein and Nalini Ambady. On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, 128(2):203–235, 2002.

[23] Aaron Elkins and Douglas Derrick. The sound of trust: Voice as a measurement of trust during interactions with embodied conversational agents. *Group. Decis. Negot.*, 22:897–913, 2013.

[24] Kexin Feng and Theodora Chaspari. A review of generalizable transfer learning in automatic emotion recognition. *Frontiers in Computer Science*, 2:9, 2020.

[25] Itziar Fernández, Pilar Carrera, Flor Sánchez, Darío Páez, and L. Candia. Differences between cultures in emotional verbal and nonverbal reactions. *Psicothema*, 12:83–92, 2000.

[26] Kerstin Fischer, Lakshadeep Naik, Rosalyn M. Langedijk, Timo Baumann, Matouš Jelínek, and Oskar Palinko. *HRI*, chapter Initiating Human-Robot Interactions Using Incremental Speech Adaptation, page 421–425. ACM, New York, NY, USA, 2021.

[27] N. Fragopanagos and J.G. Taylor. Emotion recognition in human–computer interaction. *Neural Networks*, 18(4):389–405, May 2005.

[28] Alan Fridlund. Sociality of solitary smiling: Potentiation by an implicit audience. *Journal of Personality and Social Psychology*, 60:229–240, February 1991.

[29] Alan Fridlund, John P. Sabini, Laura E. Hedlund, Julie A. Schaut, Joel I. Shenker, and Matthew J. Knauer. Audience effects on solitary faces during imagery: Displaying to the people in your head. *Journal of Nonverbal Behavior*, 14:113–137, June 1990.

[30] Mehdi Ghayoumi and Arvind K Bansal. Unifying geometric features and facial action units for improved performance of facial expression analysis, 2016. [Online] Available:arXiv:1606.00822.

[31] Helen M. Hanson. Glottal characteristics of female speakers: Acoustic correlates. *The Journal of the Acoustical Society of America*, 101(1):466–481, 1997.

[32] Akira Hayamizu, Michita Imai, Keisuke Nakamura, and Kazuhiro Nakadai. Volume adaptation and visualization by modeling the volume level in noisy environments for telepresence system. In *Proceedings of the Second International Conference on Human-Agent Interaction*, page 67–74. ACM, 2014.

[33] Valerie Hazan and Rachel Baker. Acoustic-phonetic characteristics of speech produced with communicative intent to counter adverse listening conditions. *J. Acoust. Soc*, 130:2139–52, 2011.

[34] Anna Henschel, Guy Laban, and Emily Cross. What makes a robot social? a review of social robots from science fiction to a home or hospital near you. *Curr. Robot. Rep.*, 2, 2021.

[35] Rebecca Hincks. Measures and perceptions of liveliness in student oral presentation speech: A proposal for an automatic feedback mechanism. *System*, 33(4):575–591, 2005.

[36] Tuan Vu Ho and Masato Akagi. Non-parallel voice conversion based on hierarchical latent embedding vector quantized variational autoencoder. In *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, pages 140–144, 2020.

[37] Geert Hofstede. Dimensionalizing cultures: The hofstede model in context. *Online Readings in Psychology and Culture, Unit 2*, 2(1), 2007.

[38] M-W Huang, C-W Chen, W-C Lin, S-W Ke, and C-F Tsai. Svm and svm ensembles in breast cancer prediction. *PLOS One*, 12, 2017.

[39] Angelika Hönemann and Petra Wagner. Adaptive speech synthesis in a cognitive robotic service apartment: An overview and first steps towards voice selection. In *Elektronische Sprachsignalverarbeitung (ESSV)*, 2015.

[40] Titze IR and Palaparthi A. Vocal loudness variation with spectral slope. *J. Speech Lang. Hear*, 63(1):74–82, 2020.

[41] Keith Ito and Linda Johnson. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`, 2017.

[42] Stanislav Ivanov, Ulrike Gretzel, Katerina Berezina, Marianna Sigala, and Craig Webster. *J. Hosp. Tour. Technol.*, 2019.

[43] Rachael E. Jack, Oliver G. B. Garrod, Hui Yu, Roberto Caldara, and Philippe G. Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244, 2012.

[44] S. Kiesler. Fostering common ground in human-robot interaction. In *ROMAN*, pages 729–734, 2005.

[45] Seon-Gyu Ko, Tae-Ho Lee, Hyea-Young Yoon, Jung-Hye Kwon, and Mara Mather. How does context affect assessments of facial emotion? the role of culture and age. *Psychology and Aging*, 26:48–59, 2011.

