Generating Convincing Simulation of Internalized Voices for Human-avatar Interaction

Hyejin Lee

Department of Electrical & Computer Engineering
McGill University
Montréal, Québec, Canada

August 2022

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Science.

2022/08/02

Abstract

Speech synthesis has been developed to mimic the identity of a reference speaker and generate virtual speech in that particular voice. In the medical domain, this technology has recently been utilized to simulate auditory hallucinations in the treatment of schizophrenic patients, for which there are often limited resources available to recreate a convincing simulation. In this thesis, we develop a voice modelling interface paradigm that generates a representation of the voice that lies solely in the individual's head. We first propose three exploration strategies to search for a voice sample that closely matches the user's target voice, and then optimize the output voice with two techniques: latent parameter editing and voice mixing. Through these techniques, we quantify a set of salient vocal characteristics and adjust them, or create new vocal avatars from a low-dimensional voice latent space. To evaluate our approaches and output voices, we conduct two user experiments and a cluster analysis based on multidimensional scaling. We investigate both the performance and usability of our three exploration strategies and two manipulation techniques. The main results demonstrate that our approaches not only achieve superior performance compared to existing voice morphing interfaces but are also capable of finding a convincing match of any human voice imagined by users. The voices generated and customized through our system can be easily transformed into new speech with an arbitrary text input. This way, we aim to allow the general public to depict their internalized voices and improve the quality of the auditory modality in avatar creation or voice-based applications.

Résumé Scientifique

La synthèse vocale a été développée pour établir une reproduction numérique de la voix d'un orateur de référence. En médecine, cette technologie a récemment été utilisée pour simuler les hallucinations auditives de patients atteints de schizophrénie. Cependant, les ressources disponibles pour recréer une simulation convaincante des voix entendues par les patients sont souvent limitées. Dans cette thèse, nous développons un paradigme d'interface de modélisation vocale permettant de générer une représentation satisfaisante d'une voix dont l'utilisateur est le seul à connaître les caractéristiques. Nous proposons, dans un premier temps, trois stratégies d'exploration pour trouver un échantillon qui se rapproche de la voix ciblée par l'utilisateur. Dans un second temps, la voix résultante est optimisée à l'aide de deux techniques: la modification de paramètres latents et le mélange de voix. Ces techniques permettent de retenir un ensemble de caractéristiques vocales saillantes et de les ajuster, ou de créer davantage d'avatars vocaux depuis un espace latent de voix de faible dimension. Afin d'évaluer nos approches et les avatars vocaux, nous réalisons deux études utilisateurs et une analyse de groupes produits par partitionnement multidimensionnel. Nous investiguons à la fois la performance et l'utilisabilité pour nos trois stratégies d'exploration et nos deux techniques de manipulation. Les résultats principaux démontrent que nos approches sont non seulement plus performantes que les interfaces de transformation vocale existantes, mais permettent aussi d'obtenir des voix plus proches de celles imaginées par les utilisateurs. Les voix créées et personnalisées avec nos systèmes peuvent facilement être utilisées pour générer de nouvelles phrases à partir d'entrées textuelles. De cette façon, nous permettons à un public général de créer des voix imaginées, et d'améliorer la dimension auditive des interfaces de création d'avatars, ou dans des applications vocales.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Jeremy Cooperstock, for the continuous support and guidance throughout the entire research. I am genuinely grateful to him for providing me with such critical feedback and input on the manuscripts, encouraging me, and from the first place, giving me the opportunity to conduct this research by agreeing to be my supervisor.

My completion of this dissertation would not have been accomplished without the support of many people. Many thanks to my colleague, Max Henry, for his insight and suggestion on shaping the research ideas. I appreciate Yongjae Yoo who has always generously devoted his time for research discussions and guided me with the design of user studies. My sincere thanks also goes to Antoine Weill—Duflos and Juliette Regimbal for their patient support and helping me set up the interfaces on Bach and Unicorn.

My special thanks to Clara Ducher, David Marino, and Yaxuan Li, who started this journey with me since our day one at McGill. I have been fortunate to have such wonderful friendship that is fun, supportive, and more importantly, inspiring.

I sincerely appreciate all the members of Shared Reality Lab who have been very supportive and warm-hearted. I still remember my first day, our first conversation at the campus food festival, and the warm and welcoming atmosphere of the lab. Above all, their valuable comments on my research design and manuscript have strengthened the ideas and taught me creative thinking. I would also like to thank Max Henry again for the proofreading of the thesis draft, and Clara Ducher for the French translation of the abstract.

Last, but not least, I would like to thank my family for their immense love and support at every step of my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Background

1.1 Introduction

Voice is the first instrument of human beings and is the most powerful tool for transferring knowledge. It is the principal medium through which we communicate our ideas and emotions as well as a unique identifier that projects an individual's personality. Indeed, voices are used in most communications of our daily life including the interaction with machines. Several speech processing technologies use voice as an essential or complementary modality for human-computer interaction.

For example, speech recognition allows computers to comprehend human speech. Popular AI assistants respond to our requests with appropriate solutions, and voice input on mobile phones enables us to send text messages when our hands are occupied or overburdened. Speaker verification extracts the speaker's identity from a set of speech samples. Security systems make use of this technology and employ voices as a biometric measure for personal authentication, supplementing the traditional means such as PIN number or password that can be easily forgotten or stolen.

Among other speech technologies, speech synthesis converts text into new human speech with the identity of a designated voice. Through this technology, screen readers or audio books provide abundant information when the user's eyes are busy or their vision is impaired. Other applications include virtual parties or video games that involve social interaction, which have seen a great rise in demand since the outbreak of COVID-19. In such virtual platforms, voices

can be designed to express the identity of non-player characters (NPCs). Similarly, in medical treatment, virtual speech is used to simulate auditory hallucinations of patients and assist them with controlling their symptoms through therapeutic sessions at clinics.

Then, how do we design a voice and render it in such diverse applications? To generate speech with particular vocal characteristics typically requires access to a recorded audio sample of the target voice, which may not be available, for example, when the voice only exists in someone's head. This thesis explores the reproduction of human voices when there are not any external references available, assisted by machine learning technologies. We develop three interface paradigms to search for target voice(s) by utilizing over 2000 existing voice excerpts. The voices can be refined by our system that parameterizes salient features of voices and customizes them according to the user's needs. Finally, the produced voice can be controlled by a text input and synthesized into new speech.

Through user studies, we investigate the user experience of our tools, the experience of participants using our tools, compare the performance of several techniques we propose, and discuss future work and limitations. We highlight the contributions of this research by introducing the motivational application field of our voice interface paradigms. We illustrate avatar therapy, a psychotherapy technique that treats patients with schizophrenia in a multimodal VR environment. Unlike traditional treatments that involve antipsychotic medications, the therapy encourages patients to create a virtual avatar that represents their hallucination and communicate with the avatar. This interaction is intended to help patients gain control over their imaginary thoughts through the simulation of their real-life experiences in a safe environment. For the process of avatar construction, we aim to reproduce the avatar's voice as close to the hallucinatory voice as possible to improve the level of realism of the therapy.

1.2 Background

In this section, we discuss the background knowledge for understanding the context of this thesis and the key algorithms that we used. First, we discuss the definitions of key parameters that describe the characteristics of a voice. It is then followed by an introduction of Google's multispeaker text-to-speech system, SV2TTS. In SV2TTS, the identity of a voice is represented as a 256-dimensional vector data of which size is fixed regardless of the length of the reference

speech. With the embedding vector, the system generates new speech with linguistic variations conditioned on a specific voice through deep neural networks. The chapter ends by describing two dimensionality reduction (DR) techniques that we utilized to develop our system.

1.2.1 Vocal Attributes Overview

1.2.1.1 Physical Features

In the context of voice synthesis, vocal qualities tend to be categorized into two groups: physical features and perceptual features. Just as how human faces vary with individuals, one's physical features determine an "auditory face", which depends on the size and shape of an individual's anatomical structure. The mechanism for generating a human voice involves the collaboration and subtle controls of different parts of our body. It includes the respiratory system, vocal fold vibration within the larynx,[1] and articulators such as tongue and lips. In this section, we focus on the vocal fold vibration and the vocal tract,[2] and their roles in forming the primary vocal attributes. We now introduce definitions of these vocal attributes that will be used to describe the design and findings of our research in the remainder of this thesis.

1. Frequency: Vocal fold vibration generates the raw excitation signal of a human voice. The immediate sound generated from the vocal fold is often simply described as a buzz. It is resonated and filtered as it passes through the vocal tract. This filtering can be observed by considering the spectral or 'frequency' content of a signal. Frequency is measured in hertz (Hz), which is cycles per second.

2. Fundamental frequency: As a pitched signal, the human voice is a pseudoperiodic signal that can be described in terms of a fundamental tone and a series of higher frequencies which are its integer multiples. Fundamental frequency, often denoted as f0, corresponds strongly to the perceived pitch of the voice. The size and tension of the vocal folds control the temporal variation of the f0 of a voice, which is interpreted as speech prosody.

3. Power: The amplitude of vocal fold vibrations determines the power of a voice, and is measured in decibels (Power-dB).

[1] The hollow muscular organ forming an air passage to the lungs and holding the vocal cords in humans

[2] The airway used in the production of speech, especially the passage above the larynx, including the pharynx, mouth, and nasal cavities

4. Resonance: Resonance of a human voice occurs when a harmonic from the vocal folds matches the frequency of the air in the vocal tract [1]. The resonant frequencies of the vocal tract depend on its length (inversely proportional to), and on the relative variation of its section along the propagation axis.

1.2.1.2 Perceptual Features

Apart from the attributes coming from an individual's physical traits, the perceptual features of a human voice define how the different qualities of a voice are perceived by humans. The perceptual features can not only be determined by the physical features, but can be acquired by the lifestyle or the language of speakers, such as accent or speaking rate. We present brief definitions of several perceptual features that will be useful for understanding the content of this thesis across multiple chapters.

1. Pitch: This feature is a strong perceptual correlate of fundamental frequency [2]. It is the foremost vocal attribute that describes the relative highness or lowness of a voice. In this thesis, pitch will be used interchangeably with fundamental frequency.
2. Accent: This is a distinctive mode of speaking or pronouncing words of a language [3].
3. Speech rate: This term indicates the speed of speech and is measured in words per minute (wpm) [4]. The average speech rate of English conversation is 150 wpm.
4. Loudness: This term refers to the perceptual correlate of sound intensity or level [5].
5. Monotonousness: Some people speak in a more unvarying tone than others. The term "monotonous voice" refers to a voice with low variations in pitch [6].
6. Hoarseness: A hoarse voice indicates a raspy or breathy voice that is often associated with scratchiness in the throat [7].
7. Thickness: This is highly related to the resonance frequencies [8]. A thick voice often presents a darker impression.
8. Emotional Prosody: This term refers to the non-verbal expressions found in speech that convey emotions or the speaker's personality [9]. It primarily consists of intensity, sound duration and fluctuations in fundamental frequency.

1.2.2 Speaker Verification and Voice Creation via Deep Neural Networks

1.2.2.1 Overview

In this section, we introduce a multispeaker text-to-speech synthesis system developed by Google [10], SV2TTS. The system takes an arbitrary speech input, and generates new speech with the input speaker's voice and a given text. The pipeline involves three main stages: encoder, synthesizer, and vocoder. The encoder extracts speaker embedding vectors from a provided speech input, which is a set of features extracted from a projection layer of a Deep Neural Network (DNN). Speaker embedding is used as a unique identifier of speakers that represents an individual's voice qualities. The synthesizer infers a mel spectrogram by combining a text input and a particular speaker embedding. Finally, the vocoder generates waveforms from the inferred mel spectrogram, reconstructing the voice to be audible. In this way, the system is able to produce synthesized speech for an arbitrary voice without the need to train the model again. The underlying technologies of the three models in Google's SV2TTS architecture are referred to as Generalized End-to-End Loss for speaker verification [11], Tacotron2 [12], and WaveRNN [13], respectively for encoder, synthesizer, and vocoder. The three models are independent components that can be trained separately.

1.2.2.2 Encoder

The speaker embedding vector is a 256-dimensional vector that represents the speaker identity of a voice signal. The encoder model is a three-layer long short-term memory (LSTM) network with 768 hidden nodes, follwed by a projection layer of 256 units. The model is trained with 64 speakers, with ten utterances for each speaker. The utterances are split into small time frames of 1.6 seconds overlapped by 50%, and fed into the LSTM network as inputs. The network then produces the hidden state of the last layer as output, which is the 256-dimensional speaker embedding vector.

The encoder model is designed to extract one speaker embedding from one input utterance. If there is more than one reference utterance, the resultant embedding vectors from the same speaker are averaged and L2 normalized. We denote the j-th utterance of the i-th speaker as u_{ij}, and the embedding vector of u_{ij} as e_{ij}. Then, the averaged embedding vector of all utterances from the same speaker can be defined as c_i.

The encoder model of SV2TTS was originally trained for a speaker verification task, where c_i was used to determine whether or not a new utterance belonged to speaker i. This was modified to perform a new task, multispeaker text-to-speech synthesis, where the speaker embedding vectors are utilized to produce a new series of speech samples.

1.2.2.3 Synthesizer

The synthesizer can be described as a text-to-spectrogram system that reproduces the mel spectrogram of an arbitrary speaker's speech by transcribing a given text. If we denote the original mel spectrogram of u_{ij} as x_{ij}, and the synthetically produced mel spectrogram as $\hat{x}_{ij}$, the goal of the synthesizer is to reproduce $\hat{x}_{ij}$ with a minimum discrepancy between x_{ij} and $\hat{x}_{ij}$.

The reconstruction of a speaker's voice only requires a short reference speech of minimum two seconds duration. This is achieved by employing Tacotron2, a recurrent sequence-to-sequence model that features an encoder-decoder structure (Fig. 1.1). The encoder of Tacotron2 is a separate model from the encoder of SV2TTS. It contains three convolutional layers with a Rectified Linear Unit (ReLU) and a bidirectional LSTM, which converts a series of characters from a given text into 512-dimensional embeddings of linguistic and phonetic features. Then, the decoder takes the feature embeddings as inputs and autoregressively predicts a mel spectrogram frame by frame.

In addition to the original architecture of Tacotron2, SV2TTS concatenates the speaker embedding of a given speech to the output of 2 Layer Pre-Net that contains 256 hidden ReLU units (see Fig. 1.1). The concatenation is then projected through a linear projection that skips generating the frame if a value exceeds a certain threshold. The final five convolutional layers improve the overall reconstruction of the mel spectrogram. In this way, the encoder-decoder pair generates a resultant mel spectrogram conditioned on the voice of a reference speech. The output mel spectrogram from the synthesizer will then be converted into waveforms by a vocoder. A more detailed description of the model structure of the synthesizer can be found in the authors' original paper [12].

1.2.2.4 Vocoder

Generating human speech is a complex and expensive task by nature due to the long-range temporal dependencies of audio signals. WaveNet [14] is a probabilistic deep neural network

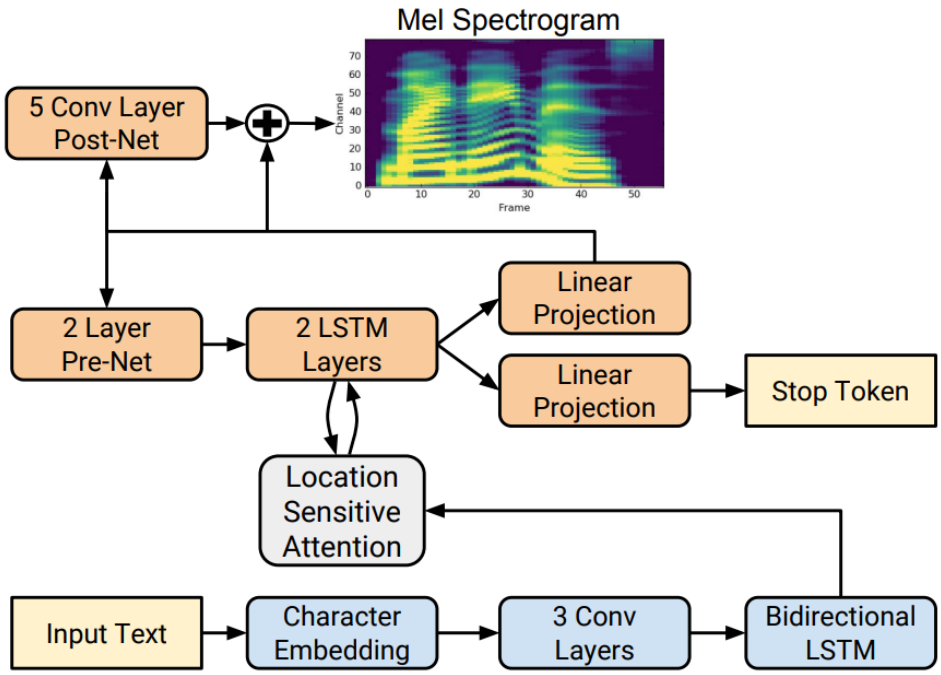


Fig. 1.1: The model architecture of Tacotron2. The blue and orange components respectively represent the structure of encoder and decoder (Reprinted, with permission, from Shen et al. © 2018 IEEE).

model to generate such waveforms. The model makes use of dilated causal convolutional layers with large receptive fields, which cover a long range of time steps and handle the temporal dependencies of audio signals. Also, it contains a fully autoregressive property, where the predictive distribution of every sample is conditioned on all previous samples. The structure of the convolutional layers is described in Fig. 1.2. The output speech of the WaveNet vocoder can be produced from a mel spectrogram that combines the embedding vector of a given speaker and an input text, both of which can be completely arbitrary and unseen during the training period.

The Wave recurrent neural network (WaveRNN) [13] is more compact than WaveNet. It improves the time efficiency to produce each sample, and facilitates speech generation even on low-power mobile CPUs. The model has a reduced computation time, fewer layers, and less overhead compared to those of WaveNet, while maintaining the same quality of output

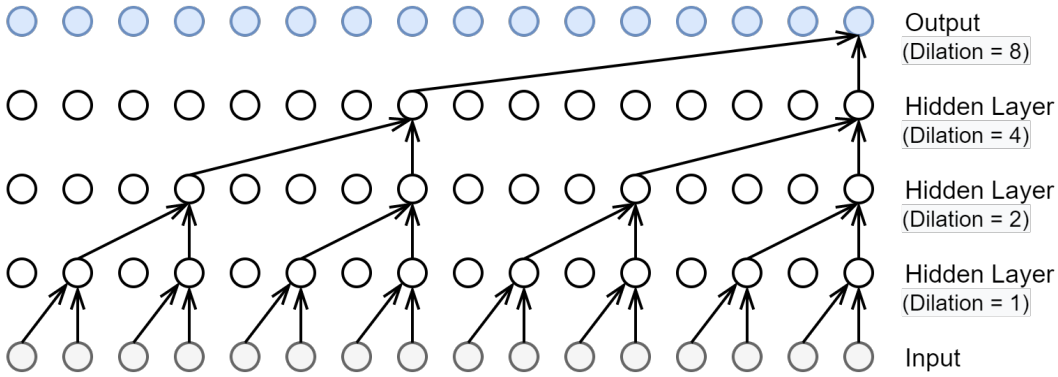


Fig. 1.2: Visualization of the dilated causal convolutions of the WaveNet model. Figure based on reference [14].

speech. In our work, we employ an open-source implementation of the WaveRNN vocoder [15] to facilitate text-to-speech synthesis with a particular speaker identity. More details on how we used the vocoder in our work can be found in Chapter 5.

