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

Hyejin Lee*
McGill University
Montréal, Québec, Canada
hyejin.lee2@mail.mcgill.ca

Ruixi Jiang
McGill University
Montréal, Québec, Canada
ruixi.jiang@mail.mcgill.ca

Yongjae Yoo
McGill University
Montréal, Québec, Canada
yongjae.yoo@mcgill.ca

Max Henry
McGill University
Montréal, Québec, Canada
max.henry@mail.mcgill.ca

Jeremy R. Cooperstock
McGill University
Montréal, Québec, Canada
jer@cim.mcgill.ca

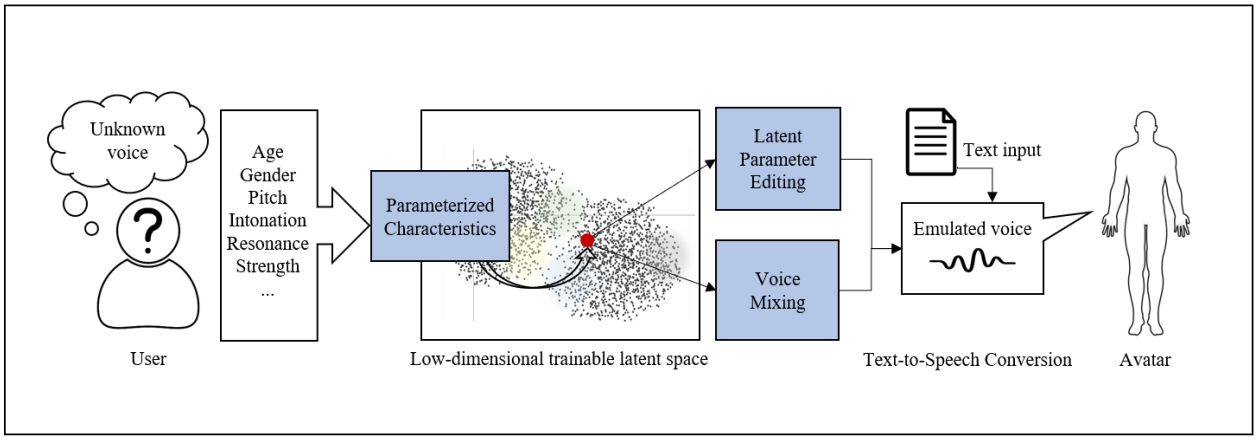


Figure 1: Overall pipeline to characterize the voice from user's head and externalize it as a resource to generate speech of a virtual avatar.

ABSTRACT

The need to generate convincing simulation of voices often arises in the context of avatar therapy, a treatment approach for disorders such as schizophrenia. This treatment involves patients interacting with simulations of the entity they imagine to be responsible for the voices they hear, for which there is often no external reference available. However, in such scenarios, there is little knowledge of how to design and reproduce these voices in a convincing manner. Existing voice manipulation interfaces are often complex to use, and highly limited in their ability to modify vocal characteristics beyond small adjustments. To address these challenges, we designed a framework that allows users to explore and select from a large set of voices, and thereafter manipulate the voice(s) to converge towards an effective match for one they have in mind. We demonstrated both the usability and superior performance of this system compared to existing voice manipulation interfaces.

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3501871

CCS CONCEPTS

• **Human-centered computing** → **Sound-based input / output**; **User studies**; *Mixed / augmented reality*.

KEYWORDS

Voice Transformation Interface, Speech Synthesis, Avatar Therapy

ACM Reference Format:
Hyejin Lee, Ruixi Jiang, Yongjae Yoo, Max Henry, and Jeremy R. Cooperstock. 2022. The Sound of Hallucinations: Toward a more convincing emulation of internalized voices. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3491102.3501871

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 [11, 14, 36]. 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 [1, 45]. 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 [17]. 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) [10].

2 RELATED WORK

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 [39]. 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 [25]. 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" [3]. 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 [44]. 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 [21]. 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 [37]. 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.

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 [43] 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 [4] 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 [31] and phase-vocoding [35] 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.

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 [26]. 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 [46], and a speaker diarization approach that identifies speakers from a recorded conversation [13]. 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 [7]. In the application of vocal synthesis for avatar therapy [23], 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.

3 VOICE MODELLING INTERFACE PARADIGM

The design of our interface (Figure 2) is intended to support voice exploration and manipulation without requiring any specialized knowledge in the audio domain. Our focus is therefore on facilitating control of perceptually meaningful qualities of human voices, based on terminology that is accessible to non-experts.