[46] Kazuhiro Kobayashi, Wen-Chin Huang, Yi-Chiao Wu, Patrick Lumban Tobing, Tomoki Hayashi, and Tomoki Toda. Crank: An open-source software for nonparallel voice conversion based on vector-quantized variational autoencoder, 2021. [Online] Available:arXiv:2103.02858.

[47] Jean Krause and Athina Panagiotopoulos. Speaking clearly for older adults with normal hearing: The role of speaking rate. *J. Speech Lang. Hearing*, 62:1–9, 2019.

[48] Jianjing Kuang, Jia Tian, and Bing'er Jiang. The effect of vocal effort on contrastive voice quality in shaoxing wu. *The Journal of the Acoustical Society of America*, 146(3):EL272–EL278, 2019.

[49] Katharina Kühne, Martin H. Fischer, and Yuefang Zhou. The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. evidence from a subjective ratings study. *Frontiers in Neurorobotics*, 14, 2020.

[50] Christa Lam and Christine Kitamura. Mommy, speak clearly: Induced hearing loss shapes vowel hyperarticulation. *Dev. Sci.*, 15(2):212–21, 2012.

[51] Sau-Lai Lee, Ivy Lau, Sara Kiesler, and Chi Yue Chiu. Human mental models of humanoid robots. In *ICRA*, pages 2767 – 2772, 2005.

[52] Alexander Lerch. *An introduction to audio content analysis : applications in signal processing and music informatics*. Wiley, 2012.

[53] D. Y. Liliana and T. Basaruddin. Review of automatic emotion recognition through facial expression analysis. In *2018 International Conference on Electrical Engineering and Computer Science (ICECOS)*, pages 231–236, 2018.

[54] Nangyeon Lim. Cultural differences in emotion: differences in emotional arousal level between the east and the west. *Integrative Medicine Research*, 5(2):105, 2016.

[55] Rui Liu, Berrak Sisman, Guanglai Gao, and Haizhou Li. Expressive tts training with frame and style reconstruction loss. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:1806–1818, Jan 2021.

[56] E Lombard. Le signe de l'élévation de la voix. *Ana. d. Mal. de L'Oreillexdu du larynx [etc]*, 37:101–119, 1911.

[57] Nichola Lubold, Erin Walker, and Heather Pon-Barry. Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 255–262, 2016.

[58] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin. Automatically detecting pain in video through facial action units. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 41(3):664, 2011.

[59] Kaixin Ma, Xinyu Wang, Xinru Yang, Mingtong Zhang, Jeffrey M Girard, and Louis-Philippe Morency. Elderreact: A multimodal dataset for recognizing emotional response in aging adults. In *2019 International Conference on Multimodal Interaction*, ICMI '19, page 349–357, New York, NY, USA, 2019. Association for Computing Machinery.

[60] Hector A. Cordourier Maruri, Sinem Aslan, Georg Stemmer, Nese Alyuz, and Lama Nachman. Analysis of contextual voice changes in remote meetings. In *Proc. Interspeech 2021*, pages 2521–2525, 2021.

[61] Akihiro Matsufuji and Angelica Lim. Perceptual effects of ambient sound on an artificial agent's rate of speech. In *Companion of HRI*, pages 67–70, 2021.

[62] D. Matsumoto and P. Ekman. The relationship among expressions, labels, and descriptions of contempt. *Journal of Personality and Social Psychology*, 87(4):529, 2004.

[63] Catherine Mayo, Vincent Aubanel, and Martin Cooke. Effect of prosodic changes on speech intelligibility. In *Proc. Interspeech 2012*, pages 1708–1711, 2012.

[64] Conor McGinn and Ilaria Torre. Can you tell the robot by the voice? an exploratory study on the role of voice in the perception of robots. In *HRI*, pages 211–221, 2019.

[65] Batja Mesquita, Michael Boiger, and Jozefien De Leersnyder. The cultural construction of emotions. *Current Opinion in Psychology*, 8:31–36, 2016.

[66] Meinard Mller. Fundamentals of music processing: Audio, analysis, algorithms, applications, 2015.

[67] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.*, 10(1):18–31, 2019.

[68] Afeefa Muhammed, Ramsi Mol, L. Revathy Vijay, S. S. Ajith Muhammed, and A. R. Shamna Mol. Facial expression recognition using support vector machine (svm) and convolutional neural network (cnn). *International Journal of Research in Engineering, Science and Management*, 3(8):574–577, Sep. 2020.

[69] Behnaz Nojavanasghari, Tadas Baltrušaitis, Charles E. Hughes, and Louis-Philippe Morency. Emoreact: A multimodal approach and dataset for recognizing emotional responses in children. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 137–144, New York, NY, USA, 2016. Association for Computing Machinery.

