Voice Chimera: Redefining Voice Identity Through Diverse Synthesis Techniques

Ju Lee
Viterbi School of Engineering
University of Southern California
mail.juyoung.lee@gmail.com

ABSTRACT

This study explores innovative voice blending techniques to create unique vocal identities by combining characteristics from multiple speakers. We investigate four methods: Linear Interpolation (LERP), Spherical Linear Interpolation (SLERP), a novel genetic algorithm-inspired approach, and Principal Component Analysis (PCA). Leveraging the VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech) model, we implement these techniques to generate natural-sounding, blended voices. The genetic algorithm-inspired method simulates genetic inheritance in voice characteristics, while PCA is used for manipulating voice embeddings in a reduced dimensional space. We evaluate the synthesized voices using both subjective Mean Opinion Scores (MOS) and objective similarity metrics, including cosine similarity, Euclidean distance, and Mel-Cepstral Distortion (MCD). Our results demonstrate the viability of each blending method, with distinct trade-offs between novelty and naturalness. LERP and SLERP show consistent performance, while the genetic approach offers higher variability and potential for unique outputs. PCA provides a balance between preserving voice characteristics and enabling extensive blending. This research contributes to the advancement of voice synthesis technologies, offering new approaches for creating diverse and natural-sounding synthetic voices. The findings have potential applications in entertainment, assistive technologies, and personalized AI interfaces, while also providing insights into the challenges of balancing voice quality, naturalness, and uniqueness in blended voices.

1. INTRODUCTION

In recent years, voice cloning technology has gained significant attention, particularly through its application in creating song covers by famous artists. This technology has enabled the production of highly convincing imitations of well-known voices, leading to a surge in AI-generated content that mimics specific performers. However, this trend has raised serious concerns regarding intellectual property rights and copyright infringement. The ability to replicate a person's voice without their consent poses ethical and legal challenges, particularly in the music and entertainment industries where an artist's voice is a crucial part of their identity and brand.

In response to these issues, voice blending technology has been emerging as a potential solution. Unlike voice cloning, which aims to replicate a specific individual's voice, voice blending focuses on creating new, unique vocal identities by combining characteristics from multiple speakers. This approach offers a creative middle ground that can produce distinctive and appealing voices without directly copying any single individual's vocal signature.

Voice blending builds upon existing research in voice conversion and multi-speaker synthesis. Notable works in this area include "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis" (Wang et al., 2018) and "SpeechSplit: A Framework for Decomposing and Recomposing Speech" (Qian et al., 2020), which have explored ways to manipulate and combine different aspects of speech. However, the specific application of these techniques to create entirely new voices as a means to address copyright concerns is a relatively new direction.

This project aims to expand on these voice-blending techniques, exploring creative ways to generate new voices by combining features from different speakers. However, it's crucial to note that the definition of a completely "new" voice in the context of voice blending is subjective and can be challenging to quantify. Some observers may easily recognize the input voices in the blended output, based on their familiarity with the original speakers or their keen auditory perception.

We posit that the concept of a "new" voice in blending techniques could be defined by the distance from its input sources. The greater the transformation of input features in the output, the harder it becomes to recognize the existing inputs, potentially leading to a perception of the blended voice as a distinct, real voice. This project, therefore, not only explores traditional linear interpolation methods but also aims to showcase more creative approaches to voice blending.

One innovative approach in this study draws inspiration from genetic principles, mimicking the universal laws of genetic inheritance in voice characteristics. This method simulates a form of "voice birth" by combining two input "parent" voices to generate a "child" voice as output. Unlike machine learning-based approaches, this genetic combination algorithm directly manipulates voice embeddings using principles inspired by genetic inheritance.

The genetic combination algorithm operates on voice embeddings, which are numerical representations of voice characteristics. It simulates genetic inheritance through several key steps. First, the algorithm implements a crossover mechanism, selecting multiple points in the embeddings and alternating segments from each parent voice. This process mimics the combination of genetic material from both parents in biological reproduction. Subsequently, the algorithm introduces small random changes to the new embedding, simulating genetic mutations that occur naturally and contribute to uniqueness. Finally, the resulting embedding undergoes normalization to ensure it remains within the valid range for voice characteristics.

This approach compels the blending process to mirror genetic inheritance in vocal traits, potentially creating more natural and diverse blended voices. The genetic combination method offers several advantages. By emulating natural genetic processes, the algorithm may produce voice blends that sound more organic and natural, enhancing biological plausibility. The crossover and mutation processes introduce controlled variability, potentially leading to more diverse and unique voice outputs. Furthermore, unlike black-box machine learning models, this algorithm's operations are transparent and can be easily understood and adjusted, improving interpretability.