1.2.3 Dimensionality Reduction

1.2.3.1 Uniform Manifold Approximation and Projection

Dimensionality reduction (DR) is the transformation of data from a high-dimensional space to a low-dimensional space. It is intended to find the lower-dimensional representation of data that best maintains the data structure of a higher-dimensional space. The goal is to minimize the information loss and maximize the interpretability of the data by reducing the number of dimensions. We now discuss Uniform Manifold Approximation and Projection (UMAP) [16], one of the DR algorithms we utilized to visualize a large size of voice data (more details can be found in Chapter 4).

There are two main steps to find a good low-dimensional data organization with UMAP: 1) construct an initial high-dimensional space, and 2) project the representation onto an optimal low-dimensional space. In the first step, UMAP builds a topological representation called “fuzzy simplicity complex” that contains the connectivity of each data. The term was derived from simplicial complex, a mathematical term that means a set composed of points and edges that

connect each point. In a simplicial complex, 0-simplex is a dot, 1-simplex is a line with two points connected by an edge, 2-simplex is a triangle with three points connected by three edges, and 3-simplex is a triangular pyramid with four points connected by six edges. Hence, simplicial complex κ can be defined as a set of simplices that satisfies :

1. Every face of a simplex from κ is also in κ.
2. The non-empty intersection of any two simplices $\sigma_1, \sigma_2 \in \kappa$ is a face of both σ_1 and σ_2.

To build a simplicial complex, n data samples on a high-dimensional space are first surrounded by n radii respectively. If two radii have a non-empty intersection, a 1-simplex is created i.e., an edge is added between the two points. Similarly, if n radii have a non-empty intersection, a $n-1$ -simplex is created. This process is called the Čech complex, a combinatorial metric that converts data samples into a simplicial complex. In this stage, the size of every radius is equivalent, which raises some problems; points in low density areas cannot be connected to each other, and points in high density areas are too connected, resulting in having unnecessarily high-dimensional simplices. To resolve this, different sizes of radii are given to each point based on the point's k-nearest neighbours, where k is a hyperparameter in UMAP that indicates the number of nearest neighbours to consider. With k-nearest neighbours, the diameter of each radius is determined by the distance between the point and its k-th nearest point. In the end, isolated points are surrounded by a larger radius than the points in populated areas. The ultimate role of the hyperparameter k is to balance the preservation of global and local structure. A bigger k leads to the projection that focuses more on preserving the global structure and loses some fine details in local areas. In contrast, a smaller k results in a projection that is more focused on the local elements.

As the name "fuzzy simplicial complex" implies, the certainty of a point belonging to a radius becomes fuzzier as it is far away from the center of a radius. Similarly, the connectivity between points is not defined by binary values, but by the distance between two points that share an intersection. Weights are given to each edge to represent the connectivity, and are later utilized to determine the distance between data points on a low-dimensional space. Since each point has its own local metric to calculate the weight of edges, there exist edges that connect the same points but have weights that are not equal. In Fig. 1.3, for example, point a is connected to its three nearest neighbours, with point b as the third nearest neighbour from a. Point b

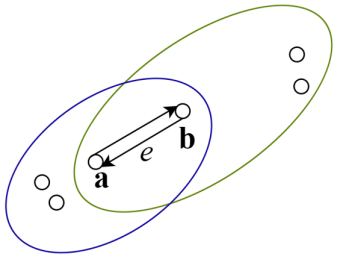


Fig. 1.3: Two points that assigned different weights for the same edge (e). The blue and green circles respectively indicate the three nearest neighbours of points a and b.

is also connected to its three nearest neighbours including a, but with a being the nearest neighbour. In this case, the weight of the edge from point a to b (w_1) is smaller than that from point b to a (w_2), as the local metric of point a did not assign a large weight to the edge with point b. To resolve this incompatibility, UMAP calculates a union of the different weights as in Equation 1.1. This process results in having only one combined edge between neighbours.

$$a + b - ab \tag{1.1}$$

From the first step, a fuzzy simplicial complex was built on a high-dimensional space with edges of different weights. The weight was found to be a good indicator of the distance between points as it represented the likelihood of the points being connected. In the second step, we use these weights to compute the optimal low-dimensional approximation of the high-dimensional representation. The best match would have to minimize the discrepancy between the two representations. We thus use binary cross-entropy (CE) as a loss function and aim to find the minimal loss. We denote E as every possible set of 1-simplices, and e as a 1-simplex that belongs to E. $w_h(e)$ is the weight of e on the high-dimensional space, and $w_l(e)$ is the weight of e on the low-dimensional space to be approximated. The CE loss function for balancing $w_h(e)$ and $w_l(e)$ can be defined as in Equation 1.2.

$$H(w_h(e), w_l(e)) \triangleq \sum_{e \in E} w_h(e) \log \frac{w_h(e)}{w_l(e)} + (1 - w_h(e)) \log \frac{1 - w_h(e)}{1 - w_l(e)} \tag{1.2}$$

In Equation 1.2, e in a populated area is assigned a low $w_h(e)$, which encourages $w_l(e)$ to be smaller to minimize the loss, pulling the points together on the low-dimensional space.

Conversely, e in an isolated area is assigned a high $w_h(e)$, resulting in a larger $w_l(e)$ and pushing the points toward each other on the low-dimensional space. With this push-pull mediation, UMAP creates the best low-dimensional manifold that balances the preservation of local and global structure.

1.2.3.2 Principal Component Analysis

Principal Component Analysis (PCA) [17] is a dimensionality reduction technique that transforms correlated data onto a new lower-dimensional subspace such that it best retains the information of the original data. The axes of the subspace are formed by k principal components, which are the eigenvectors extracted from the covariance matrix of the raw data. We denote the raw data matrix with n data points and d dimensions, as a $n \times d$ matrix X. A $d \times k$ matrix V is the loading vector with which we multiply the data matrix X to compute the matrix Z to construct the new space as in Equation 1.3. PCA can be summarized as a process to identify the k principal components to construct Z.

$$Z_{nxk} = X_{nxd} V_{dxk} \tag{1.3}$$

The process includes four main steps: data standardization, computing the covariance matrix, calculating the eigendecomposition from the covariance matrix, and data projection. The principal components are extracted in descending order of contributions to explain the data; the first principal component captures the largest variances of the data.

In the first step, the initial data is centered and normalized so that all features would have equal variances and are assigned equal weights. First, the data is shifted such that the average of the variables becomes zero. For this, the mean value is subtracted from every data point. This process does not affect the relative positions of the data points, but simplifies further computations. The centered values are then divided by the standard deviation of each feature for scaling.

The second step corresponds to calculating the covariance matrix that represents the variance of the data and covariance of each pair of n variables. The formulas for computing the variance and covariance are defined in Equation 1.4 and 1.5, where x_i and y_i are the i-th sample of variable x and y, and $\overline{x}$ and $\overline{y}$ are the corresponding mean values.

$$var(x) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \tag{1.4}$$

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1} \tag{1.5}$$

The covariance matrix can be defined as:

$$[x_1, x_2, ...x_n] = \begin{bmatrix} var(x_1) & cov(x_1,x_2) & \dots & cov(x_1,x_n) \\ cov(x_2,x_1) & var(x_2) & \dots & cov(x_2,x_n) \\ \vdots & \vdots & \ddots & \vdots \\ cov(x_4,x_1) & cov(x_4,x_2) & \dots & var(x_n) \end{bmatrix} \tag{1.6}$$

Then, the eigenvectors and eigenvalues are calculated from the covariance matrix in order to identify the principal components. An eigenvalue, often denoted as λ, is a scaling factor of the linear transformation represented by an eigenvector. The eigenvalues of a matrix A are the values that satisfy $|A - \lambda I| = 0$. The eigenvectors of the corresponding eigenvalue are the vectors v that satisfy $(A - \lambda I)v = 0$. By sorting the eigenvectors in descending order of the magnitude of the eigenvalues, we can identify the most significant k eigenvectors as k principal components, where k is smaller than d. Once we obtain the k principal components, the original data are rearranged and projected onto a new lower-dimensional space. On the new space, each principal component accounts for x% of the variance of the original data, where x is their corresponding eigenvalue divided by k.

Chapter 2

Literature Survey

2.1 Exploration of Primary Perceptual Features of Voices

The voice is a complicated signal that carries a considerable amount of information about the speaker. The mechanisms of voice production have been explored and explained from the perspective of anatomy, however, the human perception of voice is a separate question from anatomical analysis. Aside from objective, low-level features such as pitch, a voice signal carries many qualities which are subjective in nature. It is also not surprising that the prominent characteristics of voices are different from listener to listener, and research to research. It is therefore not an easy task to perform a purely quantitative assessment of voice qualities and their relative importance. In this review, we highlight studies exploring the correlation of descriptive voice qualities and human perception.

Multidimensional scaling (MDS) is a data analysis technique that uses recorded pair-wise similarity ratings to map items into a representation space. In the study of voice qualities, large sets of voices are visualized as points in such a space, based on similarity judgements from human listeners. These spaces represent the meaningful underlying perceptual dimensions for voice.

In their MDS study with 22 human listeners, Murry et al. found a first dimension correlating strongly to pitch [18]. Walden et al. also found that pitch and speech rate showed the strongest correlations, followed by the age of speakers [19]. Gelfer et al. found a first dimension that correlated to both pitch and perceived resonance of the voice, with a second dimension including

speech rate and perceived age [20]. These descriptors were evaluated based on participants' perception on a 9-point Likert scale (e.g., young — old). Other characteristics such as intensity, hoarseness or monotonousness were considered as minor factors. Pitch appeared again in another study conducted by Nolan et al., as the main perceptual parameter for judging voices [21].

The above review summarized important vocal attributes that influence our perception of different voices. In summary, a common finding from the studies was that pitch carried the most important information for all judgements. This was followed by speech rate, age, loudness, or resonance of the voice, with the order of significance varying slightly with different studies.

2.2 Human Perception of Personalities of Synthesized Voices

Nass et al. investigated the perception of two synthetic voices of opposite personalities, an extroverted voice and an introverted voice [22]. The personalities were created by changing the speech rate, intensity, pitch, and pitch variation to manifest the extroversion and introversion of the speaker. To assess the degree of such traits, they provided several adjectives as indices such as cheerful, enthusiastic, and extroverted, and asked participants to assess them on a Likert scale. Their results demonstrated that participants evaluated the same content more positively when it was spoken by a voice of their preferred personality. The research emphasizes the importance of "casting" an appropriate synthetic voice whose personality matches the role of a particular avatar, as it is highly associated with user perception and behaviour.

A recent study conducted an in-depth evaluation of the personalities of synthetic voices based on the Big-Five personality test [23]. Synthetic voices were demonstrated to express various personalities such as extroversion, conscientiousness, and agreeableness. Perceived voice naturalness appeared to have little impact on speaker personality, indicating that the formation of personality and the naturalness of synthetic speech can be treated as separate research questions. Another study revealed that speech synthesis can change perceived emotion of the speaker [24]. In their research, adjusting voice qualities (e.g., tense and lax voices, often described as having a higher or lower degree of tension in the vocal tract [25]) of the synthetic voice resulted in changing the perceived stress level of the speaker.

2.3 Speech Processing Technologies

2.3.1 Audio Feature Analysis

Audio analysis programs such as Praat [26] and Tony [27] are widely used to analyze voice sounds. In this section, we particularly discuss Praat, a free software package that is used for the analysis and synthesis of acoustic speech signals. Praat features various audio analyses such as intensity analysis, phonetic measurement, sentiment analysis, and syllable analysis.

Praat visualizes mel-spectrogram with an overlay of continuous changes of fundamental frequency (f0), loudness, and formant pitch of a given speech on a two-dimensional plot with time and f0 as the x and y axis (See Fig 2.1 for the user interface). Previous research compared the accuracy of different pitch detection algorithms, and showed that Praat scored one of the lowest error rates [28].

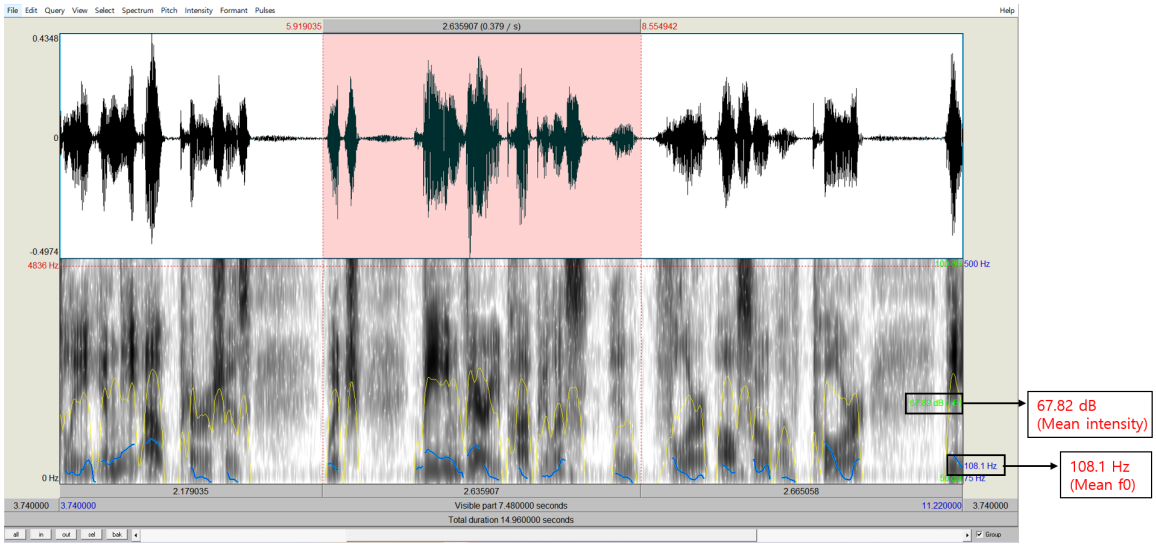


Fig. 2.1: The spectrogram analysis of the Praat software. The average fundamental frequency and intensity are extracted from the selected time frame of a speech file, respectively indicated in blue and green text on the bottom right side.

In syllable analysis, Praat provides information on the number of syllables and the speech duration of a voice file. It does not require any transcription, but counts the peaks in intensity during the duration of speaking as the number of syllables. Speech rate is calculated as the number of syllables per duration of speaking time. Previous work has confirmed this method

of syllable detection and speaking rate analysis [29].

2.3.2 Voice Morphing Interfaces

Voice changers are commonly used by gamers and on social media platforms such as Discord, Steam, Skype and Facebook messenger to disguise the voice. Conventional voice changers change vocal pitch, and add distortions to the voice through the use of simple effects. They apply the saved effect to a pre-recorded audio file or to a speaking voice in real time. We introduce two commercial voice morphing tools, MorphVOX Pro [30] from Screaming Bee incorporation, and AV Diamond Voice Changer [31] developed by AV SOFT corporation.

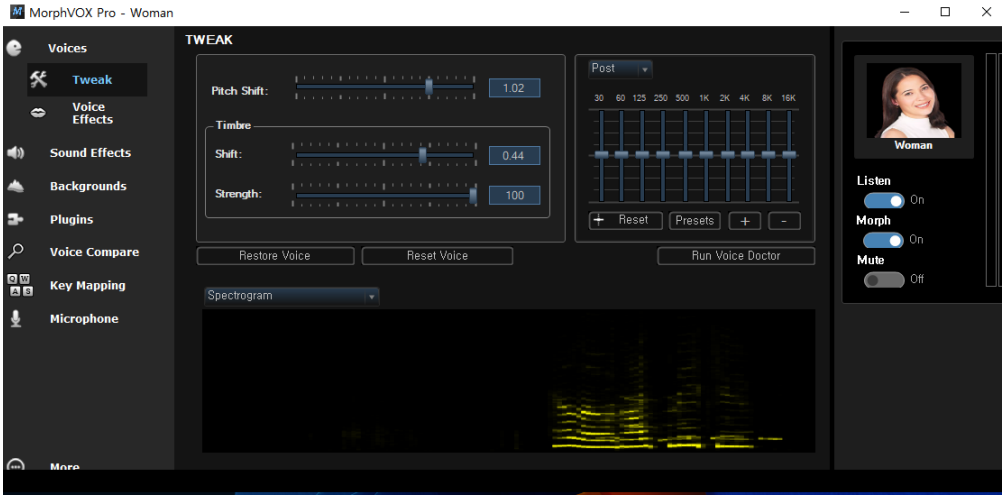


Fig. 2.2: The Graphical User Interface of MorphVOX Pro voice changing software [30] from Screaming Bee Inc.

MorphVOX Pro is a voice changing software targeted at online gamers. As seen in Fig 2.2, the three main controls are pitch shift, timbre shift, and timbre strength. Pitch shift adjusts the pitch of a voice within a range of -2 to $+2$ octaves. The timbre shift renders the voice to sound either deeper (more resonant) or more shrill (less resonant), mimicking the effect of a longer or shorter vocal tract. Lastly, timbre strength alters the intensity with which the timbre shift is applied.

AV Diamond voice changer is another example of a type of voice morphing software. In addition to pitch and timbre, the software allows for adjusting gender and even perceived age.

Fig. 2.3: The Graphical User Interface of AV Voice Changer Software Diamond [31] from AV SOFT CORP.

The GUI presents a two-dimensional slider where vertical movement alters timbre, and horizontal movement alters pitch of the voice. It also provides filters that attempt to transform the user's voice into the voice of particular celebrities, e.g., Adele and Marilyn Monroe, by adjusting the three aforementioned attributes. Other examples of voice morphing tools include Voicemod [32] and Skype Voice Changer [33].

2.4 Speech Corpus Review

We investigated six datasets that are commonly used to train machine learning models. From these, we selected two for use in our interface (discussed in Chapter 4 and 5). The six datasets are LibriSpeech [34], Speech-accent Archive [35], VoxCeleb series [36] [37], Common Voice [38], CSTR VCTK [39], and TIMIT [40]. All datasets contain multispeaker English speech samples and are publicly available. A summary of our comparison can be found in Table 2.1.

1. LibriSpeech: LibriSpeech consists of 2484 clear voices reading English audiobooks from the public domain. This dataset is used for training speech recognition and speech synthesis models due to the long spoken time per speaker.

2. The speech-accent archive: This dataset contains 2953 voices reading a uniform transcription with phonetically rich sentences. The speakers have a variety of language backgrounds.

3. VoxCeleb: VoxCeleb is an audio-visual dataset that was extracted from celebrity interview videos on Youtube.

4. Common Voice: This dataset is known for its immense size and it continues to grow as more participants contribute to the project. The samples recorded from individual participants are submitted and verified by other participants to guarantee the quality.

5. CSTR VCTK: It consists of 110 speakers reading about 400 sentences. The corpus was originally aimed for speech synthesis and speaker adaptation technologies.

6. TIMIT: The TIMIT database contains 630 speakers with eight major dialects of American English. The samples were carefully transcribed and recorded in a regulated laboratory environment.