3.1 System Overview

Figure 3 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 voice characteristics available. After users have selected and refined their target voice, an external TTS module, such as the Wavenet neural vocoder [32] 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.

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 [24], as shown in Figure 3. 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 [44]. We then applied the Uniform Manifold Approximation Projection (UMAP) [29] 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 conventional 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.

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 [9], and our perception of the effects of these parameters in

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

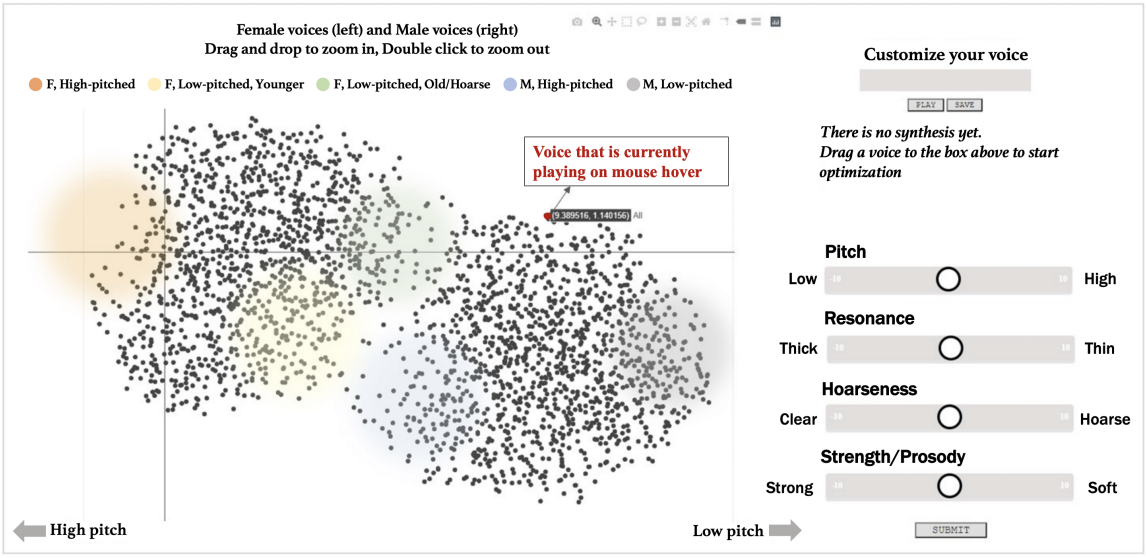


Figure 2: The GUI of overall interface. The voice map exploration is displayed on the left side and latent parameter editing is on the right side. The map is represented as a lower-dimensional manifold of a large set of voice samples. In theory, the axes do not have any physical meaning, but indicate 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.

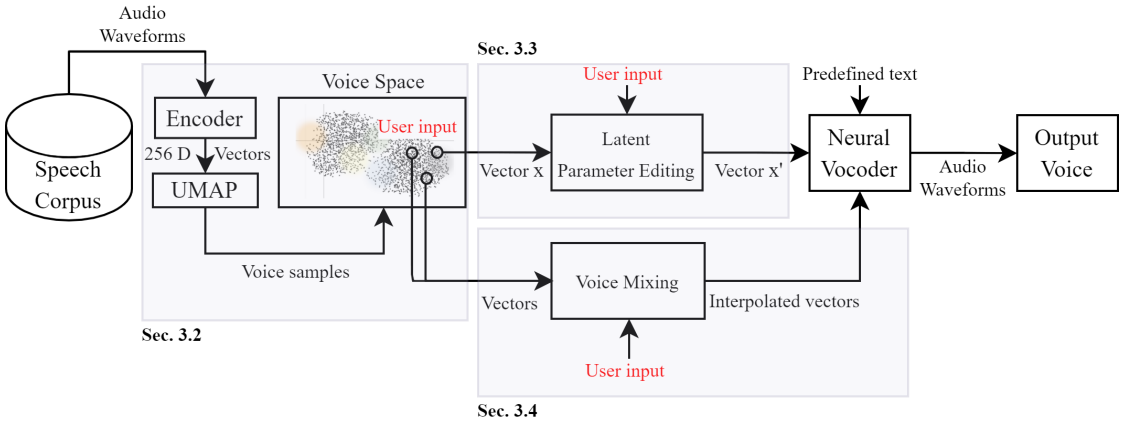


Figure 3: The overall procedure of voice modelling through latent parameter editing (Section 3.3) and voice mixing (Section 3.4).