The genetic algorithm-inspired method represents a novel approach in voice blending technology, offering a balance between controlled manipulation and natural variation. Its potential for creating diverse yet biologically plausible voice outputs makes it a promising avenue for further research in voice synthesis and transformation.

It's important to note that this genetic-inspired approach makes some simplifying assumptions. The model focuses primarily on the genetic factors influencing voice characteristics, deliberately ignoring environmental factors that can affect voice development. While this is a limitation, we believe that genetic factors play a crucial role in determining voice characteristics, making this a valuable avenue for exploration in voice blending technology.

In addition to the genetic algorithm-inspired method, this project also explores the use of Principal Component Analysis (PCA) for voice blending. PCA is a dimensionality reduction technique that can be used to identify and manipulate the principal components of voice embeddings. By applying PCA to voice embeddings, we aim to create blended voices that capture the most significant characteristics of the input voices while potentially generating novel combinations.

By exploring this genetics-inspired method alongside other mathematical approaches to blend speaker identities, this project aims to contribute to the development of ethical and creative voice synthesis technologies. The potential to create an infinite variety of new, natural-sounding voices could have far-reaching implications across multiple industries, while addressing the legal and ethical concerns raised by direct voice cloning.

To implement our voice blending techniques, including the genetic combination algorithm, we build upon the VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech) model. VITS, introduced by Kim et al. (2021), is a state-of-the-art end-to-end text-to-speech synthesis model that combines variational inference with adversarial learning. Its ability to generate high-quality speech and its flexible architecture makes it an ideal foundation for our voice-blending experiments.

By leveraging the VITS model and incorporating our genetic combination algorithm, we aim to push the boundaries of voice synthesis technology. This approach allows us to explore innovative ways of creating new voices that are both unique and natural-sounding while addressing the ethical and legal concerns associated with direct voice cloning.

2. BACKGROUND

The field of speech synthesis has undergone a remarkable transformation in recent years, driven by significant advancements in deep learning and neural network architectures. This evolution has fundamentally reshaped text-to-speech (TTS) technologies, voice conversion techniques, and most crucially for our research, speaker embedding methods, ultimately paving the way for innovative approaches in voice blending.

Text-to-speech synthesis has progressed from early rule-based systems to sophisticated neural models. Landmark developments in this journey include WaveNet (Oord et al., 2016), which significantly improved synthesized speech quality through its deep generative model for raw audio waveforms. This was followed by Tacotron 2 (Shen et al., 2018), an end-to-end neural TTS system that combined a sequence-to-sequence model with a modified WaveNet vocoder, achieving near-human naturalness in synthesized speech. These advancements set the stage for more complex voice manipulation techniques.

Central to these advancements are speaker embeddings, compact numerical representations of speaker characteristics. The x-vector embeddings proposed by Wan et al. (2018) have become crucial in multi-speaker TTS systems and voice conversion applications. These embeddings capture essential voice features such as pitch, timbre, and speaking style in a high-dimensional space, enabling more nuanced manipulation of voice characteristics. The ability to extract, modify, and combine these embeddings forms the foundation of modern voice synthesis and transformation techniques.

Voice conversion, which modifies a source speaker's voice to mimic a target speaker, has seen remarkable progress largely due to the manipulation of speaker embeddings. Recent advancements in this field have been driven by deep-learning approaches. Notable among these is the CycleGAN-VC model introduced by Kaneko and Kameoka (2018), which demonstrated impressive results in non-parallel voice conversion. This technology has further evolved to include voice cloning, a specialized form that focuses on replicating voices from limited samples. In this area, Jia et al. (2018) made significant strides, showcasing the ability to efficiently learn and reproduce speaker embeddings from minimal input data, thus enabling the creation of highly convincing voice clones. These developments in voice conversion and cloning techniques have paved the way for more advanced voice manipulation methods, including the voice blending approaches explored in our research.

Our research explores innovative approaches to manipulating speaker embeddings, including a method inspired by genetic algorithms and Principal Component Analysis (PCA). Genetic algorithms, originally developed in the field of computational biology, mimic the processes of natural selection and genetic inheritance. While traditional genetic algorithms involve evolutionary processes over multiple generations, our approach draws inspiration from key genetic mechanisms such as crossover and mutation. We apply these concepts to the domain of voice embeddings, allowing us to create new voice characteristics by combining and altering features from parent voices in a manner analogous to genetic recombination and mutation in biological systems.