Below is a detailed description of the terminologies used in Table 2.1.

- Popular usage: This indicates the popular use of the dataset in previous research.
- Primary source: This indicates the way the samples were recorded or created. The samples produced from a regulated lab environment have higher greater intelligibility than the samples recorded by individual participants.
- Speakers: The number of speakers that the dataset contains.
- Average speech length: The average length of recordings per speaker. Some entries are written by the approximate time length, and some are written by the number of sentences.
- Gender: Gender distribution of the recordings.
- Percentage of North American English: What percentage of North American English accents are included in the dataset, by number of speakers. Only the Common Voice data set includes speech samples of other languages.

Table 2.1: Comparison table of six speech corpora.

	Librispeech	Speech-accent Archive	VoxCeleb	Common Voice	CSTR VCTK	TIMIT
Popular usage	Speech recognition, Speech synthesis	Speech recognition, Accent analysis	Speaker identification, emotion recognition	Speech recognition	Speech synthesis	Speech recognition
Primary source	Derived from audiobooks	Recorded by participants and verified by researchers	Extracted from Youtube videos	Recorded by participants and verified by researchers	Recorded by participants from a well-regulated lab environment	Recorded by participants from a well-regulated lab environment
Speakers	2484	2953	7363	66,173	110	630
Average speech length	Approximately 24 minutes	Approximately 30 seconds	Approximately 17 minutes	1.5 minutes	4 sentences	10 sentences, 30 seconds for each
Gender	51.7% Male, 48.3% Female	51.5% Male, 48.5% Female	61% Male, 39% Female	47% Male, 15% Female, 38% Unlabelled	57.5% Female, 42.5% Male	70% Male, 30% Female
Percentage of North American English	Unknown, but directed towards North American English accent	17.4% American, 2% Canadian	29% American	24% American, 3% Canadian	20.2% American, 7% Canadian	100% (8 different American dialects)
Speaker metadata	Gender	Gender, Native language, Age, Country	Gender, Country	N/A	Gender, Accent, Age, Region	Gender, Dialect, Birth date, Race, Height
Recent release	2015	2021-06-19	2019	2020-12-11	2019	1988

- Speaker metadata: Relevant Speakers demographic information that was collected by the creators and included as labels in the dataset.
- Recent release: Date of the most recent release of the dataset as of the time of this writing.

Chapter 3

Avatar Therapy in a Multimodal Virtual Reality Environment

In this chapter, we introduce avatar therapy, a medical application where patients can interact with a virtual avatar and overcome their auditory verbal hallucinations. Our goal is to recreate the voice of their hallucinations with a high level of fidelity and achieve a realistic simulation for the therapy. We provide an overview of the environment, discuss the construction of an avatar and its voice, and describe how the voice will be used to reduce the frequency and intensity of the psychotic episodes of patients.

3.1 Introduction

Avatar therapy is an effective and safe psychotherapy technology for treating patients with schizophrenia. The therapy allows patients to create a virtual avatar, and to interact with the avatar in a multimodal virtual reality (VR) environment. The avatar's appearance and voice are rendered to match the patient's visual and auditory hallucinations. From the interaction with the avatar, patients learn to gain control over their imaginary thoughts, and to further reduce the associated distress and helplessness.

Auditory verbal hallucination accounts for approximately 60 to 70% of all sensory hallucinations in schizophrenia [41]. The level of fidelity of the avatar's voice, which gives a sense of realism that is comparable to the patient's experience, has been found to significantly contribute to the delivery of therapy [42]. Thus, it is important to find a voice representation that closely matches the tone and timbre of the voice that the patient hears. In this chapter, we introduce the concept of avatar therapy in more detail, and discuss its efficacy as validated from previous work. Then, we present the architecture and implementation of our system.

The research described in this thesis was conducted as part of a project that simulates hallucinations using different modalities and investigates their effects in avatar therapy. The author's contribution is found in the integration of a 3D avatar model (Section 3.3.1), the implementation of avatar locomotion (Section 3.4), and avatar lip animation (Section 3.5). The author was not involved in the design and implementation of the haptic technology (Section 3.3.2) and physiological measurement (Section 3.3.3).

3.2 Background

Schizophrenia is a psychiatric diagnosis that interferes with basic human perception, emotion, and the ability to interpret reality [43]. Major symptoms include anxiety and paranoia associated with visual, auditory, and tactile hallucinations. Schizophrenia affects approximately 1% of the Canadian population, of whom 40 to 60% have attempted suicide [44]. Unfortunately, about one in three patients shows resistance to conventional treatment methods involving antipsychotic medications, and continues to experience hallucinations [45] [46].

Cognitive Behavioural Therapy (CBT) is a non-pharmacological mental treatment that helps patients to challenge their delusions and cognitive distortions. Avatar therapy is a CBT-inspired

approach for schizophrenia, employing a face-to-face conversation between patients and an avatar representing their hallucination in a VR environment. The purpose of this interaction is to provide patients with the power to control their imaginary thoughts over the course of therapy. During the session, clinicians monitor the physiological responses of the patients to measure their stress level, and evaluate the progress of the therapy. Clinicians also assist with the avatar's interaction by controlling the avatar's locomotion and speech. Previous research has established the effectiveness of this therapy as a viable treatment for schizophrenia, particularly for medication-resistant auditory hallucinations [47] [48] [49].

3.3 Environment Overview

In our pipeline of avatar therapy, patients participate in an initial avatar creation step, where they modify the avatar's face and voice. The avatar is then made into a 3D full-body model, and displayed to the patients in a virtual environment. As the patients speak to the avatar, clinicians respond to their words. Their speech is transformed into text by automatic speech recognition, which is then synthesized into the avatar's speech with the voice initially created by the patients. Finally, the avatar's speech is synchronized with its body movements and lip animation for interacting with the patients.

3.3.1 Avatar Construction

Fig. 3.1: Realistic faces of the 3D avatars rendered from a single image.

Building an avatar requires the integration of multiple components—its visual appearance, voice, and locomotion. To generate the avatar appearance, we utilized an avatar creation tool from previous work based on Generative Adversarial Network (GAN) [50]. The tool generates 2D images of photorealistic synthetic faces from a sequential editing of facial components. We converted the 2D image into a 3D human head model through an image-based modelling provided by Avatar Maker [51]. This process both maintained the photorealistic quality of the avatar's appearance, and generated the 3D model in a short amount of time (See Figure 3.1). Then, we attached a body to the head model as shown in Figure 3.2. The 3D full-body model was produced with the Unity Multipurpose Avatar (UMA) asset [52]. The current challenges include enhancing the unrealistic appearance of the 3D model, which arises from several elements such as hair style or the magnitude of each body part. In order to render the avatar voice, we attached an audio source to the assembled 3D avatar by using the Unity game engine that allowed for spatial audio rendering.

3.3.2 Haptic Sensation

While this work has not yet been implemented, our group plans to recreate haptic sensations for interaction with an avatar. Among different types of sensory hallucinations, tactile hallucinations occur in about 10 to 20% of patients with schizophrenia, often as a false sensation of touch or movement on the skin [53]. Targeted sensations include tapping, stroking, tickling or hugging. To generate such stimuli over the patient's body, we use a full-body haptic VR suit such as Teslasuit [54] or Tactsuit [55], and design multiple actuators for the body parts where the majority of human touch happens — arms and torso. Besides vibration or force feedback, technologies such as pneumatic actuation can also be utilized to augment natural human touch by producing tactile feedback through compressed air [56].

Fig. 3.2: Image of a 3D full-body avatar. The avatar was animated to perform basic motions such as walking.

3.3.3 Physiological Measurement

Avatar therapy involves continuous data collection of the patient's physiological response to the avatar, and monitors how it evolves over multiple sessions of the therapy. The current design employs the uses of wrist-worn photoplethysmography (PPG) [57] and electrodermal activity (EDA) [58] sensors as physiological indicators. The sensors monitor heart rate variability and skin conductivity respectively. These metrics serve as indices of patient autonomic arousal in response to interactions with the avatar. The measurement also involves recording the patient speech. Patient speech is analyzed for changes in prosody and mean fundamental frequency, as well as other perceptual features such as intensity and the emotional timbre of voices. The collected data can be used to assess the overall progress of the therapy, and as a safety measure to ensure that patients are not overly distressed throughout the administration of the therapy.

3.4 Avatar Locomotion

Avatar locomotion refers to any movements of a virtual avatar. Locomotion includes full-body movements, lip animation, and facial expressions such as eye blinking. Realistic avatar locomotion brings the avatars to life, by letting them display emotions and have distinctive personalities. In Unity, avatar movements are rendered by attaching an animation clip, a designated pattern of animations, to the avatar. In our virtual environment, we used animation clips to construct several body movements such as walking, turning heads, and hand gestures while talking. To animate the face, we used SALSA LipSync Suite [59] that provided a preset of lip shapes for lip synchronization and subtle movements of the eyes. This has a limitation in that the lip movements are constrained to the elements in the preset, and further motions require additional 3D modelling.

3.5 Lip Animation Synchronized with Rendered Voice

3.5.1 Literature Review

Lip synchronization is a technical term for matching human lip movements with singing or speech signals, aiming to achieve a perfect synchrony of the motion and sound. Many video games and computer-animated cartoons make extensive use of lip synchronization techniques.

Several methodologies have been developed in order to achieve timely and accurate lip animation. The most frequently used methods include “audio-driven”, “viseme-based”, and “motion-based” lip synchronization. The audio-driven technique produces lip motions based on analyzing the incoming speech signal. For example, Kumar et al. extracted lip landmarks from an image, and used a long short-term memory (LSTM) network to predict 2D lip animations, using audio from a Barack Obama speech [60]. Viseme-based lip synchronization converts incoming speech into phonemes (the smallest unit of pronunciation), and maps them to lip patterns. This technique has been utilized in many previous works and can achieve high accuracy in finely detailed lip movements [61, 62, 63]. Lip synchronization can also be driven by motion tracking, which captures the user’s head motion, lip motion, and facial expressions in real time [64].

3.5.2 Implementation

We selected the viseme-based technique to implement lip movement in our avatar engine. Our approach does not require any model training to predict lip animation as in the audio-driven approach, and is able to generate movements even in the absence of a driver for motion tracking. The following sections describe the pipeline and implementation of our method from collecting the therapist’s speech to generating avatar lip animation.

3.5.2.1 Speech-to-text conversion

In our system, the therapist’s speech is analyzed in the following signal chain to allow for lip animation. By using the Google Cloud Speech-to-text API, we transcribed the speech into a series of written text, and tracked the start and end time of each word being uttered. The time data were later used to enhance the level of synchrony in the lip motion and incoming speech.

3.5.2.2 Text-to-Phoneme Mapping

We used the Carnegie Mellon University (CMU) dictionary [65] to analyze the transcribed text. The CMU dictionary is an open source glossary for the pronunciation of North American English. It is based on ARPAbet [66], a common phoneme set developed for speech recognition research, and contains 134,000 words and their pronunciation. Our program searches each

word from the dictionary and returns the corresponding phoneme set, assuming that the word exists in the dictionary. If the word is not in the list, it returns the phoneme set of the nearest word.

3.5.2.3 Phoneme-to-Viseme Implementation

The transcribed phoneme set is sent to the Unity environment in real time, using Transmission Control Protocol (TCP) socket communication. The phonemes trigger visemes (lip shapes) of the avatar, produced by SALSA LipSync Suite [59]. While visemes are meant to be a visual equivalent of the phoneme, they do not always have a one-to-one correspondence. To compensate for this discrepancy, the team collected a phoneme-to-viseme mapping based on the mappings presented in previous literature [67, 68]. We mapped 39 phonemes from the CMU dictionary to fifteen viseme classes, as shown in Table 3.1. Figure 3.3 shows several examples of visemes attached to our 3D avatar.

Table 3.1: A mapping table between phonemes and visemes used in the lip synchronization. The set of phonemes were extracted from the CMU phoneset [65] and viseme classes were produced from SALSA LipSync Suite [59].

Viseme ID	SALSA Viseme Classes	CMU Phoneset	Examples
1	IH	IH, EH, AE, EY, AY	bit, bet, bat, bait, bite
2	OH	UH, UW, AH, ER	book, boot, way, but, hurt
3	AA	AA	bott
4+5	OH + OU	AW, AO, OW	bought, boat, cow
6	EE	Y, IY	yacht, beet
4 7	TH	DH, TH	then, thin
8	D	T, D, L	tea, day, lay
9	F	F, V	fin, van
10	K	K, HH, G, NG	key, hay, gay, sing
11	N	N	noon
12	P	B, P, M	bee, pea, mom
13	R	R, W	ray, we
14	S	S, Z	sea, zone
15	CH	CH, ZH, SH, JH	choke, azure, she, joke

Each phoneme is used to activate one viseme for a certain amount of time. This duration is calculated according to Equation 3.1. Here, T indicates for how long each lip motion is

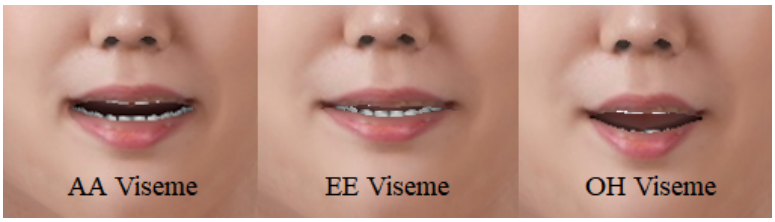


Fig. 3.3: Example visemes from a generated 3D model.

activated. The *start time* and *end time* variables indicate the desired start and end time of the word being pronounced, and are measured in microseconds. The n variable represents the number of phonemes included in each word.

$$T = (endtime - starttime)/n \tag{3.1}$$

3.6 Conclusion

This chapter provided a brief overview of the motivational application of the work of this thesis. In avatar therapy, patients enter into a virtual environment to confront their hallucination, and learn to cope with their anxiety and paranoid thoughts. Our virtual environment for avatar therapy featured a 3D avatar that simulates the visual and auditory hallucination of the patient. The author implemented a 3D full-body avatar, and constructed its body movements and lip synchronization. The lips were animated in response to incoming speech for which the content is transcribed into text, divided into phonemes, and then mapped to a series of lip shapes. This application raised the need for rendering the avatar voice to match the patient's auditory hallucination, and generating its speech in that voice.

Chapter 4

Interface Paradigms for Navigating Voice Space

Preface

In this chapter, we develop three interface paradigms that help a user find a voice to match a reference in their head. We build three different voice maps: a plane based on two vocal features, and two maps organized by a dimension reduction algorithm. As one of the attributes, we utilize speaker's vector data to arrange a large collection of voices based on their speaker identities. Many of the following techniques rely on this speaker vector, a (X-dimensional) vector corresponding to an individual speaker, that is learned by a neural network trained on a large corpus of speech samples. We evaluate the performance, usability, and user experience of these interfaces through a user study.

Author's Contribution and Acknowledgement

Hyejin Lee designed and implemented the voice exploration interfaces, conducted user studies, analyzed results, and wrote this chapter. Professor Cooperstock supervised the research and edited the manuscript. We thank Clara Ducher for the suggestion of developing the Voice UMAP and Nicolas Bieber for his support in implementing the user interfaces. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), MEDTEQ, iMD Research and IA Précision Santé Mentale. McGill REB #20-08-023.

4.1 Introduction

It is not an easy task to describe a voice that only exists in one's mind. It is even difficult to describe a voice in objective terms; some might refer to traits such as gender, age and ethnicity. Others might use more voice-specific features such as pitch, loudness and tone. Even when we think we have precisely described a voice to others, it is almost impossible to ensure that the voice in our head matches the one being described. A search mechanism should present voices in a way that facilitates exploration. This requires a large number of voices so that users can find a convincing match to their target voice among the given voices. An efficient search paradigm avoids use of technical jargon in the implemented user interface.

We propose three interface paradigms that each make use of different visualization techniques and machine learning algorithms. To this end, we utilize 2140 speech recordings and their speaker-related metadata from the speech-accent archive [35]. As additional features, we convert the voice waveforms into mel-spectrograms and analyze the pitch and speech rate of the voices by using Praat [26]. From these features, we create a visualization that represents the distances between voices based on their abstract qualities. Our analysis generates high-dimensional speaker vectors, so for two of the three interfaces, we apply dimension reduction techniques to map the voices to a two-dimensional space. In this chapter, we describe the design and implementation of the interfaces, present our experimental protocol for their evaluation, and discuss the results.

4.2 Interface Paradigms

4.2.1 Traditional 2D Exploration

First, we took a naive approach, arranging voices according to two features on a simple 2D plot.[1] Labels included three speaker-specific traits (age, gender, and language) provided by the metadata, and two primary acoustic features (mean f0 and speaking rate) we extracted by using Praat's [26] mel-spectrogram analysis and syllable analysis. The dataset contained 2140 speech files (MPEG Audio, 44.1 kHz, 1 channel) recorded in 214 different first languages of the speakers, and we classified the languages into 34 accent groups based on their language

[1] The visualization was built with Plotly.js package [69].

families. Mean f0 and speaking rate were respectively labelled as pitch and speed to avoid technical jargon and provide familiar terms to non-expert users.

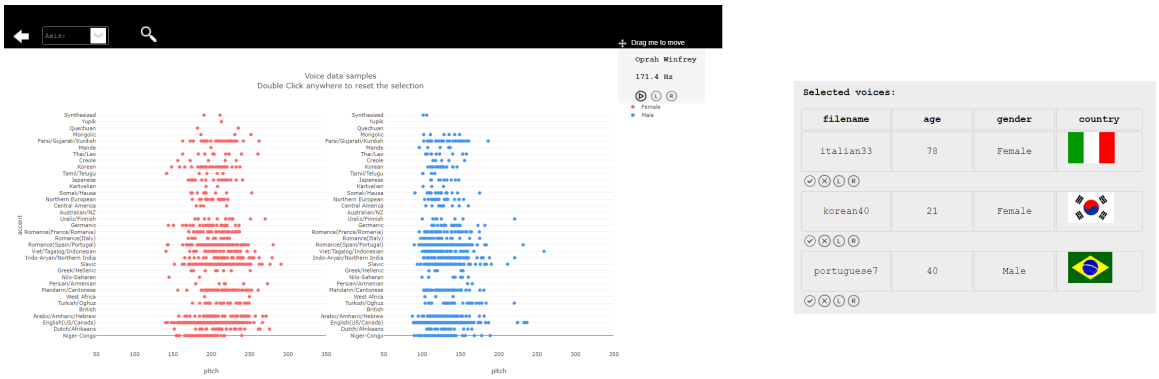


Fig. 4.1: Design of the Traditional 2D Plot. Voices are displayed on an interactive plot where the x and y axes respectively correspond to mean pitch and accent of the speaker.

The design of the interface is illustrated in Figure 4.1. Users can zoom in or out, and pan to interact with the 2D plot. They can listen to the speech samples on mouse hover, and save it to a list on mouse click. Two buttons (marked as L and R) were provided for users to audition two speakers at a time, in the left and right ears. Additionally, the system allows users to upload their voice recording in case they have a suitable reference of their target voice. Once a user uploads such a recording, the system computes the mean pitch of the uploaded file and displays the result in Hz on the top right corner of the screen. Based on this information, users can learn the location of voices that are close to their target voice in terms of voice pitch.