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 [16, 41]. "Hoarseness" refers to the speaker's voice quality, in line with a raspy, husky voice [9], and the prosodic qualities may be described as "confidence" [8, 38]. These features correlated with findings from the prior literature regarding the characteristics most important to human perception of voice [18, 19, 28]. 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 principal components in the user interface.

3.4 Voice Mixing

The mixing technique recommends new design directions with a minimal amount of user effort (Figure 4). Once the user selects two voices from the map, the system automatically calculates speaker feature vectors of five interpolated points between these selected

Figure 4: The GUI of the voice mixing interface

voices, based on the L2 norm (See Figure 3). 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.

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.

4.1 Study 1: Latent Parameter Editing vs. Mixing

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

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.

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.

4.1.3 Statistical Analysis. We compared the performance within the three conditions: latent parameter editing (Section 3.3), voice mixing (Section 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.

4.2 Study 1: Results

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, 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).

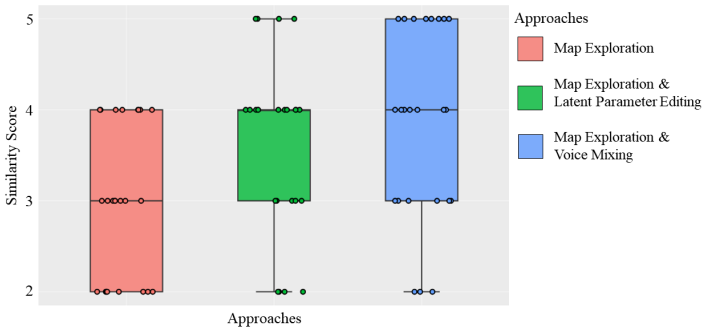


Figure 5: Comparison of three groups of voices on how similarly they matched to the target voices of participants.

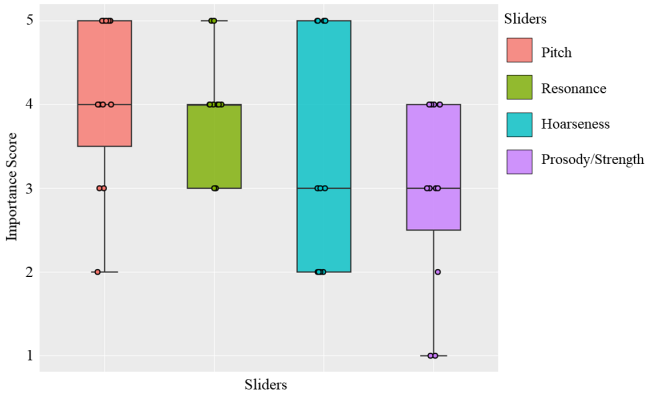


Figure 6: Comparison of effectiveness of the four latent parameters in reproducing target voices of participants.

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 (σ = 4.83) for pitch, 3.67 times (σ = 2.56) for resonance, 2.7 times (σ = 1.46) for hoarseness, and 3 times (σ = 1.62) for prosody, for each target voice. Figure 7 illustrates the overall tendency of participants adjusting the four parameters 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 6).

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, σ = 0.65 on a 5-point Likert scale). Based on the generation of such voices, participants also had positive assessments of the effectiveness of

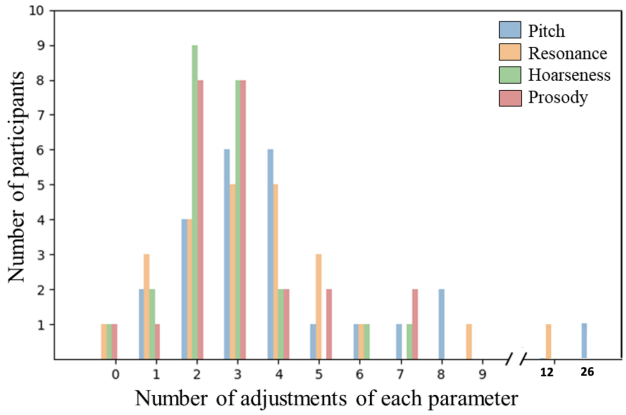


Figure 7: The number of adjustments made on each parameter in the experimental trials.

the mixing technique to achieve better matches to their mental representation ($\bar{x}$ = 4.0, σ = 0.7).

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 8, 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%**

Figure 8: Comparison of participants' preference among the three approaches.

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 Section 5.1.2.

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.

4.3 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.