Principal Component Analysis (PCA) is a dimensionality reduction technique widely used in various fields of data analysis and machine learning. It works by identifying the principal components of variation in high-dimensional data, allowing for the representation of complex datasets in a lower-dimensional space while retaining the most significant features of the original data. This technique can be particularly useful when dealing with high-dimensional data.

The VITS model (Kim et al., 2021) marks a significant advancement in end-to-end TTS technology and serves as the foundation for our research. Its architecture, combining variational inference with adversarial learning, provides a robust platform for voice synthesis and blending experiments. VITS's ability to generate high-quality speech directly from text input, coupled with its flexible architecture for handling speaker embeddings, makes it an ideal base for exploring innovative voice blending techniques.

As these technologies evolve, they raise important ethical and legal considerations, particularly regarding voice rights and potential misuse. The challenge lies in balancing technological advancement with the need to protect individual privacy and prevent unauthorized voice replication, especially when working with manipulable speaker embeddings.

This background sets the stage for our exploration of innovative voice blending techniques, leveraging the VITS model and incorporating approaches inspired by genetic algorithms and PCA to manipulate speaker embeddings. Our research aims to contribute to the ongoing development of voice synthesis technologies, exploring new methods for creating diverse and natural-sounding synthetic voices while addressing the ethical and legal concerns associated with direct voice cloning.

3. PROJECT DESCRIPTION

3.1 Objectives

This study aims to develop and evaluate innovative voice blending techniques that combine characteristics from multiple parent voices to create unique, natural-sounding synthetic voices. Our primary objective is to implement and assess a genetic algorithm-inspired method for voice blending, exploring its potential to generate diverse and natural voice combinations. Simultaneously, we seek to investigate the application of Principal Component Analysis (PCA) in voice blending, examining its efficacy in capturing and manipulating key voice characteristics. The study also aims to evaluate the effectiveness of linear interpolation (LERP) and spherical linear interpolation (SLERP) methods for voice blending, providing a comprehensive comparison of different approaches. Through both subjective and objective metrics, we will evaluate the quality, naturalness, blend quality, and similarity of the synthesized voices to their parent voices. Ultimately, this research aims to address ethical and legal concerns associated with direct voice cloning by proposing voice blending as a viable alternative, potentially opening new avenues for ethical voice synthesis in various applications.

3.2. Methodology

3.2.1 Linear Interpolation (LERP) Method

The Linear Interpolation (LERP) method for voice blending begins with the extraction of embeddings from the parent voices. These embeddings are numerical representations of voice characteristics. The core of the LERP method lies in its application of linear interpolation to these embeddings. This process is mathematically represented as:

$$e_{blended} = (1 - \alpha) \cdot e_1 + \alpha \cdot e_2$$

Where e_1 and e_2 are the parent embeddings, and α is the blend ratio. In our experiments, α is fixed at 0.5 to achieve an equal blend of both parent voices. The resulting blended embedding is then used to synthesize the new voice using the VITS model.

3.2.2 Spherical Linear Interpolation (SLERP) Method

The Spherical Linear Interpolation (SLERP) method follows a similar initial step of extracting voice embeddings from the parent voices. However, SLERP applies a more sophisticated interpolation technique that considers the spherical nature of the embedding space. The SLERP formula is defined as:

$$e_{blended} = \frac{\sin((1-t)\omega)}{\sin \omega} e_1 + \frac{\sin(t\omega)}{\sin \omega} e_2$$

In this equation, e_1 and e_2 represent the normalized parent embeddings, t is the interpolation parameter (set to 0.5 in our experiments for equal blending), and ω denotes the angle between e_1 and e_2. The blended embedding resulting from this process is subsequently used to synthesize the new voice through the VITS model.

3.2.3 Genetic Algorithm-Inspired Method

The Genetic Algorithm-Inspired Method for voice blending draws inspiration from biological processes of genetic inheritance to create novel voice embeddings. This approach simulates the genetic recombination and mutation that occur in nature, applying these concepts to the domain of voice characteristics.

At the core of this method is a three-stage process: crossover, mutation, and normalization. The crossover stage mimics genetic recombination by combining segments of two parent voice embeddings.