The voice data were originally plotted on two-dimensional plots in three predetermined combinations we thought might be useful, presented to the user one at a time. The combinations included a pitch-accent plot, an age-accent plot, and a pitch-speed plot. We observed that participants from our pilot study most frequently chose to visualize the voice data using the pitch-accent plot and thus, we selected this plot to investigate through our main experiment.

Despite its conciseness and ease of use, this visualization has a few drawbacks. Users from our pilot study found it difficult to find a voice with a desired accent, since accent is a non-continuous attribute that cannot be arranged in ascending/descending order. To fix this issue, we designed a new visualization tool that compresses multiple dimensions into a two-dimensional space by using a dimension reduction algorithm, which is discussed in the next

section.

4.2.2 Voice UMAP

To efficiently search for a target voice, we built a voice map based on four features: age, gender, speech rate, and mean pitch. A dimension reduction technique, UMAP,[2] was utilized to compress these attributes into an optimal two-dimensional representation. The design of the interface is illustrated in Figure 4.2.

The points on the map are colorized according to pitch as follows: higher than 250 Hz, 220 Hz to 250 Hz, 190 Hz to 220 Hz, 160 Hz to 190 Hz, 130 Hz to 160 Hz, 100 Hz to 130 Hz, and lower than 100 Hz. Similarly, speech rate was divided into four categories: slow, slow-moderate, moderate-fast, and fast. Speaker age was divided by decade into nine groups, from teenage speakers to speakers in their 90s. The new categories are displayed on the top right side of the plot, allowing users to click on the category to include in or exclude from the search. Accent was excluded from the dimensions since it is not an ordinal or continuous variable. Instead, accent data are selected for separately via a world map, where selecting a continent filters speakers.

To achieve the best visualization, we examined two: hyperparameters of the UMAP algorithm the number of k nearest neighbors, and the minimum distance. Nearest neighbors determines the balance between global and local structures (See Figure 4.1). For example, when k is too large, local structure is lost as a result of averaging over too many samples. This can be observed when $k = 500$ in Table 4.1. Minimum distance defines the compactness of the plotted data points. Selecting a too small number creates a highly dense plot that prevents users from effectively selecting individual points with their cursor. Hence, we selected 100 nearest neighbors and 0.75 minimum distance as our hyperparameters.

The final map forms two large clusters according to gender, a male cluster on the left and a female cluster on the right side. Pitch is arranged horizontally on the clusters. The three highest pitch groups, coloured as red, orange and yellow, only appear in the female cluster. Likewise, the two lowest pitch groups, coloured as navy and purple, only appear in the male cluster.

While the map is colored according to pitch by default, the interface provides buttons to

[2] Scikit-learn Python package was used to calculate the voice UMAP [70].

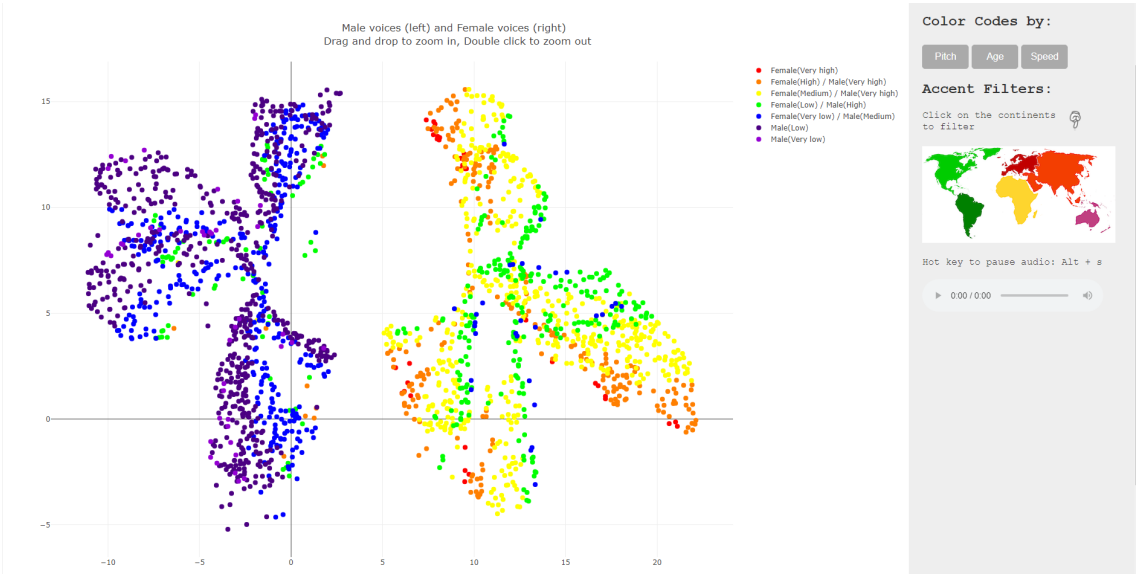


Fig. 4.2: Design of the Voice UMAP. By default, the map was set up to show pitch information of the voices, marked by different colors. The axes do not have any physical meaning, but represent the relative proximity of the voices based on four dimensions (age, gender, speech rate, and mean pitch) as a result of UMAP calculation.

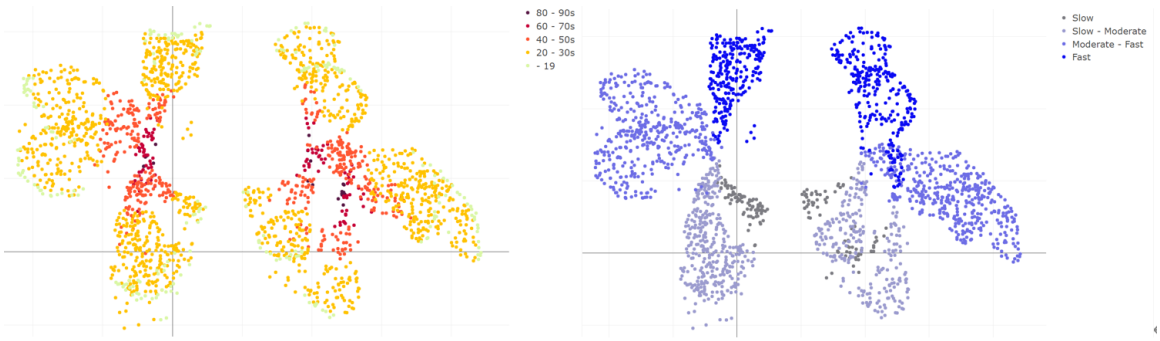
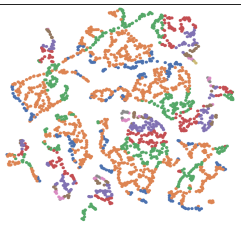
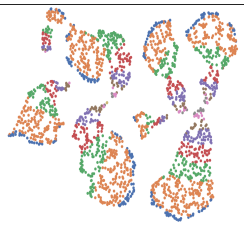
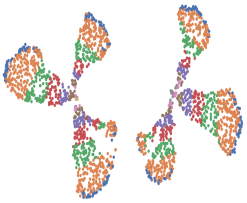
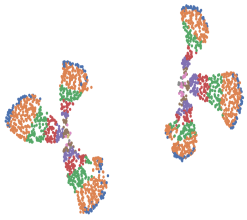
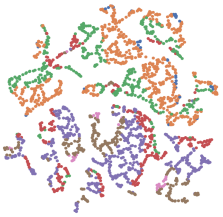
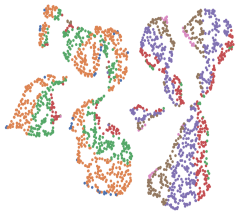
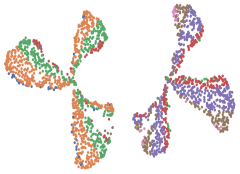
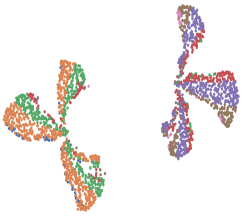
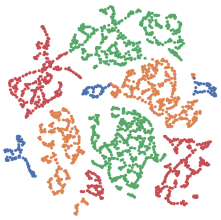
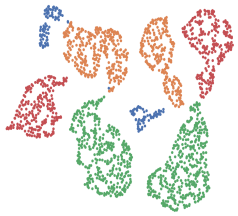
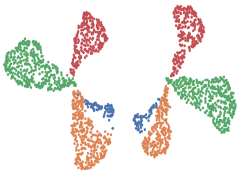
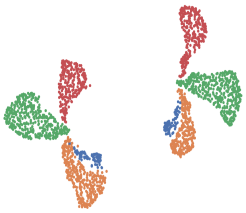
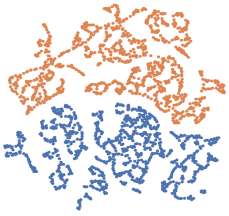
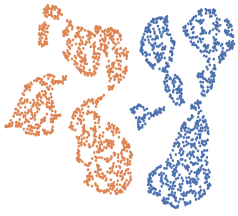
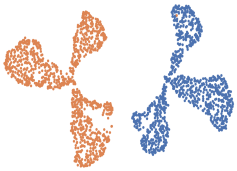
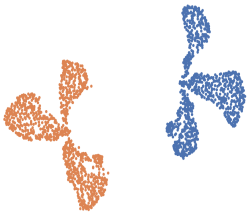


Fig. 4.3: Visualization of the same data organization with age labels (left) and speed labels (right), based on the same axes as in Figure 4.2.

switch to color coding by age, and by speech rate. Figure 4.2 shows the default plot displayed with pitch labels. The plots in Figure 4.3 correspond to the same set of voices labelled with age (left) and speech rate (right). Vertical position of points within a cluster varies by age. Speech rate creates four local clusters within each gender group, slow speech samples being located

at the bottom and fast samples on the top.

Table 4.1: UMAP of voice dataset with different number of nearest neighbors (min_dist = 0.75) Each column includes the same figures with four different types of labels: age, pitch, speech rate, and gender.

	n_neighbor = 5	n_neighbor = 10	n_neighbor = 100	n_neighbor = 500
10s 20s 30s 40s 50s 60s 70s 80s 90s				
-100Hz 101-130Hz 131-160Hz 161-190Hz 191-220Hz 221-250Hz 251Hz-				
Slow Slow-Moderate Moderate-Fast Fast				
Female Male				

4.2.3 Waveform Similarity Map

Previous UMAP-based approach suffered a limitation in reflecting abstract features of the human voice, which are often described as timbre. Besides arranging voices by their acoustic features and speaker information, we designed another voice map that presents the similarity of voices based on their timbre. For a second voice map, rather than use off-the-shelf features, we sorted speakers by similarity according to a learned feature space. We computed the 256-dimensional speaker embedding vectors of all voices by utilizing an open-source framework [71] built based on the SV2TTS network developed by Google [10]. We then applied UMAP to reduce the number of dimensions into a two-dimensional representation. We built two large clusters based on genders using binary logistic regression. As a result, the distance between the voices indicated the similarity of their timbres. The design of the map is shown in Figure 4.4.

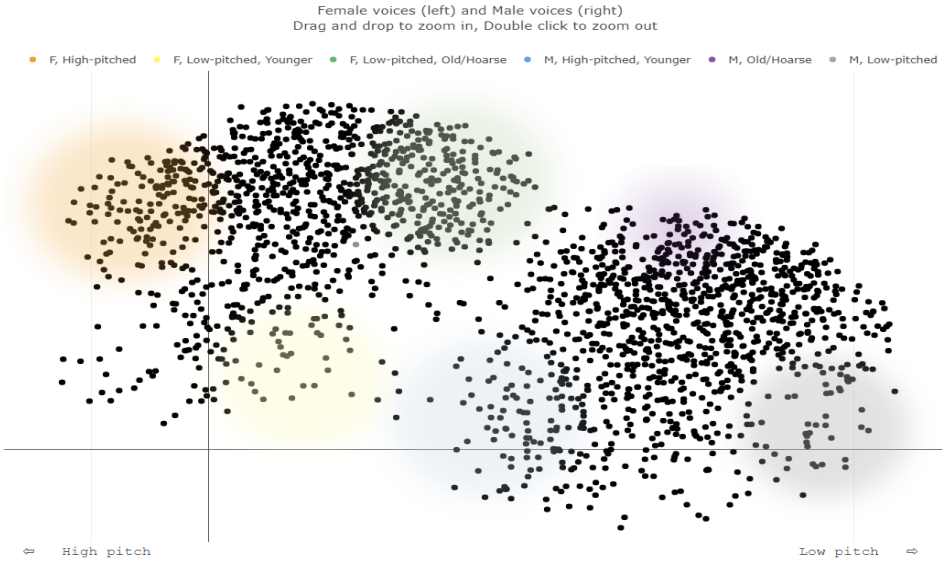


Fig. 4.4: Design of the waveform similarity map. The voices were organized based on the vectorized speaker representations extracted from the waveforms. The six circular regions indicate different groups of speakers.

On the resulting UMAP space, the voices arranged primarily by pitch, plotted from the left to the right side of the map in descending order. We found that low-pitched female voices and high-pitched male voices tend to gather at the intersection of the two clusters. According to the author's own observations, some abstract qualities such as hoarseness or nasality were observed

to form small local clusters . Hoarse voices, often found on the top regions of both clusters, were inferred to be uttered by older speakers due to the age-related alterations of voices. Based on these observations, we marked the map with six different color labels corresponding to the approximate positions of six speaker clusters. The clusters include high-pitched female voices, low-pitched older female voices, low-pitched younger female voices, high-pitched male voices, low-pitched male voices, and older male voices.

4.3 Experiment

We conducted a user experiment to assess the performance of our three proposed approaches based on the perceptual domain. The experiment mainly involved a comparison between approaches to find a voice that is the most similar to the target voice, and evaluation of the level of similarity on a five-point Likert scale.

4.3.1 Participants

We recruited eighteen participants, comprising nine females and nine males. Ten out of the eighteen participants self-reported to be a native English speaker or an English speaker whose proficiency level is comparable with that of a native speaker. Participants were recruited from the general population through online advertisements, namely the McGill Marketplace for Research and Surveys. All participants provided both written and oral consent, and were compensated $10 for the session.

4.3.2 Protocol

The experiment sessions were conducted via video conference, and lasted approximately an hour. Participants shared their screen and audio, and this stream was recorded for later analysis. Participants were asked to find the most similar voice to a celebrity of their choice, using the three proposed UIs: the traditional 2D exploration, Voice UMAP, and the waveform similarity map. For the target voices, participants selected one male and one female celebrity whose voices they have often heard and were familiar with.

To prevent any possible order effects, we used a Latin Square design to order the presentation of the three UIs. To further prevent participants' own-gender bias (i.e., better identification

for voices of same gender than different gender), half of them started with the target celebrity of their same gender, and the other half started with the different gender. Participants were asked to save the most similar voice for each of their two target celebrities, and repeat this task on each interface, for a total of six selections. Since comparing multiple voice samples was a fundamental task in this experiment, participants were asked to listen to two voice samples of their selection simultaneously in their left and right ears. Following the UI exploration, participants completed a post-test questionnaire, evaluating the UIs and the voice files they had saved from the experiment.

4.3.3 Measurements

For every participant, we recorded the start time and end time of each trial and the number of voice samples they saved in the list for comparison. The post-test questionnaire assessed three main items: voice similarity scores, usability of the UIs, and participants' perceptual importance of vocal attributes.

The voice similarity score was meant to investigate performance on the main task—how close the selected voices were to the target voices. We asked participants to make an overall evaluation of the resemblance of the voices according to their perception on a 5-point Likert scale (1: The two voices did not sound identical at all, 5: the voices were completely indistinguishable). Questions on the usability of UIs included subjective factors such as their preference, ease of use, and perceived time efficiency (how long they felt it had taken to complete each task). Finally, we presented a set of common vocal features, and asked participants to rate the importance of each feature in their subjective evaluation of vocal similarities. The characteristics included some features used in the proposed UIs, such as accent, pitch (f0), age, and speech rate — and other more general qualities such as perceived hoarseness and monotonousness.

4.4 Results

To compare the performance and usability of the three UIs, we performed the Kruskal-Wallis H Test with the eta-squared effect size based on the H-statistic. We utilized the same analysis technique to investigate the importance of multiple vocal characteristics. For the above tests,

we used Dunn's Test with a series of pairwise comparisons for post hoc analysis. Also, we used the Mann-Whitney U Test to analyze different user categories and their performance on searching voices.

4.4.1 Voice Similarity

Eighteen participants rated 108 voices from the six trials of their voice exploration (three UIs × two target voices). Across all interfaces, selected voices had a mean similarity-to-target rating of 3.55 (SD = 0.99), and we did not find any significant difference in the score among the three interfaces (p = 0.75). Four out of the eighteen participants selected the same voice in more than one interface. One participant (P01) found the same voice from the three UIs, and graded the similarity of the voice as five for all the UIs, showing their confidence in the result. Three other participants (P04, P06, P11) selected the same voice from two out of the three UIs, and rated the similarity highly for the UIs where they found the common voices (avg = 4.33, SD = 0.47). Participants' voice sample and their target voice can be found in Appendix A.2.

4.4.2 User Preference

There was a noticeable difference in preference between the three UIs despite the marginal difference in performance. Over 50% of the participants preferred the traditional 2D exploration tool the most. This was followed by Voice UMAP (33%), and the waveform similarity map (11%) as shown in Figure 4.5.

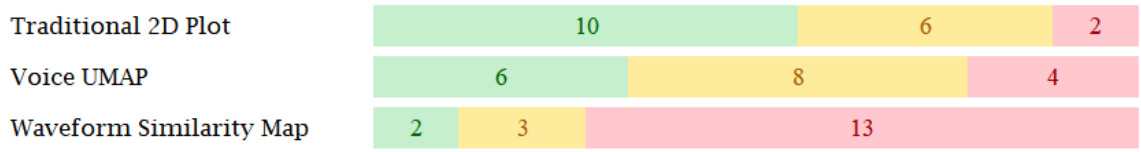


Fig. 4.5: Comparison of subjective preference between the three UIs (n=18). Green corresponds to the most preferred, yellow corresponds to neutral, and red corresponds to the least preferred UI.

4.4.3 Interfaces: Quantitative Evaluation

4.4.3.1 Usability

We found a highly significant difference in usability of the three UIs ($p < 0.0001$, $\eta_H^2 = 1.19$), based on participants' subjective ratings. A post hoc Dunn's test showed a significant difference between the traditional 2D plot and the waveform similarity map ($p < 0.0001$). The ratings for the traditional 2D plot and Voice UMAP were not significantly different ($p=0.08$); likewise, the ratings for the Voice UMAP were not significantly different from the ratings for the waveform similarity map ($p=0.07$) (See Figure 4.6).