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 (Section 3.3), voice mixing (Section 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.

4.3.2 Participants. Twelve participants (6F, 6M) with an average age of 24.7 (σ = 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.

4.3.3 Procedure. We conducted cluster analysis to evaluate voices from the preliminary session based on multidimensional scaling (MDS). A Windows application (Figure 9) 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.

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 [2, 33, 40].

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 to calculate the dissimilarity matrix D for every non-diagonal component, as in Equation 1. Every

[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/

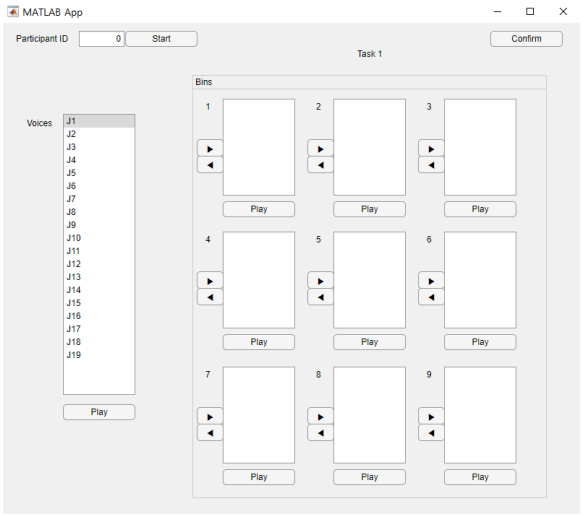


Figure 9: The GUI of the Windows application developed for conducting the cluster sorting task.

diagonal component was set to zero.

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

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.

4.4 Study 2: Results

4.4.1 Evaluation Criteria. Results from our open-ended post-test questionnaire (Figure 10) 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.

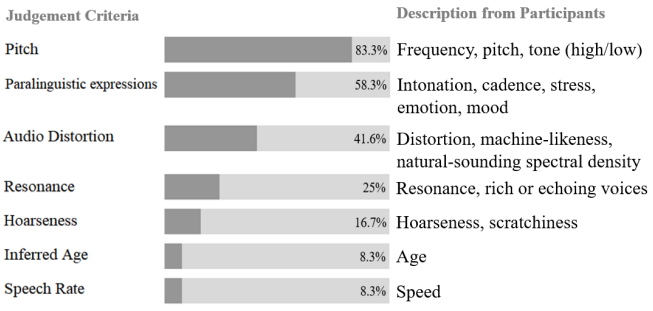
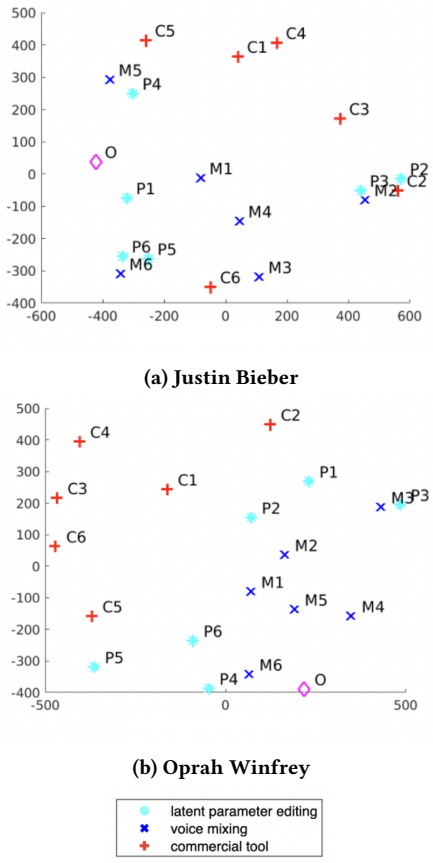


Figure 10: Main factors considered in the voice classification task reported by participants



(a) Justin Bieber

(b) Oprah Winfrey

Figure 11: 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 Figure 11b, M4 is perceived to be roughly twice as similar to O3 as M3.

4.4.2 Multidimensional Scaling Results. The two-dimensional MDS results show that both latent parameter editing and voice mixing yielded significantly higher similarity than the commercial tool (Figure 11). 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.

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 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 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

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 DISCUSSION

5.1 Voice Space Exploration

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.

5.1.2 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 [42]. 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.

5.2 Synthesis Techniques

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 [12, 38]. 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 [5, 6, 27]. 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.

5.2.2 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 7. 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.

5.2.3 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 [22] 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.

5.3 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.