Given two parent embeddings e_1 and e_2, we define a set of crossover points that divide the embeddings into segments. The new embedding is then constructed by alternating these segments from each parent:

$$e_{new}[j] = \begin{cases} e_1[j] & \text{if } j \in [c_{2i}, c_{2i+1}) \text{ for some } i \\ e_2[j] & \text{if } j \in [c_{2i+1}, c_{2i+2}) \text{ for some } i \end{cases}$$

Following the crossover, a mutation stage introduces small, random variations to the new embedding. This step is crucial for maintaining diversity and exploring the voice characteristic space beyond the direct combinations of the parent embeddings. The mutation is applied probabilistically to each element of the embedding:

$$e_{mutated}[j] = e_{new}[j] \cdot (1 + I_{mut} \cdot m \cdot s)$$

Here, I_mut is an indicator function that determines whether mutation occurs for a given element, m is drawn from a standard normal distribution to provide the direction and magnitude of mutation, and s is a scalar parameter controlling the overall strength of mutations.

The final stage involves normalizing the mutated embedding to ensure consistency in magnitude across different blended voices:

$$e_{final} = \frac{e_{mutated}}{||e_{mutated}||_2}$$

This normalization step is crucial for maintaining the stability of the voice synthesis process, ensuring that the blended embedding remains within the expected range of the voice synthesis model.

The genetic algorithm-inspired method offers a unique approach to voice blending that balances the preservation of parent voice characteristics with the introduction of novel variations. By adjusting parameters such as the number and position of crossover points, mutation probability, and mutation strength, researchers can fine-tune the balance between fidelity to parent voices and the generation of novel voice characteristics.

This method stands in contrast to more deterministic approaches like principal component analysis or simple interpolation. While it introduces greater variability in results, it also offers the potential for generating a wider range of unique, blended voices from the same set of parent embeddings. This variability may necessitate multiple generation attempts or careful parameter tuning to achieve desired outcomes, but it also opens up possibilities for creating diverse and unexpected voice blends.

The resulting final embedding, e_final, encapsulates a unique combination of voice characteristics derived from both parent voices, enhanced by controlled random variations. When used in conjunction with the VITS model, this embedding facilitates the synthesis of a blended voice that may exhibit both familiar and novel characteristics, potentially pushing the boundaries of voice synthesis technology.

3.2.4 Principal Component Analysis (PCA) Method

The Principal Component Analysis (PCA) method for voice blending leverages dimensionality reduction techniques to combine voice characteristics in a lower-dimensional space. This approach allows for the identification and manipulation of the most salient features of voice variation across a population of speakers. At the core of this method is the following equation:

$$e_{blended} = W^T \cdot \frac{1}{2}(W \cdot e_1 + W \cdot e_2)$$

Here, e_1 and e_2 represent the original embeddings of the two parent voices, while W denotes the PCA transformation matrix. This matrix W is derived from applying PCA to a comprehensive set of speaker embeddings, encompassing a diverse range of voices beyond just the parent voices under consideration.

The blending process involves three key steps. First, the parent embeddings are projected into the PCA space through the transformation $W \cdot e_1$ and $W \cdot e_2$. This step maps the voice characteristics into a lower-dimensional space where the axes correspond to the principal components of variation across the speaker population. Second, these transformed embeddings are averaged in the PCA space, effectively combining the voice characteristics in a dimension-reduced representation. Finally, this averaged embedding is projected back to the original embedding space through multiplication with W^T, resulting in the blended embedding e_blended.

The dimensionality of the PCA space, determined by the number of principal components retained in W, serves as a crucial parameter in this method. It allows researchers to balance between preserving fine-grained voice characteristics and enabling more extensive blending. A higher number of components tends to preserve more detailed voice features, while a lower number facilitates more substantial blending by focusing on the most significant variations.

This PCA-based approach offers several advantages in the context of voice blending. Primarily, it captures and utilizes the most significant variations in voice characteristics across a diverse speaker population. By operating in a reduced space where key voice features are emphasized, the method allows for a sophisticated blending process that focuses on the most perceptually relevant aspects of voice variation.

Moreover, the PCA method provides a deterministic and interpretable approach to voice blending. Unlike stochastic methods such as the genetic algorithm-inspired approach, PCA offers consistent results for a given set of inputs and parameters. This consistency facilitates systematic experimentation and analysis of the blending process.

The resulting blended embedding, e_blended, represents a novel combination of voice characteristics derived from both parent voices, mediated through the PCA transformation. When utilized in conjunction with the VITS model, this embedding enables the synthesis of a blended voice that incorporates key features from both parent voices in a controlled and interpretable manner.

In conclusion, the PCA method for voice blending presents a powerful tool for researchers in voice synthesis. It offers a balance between preserving individual voice characteristics and creating novel voice combinations, all within a framework that emphasizes the most significant aspects of voice variation. The method's tunable nature, through adjustment of the number of principal components, provides flexibility in tailoring the blending process to specific research objectives or desired voice outcomes.