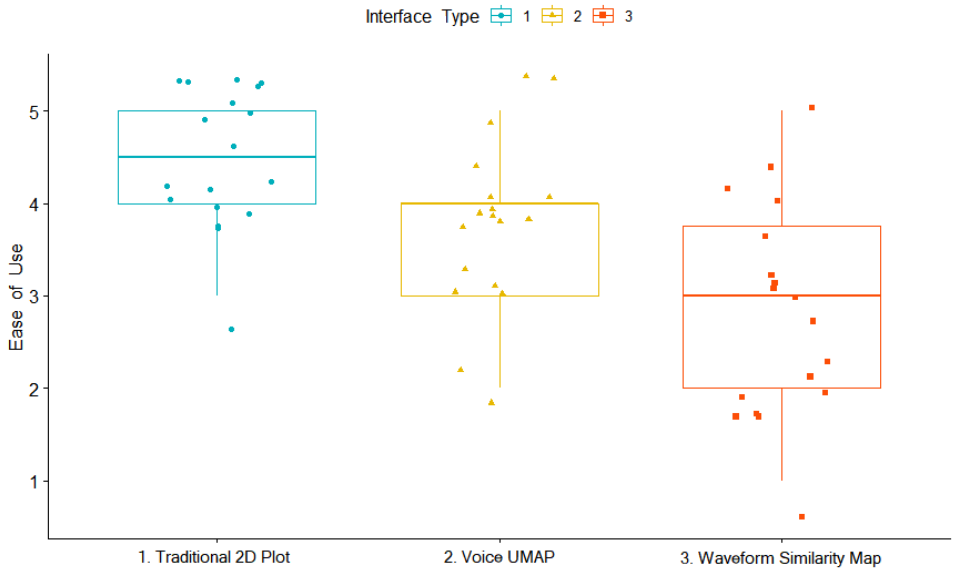


Fig. 4.6: Comparison of usability of the three UIs assessed by eighteen participants.

4.4.3.2 Time

Systems yielded a significant difference in perceived time efficiency ($p < 0.001$, $\eta_H^2 = 0.8$) Once again, the significant difference lied between the traditional 2D plot and the waveform similarity map, with a p-value less than 0.01 from our post hoc analysis. The other two pairs of comparisons did not yield a significant result. See Figure 4.7 for the median and standard deviation of the data.

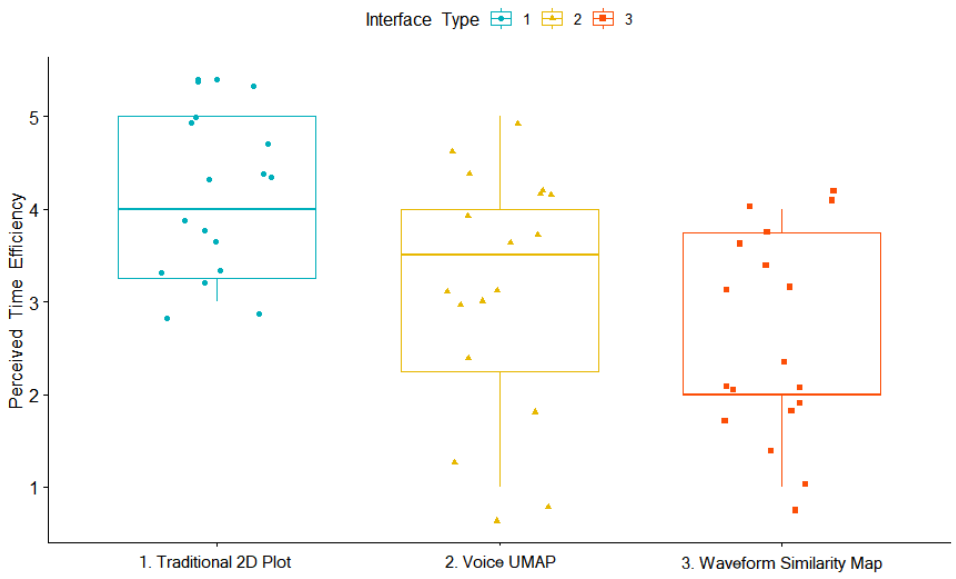


Fig. 4.7: Comparison of perceived time efficiency of the three UIs assessed by eighteen participants.

In comparison to the perceived time efficiency, there was no significant difference in the actual recorded time spent on each of the tools ($p = 0.32$) (see Table 4.2).

Table 4.2: Time spent on each of the three user interfaces to find a match to the target voice.

Interface Name	Average Time (s)	Longest Time (s)	Shortest Time (s)
Average of three interfaces	370	1330	40
Traditional 2D plot	331	1138	40
Voice UMAP	398	1140	47
Waveform similarity map	382	1330	122

4.4.3.3 Features for Exploration

Participants found the colour labels helpful for understanding the visualization and conducting their search (avg = 4.11, SD = 0.93, Likert scale, 1: Not helpful at all~5: Very helpful). We also asked participants how well the distances between points on the map reflected their vocal similarity. For this question, participants answered 3.39 on average with SD = 1.2 on a 5-point

Likert scale (1: Not meaningful at all∼5: Very meaningful).

4.4.3.4 Ease of Comparison

Participants added a set of reasonably similar voices to a list in the UIs during our experiment. The average number of samples saved in the list was 3.1 from all three UIs (SD = 2), with the largest pool containing eight voices and the smallest pool containing one voice. We also collected preference data on the feature that allows participants to listen to two samples in left and right ears simultaneously. Participants assessed the usefulness of the feature to be slightly lower than neutral (avg = 2.72, SD = 0.93) on a 5-point Likert scale, where five indicated very helpful and one indicated not helpful at all.

4.4.3.5 Performance by User Categories

We performed a Mann-Whitney U test to explore the difference in similarity scores by two participant subsets: first, by biological gender of the participant and second, by whether the participant was a native English speaker. We did not find a significant difference in either analysis, with p-values observed to be of 0.99 and 0.64 respectively. Also, there was no correlation between the participant's opinion on the importance of accent of a voice and whether the participant was a native speaker (p = 0.42).

4.4.4 Interfaces: Qualitative Feedback from Interview

In accordance with our quantitative analysis, participants gave the most positive feedback on the traditional 2D plot, particularly for its simplicity and straightforwardness. One participant (P17) said "It is nicely ordered in a linear fashion. The pitch got higher to the right so I could just go through them." Another participant (P29) opted for Emma Watson's voice as the target voice, and mentioned the ease of approximating to the voice by accent. "I found it easy to focus on a specific accent to find my target voice. It is clear and straightforward." On the other hand, another participant (P25) did not prefer the interface for the same reason, saying that "I don't think organizing voices by accent is the best idea. The quality of the voice is actually what I am looking for."

Participants who most preferred Voice UMAP appreciated the different search options provided by various filters. One participant (P30) said "I found it the most interesting because there were a lot of filters to check. It was appealing." The visualization constructed by UMAP was selected as the most favourite by some participants, as well as the least favourite by others. One participant (P06) found that it was "fun" to explore the different parts of this map, and that they could select the accent of voices by selecting a region from the world map. Opposing this, another participant (P17) said "There were unnecessarily ample spaces between the samples and the organization was very irregular. I don't really know if it actually helped me."

Lastly, two participants who preferred the waveform similarity map found it useful to refer to the coloured labels that marked specific speaker regions when approximating to their target speaker's voice. However, the interface was not preferred by most participants, because there were not enough labels that describe the speaker's information as in the other two interfaces. One participant (P14) mentioned the difficulty of finding a good strategy to search a voice, "I found it most confusing because it's like clicking around. It was hard to zoom into particular accents. Broad searches were kind of difficult for me."

4.5 Discussion

4.5.1 Summary of Results

Overall, our users were able to acquire satisfactory matches of their target voices by using the three UIs. Voice UMAP received the highest mean score of similarity-to-target, 3.64 on a five-point Likert scale (SD = 1.08). Although the three UIs did not show a meaningful difference in similarity scores, we found a significant difference in user preference. Our user interview showed that participants appreciated the conciseness and ease of use of the traditional 2D plot, whereas the waveform similarity map was considered to suffer from a comparatively complicated organization of voices.

Our participant interviews suggest that the same visualization can be perceived differently by different users. Some features that are specifically liked by some participants appear to rather interfere with the user experience of other participants, and vice versa, implying the difficulties of designing a voice searching tool that is efficient and satisfying for every user.

4.5.2 Limitations and Future Work

The construction of the traditional 2D plot and Voice UMAP solely relies on the data description provided by the dataset (Speech-accent archive [35]). Thus, the challenge lies in reducing the dimensionality of the data and presenting it on a plot with only two or three attributes. To address this constraint, we investigated the most salient features of the human voice, such as mean f0 and accent, and used these to map the plot. Beyond our research scope, future work might explore other features such as intensity or monotonousness, and their performance in reducing the data dimension. Other future work could investigate the correlation between data size and search efficiency, by employing a larger and more diverse dataset.

4.6 Conclusion

In this work, we tackled the problem of finding a similar match to a particular voice in the absence of any external reference. To do so, we developed three exploration strategies to find a voice from a large collection of existing sound clips. We utilized the speech-accent archive that contains 2140 different voices and rich speaker metadata. We explored different arrangements of voices from a traditional 2D plane to a map computed by a dimension reduction algorithm. Results showed that users preferred the interface with simpler user interactions, regardless of the level of similarity they could achieve. Our results demonstrated that users were able to find a voice that they rated as moderately similar or better (average of 3.55 out of 5).

Chapter 5

Sound of Hallucinations: Toward a more convincing emulation of internalized voices

Preface

This chapter presents a manuscript accepted at a peer-reviewed conference. In this work, we build a voice modelling interface to directly manipulate a voice selected from Chapter 4 and improve its similarity to the target voice. Our technique utilizes principal component analysis (PCA) to quantify perceptually meaningful characteristics of a voice and modifies them to create novel voices. The performance and usability of our interface were demonstrated by two user studies, user's subjective judgement and a blind test with multidimensional scaling analysis on perceptual similarity of different voices.

Author's Contribution and Acknowledgement

Hyejin Lee designed and implemented the voice manipulation techniques and the user interface, conducted user experiments, analyzed the results and wrote this manuscript. Professor Cooperstock supervised the research and edited the manuscript. Cecilia Jiang participated in running user studies and implementing the user interface. Yongjae Yoo supervised the design of user studies. Max Henry suggested the development of PCA-based voice morphing tech-

nique. Every person mentioned above edited the manuscipt as co-authors. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), MEDTEQ, iMD Research and IA Précision Santé Mentale. McGill REB #20-08-023.

5.1 Introduction

Avatar therapy offers a communication environment in which patients can enter into a face-to-face dialogue with an avatar representing their auditory hallucinations [72, 73, 42]. The avatar is typically voiced by clinicians who talk to the patients in their own voice or through a voice transformer to match the vocal characteristics of the patient's auditory hallucinations. Since most hallucinations among schizophrenic patients are auditory, the avatar's vocal characteristics are considered to be central to the success of the therapy [74, 75]. However, achieving convincing matches for the voice properties that may be present only in the mind of a patient remains challenging. Although there exist powerful tools to facilitate the design of graphical avatars, e.g., character creation interfaces for video games, the same is unfortunately not true for the auditory domain.

In this paper, we develop a voice modelling paradigm that assists the design of an avatar's voice for non-expert users. At its core, we identify the dominant effect vectors from the speaker embeddings by the usage of principal component analysis (PCA) and utilize them as tuning parameters. By leveraging a large corpus of speech samples and our two refinement techniques, we allow users to easily obtain a convincing emulation of their internalized voices. To be effective, the created voice should match specific traits of the intended voice, such as age, pitch, resonance, and prosody.

The interaction techniques we investigated for this purpose included:

- Voice Space Navigation: Exploration of a voice map to find an initial estimation to the target voice the user has in mind.
- Latent Parameter Editing: Modulating the parameters of pitch, resonance, hoarseness, and emotional prosody, which we determined to be highly salient to one's perception of voice properties.
- Voice Mixing: Synthesizing new voices by interpolating two selected samples from the space of existing voices.

Our main contributions include the development of a user interaction paradigm for voice design that is efficient and accessible to novice users. Our results indicate that the latent parameter editing can generate voices that are highly similar to the voices in the user's head.

The voice blending technique demonstrates comparable accuracy, preferred by users due to its simplicity. An additional blind assessment validated that voices produced with both techniques achieved higher similarity scores than the voices generated from a commercial voice morphing tool. The voices designed through our system can subsequently be used in conjunction with a text-to-speech (TTS) tool to produce any output speech in the target voice. In this manner, we seek to enhance the degree of realism for auditory psychotherapy [76]. Beyond the initial target of avatar therapy for people who suffer from auditory hallucinations, our work may benefit a variety of other such therapeutic applications that similarly rely on voice stimuli, for example, autism spectrum disorder (ASD), bipolar disorder, and post-traumatic stress disorder (PTSD) [77].

5.2 Related Work

5.2.1 Speaker-based Voice Transformation

Voice conversion (VC) is a technology that adapts the speech of a source speaker to that of a target speaker while keeping the linguistic content unchanged [78]. Early work in VC involves decomposition of the signal into excitation (i.e., pitch and prosody) and spectral (i.e., voice timbre) components using linear predictive coding [79]. In this approach, the authors make use of a Gaussian mixture model (GMM), wherein the spectral parameters of a source voice are made to predict the spectral parameters of a target voice. With the advancement of deep learning, state-of-the-art VC algorithms achieve natural speech transformation with a higher degree of fidelity, reported in recent Voice Conversion Challenges (www.vc-challenge.org). Such approaches in VC, however, often require a large amount of speech data both from source and target speakers and are not able to generate the voice of a speaker that is unseen in training data.

Another recent work has seen the development of vocal synthesis methods conditioned on low-dimensional speaker representation of arbitrary voices, which is often denoted as “speaker embedding” [80]. While these systems are intended to recreate known voices (i.e., those whose embeddings can be captured from input recordings), we expand on this speaker embedding technology to generate novel vocal avatars by directly manipulating the embedding space. Speaker embeddings were initially developed to identify unique voices, and have been shown

to perform strongly in this domain with an accuracy of above 95% among 1000 speakers [11]. Yet, it is unclear which vocal characteristics are encoded in speaker embedding and how they are mapped to meaningful properties that account for human perception of voices. To the best of our knowledge, no study has demonstrated the alteration of voices by retrieving quantifiable acoustic qualities from the high-dimensional feature vector and evaluated its effects by human listeners. Beyond learning from input speakers, our work investigates new strategies to optimize the feature vector and produce user's desired voice.

The authors note that, when extracting embeddings from speaker recordings, it is possible that other acoustical factors, such as speaker-distance from the microphone, or extraneous environmental sound, influence the resulting representation. Some work in speaker recognition attempts to capitalize on this fact [81]. In a related sense, it may be desirable to model and replicate the acoustic environment of the imagined target speaker by the well-established process of so-called room modelling [82]. However, we are unaware of any research on the interaction between the acoustic environment and its effectiveness in avatar therapy; such pursuits fall outside of the scope of the current study.

5.2.2 Voice Morphing Software for End Users

Apart from research-based technologies to morph one's speech, commercial voice morphing tools appear to be an accessible approach to the general public. However, existing commercial tools are difficult to use, limited in their ability to attain satisfactorily close matches to target voices, and often result in significant distortion or output voices that sound mechanical. This is because such tools mainly focus on generating alien or non-human audio effects for entertainment purposes. Tools such as Voicemod Pro [32] and MorphVOX Pro [30] allow for easy manipulation of basic features such as pitch and some elements of timbre, but are not capable of transforming one voice into that of another person without turning the voice into a robotic voice. The AV Voice Changer Software Diamond [31] provides a built-in voice library with approximately 100 preset voices added on the basic control, however, it still suffers from the same mechanization or degradation in audio quality. While many of these tools do not disclose their signal processing methods, they may make use of classic techniques such as pitch-synchronous overlap add [83] and phase-vocoding [84] to manipulate pitch and formant structure; these may be used in tandem with standard processing technique that requires

domain-specific knowledge like dynamic compression and equalization.

5.2.3 Dimensionality Reduction for Speech Transformation

Principal component analysis (PCA) is a dimensionality reduction technique that projects high-dimensional data on a lower dimension such that most of the information is efficiently contained in a small set of dominant features. Using this technique, previous work introduced the concept of eigenvoice, a combination of basis vectors extracted from Hidden Markov Models trained on speech data [85]. The basis vectors are determined using PCA, and each component reflects an important dimension of variation in the timbre of the reference voices.

The eigenvoice approach has been expanded upon in numerous speech-related tasks such as speaker recognition, in which the span of subspaces specific to different speakers was characterized [86], and a speaker diarization approach that identifies speakers from a recorded conversation [87]. Beyond speaker recognition, applications were also found in speech reconstruction, in which the goal is to enhance the quality of the input audio with minimum distortion in the original signal by removing the least important eigenvoices, resumed to be associated with noise components [88]. In the application of vocal synthesis for avatar therapy [89], authors took a composite approach by applying dimensionality reduction to GMMs trained on speech data. The principal components permit the direct manipulation of spectral mappings in voice conversion, however, this trend has moved away from brute-force spectral manipulation, leaving the subtleties of voice conversion to more expressive neural networks.

Along with a range of applications based on the eigenvoice, our work investigates the effects of voice feature adaptation, assisted by PCA, and the corresponding user experience in supporting the creation of new speaker identities. This research topic is relatively less explored than the areas of speaker recognition or speech reconstruction, yet represent an interesting research problem, with numerous possible applications of manipulation of the acoustic features and design of voice personas.

5.3 Voice Modelling Interface Paradigm

The design of our interface (Figure 5.1) is intended to support voice exploration and manipulation without requiring any specialized knowledge in the audio domain. Our focus is therefore

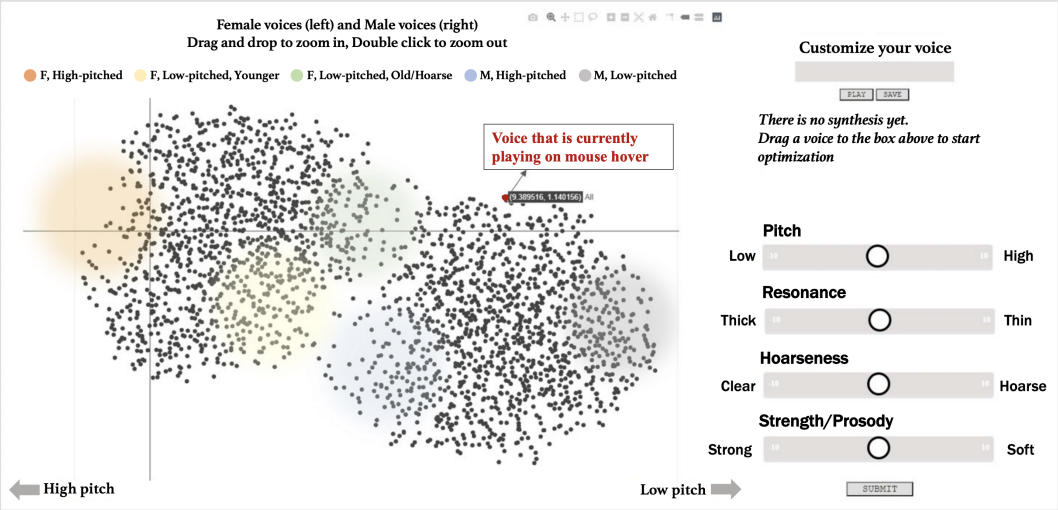


Fig. 5.1: The GUI of overall interface. The voice map exploration is displayed on the left side and latent parameter editing is on the right side. The map is represented as a lower-dimensional manifold of a large set of voice samples. In theory, the axes do not have any physical meaning, but indicate the relative proximity of the timbre of the voices based on Euclidean distance. However, we observed that the x-axis was primarily associated with pitch.

on facilitating control of perceptually meaningful qualities of human voices, based on terminology that is accessible to non-experts.