5.3.1 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 [20]. The impression of extroversion or perceived charisma of the speaker arises from a collective judgement of speech rate, pitch variation, and loudness [8, 34]. 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 [15].

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.

5.3.2 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) [24]. 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 [47], 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.

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.

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.

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

ACKNOWLEDGMENTS

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

REFERENCES

[1] Ben Alderson-Day, Angela Woods, Peter Moseley, Stephanie Common, Felicity Deamer, Guy Dodgson, and Charles Fernyhough. 2021. Voice-hearing and personification: characterizing social qualities of auditory verbal hallucinations in early psychosis. *Schizophrenia bulletin* 47, 1 (2021), 228–236.

[2] Kirsteen M Aldrich, Elizabeth J Hellier, and Judy Edworthy. 2009. What determines auditory similarity? The effect of stimulus group and methodology. *Quarterly journal of experimental psychology* 62, 1 (2009), 63–83.

[3] Sercan Arik, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. 2017. Deep Voice 2: Multi-Speaker Neural Text-to-Speech. arXiv:1705.08947 [cs.CL]

[4] AV Voice Changer 2021. AV Voice Chaner Software Diamond. https://www.audio4fun.com/voice-changer.htm.

[5] Alice Baird, Stina Hasse Jørgensen, Emilia Parada-Cabaleiro, Nicholas Cummins, Simone Hantke, and Björn Schuller. 2018. The perception of vocal traits in synthesized voices: age, gender, and human likeness. *Journal of the Audio Engineering Society* 66, 4 (2018), 277–285.

[6] Alice Baird, Emilia Parada-Cabaleiro, Simone Hantke, Felix Burkhardt, Nicholas Cummins, and Björn Schuller. 2018. The perception and analysis of the likeability and human likeness of synthesized speech. (2018).

[7] Sangita Bavkar and Shashikant Sahare. 2013. PCA based single channel speech enhancement method for highly noisy environment. In *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 1103–1107.

[8] Stephanie Berger, Oliver Niebuhr, and Benno Peters. 2017. 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*. 1454–1457.

[9] Tarika Bhuta, Linda Patrick, and James D Garnett. 2004. Perceptual evaluation of voice quality and its correlation with acoustic measurements. *Journal of voice* 18, 3 (2004), 299–304.

[10] MM Bohlken, K Hugdahl, and IEC Sommer. 2017. Auditory verbal hallucinations: neuroimaging and treatment. *Psychological Medicine* 47, 2 (2017), 199–208.

[11] Tom KJ Craig, Mar Rus-Calafell, Thomas Ward, Julian P Leff, Mark Huckvale, Elizabeth Howarth, Richard Emsley, and Philippa A Garety. 2018. AVATAR therapy for auditory verbal hallucinations in people with psychosis: a single-blind, randomised controlled trial. *The Lancet Psychiatry* 5, 1 (2018), 31–40.

[12] Laurence Devillers and Laurence Vidrascu. 2006. Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In *Ninth International Conference on Spoken Language Processing*.

[13] Mireia Diez, Lukás Burget, and Pavel Matejka. 2018. Speaker Diarization based on Bayesian HMM with Eigenvoice Priors.. In *Odyssey*. 147–154.

[14] Olivier Percie du Sert, Stéphane Potvin, Olivier Lipp, Laura Dellazizzo, Mélanie Laurelli, Richard Breton, Pierre Lalonde, Kingsada Phraxayavong, Kieron O'Connor, Jean-François Pelletier, et al. 2018. Virtual reality therapy for refractory auditory verbal hallucinations in schizophrenia: a pilot clinical trial. *Schizophrenia research* 197 (2018), 176–181.

[15] Stanley Feldstein and Ronald N Bond. 1981. Perception of speech rate as a function of vocal intensity and frequency. *Language and Speech* 24, 4 (1981), 387–394.

[16] W Tecumseh Fitch and Jay Giedd. 1999. Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America* 106, 3 (1999), 1511–1522.

[17] Edna B Foa and Michael J Kozak. 1986. Emotional processing of fear: exposure to corrective information. *Psychological bulletin* 99, 1 (1986), 20.

[18] Marylou Pausewang Gelfer. 1988. Perceptual attributes of voice: Development and use of rating scales. *Journal of Voice* 2, 4 (1988), 320–326.

[19] Marylou Pausewang Gelfer. 1993. A multidimensional scaling study of voice quality in females. *Phonetica* 50, 1 (1993), 15–27.