3.3 Experimental Setup

3.3.1 Voice Embedding Selection

To ensure consistency across all blending methods, we selected two parent voice embeddings (e_1 and e_2) from the VITS model's embedding space, corresponding to speaker IDs 4 and 100. These embeddings served as the basis for all blending operations.

3.3.2 Blending Parameters

For LERP and SLERP, we fixed the blending ratio α at 0.5 to achieve an equal contribution from both parent voices. This decision allows us to focus on the inherent characteristics of each blending method rather than the effects of varying blend ratios.

For the Genetic Algorithm-inspired method, we explored a range of parameters:

Crossover points: 2, 5, and 25

Mutation rates: 0.01, 0.1, and 0.2

Mutation strengths: 0.05, 0.1, and 0.2

For the PCA method, we tested different dimensionalities by varying the number of principal components: 16, 32, 64, and 128.

3.4 Implementation

We implemented all blending methods using Python, leveraging libraries such as NumPy for numerical operations, Scikit-learn for PCA, and PyTorch for deep learning tasks. The VITS model served as the foundation for voice synthesis, allowing us to generate audio from the blended embeddings.

3.5 Evaluation Metrics

We employed both subjective and objective metrics to evaluate the performance of each blending method:

3.5.1 Subjective Evaluation

We conducted a Mean Opinion Score (MOS) study with five participants. Each participant rated the blended voices on three aspects:

Sample Quality (Q): Overall quality of the synthesized speech

Blend Quality (B): How well the sample blends features from both parent voices

Naturalness (N): How natural or human-like the speech sounds

Ratings were given on a 5-point Likert scale, where 1 represents the lowest quality and 5 the highest.

3.5.2 Objective Evaluation

To quantitatively assess the similarity between the blended voice and its parent voices, we employed three objective metrics. These metrics provide insights into both the uniqueness of the blended voice and its fidelity to the parent voices:

Cosine Similarity (CS):

The cosine similarity measures the cosine of the angle between two vectors, providing a measure of directional similarity. It is defined as:

$$CS = \frac{1}{2} \sum_{i=1}^{2} \frac{e_{\text{blended}} \cdot e_{\text{parent}_i}}{||e_{\text{blended}}|| \, ||e_{\text{parent}_i}||}$$

Where e_blended is the blended voice embedding and e_parent_i are the parent voice embeddings.

Euclidean Distance (ED):

The Euclidean distance measures the straight-line distance between two points in Euclidean space. It is calculated as:

$$ED = \frac{1}{2} \sum_{i=1}^{2} ||e_{blended} - e_{parent_i}||_2$$

Mel-Cepstral Distortion (MCD):

MCD is a measure of the spectral difference between two speech signals, defined as:

$$MCD = \frac{1}{2} \sum_{i=1}^{2} \frac{10}{\ln 10} \sqrt{2 \sum_{d=1}^{D} (c_{d,parent_i} - c_{d,blended})^2}$$

Where c_d,parent_i and c_d,blended are the d-th mel-cepstral coefficients of the parent and blended voices, respectively, and D is the total number of coefficients.

These metrics serve dual purposes in our evaluation. A greater distance or dissimilarity (lower CS, higher ED and MCD) between the blended voice and its parents may indicate a more unique or creative blend, potentially representing a novel voice. Conversely, higher similarity (higher CS, lower ED and MCD) might suggest a more faithful blend that effectively captures characteristics of both parent voices.

By calculating these metrics between the blended voice and each parent voice, then averaging the results, we obtain a comprehensive measure of how the blended voice relates to its parent voices in the embedding space and in terms of spectral characteristics. This approach allows us to quantify the balance between creativity and fidelity in our voice blending techniques, providing valuable insights into the effectiveness of each method.