[70] Nathan Oesch. Music and language in social interaction: Synchrony, antiphony, and functional origins. *Frontiers in Psychology*, 10, 2019.

[71] National Academies of Sciences Engineering and Medicine. *How People Learn II: Learners, Contexts, and Cultures*, chapter 2: Context and Culture, page 21–34. The National Academies Press, Washington, DC, USA, 2018.

[72] Kyung-Geune Oh, Chan-Yul Jung, Yong-Gyu Lee, and Seung-Jong Kim. Real-time lip synchronization between text-to-speech (tts) system and robot mouth. In *ROMAN*, pages 620–625, 2010.

[73] Yusuke Okuno, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. Providing route directions: Design of robot's utterance, gesture, and timing. In *HRI*, pages 53–60, 2009.

[74] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[75] David Pelegrin-Garcia, Bert Smits, Jonas Brunskog, and Cheol-Ho Jeong. Vocal effort with changing talker-to-listener distance in different acoustic environments. *J. Acoust. Soc*, 129 4:1981–90, 2011.

[76] Trinh Thi Doan Pham, Sesong Kim, Yucheng Lu, Seung-Won Jung, and Chee-Sun Won. Facial action units-based image retrieval for facial expression recognition. *IEEE Access*, 7:5200–5207, 2019.

[77] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, Heysem Kaya, Maximilian Schmitt, Shahin Amiriparian, Nicholas Cummins, Denis Lalanne, Adrien Michaud, Elvan Ciftçi, Hüseyin Güleç, Albert Ali Salah, and Maja Pantic. Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, AVEC'18, page 3–13, New York, NY, USA, 2018. Association for Computing Machinery.

[78] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, Siyang Song, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, Mohammad Soleymani, and Maja Pantic. Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, AVEC '19, page 3–12, New York, NY, USA, 2019. Association for Computing Machinery.

[79] L.M. Romanski. 3.20 - specialization of primate ventrolateral prefrontal cortex for face and vocal processing: Precursor to communication. In Jon H. Kaas, editor, *Evolution of Nervous Systems (Second Edition)*, pages 357–370. Academic Press, Oxford, second edition edition, 2017.

[80] Paul Rozin, Laura Lowery, Sumio Imada, and Jonathan Haidt. The cad triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology*, 76(4):574–586, January 1999.

[81] Klaus R Scherer, Heiner Ellgring, Anja Dieckmann, Matthias Unfried, and Marcello Mortillaro. Dynamic facial expression of emotion and observer inference. *Frontiers in Psychology*, 10:508, 2019.

[82] Anna Schouten, Michael Boiger, Alexander Kirchner-Häusler, Yukiko Uchida, and Batja Mesquita. Cultural differences in emotion suppression in belgian and japanese couples: A social functional model. *Frontiers in Psychology*, 11:1048, 2020.

[83] Juergen Schroeter, A. Conkie, A. Syrdal, Mark Beutnagel, M. Jilka, V. Strom, Yeon-Jun Kim, Hong-Goo Kang, and D. Kapilow. A perspective on the next challenges for tts research. In *Speech Synth.*, pages 211 – 214, 2002.

[84] A. D. Sergeeva, A. V. Savin, V. A. Sablina, and O. V. Melnik. Emotion recognition from micro-expressions: Search for the face and eyes. In *2019 8th Mediterranean Conference on Embedded Computing (MECO)*, pages 1–4. IEEE, 2019.

[85] Garima Sharma and Abhinav Dhall. A survey on automatic multimodal emotion recognition in the wild. In G. Phillips-Wren, A. Esposito, and L.C. Jain, editors, *Advances in Data Science: Methodologies and Applications*. Springer, Cham, 2020.

[86] M. Singh, B. B. Naib, and A. K. Goel. Facial emotion detection using action units. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, pages 1037–1041, 2020.

[87] Henrique Siqueira, Sven Magg, and Stefan Wermter. Efficient facial feature learning with wide ensemble-based convolutional neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5800–5809, Apr 2020.

[88] Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. An overview of voice conversion and its challenges: From statistical modeling to deep learning, 2020. [Online] Available:arXiv:2008.03648.

[89] Berrak Sisman, Mingyang Zhang, Minghui Dong, and Haizhou Li. On the study of generative adversarial networks for cross-lingual voice conversion. In *ASRU*, pages 144–151, 2019.

[90] Daisy Stanton, Yuxuan Wang, and RJ Skerry-Ryan. Predicting expressive speaking style from text in end-to-end speech synthesis. In *SLT*, pages 595–602, 2018.