[20] James D Harnsberger, Rahul Shrivastav, WS Brown Jr, Howard Rothman, and Harry Hollien. 2008. Speaking rate and fundamental frequency as speech cues to perceived age. *Journal of voice* 22, 1 (2008), 58–69.

[21] Yosuke Higuchi, Masayuki Suzuki, and Gakuto Kurata. 2020. Speaker Embeddings Incorporating Acoustic Conditions for Diarization. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7129–7133. https://doi.org/10.1109/ICASSP40776.2020.9054273

[22] Guang Hua, Jiwu Huang, Yun Q Shi, Jonathan Goh, and Vrizlynn LL Thing. 2016. Twenty years of digital audio watermarking—a comprehensive review. *Signal processing* 128 (2016), 222–242.

[23] Mark Huckvale, Julian Leff, and Geoff Williams. 2013. Avatar therapy: an audio-visual dialogue system for treating auditory hallucinations. In *Proc. Interspeech 2013*. 392–396. https://doi.org/10.21437/Interspeech.2013-107

[24] Ye Jia, Yu Zhang, Ron J Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *arXiv preprint arXiv:1806.04558* (2018).

[25] Alexander Kain and Michael W Macon. 1998. Spectral voice conversion for text-to-speech synthesis. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, Vol. 1. IEEE, 285–288.

[26] Roland Kuhn, J-C Junqua, Patrick Nguyen, and Nancy Niedzielski. 2000. Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing* 8, 6 (2000), 695–707.

[27] Katharina Kühne, Martin H Fischer, and Yuefang Zhou. 2020. The Human Takes It All: Humanlike Synthesized Voices Are Perceived as Less Eerie and More Likable. Evidence From a Subjective Ratings Study. *Frontiers in neurorobotics* 14 (2020), 105.

[28] Hiroshi Matsumoto, Shizuo Hiki, Toshio Sone, and Tadamoto Nimura. 1973. Multidimensional representation of personal quality of vowels and its acoustical correlates. *IEEE Transactions on Audio and Electroacoustics* 21, 5 (1973), 428–436.

[29] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[30] MorphVOX Pro 2021. MorphVOX Pro voice changer. https://screamingbee.com/.

[31] Eric Moulines and Francis Charpentier. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication* 9, 5-6 (1990), 453–467.

[32] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[33] Jerome Pasquero, Joseph Luk, Shannon Little, and Karon MacLean. 2006. 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*. IEEE, 437–444.

[34] Jeff Pittam. 1994. *Voice in social interaction*. Vol. 5. Sage.

[35] Miller Puckette. 1995. Phase-locked vocoder. In *Proceedings of 1995 Workshop on Applications of Signal Processing to Audio and Accoustics*. IEEE, 222–225.

[36] Mar Rus-Calafell, Thomas Ward, Xiao Chi Zhang, Clementine J Edwards, Philippa Garety, and Tom Craig. 2020. The role of sense of voice presence and anxiety reduction in AVATAR therapy. *Journal of Clinical Medicine* 9, 9 (2020), 2748.

[37] Lauri Savioja. 1999. Modeling techniques for virtual acoustics. *Simulation* 45, 10 (1999), 10.

[38] Klaus R Scherer, Harvey London, and Jared J Wolf. 1973. The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality* 7, 1 (1973), 31–44.

[39] Berrak Sisman, Junichi Yamagishi, Simon King, and Haizhou Li. 2020. An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2020).

[40] David A Stevens, Rebecca F Smith, and Harry T Lawless. 2006. Multidimensional scaling of ferrous sulfate and basic tastes. *Physiology & Behavior* 87, 2 (2006), 272–279.

[41] Johan Sundberg and RT Sataloff. 2005. Vocal tract resonance. (2005).

[42] Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. 2017. Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. (2017).

[43] Voicemod 2021. Voicemod: Free Real-time Voice Changer. https://www.voicemod.net/.

[44] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4879–4883.

[45] Thomas Ward, Rachel Lister, Miriam Fornells-Ambrojo, Mar Rus-Calafell, Clementine J Edwards, Conan O'Brien, Tom KJ Craig, and Philippa Garety. 2021. The role of characterisation in everyday voice engagement and AVATAR therapy dialogue. *Psychological medicine* (2021), 1–8.

[46] Lu Xiao-chun, Yin Jun-xun, and Hu Wei-ping. 2012. 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. IEEE, 255–260.

[47] Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, RJ Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. 2019. Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning. *arXiv preprint arXiv:1907.04448* (2019).