5.3.1 System Overview

Figure 5.2 illustrates the pipeline of selecting an initial voice and applying techniques to transform the vocal features. The voice generation process begins with a voice similarity map, a low-dimensional representation of 2484 existing voice samples collected from the LibriSpeech corpus.[1] The map interface visualizes the set of voices and allows users to search for one or more samples similar to their target, selecting them for playback on demand. Once suitable samples have been selected, the system allows for manipulating the selected voices, by an additional fine-tuning of the latent parameters computed by PCA or by voice mixing. In voice mixing, the system further automatically synthesizes a number of new voices, interpolated in the latent feature space between any two selected voices, thereby expanding the diversity of

[1]Librispeech: https://www.openslr.org/12

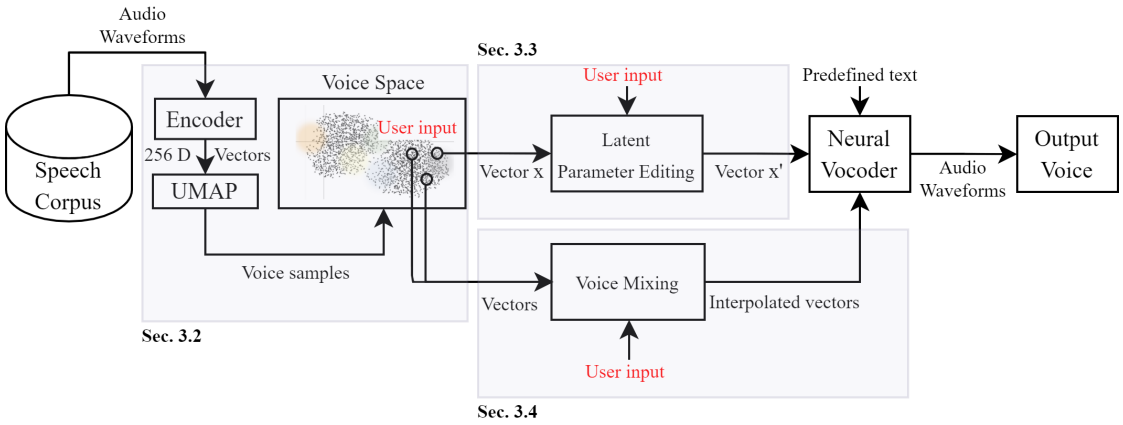


Fig. 5.2: The overall procedure of voice modelling through latent parameter editing (subsection 5.3.3) and voice mixing (subsection 5.3.4).

voice characteristics available. (Note that the terminology "voice mixing" in this dissertation does not refer to the linear combination of sound signals, but describes the proposed technique that combines the timbre of two different voices.) After users have selected and refined their target voice, an external TTS module, such as the Wavenet neural vocoder [14] can be used to render arbitrary speech input in that voice. Users may also save the output voice samples for later use in conjunction with other tools.

5.3.2 Navigating the Voice Space

To create an initial voice space with sufficient diversity, we trained 256 speaker feature vectors (i.e., speaker embeddings) from raw waveforms of 2484 speakers by using the encoder of a multispeaker TTS system [10], as shown in Figure 5.2. The encoder extracts a sequence of log-mel spectrograms from multiple time frames of each audio sample, which is then provided to a 3-layer long short-term memory (LSTM) network of 768 hidden nodes and a projection of size 256. This outputs a 256-dimensional vector per time frame, and all these vectors are then L2 normalized to obtain the speaker embedding that represents the unique timbre of each individual's voice, independent of speech content and background noise [11]. We then applied the Uniform Manifold Approximation Projection (UMAP) [16] on the resultant speaker embeddings and created a 2D projection. The obtained manifold was used as an initial map to search for an approximation of a target voice within the large pool of voice samples, using conven-

tional panning and zoom interaction techniques. The displayed voices are played automatically on mouse hover and saved on mouse click to minimize the required user interaction.

The constructed map was organized primarily by pitch, progressing from high-pitched voices on the left to low-pitched voices on the right, approximately forming two clusters of female and male voices. Interestingly, the map formed a few local clusters that contained abstract qualities of the voices such as hoarseness or the speaker's age. Hoarse voices were often found on the top regions of both clusters, and were inferred to be uttered by older speakers. According to such observations, we marked the map with five different color labels as local indicators of the speaker clusters. The clusters included high-pitched female voices, low-pitched older female voices, low-pitched younger female voices, high-pitched male voices, and low-pitched male voices.

5.3.3 Latent Parameter Editing

To parameterize particular qualities of a voice and enable controlling them, we performed PCA to obtain a manageable, small set of the most important latent variables from the speaker feature vectors of the LibriSpeech corpus voice samples. Based on a literature review of measurement and perceptual evaluation of voice parameters [90], and our perception of the effects of these parameters in preliminary synthesis experiments, each author proposed several descriptive names for the first four latent parameters. The identified parameters included pitch (high-low), resonance (resonant-shrill), hoarseness (clear-hoarse), and certain characteristics of prosody. "Pitch" relates to the perceived frequency of the voice. "Resonance," also attributed terms such as "deepness" or "thickness," agrees with an established dimension of variation in voices, given that voice depth is perceived differently based on its resonance inside a vocal tract of which the shape differs across individuals and for a given phoneme [91, 1]. "Hoarseness" refers to the speaker's voice quality, in line with a raspy, husky voice [90], and the prosodic qualities may be described as "confidence" [92, 93]. These features correlated with findings from the prior literature regarding the characteristics most important to human perception of voice [94, 20, 95]. To validate the suggested naming of these parameters, three non-author team members completed a brief questionnaire, assessing how helpful these names were for understanding the variables, and in conducting the voice editing task. Given the unanimous agreement between these team members, we included sliders for adjusting these top four prin-

cipal components in the user interface.

5.3.4 Voice Mixing

Fig. 5.3: The GUI of the voice mixing interface.

The mixing technique recommends new design directions with a minimal amount of user effort (Figure 5.3). Once the user selects two voices from the map, the system automatically calculates speaker feature vectors of five interpolated points between these selected voices, based on the L2 norm (See Figure 5.2). Specifically, we interpolate in L2 space in a two-step process: first we calculate the element-wise linear interpolation between the 256-dimensional vectors, and then normalize the resulting vector so that it has unit magnitude in L2 space. Elements of the two original voices determine the upper and lower bounds of the timbre properties to be interpolated, whereas the timbre of the middle (third) interpolated voice is half-way between the two selected voices. We opted to generate this number of interpolated voices as a compromise between distinctiveness of outputs, computation time to generate the interpolated samples, and the demands on short-term memory of the user to keep track of the differences between the samples.

5.4 User Studies

We investigate the effectiveness of our proposed approaches to support voice generation to match the user's assessment of their desired voice. Although usability of the interface is also an important factor, our focus in this work is on the perceptual domain and the performance of the system, rather than the interface itself. Accordingly, for our first study, we compared the perceptual performance of the proposed voice synthesis methods, and then evaluated the impact of latent parameters. In the second study, we assessed the performance of our system through a comparative research with an external voice morphing tool. The studies were conducted under the approval of McGill University's Research Ethics Board, REB #20-08-023.

5.4.1 Study 1: Latent Parameter Editing vs. Mixing

5.4.1.1 Participants

We recruited twelve participants (6F, 6M) with an average age of 26.7 ($\sigma = 2.6$) from the university population via online advertisement. All participants provided informed consent, and received monetary compensation of $10 for their time.

5.4.1.2 Procedure

The sessions took place by video conference, with audio and screen recording for later analysis. Participants were shown a brief tutorial video on how to interact with the UI components, and were then instructed to select as targets one male and one female celebrity with a North American English accent, with whose voices they were familiar. The experiment involved a comparison between strategies to create synthesized approximations to these target voices. First, participants carried out voice map exploration to select an initial voice sample for each celebrity, since this was a prerequisite to both of the refinement techniques. Participants were then presented with the latent parameter editing (LPE) and voice mixing refinement conditions in counterbalanced order. Audio details such as format, sample rate, and number of channels remained unchanged from the original recordings collected from the LibriSpeech corpus (FLAC, 16 kHz, 1 channel).

Following the experiment, participants completed a post-test questionnaire, evaluating the usability of the interface and their judgement of similarity between the target voices and samples they were able to produce using the different experimental conditions. The study concluded with a debriefing interview to elicit participant-specific information regarding their observed behavior. The post-test questionnaire consisted of the following questions (Q1-Q8: 5-point Likert scale, Q9: ranking question, Q10: open-ended):

Q1. (Map) How easy was it to understand the arrangement of the map?

Q2. (Map) How useful were the color labels to understand the arrangement of the map?

Q3. (Map) How close was your final voice to the celebrity's voice with regard to the overall similarity?

Q4. (LPE, for each of the four sliders) How effective was the [n-th] slider for morphing the voice to match your target celebrity's voice?

Q5. (LPE) How close was your final voice to the celebrity's voice with regard to the overall similarity?

Q6. (Mixing) How effective was this feature for obtaining the voice that is more similar to your celebrity's voice?

Q7. (Mixing) In your perception, did the five voices possess reasonably mixed qualities of the voices you mixed?

Q8. (Mixing) How close was your final voice to the celebrity's voice with regard to the overall similarity?

Q9. Please rate your overall preference.

Q10. Please share any other comments on your experience with our tool.

5.4.1.3 Statistical Analysis

We compared the performance within the three conditions: latent parameter editing (subsection 5.3.3), voice mixing (subsection 5.3.4), and not applying any syntheses. We evaluated both subjective preferences and subjective similarities between the target and synthesized voices, the latter as ranked by participants on a Likert scale, ranging from 1 (voices did not sound at all identical) to 5 (voices were completely indistinguishable). Since the data did not follow a normal distribution, we applied the Kruskal-Wallis H Test, with effect size indicated by the eta-squared ($\eta^2[H]$) value. Given the non-normal data distribution, we performed post-hoc analysis with Dunn's Test with Bonferroni correction (significance at $\alpha = 0.05/2$), finding statistically significant differences between the three conditions.

5.4.2 Study 1: Results

5.4.2.1 Overall Performance

Our results show that both refinement techniques significantly improved the fidelity of the voices selected from map exploration, with a medium effect size ($\eta^2 \approx 0.1$). As seen in Figure 5.4, without applying any refinements, the selected voices were evaluated as moderately similar to the targets ($\bar{x} = 3.0, \sigma = 0.82$) on the 5-point Likert scale. After applying the latent parameter editing, the mean score significantly improved ($\bar{x} = 3.83, \sigma = 1.03$). Similar improvements were observed from the voice mixing refinement ($\bar{x} = 3.63, \sigma = 0.95$). Dunn's test suggests that only the improvement from the latent parameter editing condition was significant (p = 0.007, Z = 2.835), while that of mixing condition was not (p = 0.045, Z = 2.167, $\alpha = 0.025$). No significant difference was observed between the mixing and latent parameter editing conditions (p = 0.756, Z = 0.668).

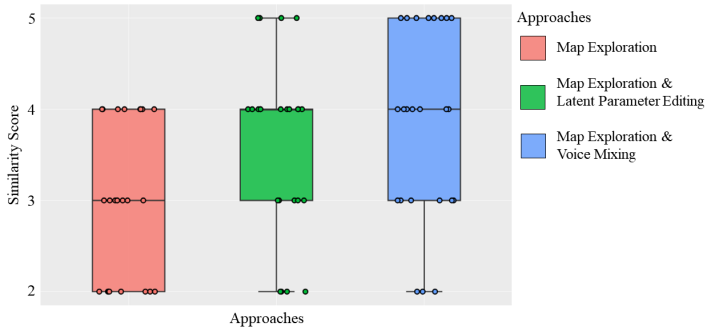


Fig. 5.4: Comparison of three groups of voices on how similarly they matched to the target voices of participants.

5.4.2.2 Latent Parameter Editing

We recorded the number of times participants optimized the four attributes. On average, participants moved the sliders 4.67 times ($\sigma = 4.83$) for pitch, 3.67 times ($\sigma = 2.56$) for resonance, 2.7 times ($\sigma = 1.46$) for hoarseness, and 3 times ($\sigma = 1.62$) for prosody, for each target voice. Figure 5.6 illustrates the overall tendency of participants adjusting the four parameters

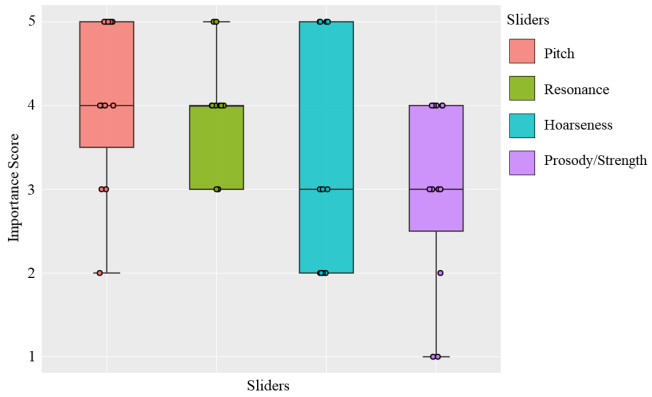


Fig. 5.5: Comparison of effectiveness of the four latent parameters in reproducing target voices of participants.

at each trial. It was observed that most adjustments were made less than five times and a few participants explored the features between five to ten times, which led to increased standard deviation. We did not find a statistically significant difference in the number of adjustments of the four latent parameters ($p = 0.39$). Participants' evaluation on the importance of the parameters also did not show a significant difference ($p = 0.09$), however, was observed to have a large effect size ($\eta^2[H] = 0.42$) (See Figure 5.5).

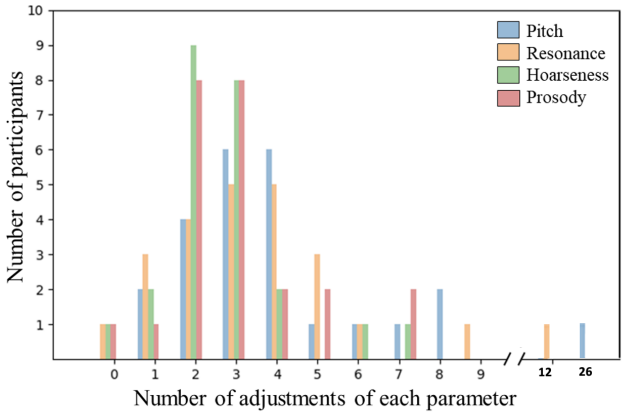


Fig. 5.6: The number of adjustments made on each parameter in the experimental trials.

5.4.2.3 Voice Mixing

Participants' responses indicated that they considered the synthesized voices to exhibit suitable qualities, representing a mixture of the two original voices ($\bar{x} = 4.5, \sigma = 0.65$ on a 5-point Likert scale). Based on the generation of such voices, participants also had positive assessments of the effectiveness of the mixing technique to achieve better matches to their mental representation ($\bar{x} = 4.0, \sigma = 0.7$).

5.4.2.4 Time and User Preference

The computation time to transform a 5 s speech sample was within 12 s during the experiment sessions. As seen in Figure 5.7, despite this delay, three quarters of the participants preferred using the voice mixing approach to obtain a similar voice to their target, and approximately a further 17% preferred latent parameter editing, compared to the condition in which they could not modify the voices they selected from the map.

	1st	2nd	3rd
Voice Mixing	**75%**	16.7%	8.3%
Latent Parameter Editing	16.7%	**66.7%**	16.7%
None	8.3%	25%	**66.7%**

Fig. 5.7: Comparison of participants' preference among the three approaches.

5.4.2.5 User Behavior and Experience

Direct observations of user experience with our tool were made during remote studies. Our findings regarding user behaviour using the latent parameter editing include:

Pitch. A high degree of pitch adjustment sometimes resulted in change of gender. This was utilized by a user who opted for an initial voice of different gender but similar timbre to their target speaker

Resonance. Some users required a few trials to understand the concept of resonance. They tended to test two opposite ends of the slider space to explore the permitted range of manipulation.

Hoarseness. Some users used this feature to reproduce the hoarseness of their target voice, while others considered the level of intelligibility of speech.

Prosody. Although the number of adjustments did not differ significantly from other parameters, participants only created speech of a neutral or slightly tweaked emotion to reproduce the identity of their celebrity, rather than making a dramatic alteration in prosody.

Since all voices presented on the map were synthesized by the computer, users occasionally encountered unnatural voices, describing them as the sound of a "ghost" or a "turkey". Additionally, users found that younger voice samples were more sparse in the female group than in the male group presented on the map. This made it particularly challenging for our users to recreate young female celebrities' voices, as we describe in subsubsection 5.5.1.2.

5.4.2.6 User Interview

Our user interview suggested that participants appreciated the ease and straightforwardness of the voice mixing approach. One participant remarked that the technique was easy to use since it only required selecting two voice samples. According to another participant, blending voices provided an additional benefit; it was helpful to make a decision between two voices.

With regard to the latent parameter editing, participants were in favour of having control on particular features of voices. However, this approach was perceived to be more difficult than expected by most users, with one participant mentioning that it was hard to capture which characteristics are being changed when adjusting multiple parameters back and forth.

5.4.3 Study 2: Proposed Voice Editing Approaches vs. Commercial Software

In this study, we evaluated the voices generated from our two synthesis techniques and a commercial voice morphing tool with respect to their ability to generate similar matches to the target voice. To select the commercial tool, we initially compared five commercially available

options: Voicemod,[2] MorphVOX Pro,[3] Skype Voice Changer,[4] ClownFish Voice Changer,[5] and AV Voice Changer Software Diamond.[6] We eliminated from consideration three tools that did not support uploading of a voice file, but rather, real-time recording of the user's own voice by microphone, since these were unsuitable for our intended use case. This left us with two tools that were evaluated by three non-author members of our research team. The evaluation criteria were expressivity to enable a variety of modulations, and minimization of sound distortion. In these respects, we found the AV Voice Changer to be the most compelling; this tool features a 2D pitch-timbre plane and supports adjustment of other elements such as frequency ranges by using bandpass filtering. Although we did not consider product price in our criteria, the selected system appeared to be the most expensive among those we evaluated. MorphVOX Pro offered limited capacity to modify features beyond pitch and a small degree of timbre adjustment, and was therefore excluded from the formal experiment we describe below.

5.4.3.1 Preliminary Sessions

Six researchers from the project team were involved in the preliminary session, each reproducing voices of two celebrities, Oprah Winfrey and Justin Bieber. Samples of both celebrities' voices were extracted from the VoxCeleb[7] data set, with each sample approximately 4 s in duration. To reproduce the given speech files, each member generated two pairs of voices under three counterbalanced conditions: latent parameter editing (subsection 5.3.3), voice mixing (subsection 5.3.4), and AV Voice Changer Diamond. To avoid potential bias, every voice that could be explored or generated through these interfaces was adjusted to output the same content with the speech of the celebrities. This procedure resulted in two pairs of 18 synthesized voices for the two target celebrities, which were then evaluated in the following experiment.