[91] Guangzhi Sun, Yu Zhang, Ron J. Weiss, Yuan Cao, Heiga Zen, Andrew Rosenberg, Bhuvana Ramabhadran, and Yonghui Wu. Generating diverse and natural text-to-speech samples using a quantized fine-grained vae and auto-regressive prosody prior, 2020. [Online]. Available: arXiv:2002.03788.

[92] Johan Sundberg and Maria Nordenberg. Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech. *J. Acoust. Soc.*, 120 1:453–7, 2006.

[93] Selina Jeanne Sutton, Paul Foulkes, David Kirk, and Shaun Lawson. Voice as a design material: Sociophonetic inspired design strategies in human-computer interaction. In *ACM*, page 1–14, 2019.

[94] Wout Swinkels, Luc Claesen, Feng Xiao, and Haibin Shen. Svm point-based real-time emotion detection. In *2017 IEEE Conference on Dependable and Secure Computing*, pages 86–92. IEEE, 2017.

[95] Marie Tahon, Gilles Degottex, and Laurence Devillers. Usual voice quality features and glottal features for emotional valence detection. In *SP-2012*, volume 2, 01 2012.

[96] Akihiro Tanaka, Ai Koizumi, Hisato Imai, Saori Hiramatsu, Eriko Hiramoto, and Beatrice de Gelder. I feel your voice. cultural differences in the multisensory perception of emotion. *Psychol Sci*, 21:1259–1262, 2010.

[97] Ilaria Torre, Jeremy Goslin, Laurence White, and Debora Zanatto. Trust in artificial voices: A "congruency effect" of first impressions and behavioural experience. In *the Technology, Mind, and Society (TMS)*, 2018.

[98] Ilaria Torre, Adrian Benigno Latupeirissa, and Conor McGinn. How context shapes the appropriateness of a robot's voice. In *ROMAN*, pages 215–222, 2020.

[99] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 149–149, 2006.

[100] Rik van den Brule, Ron Dotsch, Gijsbert Bijlstra, Daniel Wigboldus, and Pim Haselager. Do robot performance and behavioral style affect human trust?: A multi-method approach. *Int. J. Soc. Robot*, 6:519–531, 2014.

[101] Ravichander Vipperla, Sangjun Park, Kihyun Choo, Samin Ishtiaq, Kyoungbo Min, Sourav Bhattacharya, Abhinav Mehrotra, Alberto Gil C. P. Ramos, and Nicholas D. Lane. Bunched lpcnet : Vocoder for low-cost neural text-to-speech systems, 2020. [Online] Available:arXiv:2008.04574.

[102] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[103] T. Vo, G. Lee, H. Yang, and S. Kim. Pyramid with super resolution for in-the-wild facial expression recognition. *IEEE Access*, 8:131988–132001, 2020.

[104] Xin Wang, Shinji Takaki, Junichi Yamagishi, Simon King, and Keiichi Tokuda. A vector quantized variational autoencoder (vq-vae) autoregressive neural f_0 model for statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:157–170, 2020.

[105] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram, 2020. [Online]. Available: arXiv:1910.11480.

[106] P. Yang, Q. Liu, and D. N. Metaxas. Boosting coded dynamic features for facial action units and facial expression recognition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, page 1, 2007.

[107] Yusuke Yasuda, Xin Wang, and Junichi Yamagishd. End-to-end text-to-speech using latent duration based on vq-vae. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5694–5698, 2021.

[108] Siyang Yuan, Pengyu Cheng, Ruiyi Zhang, Weituo Hao, Zhe Gan, and Lawrence Carin. Improving zero-shot voice style transfer via disentangled representation learning, 2021. [Online] Available:arXiv:2103.09420.

[109] Yi Zhao, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, Tomi Kinnunen, Zhenhua Ling, and Tomoki Toda. Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion, 2020. [Online] Available:arXiv:2008.12527.

Appendix A

Supplementary Material

Data collection protocol for study 2 (chapter 3) can be found at the link.

Script 1 (Fine Dining):link.
Script 2 (café):link.
Script 3 (Lively Restaurant):link.
Script 4 (Quiet Bar):link.
Script 5 (Noisy Bar):link.
Script 6 (Night Club):link.
Script 7 (Baseline):link.

Appendix B

Code

Code for study 1 (chapter 2) can be found at this link

Code for study 2 (Chapter 3) can be found at: CRANK, ParallelWaveGAN, Feature Extraction (Voice Toolbox),dataset, and Perception study questionnaire

Note: Dataset for study 2 does not contain raw audio files. These will be made available at a later time on FRDR as it contains identifiable human participant data.