3.6 Experimental Procedure

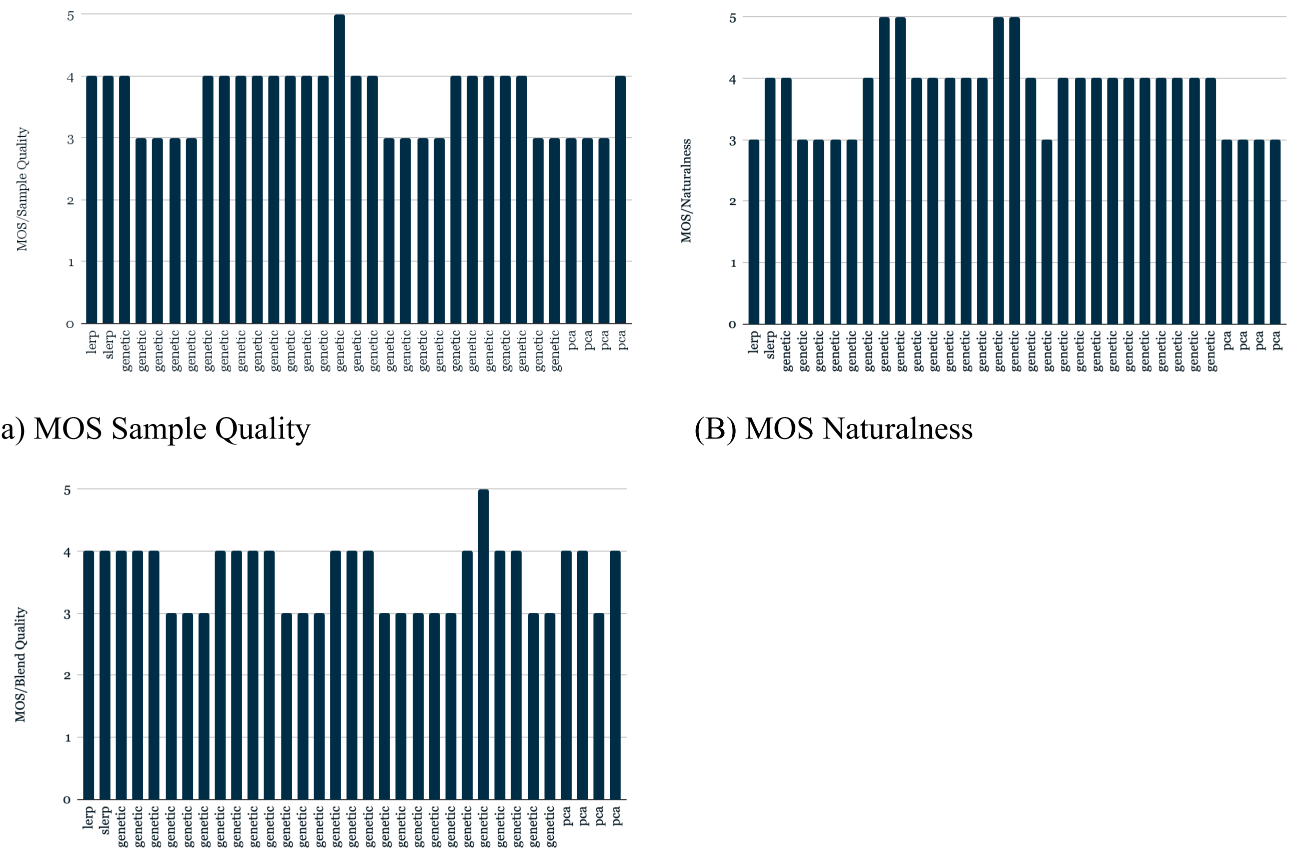
1. For each blending method and parameter combination, we generated a blended voice embedding.
2. We synthesized speech using the VITS model with the blended embedding.
3. Five participants evaluated the synthesized speech using the MOS criteria.
4. We calculated the objective similarity metrics between the blended embedding and each parent embedding.
5. Results were recorded and analyzed to compare the performance of different blending methods and parameter combinations.

This experimental design allows for a comprehensive evaluation of the four voice blending techniques, considering both perceptual quality and objective similarity to parent voices. The use of fixed parent embeddings and a consistent evaluation process ensures comparability across all methods and parameter combinations.

3.7 Ethical and Legal Considerations

This project acknowledges the ethical and legal concerns surrounding voice synthesis technology. By focusing on voice blending rather than direct voice cloning, the project aims to create unique voices that do not directly replicate any individual's voice, thus mitigating potential misuse and intellectual property issues. Additionally, the project will adhere to privacy guidelines and seek informed consent from all participants providing voice samples.