[2]Voicemod: https://www.voicemod.net/
[3]MorphVOX Pro: https://screamingbee.com/morphvox-voice-changer
[4]Skype Voice Changer: https://skypevoicechanger.net/
[5]Clownfish Voice Changer: https://clownfish-translator.com/voicechanger/
[6]AV Voice Changer: https://www.audio4fun.com/voice-changer.htm
[7]VoxCeleb, A large scale audio-visual data set of human speech: https://www.robots.ox.ac.uk/~vgg/data/voxceleb/

5.4.3.2 Participants

Twelve participants (6F, 6M) with an average age of 24.7 ($\sigma = 2.6$) were recruited from the general population. All participants volunteered to participate in this study and provided both oral and written consent. No participant reported any hearing impairment or cognitive disorders.

5.4.3.3 Procedure

We conducted cluster analysis to evaluate voices from the preliminary session based on multidimensional scaling (MDS). A Windows application (Figure 5.8) was developed for this purpose, which participants ran on their personal computers. Participants were provided with a brief user manual for how to interact with the system. The main task involved classification of 19 voice samples for each celebrity—the 18 voice files selected by team members from the previous session, plus the original speech file of the celebrity—into different numbers of bins (3, 5, 7, and 9), randomly ordered throughout four trials. Participants were not provided with any specific features as evaluation criteria but instructed to judge similarity as they saw fit. No limits were placed on the number of times a voice sample could be replayed. The study took approximately one hour and concluded with a post-test questionnaire investigating the main factors that impacted evaluation of the voices.

5.4.3.4 Perceptual Dissimilarity Analysis

To evaluate the perceptual similarity of voice samples, we follow the general methods of cluster analysis and multidimensional scaling (MDS). These methods are often used to quantify and visualize similarity of different sensory stimuli in the perceptual domain such as sound, taste or haptic sensations [96, 97, 98].

First, we calculated a pairwise similarity matrix S based on the results of voice clustering. At each trial, every item in each bin received a pairwise similarity score of which the value was equal to the number of bins of the trial. For example, if the $1st$ and $2nd$ voice were classified in the same bin from a trial with five bins, five was added to the $(1,2)$ cell of the similarity matrix. Since there were four trials, with three, five, seven, and nine bins, respectively, the theoretical maximum value of similarity was 24 $(= 3+5+7+9)$. We then inverted the similarity matrix

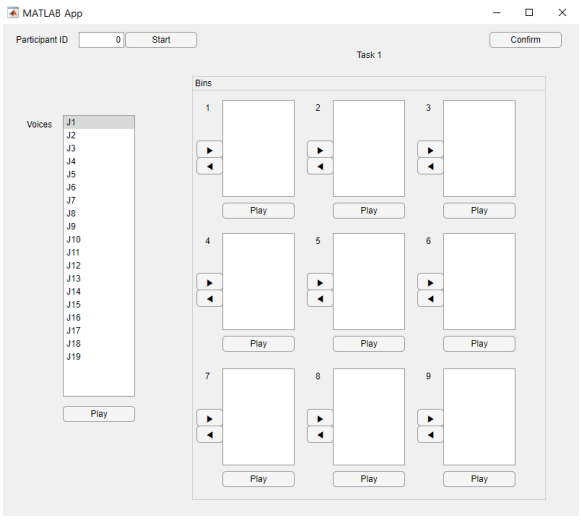


Fig. 5.8: The GUI of the Windows application developed for conducting the cluster sorting task.

to calculate the dissimilarity matrix D for every non-diagonal component, as in Equation 5.1. Every diagonal component was set to zero.

$$D(i,j) = 1000 \times \left\{1 - \frac{S(i,j)}{24}\right\} \tag{5.1}$$

These dissimilarity scores, for each cell of the matrix, were averaged over the twelve participants. We then conducted MDS on the resulting pairwise dissimilarity matrix D to project the values onto a two-dimensional diagram, representing the relative similarity of the auditory stimuli; nearby voices were perceived as similar, while distant voices were perceived as different.

5.4.4 Study 2: Results

5.4.4.1 Evaluation Criteria

Results from our open-ended post-test questionnaire (Figure 5.9) show the main factors that affected participants' metrics to classify voices. A set of common characteristics were found in the responses, including the primary vocal characteristic (pitch), human-like expression, and the level of audio distortion.

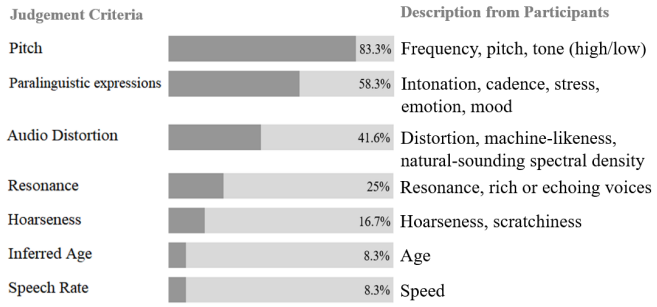


Fig. 5.9: Main factors considered in the voice classification task reported by participants.

5.4.4.2 Multidimensional Scaling Results

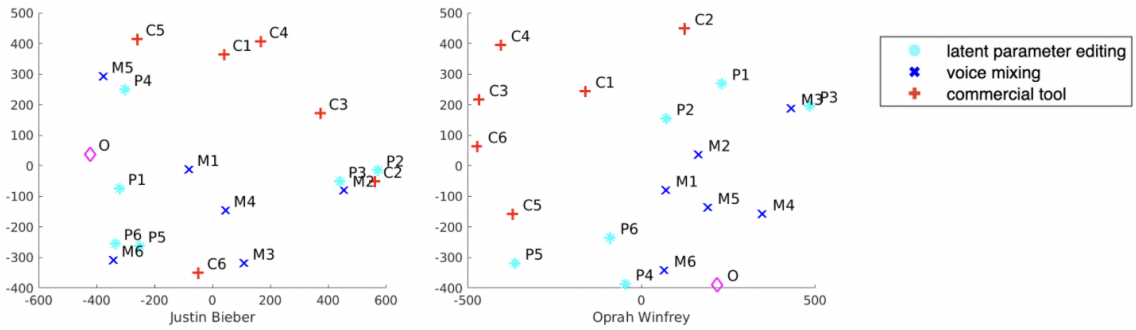


Fig. 5.10: Two-dimensional MDS results for reproducing the voices of the two celebrities, Justin Bieber (left) and Oprah Winfrey (right), with samples marked as P, M, C, O for the latent (P)arameter editing, voice (M)ixing, (C)ommercial tool (AV Voice Changer Diamond), and the (O)riginal speech samples. The X and Y axes are dimensionless; the Euclidean distance between points indicates perceived dissimilarity calculated from the study, e.g., in the left plot, M4 is perceived to be roughly twice as similar to O as M3.

The two-dimensional MDS results show that both latent parameter editing and voice mixing yielded significantly higher similarity than the commercial tool (Figure 5.10). In general, the two synthesized voices from our two methods are observed to be closest to the original speech sample, while voices from the commercial tool are further away. For Justin Bieber's voice (left plot), latent parameter editing appeared to result in the closest matches, while for Oprah Winfrey's (right), voice mixing performed better. On the left plot, a small cluster of a few samples (P3, P2, M2, C2) is formed in a distant location from the original voice due to their

low intelligibility of speech caused by the TTS synthesis. Kruskal's stress of MDS was found to be 0.19 for Oprah Winfrey and 0.20 for Justin Bieber's voice.

5.4.4.3 Statistical Tests on Dissimilarity

Our results showed that both techniques outperformed the commercial voice morphing tool. We ran a two-way repeated measures ANOVA on the averaged dissimilarity scores compared to the original speech, $D(i, Original)$, with two independent variables of *Method* and *Voice*. The values passed the Shapiro-Wilk Normality test ($W = 0.965$, $p = 0.303$) and Mauchly's sphericity test ($W = 0.807$, $p = 0.651$ for *Method* and $W = 0.691$, $p = 0.478$ for *Method*×*Voice*). As seen in Table 5.1, the effect of *Method* was significant on the similarity scores, while *Voice* and the interaction term were not significant. Tukey's HSD posthoc test indicated that both voice mixing and latent parameter editing showed significantly smaller average dissimilarity values compared to the commercial tool's average (voice mixing and commercial tool: $\bar{x} = 110.8$ and $p = 0.003$, Latent parameter editing and commercial tool: $\bar{x} = 104.2$ and $p = 0.005$). There was no significant difference of dissimilarity values between the two synthesis techniques (voice mixing and latent parameter editing: $\bar{x} = -6.65$, $p = 0.974$).

Table 5.1: Two-way ANOVA results of dissimilarity values to the original voice.

Factor	Statistics	p-value	Effect size (η^2)
Voice	$F(1,5) = 0.001$	0.971	0.0004
Method*	$F(2,10) = 8.387$	0.0128*	0.3340
Voice×Method	$F(2,10) = 1.715$	0.197	0.0683

5.4.5 Summary of Results

Both the quantitative measures from the MDS analysis, using perceptual dissimilarity metrics, and the qualitative responses to the post-test questionnaire indicate advantages of our approaches to voice synthesis. The synthesized voices generated by latent parameter editing and voice mixing approaches were judged to be more similar to the target than were the outputs of a traditional voice manipulation tool.

Although our participants had access to samples of the target voices, as necessary for a within-subjects design, the ability to find reasonably close matches to these targets suggests

the possibility of also doing so in the absence of such references, i.e., when the voice exists only in the user's mind. We also observed that the participants spent most of their time adjusting parameters they felt had the most influence and importance.

5.5 Discussion

5.5.1 Voice Space Exploration

5.5.1.1 Integration of Voice Exploration and Voice Editing

Existing voice morphing tools have shown to be successful in offering a variety of audio effects. Despite their success, several areas remain for potential improvement. First, these tools do not provide a strategy for searching among potential voice recordings, nor do they support exploration over a wide array of vocal characteristics. Second, they are prone to introduce distortion or "mechanical sounding" voices, unless the user is skilled and knowledgeable in the manipulation of the relevant controls. This results in two main limitations: the large number of required adjustments for users to change a voice that differs significantly from their own and the cognitive effort this entails; and the resulting distortion or mechanization of the output voices. To resolve the first issue, some tools allow users to provide recorded speech of a target speaker, as an alternative to searching or exploring within a vocal database. This approach works when the target voice can be recorded, but this is not always the case. We overcome these limitations by integrating a similarity map of voices that can be explored, and then operated on with synthesis techniques. Our results demonstrate that it was possible to select voices directly from the similarity map that were perceived to be reasonably close to the target, and subsequently, to improve upon the quality of vocal match using either of the two types of modification interfaces.

5.5.1.2 Demographic Imbalance

Our voice similarity map was built on Librispeech, a massive speech database derived from the Librivox project, which contains approximately 8000 audiobooks recorded by volunteer readers [39]. Although this database ensures a reasonable gender balance (52% M, 48% F), we observed an imbalance in speaker age, which skewed towards older volunteers. Indeed,

two participants from Study 1 mentioned that it was difficult to search for their younger target celebrity, reporting that there were more "old woman voices" than younger voices. This result suggests a potential benefit from using a speech database with a higher demographic diversity and well-documented speaker metadata.

5.5.2 Synthesis Techniques

5.5.2.1 Degree of Human Likeness

In the post-experiment questionnaire of Study 2, we investigated the main factors determinative of how participants sorted the voices. After the primary factor of pitch, paralinguistic expression (e.g., stress, fluctuation, or emotional prosody) and distortion of sound were considered as the most significant factors, appearing in 60% and 41.7% of the responses, respectively. This is consistent with previous findings that paralinguistic expression is the primary acoustic cue to infer the emotional state and personality of a speaker [99, 93]. In our system, expression is modulated to a certain extent by the last latent parameter (named "strength/prosody"); changes in the positive direction produced faster, louder, and more powerful speech. Along with naturalness of sound, paralinguistic features often determine the degree of human likeness of synthetic speech, since they mimic various human emotions and identities [100, 101, 102]. In light of these factors, participants evaluated voices generated from our interface as more similar to the original speech of celebrities than the voices from an existing voice morphing tool, as demonstrated by the MDS analysis and statistical tests on dissimilarity. Our results suggest that the system not only matches the vocal characteristics of the target speaker, but also creates a more convincing artifact that is closer to natural human speech and real-world expressions.

5.5.2.2 Perceptual Importance of Latent Parameters

In Study 1, we investigated the number of times participants adjusted each latent parameter and the subjective importance of the parameters. Although we did not find a statistically meaningful result, the effect size appeared to be very large ($\eta^2[H] = 0.42$). During the observation, we found several factors that were difficult to control. For example, participants expressed different levels of satisfaction with their output and some participants interacted

with the interface much more than others. Another factor was the perceptual gap between target and initial voices that the participants had selected from the map. Indeed, we observed a large variation in the number of adjustments by participants as shown in Figure 5.6. Based on the given factors, the large effect size may imply a reasonable correlation between each parameter's subjective importance and the number of adjustments made on the parameter.

5.5.2.3 Potential Harm and Implication of Voice Replication

The rapid technological advances of "deep fakes", able to produce persuasive reproductions of the appearance and vocal characteristics ("voice cloning") of arbitrary individuals, raises several ethical problems that society must confront. The most obvious concern is the potential violation of one's identity by generating fake speech in that person's voice. The utilization of copyright protection technologies such as audio watermarking [103] represents a possible safeguard. These techniques were originally designed to secure and authenticate digital audio by adding a signal—imperceptible to the human ear—to an audio file that enables a computer to identify the result by analyzing its spectrogram. However, this has its limitations in that it is subject to voluntary adoption by those producing the deep fakes.

5.5.3 Limitations and Future Work

We note several limitations of our present system. First, a consequence of our latent parameter editing approach is that a single controllable parameter can affect several characteristics of the resulting output voice. Additionally, due to a comparatively small corpus of international speech samples, the system is currently limited to text-to-speech (TTS) synthesis with a North American English accent. This limitation arises from the dependency of of the TTS model on the dataset on which it was trained. To expand the usage of the proposed techniques, model training may be required as a future task to accommodate different accents or languages.

5.5.3.1 Multidimensionality of Human Voice Perception

Modelling a human voice involves consideration of multiple acoustic features that are unavoidably intertwined with one another. Humans infer the speaker's age based on a multitude of cues such as pitch, speech rate, and hoarseness: a low, hoarse voice with a slow speech rate is often

perceived as older [104]. The impression of extroversion or perceived charisma of the speaker arises from a collective judgement of speech rate, pitch variation, and loudness [105, 92]. Moreover, changes of any single characteristic may affect how another characteristic is perceived: speech produced with a high-pitched voice is often considered faster than a low-pitched voice uttered for an identical duration [106].

We observe this phenomenon in the entangled latent variables extracted from PCA of speaker embeddings. This occurs because a dominant pattern obtained from dimensionality reduction is not always perceived by the listener as a single feature. This is particularly evident for the latent parameter of emotional prosody, which subsumes the changes of speech rate, intensity, and intonation, jointly represented in one dimension. Due to this entanglement, the manipulation of a single variable may result in undesired changes of other (coupled) qualities.

5.5.3.2 Accent Variations

The neural network that we used for TTS synthesis was trained with two public speech databases, VCTK[8] and Librispeech,[9] in which the predominant accent is American (approximately 1200 speakers) followed by British (100 speakers) [10]. Given this training data, the model was not capable of reproducing the wide variety of accents of non-American, non-British speakers. To render multiple accents with synthetic speech, related work introduced a new system called language embedding [107], a three-dimensional vector that represents the way words are pronounced in different accents. This does not involve any adaptation in the speaker feature vectors, but simply concatenates the language embedding to the speaker embedding. It also creates speech in multiple languages in the same way it facilitates various accents, containing language-specific information i.e., tone embedding for certain languages such as Mandarin and stress embeddings for English or Spanish. Future work might include combining the language embedding technology with speaker feature vectors that can be optimized from our system through the two proposed synthesis techniques.

[8] CSTR VCTK Corpus: https://datashare.ed.ac.uk/handle/10283/3443
[9] Librispeech Corpus: https://www.openslr.org/12

5.5.3.3 Computation Time

The main computational bottleneck of the current system at present is in the vocoding portion of the speech synthesis. The vocoder performs batched sampling to generate a series of time segments of audio waveforms, where the number of segments increases the computation time. As future work, we plan to pre-synthesize every possible combination of latent parameters with particular intervals at which the current latent parameter editing is being performed. This may require a simple retrieval of the stored data upon user interaction, significantly reducing the response time, enabling the voice editing experience to be closer to real time.

5.6 Conclusion

Generating artificial speech in a particular voice often requires one or more reference recordings to learn the voice identity. In this work, we developed a novel approach to externalize a voice that only exists in the user's head, and synthesize new speech in that voice without any reference data. We combined speaker embedding technology with a dimensionality reduction algorithm on an existing set of voices, and provided a direct manipulation on the low-dimensional representation of feature vectors through two voice editing techniques. The editing techniques, in conjunction with a voice exploration map, allowed our users to either create fictitious voice identities or manifest perceptually meaningful characteristics of a selected voice. Through user studies, we evaluated the performance of our two techniques and compared them with an external voice morphing tool that we found the most promising from our literature review. Our results demonstrated that the system is capable of generating a convincing match to a target voice with both techniques significantly enhancing the level of fidelity of voices compared to the existing technology. Returning to the motivating use case, our hope is that this system will lower the barriers for schizophrenic patients to engage actively in the avatar creation step, reproducing a convincing emulation of the sound of hallucinations they hear.

5.7 Acknowledgement

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), MEDTEQ, iMD Research and IA Précision Santé Mentale.

Chapter 6

Conclusion

This thesis explored how to generate a convincing simulation of a voice that only exists in one's head. We started by presenting our motivational use case, avatar therapy, a treatment approach for schizophrenia that involves patients interacting with an entity that represents their auditory hallucination. In contrast to traditional avatar therapy, where clinicians use their voice to talk to the patients, we propose a system that creates synthetic speech in the specific voice of their hallucination.

Our system consists of two stages; a search from a large dataset to find a pre-existing voice that is closest to a target voice; and a customization (fine-tuning) process. The voice searched and manipulated through our system can be used to generate new speech content, combined with text-to-speech synthesis. Our results demonstrated that the system outperformed existing technologies (e.g., voice morphing tools) in generating a similar match to voices imagined by users.

Our current system supports speech synthesis only with a North American English accent. This limitation could be addressed in future work through additional model training and expansion of the number of languages and accents to which our techniques can be applied. Other future work could involve applying the proposed voice manipulation techniques to non-human voices such as monsters or animals, and use them for the treatment of patients who suffer from such hallucinations [108, 109]. The use of our system in avatar therapy could also encourage researchers to investigate the level of realism of the reproduced voices, for the optimal delivery of the therapy.

Our research suggests that the primary features of a human voice may be quantifiable and controllable, and can be used to create novel virtual voices. It should be noted that in the scope of this thesis, we only discussed the treatment of auditory verbal hallucinations as a possible application field; however, our work may find applications in other fields that involve similar voice stimuli or avatar design in the absence of suitable voice references. In light of the fact that our fundamental technology was developed from a speaker verification technology, it can also be used for voice reconstruction to support forensic investigations.

Appendix A

Audio and Video Material

A.1 Interface Demonstration Videos

This section includes demonstration videos of the interfaces developed in this research: three interfaces for voice exploration and one interface for voice manipulation.

1. Voice Exploration Interfaces from Chapter 4:
   Traditional 2D Plot, Voice UMAP, Waveform Similarity Map
2. Voice Manipulation Interface from Chapter 5: Latent Parameter Editing and Voice Mixing

A.2 Examples of Experimental Audio Files

This section includes example audio files that were selected or edited by participants from our experiments. The closest match of Emma Watson's voice was searched from voice exploration experiment discussed in Chapter 4. The closest matches of Justin Bieber and Oprah Winfrey's voices were generated from voice manipulation experiment discussed in Chapter 5, respectively with latent parameter editing and voice mixing.

1. Emma Watson: Original, Closest Match
2. Justin Bieber: Original, Closest Match
3. Oprah Winfrey: Original, Closest Match

References

[1] J. Sundberg and R. Sataloff, "Vocal tract resonance," *Plural Publishing San Diego, California*, 2005.

[2] B. Hammarberg, B. Fritzell, J. Gaufin, J. Sundberg, and L. Wedin, "Perceptual and acoustic correlates of abnormal voice qualities," *Acta oto-laryngologica*, vol. 90, no. 1-6, pp. 441–451, 1980.

[3] J. C. Wells and J. C. Wells, *Accents of English*, vol. 1. Cambridge University Press, 1982.

[4] N. Campbell, "Measuring speech-rate in the spoken english," *Theory and practice in corpus linguistics*, no. 4, p. 61, 1990.

[5] H. L. Lane, A. C. Catania, and S. S. Stevens, "Voice level: Autophonic scale, perceived loudness, and effects of sidetone," *The Journal of the Acoustical Society of America*, vol. 33, no. 2, pp. 160–167, 1961.

[6] J. P. L. Brokx and S. G. Nooteboom, "Intonation and the perceptual separation of simultaneous voices," *Journal of Phonetics*, vol. 10, no. 1, pp. 23–36, 1982.

[7] C. A. Rosen, D. Anderson, and T. Murry, "Evaluating hoarseness: keeping your patient's voice healthy," *American family physician*, vol. 57, no. 11, p. 2775, 1998.

[8] E. Sapir, "Speech as a personality trait," *American Journal of Sociology*, vol. 32, no. 6, pp. 892–905, 1927.

[9] S. A. Kotz and S. Paulmann, "When emotional prosody and semantics dance cheek to cheek: Erp evidence," *Brain research*, vol. 1151, pp. 107–118, 2007.

[10] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno, *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *arXiv preprint arXiv:1806.04558*, 2018.

[11] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4879–4883, IEEE, 2018.

[12] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, IEEE, 2018.

[13] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *International Conference on Machine Learning*, pp. 2410–2419, PMLR, 2018.

[14] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[15] O. McCarthy, "Github repository (https://github.com/fatchord/wavernn), wavernn." https://github.com/fatchord/WaveRNN, 2019. Accessed: 2022-07-23.

[16] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[17] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[18] T. Murry and S. Singh, "Multidimensional analysis of male and female voices," *The Journal of the Acoustical Society of America*, vol. 68, no. 5, pp. 1294–1300, 1980.

[19] B. E. Walden, A. A. Montgomery, G. J. Gibeily, R. A. Prosek, and D. M. Schwartz, "Correlates of psychological dimensions in talker similarity," *Journal of Speech and hearing Research*, vol. 21, no. 2, pp. 265–275, 1978.

[20] M. P. Gelfer, "A multidimensional scaling study of voice quality in females," *Phonetica*, vol. 50, no. 1, pp. 15–27, 1993.

[21] F. Nolan, K. McDougall, and T. Hudson, "Some acoustic correlates of perceived (dis) similarity between same-accent voices.," in *ICPhS*, pp. 1506–1509, 2011.

[22] C. Nass and K. M. Lee, "Does computer-generated speech manifest personality? an experimental test of similarity-attraction," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 329–336, 2000.

[23] M. P. Aylett, A. Vinciarelli, and M. Wester, "Speech synthesis for the generation of artificial personality," *IEEE transactions on affective computing*, vol. 11, no. 2, pp. 361–372, 2017.

[24] C. G. Buchanan, M. P. Aylett, and D. A. Braude, "Adding personality to neutral speech synthesis voices," in *International Conference on Speech and Computer*, pp. 49–57, Springer, 2018.

[25] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[26] P. Boersma and V. Van Heuven, "Speak and unSpeak with PRAAT," *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.

[27] M. Mauch, C. Cannam, R. Bittner, G. Fazekas, J. Salamon, J. Dai, J. Bello, and S. Dixon, "Computer-aided melody note transcription using the Tony software: Accuracy and efficiency," 2015.

[28] D. Jouvet and Y. Laprie, "Performance analysis of several pitch detection algorithms on simulated and real noisy speech data," in *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 1614–1618, IEEE, 2017.

[29] N. H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior Research Methods*, vol. 41, no. 2, pp. 385–390, 2009.

[30] "Morphvox pro voice changer produced by screaming bee inc." https://screamingbee.com/. Accessed: 2021-06-30.

[31] "AV Voice Changer Software Diamond produced by AVSOFT CORP." https://www.audio4fun.com/voice-changer.htm. Accessed: 2021-06-30.

[32] "Voice maker and effect generator produced by Voicemod." https://www.voicemod.net/. Accessed: 2021-07-26.

[33] "Skype voice changer pro." https://skypevoicechanger.net/. Accessed: 2021-07-26.

[34] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, IEEE, 2015.

[35] S. Weinberger, "Speech accent archive." https://accent.gmu.edu/, 2019.

[36] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

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

[38] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[39] C. Veaux, J. Yamagishi, K. MacDonald, *et al.*, "Superseded-CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit," 2017.

[40] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: Timit and beyond," *Speech communication*, vol. 9, no. 4, pp. 351–356, 1990.

[41] F. Waters, P. Allen, A. Aleman, C. Fernyhough, T. S. Woodward, J. C. Badcock, E. Barkus, L. Johns, F. Varese, M. Menon, *et al.*, "Auditory hallucinations in schizophrenia and nonschizophrenia populations: a review and integrated model of cognitive mechanisms," *Schizophrenia bulletin*, vol. 38, no. 4, pp. 683–693, 2012.

[42] M. Rus-Calafell, T. Ward, X. C. Zhang, C. J. Edwards, P. Garety, and T. Craig, "The role of sense of voice presence and anxiety reduction in avatar therapy," *Journal of Clinical Medicine*, vol. 9, no. 9, p. 2748, 2020.

[43] C. A. Ross, R. L. Margolis, S. A. Reading, M. Pletnikov, and J. T. Coyle, "Neurobiology of schizophrenia," *Neuron*, vol. 52, no. 1, pp. 139–153, 2006.

[44] Health Canada, "A report on mental illnesses in Canada," *Ottawa, Canada*, 2002.

[45] J. M. Kane, "Treatment-resistant schizophrenic patients.," *The Journal of Clinical Psychiatry*, vol. 57, pp. 35–40, 1996.

[46] A. Demjaha, J. Lappin, D. Stahl, M. Patel, J. MacCabe, O. Howes, M. Heslin, U. Reininghaus, K. Donoghue, B. Lomas, *et al.*, "Antipsychotic treatment resistance in first-episode psychosis: prevalence, subtypes and predictors," *Psychological Medicine*, vol. 47, no. 11, pp. 1981–1989, 2017.

[47] J. Leff, G. Williams, M. A. Huckvale, M. Arbuthnot, and A. P. Leff, "Computer-assisted therapy for medication-resistant auditory hallucinations: Proof-of-concept study," *The British Journal of Psychiatry*, vol. 202, no. 6, pp. 428–433, 2013.

[48] M. A. Huckvale, J. Leff, and G. Williams, "Avatar therapy: an audio-visual dialogue system for treating auditory hallucinations," in *INTERSPEECH*, pp. 392–396, 2013.

[49] T. K. Craig, M. Rus-Calafell, T. Ward, M. Fornells-Ambrojo, P. McCrone, R. Emsley, and P. Garety, "The effects of an audio visual assisted therapy aid for refractory auditory hallucinations (avatar therapy): Study protocol for a randomised controlled trial," *Trials*, vol. 16, no. 1, pp. 1–9, 2015.

[50] C. Ducher, "Gan-based interaction paradigms for photorealistic avatar creation," Master's thesis, McGill University, Department of Electrical and Computer Engineering, Montreal, Canada, 2021.

[51] "Avatar Maker Pro, 3D avatar from a single selfie." https://avatarsdk.com//. Accessed: 2021-07-20.

[52] "Unity multipurpose avatar, an open avatar creation framework." http://umawiki.secretanorak.com/Main_Page. Accessed: 2021-07-20.

[53] A. Lim, H. W. Hoek, M. L. Deen, J. D. Blom, R. Bruggeman, W. Cahn, L. De Haan, R. S. Kahn, C. J. Meijer, I. Myin-Germeys, *et al.*, "Prevalence and classification of hallucinations in multiple sensory modalities in schizophrenia spectrum disorders," *Schizophrenia research*, vol. 176, no. 2-3, pp. 493–499, 2016.

[54] V. E. L. Teslasuit, "Teslasuit, a full body haptic feedback and motion capture tracking VR suit." https://teslasuit.io/. Accessed: 2021-07-23.

[55] bHaptics Inc., "Tactsuit, most advanced full body haptic suit." https://www.bhaptics.com/tactsuit. Accessed: 2021-07-23.

[56] A. Talhan and S. Jeon, "Pneumatic actuation in haptic-enabled medical simulators: A review," *IEEE Access*, vol. 6, pp. 3184–3200, 2017.

[57] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological Measurement*, vol. 28, no. 3, p. R1, 2007.

[58] M. Benedek and C. Kaernbach, "A continuous measure of phasic electrodermal activity," *Journal of Neuroscience Methods*, vol. 190, no. 1, pp. 80–91, 2010.

[59] L. Crazy Minnow Studio, "Salsa lipsync suite v2." https://crazyminnowstudio.com/unity-3d/lip-sync-salsa/, 2019.

[60] R. Kumar, J. Sotelo, K. Kumar, A. de Brébisson, and Y. Bengio, "Obamanet: Photo-realistic lip-sync from text," *arXiv preprint arXiv:1801.01442*, 2017.

[61] P. Edwards, C. Landreth, E. Fiume, and K. Singh, "Jali: an animator-centric viseme model for expressive lip synchronization," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–11, 2016.

[62] W. Mattheyses, L. Latacz, and W. Verhelst, "Comprehensive many-to-many phoneme-to-viseme mapping and its application for concatenative visual speech synthesis," *Speech Communication*, vol. 55, no. 7-8, pp. 857–876, 2013.

[63] Y. Fu, R. Li, T. S. Huang, and M. Danielsen, "Real-time multimodal human–avatar interaction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 4, pp. 467–477, 2008.

[64] L. Wang, W. Han, and F. K. Soong, "High quality lip-sync animation for 3D photo-realistic talking head," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4529–4532, IEEE, 2012.

[65] K. Lenzo, "The CMU Pronouncing Dictionary, 2017," 2014.

[66] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Prentice-Hall, Inc., 1993.

[67] E. Bozkurt, C. E. Erdem, E. Erzin, T. Erdem, and M. Ozkan, "Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation," in *2007 3DTV Conference*, pp. 1–4, IEEE, 2007.

[68] P. Lucey, T. Martin, and S. Sridharan, "Confusability of phonemes grouped according to their viseme classes in noisy environments," in *Proc. of Australian Int. Conf. on Speech Science & Tech*, pp. 265–270, Citeseer, 2004.

[69] "Plotly technologies inc. collaborative data science." https://plot.ly, 2015.

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

[71] C. Jemine, "Github repository, real-time-voice-cloning." https://github.com/CorentinJ/Real-Time-Voice-Cloning, 2018.

[72] T. K. Craig, M. Rus-Calafell, T. Ward, J. P. Leff, M. Huckvale, E. Howarth, R. Emsley, and P. A. Garety, "Avatar therapy for auditory verbal hallucinations in people with psychosis: a single-blind, randomised controlled trial," *The Lancet Psychiatry*, vol. 5, no. 1, pp. 31–40, 2018.

[73] O. P. du Sert, S. Potvin, O. Lipp, L. Dellazizzo, M. Laurelli, R. Breton, P. Lalonde, K. Phraxayavong, K. O'Connor, J.-F. Pelletier, *et al.*, "Virtual reality therapy for refractory auditory verbal hallucinations in schizophrenia: a pilot clinical trial," *Schizophrenia research*, vol. 197, pp. 176–181, 2018.

[74] T. Ward, R. Lister, M. Fornells-Ambrojo, M. Rus-Calafell, C. J. Edwards, C. O'Brien, T. K. Craig, and P. Garety, "The role of characterisation in everyday voice engagement and avatar therapy dialogue," *Psychological medicine*, pp. 1–8, 2021.

[75] B. Alderson-Day, A. Woods, P. Moseley, S. Common, F. Deamer, G. Dodgson, and C. Fernyhough, "Voice-hearing and personification: characterizing social qualities of auditory verbal hallucinations in early psychosis," *Schizophrenia Bulletin*, vol. 47, no. 1, pp. 228–236, 2021.

[76] E. B. Foa and M. J. Kozak, "Emotional processing of fear: exposure to corrective information.," *Psychological bulletin*, vol. 99, no. 1, p. 20, 1986.

[77] M. Bohlken, K. Hugdahl, and I. Sommer, "Auditory verbal hallucinations: neuroimaging and treatment," *Psychological Medicine*, vol. 47, no. 2, pp. 199–208, 2017.

[78] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2020.

[79] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, vol. 1, pp. 285–288, IEEE, 1998.

[80] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, vol. 30, p. 2966–2974, 2017.

[81] Y. Higuchi, M. Suzuki, and G. Kurata, "Speaker embeddings incorporating acoustic conditions for diarization," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7129–7133, 2020.

[82] L. Savioja, "Modeling techniques for virtual acoustics," *Simulation*, vol. 45, no. 10, p. 10, 1999.

[83] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[84] M. Puckette, "Phase-locked vocoder," in *Proceedings of 1995 Workshop on Applications of Signal Processing to Audio and Accoustics*, pp. 222–225, IEEE, 1995.

[85] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, 2000.

[86] L. Xiao-chun, Y. Jun-xun, and H. Wei-ping, "A text-independent speaker recognition system based on probabilistic principle component analysis," in *2012 3rd International Conference on System Science, Engineering Design and Manufacturing Informatization*, vol. 1, pp. 255–260, IEEE, 2012.

[87] M. Diez, L. Burget, and P. Matejka, "Speaker diarization based on bayesian HMM with eigenvoice priors.," in *Odyssey*, pp. 147–154, 2018.

[88] S. Bavkar and S. Sahare, "Pca based single channel speech enhancement method for highly noisy environment," in *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1103–1107, IEEE, 2013.

[89] M. Huckvale, J. Leff, and G. Williams, "Avatar therapy: an audio-visual dialogue system for treating auditory hallucinations," in *Proc. Interspeech 2013*, pp. 392–396, 2013.

[90] T. Bhuta, L. Patrick, and J. D. Garnett, "Perceptual evaluation of voice quality and its correlation with acoustic measurements," *Journal of Voice*, vol. 18, no. 3, pp. 299–304, 2004.

[91] W. T. Fitch and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1511–1522, 1999.

[92] S. Berger, O. Niebuhr, and B. Peters, "Winning over an audience–a perception-based analysis of prosodic features of charismatic speech," in *Proc. 43rd Annual Conference of the German Acoustical Society, Kiel, Germany*, pp. 1454–1457, 2017.

[93] K. R. Scherer, H. London, and J. J. Wolf, "The voice of confidence: Paralinguistic cues and audience evaluation," *Journal of Research in Personality*, vol. 7, no. 1, pp. 31–44, 1973.

[94] H. Matsumoto, S. Hiki, T. Sone, and T. Nimura, "Multidimensional representation of personal quality of vowels and its acoustical correlates," *IEEE Transactions on Audio and Electroacoustics*, vol. 21, no. 5, pp. 428–436, 1973.

[95] M. P. Gelfer, "Perceptual attributes of voice: Development and use of rating scales," *Journal of Voice*, vol. 2, no. 4, pp. 320–326, 1988.

[96] K. M. Aldrich, E. J. Hellier, and J. Edworthy, "What determines auditory similarity? The effect of stimulus group and methodology," *Quarterly Journal of Experimental Psychology*, vol. 62, no. 1, pp. 63–83, 2009.

[97] J. Pasquero, J. Luk, S. Little, and K. MacLean, "Perceptual analysis of haptic icons: an investigation into the validity of cluster sorted MDS," in *2006 14th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, pp. 437–444, IEEE, 2006.

[98] D. A. Stevens, R. F. Smith, and H. T. Lawless, "Multidimensional scaling of ferrous sulfate and basic tastes," *Physiology & Behavior*, vol. 87, no. 2, pp. 272–279, 2006.

[99] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," in *Ninth International Conference on Spoken Language Processing*, 2006.

[100] K. Kühne, M. H. Fischer, and Y. 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*, vol. 14, p. 105, 2020.

[101] A. Baird, S. H. Jørgensen, E. Parada-Cabaleiro, N. Cummins, S. Hantke, and B. Schuller, "2018-04 the perception of vocal traits in synthesized voices: age, gender, and human likeness," *Journal of the Audio Engineering Society*, vol. 66, no. 4, pp. 277–285, 2018.

[102] A. Baird, E. Parada-Cabaleiro, S. Hantke, F. Burkhardt, N. Cummins, and B. Schuller, "2018-09 the perception and analysis of the likeability and human likeness of synthesized speech," 2018.

[103] G. Hua, J. Huang, Y. Q. Shi, J. Goh, and V. L. Thing, "Twenty years of digital audio watermarking—a comprehensive review," *Signal processing*, vol. 128, pp. 222–242, 2016.

[104] J. D. Harnsberger, R. Shrivastav, W. Brown Jr, H. Rothman, and H. Hollien, "Speaking rate and fundamental frequency as speech cues to perceived age," *Journal of Voice*, vol. 22, no. 1, pp. 58–69, 2008.

[105] J. Pittam, *Voice in social interaction*, vol. 5. Sage, 1994.

[106] S. Feldstein and R. N. Bond, "Perception of speech rate as a function of vocal intensity and frequency," *Language and Speech*, vol. 24, no. 4, pp. 387–394, 1981.

[107] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. Skerry-Ryan, Y. Jia, A. Rosenberg, and B. Ramabhadran, "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning," *arXiv preprint arXiv:1907.04448*, 2019.

[108] T. H. Nayani and A. S. David, "The auditory hallucination: a phenomenological survey," *Psychological medicine*, vol. 26, no. 1, pp. 177–189, 1996.

[109] F. J. Prerost, D. Sefcik, B. D. Smith, *et al.*, "Differential diagnosis of patients presenting with hallucinations," *Osteopathic Family Physician*, vol. 6, no. 2, 2014.