4. RESULTS AND DISCUSSION

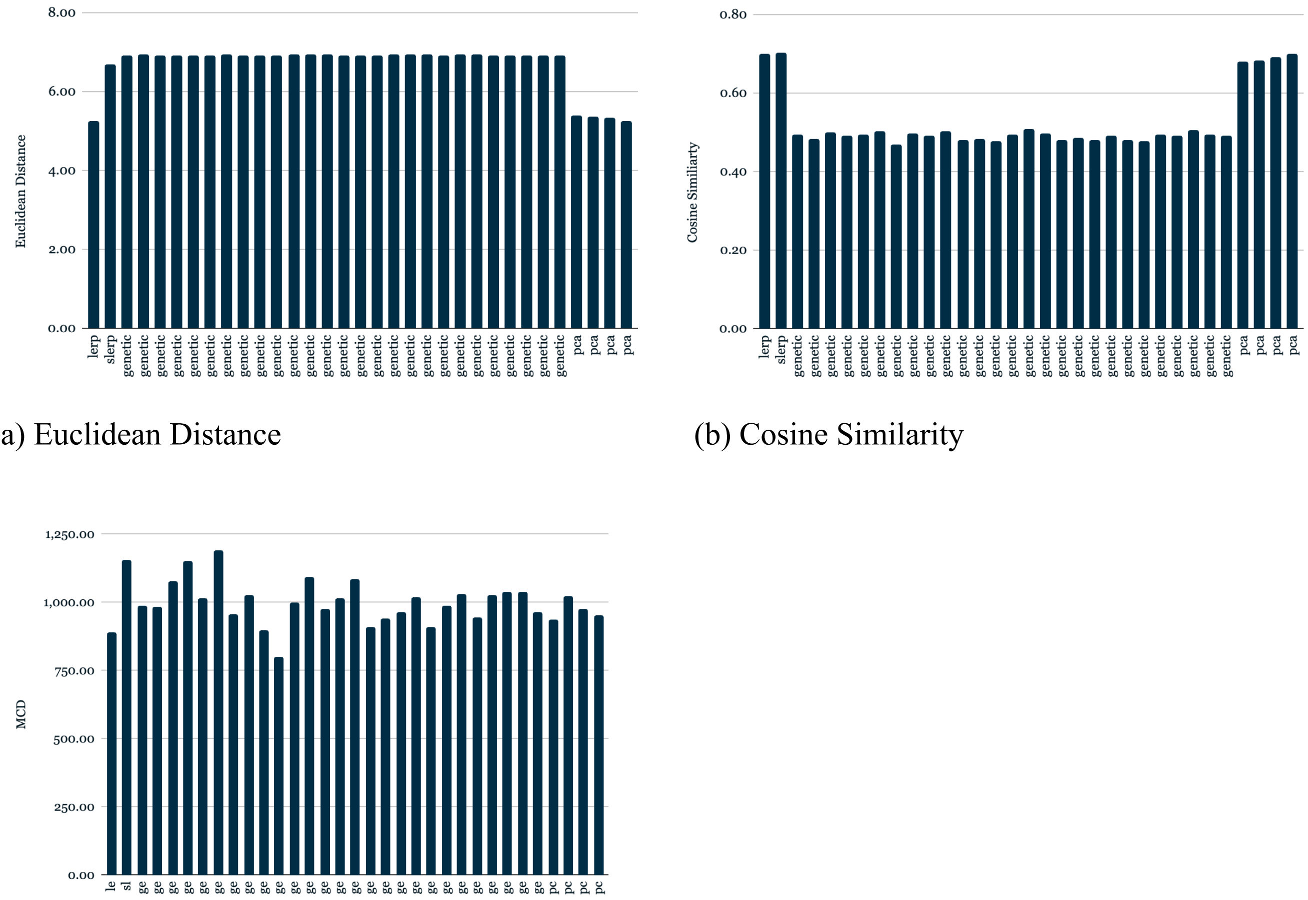


(a) MOS Sample Quality

(B) MOS Naturalness

(c) MOS Blend Quality

FIG. 1. Comparison of voice blending methods across multiple mean opinion score (MOS) metrics



(a) Euclidean Distance

(b) Cosine Similarity

(c) Mel-Cepstral Distortion (MCD)

FIG. 2. Objective similarity metrics for voice blending methods

4.1 Overview of Experimental Results

Our experiment compared four voice blending methods: Linear Interpolation (LERP), Spherical Linear Interpolation (SLERP), Genetic Algorithm-inspired method, and Principal Component Analysis (PCA). We evaluated these methods using Mean Opinion Scores (MOS) for sample quality, blend quality, and naturalness, as shown in FIG. 1, as well as objective similarity metrics including Cosine Similarity, Euclidean Distance, and Mel-Cepstral Distortion (MCD), illustrated in FIG. 2.

4.2 Comparative Analysis of Blending Methods

4.2.1 LERP and SLERP

As evident in FIG. 1(a) and FIG. 1(b), LERP and SLERP demonstrated comparable performance in terms of MOS scores, with LERP slightly outperforming SLERP in sample quality and naturalness (LERP: 4/4/3, SLERP: 4/4/4). FIG. 2(b) shows that both methods achieved consistent blending, as evidenced by their balanced average cosine similarities (LERP: 0.70, SLERP: 0.70). However, FIG. 2(a) and FIG. 2(c) indicates that SLERP resulted in higher Euclidean distance and MCD values, suggesting it produces more diverse outputs compared to LERP.

4.2.2 Genetic Algorithm-inspired Method

The genetic approach showed high variability in results, which is expected due to its stochastic nature. Notably, FIG. 1(b) shows that some configurations achieved high naturalness scores (e.g., test 9 with a score of 5), indicating the potential for creating unique yet natural-sounding voices. However, FIG. 2(b) reveals that cosine similarities often favored one input over the other, suggesting a tendency for the method to emphasize characteristics from one parent's voice.

4.2.3 PCA Method

PCA demonstrated the most consistent performance across different numbers of components (16, 32, 64, 128), as seen in FIG. 1. As the number of components increased, results became more similar to LERP. FIG. 2(a) shows that PCA generally resulted in lower Euclidean distances compared to other methods, indicating closer proximity to original voices. However, FIG. 1 indicates that MOS scores for PCA were slightly lower than for LERP and SLERP, suggesting a trade-off between consistency and perceived quality.

4.3 Impact of Parameter Variation

In the genetic algorithm approach, varying the crossover points, mutation rate, and mutation strength significantly affected the results. For instance, FIG. 1 shows that configurations with higher mutation rates (0.2) tended to produce more diverse voices but at the cost of lower quality scores. The number of crossover points didn't show a clear trend, suggesting that the optimal number may depend on the specific voice characteristics being blended.

For PCA, increasing the number of components generally led to results closer to LERP, with 128 components producing nearly identical cosine similarities to LERP, as shown in FIG. 2(b). This suggests that higher-dimensional PCA captures more of the original voice characteristics, potentially at the expense of generating novel voice features.

4.4 Quality-Novelty Trade-off

A key observation from our results is the apparent trade-off between voice quality/naturalness and novelty. Methods that produced more novel voices (as indicated by higher MCD in FIG. 2(c) and more varied cosine similarities in FIG. 2(b)) often scored lower on quality and naturalness MOS in FIG. 1(a) and FIG. 1(b). This highlights the challenge in creating voices that are both unique and high-quality.

4.5 Implications for Voice Blending Technology

Our results suggest that while more complex methods like the genetic algorithm can produce interesting and potentially more unique voices, simpler methods like LERP remain competitive in terms of perceived quality and naturalness. The challenge lies in finding the right balance between creating novel voices and maintaining high quality.

The consistency of PCA results across different configurations suggests it could be a reliable method for voice blending, particularly in applications where stability is crucial. The genetic algorithm's ability to occasionally produce highly natural voices indicates its potential for creating diverse voice portfolios, though it may require more fine-tuning and potentially multiple generation attempts to achieve desired results.

4.6 Limitations and Future Work

This study was limited to blending two parent voices and used a fixed blend ratio for LERP and SLERP. Future work could explore blending multiple voices and varying blend ratios. Additionally, while our MOS evaluations provide valuable insights, a larger-scale perceptual study could offer more robust conclusions about the perceived quality and uniqueness of the blended voices.

Further research could also investigate hybrid approaches, such as using PCA for initial blending followed by genetic algorithm fine-tuning, which might combine the stability of PCA with the creative potential of genetic algorithms.

In conclusion, our results demonstrate the viability of various voice blending techniques in creating new voices, each with its own strengths and trade-offs. These findings contribute to the ongoing development of ethical and creative voice synthesis technologies, offering alternatives to direct voice cloning that can address associated legal and ethical concerns.

5. CONCLUSION

This study has investigated four innovative voice blending techniques: Linear Interpolation (LERP), Spherical Linear Interpolation (SLERP), a genetic algorithm-inspired method, and Principal Component Analysis (PCA). By leveraging the VITS model and manipulating speaker embeddings, we have demonstrated the feasibility of creating unique, natural-sounding synthetic voices that combine characteristics from multiple speakers.

Our results reveal distinct trade-offs between novelty and naturalness across the different methods. LERP and SLERP showed consistent performance with balanced cosine similarities, suggesting their reliability in producing blended voices. However, SLERP's higher Euclidean distance and MCD values indicate its potential for generating more diverse outputs compared to LERP.

The genetic algorithm-inspired approach exhibited high variability in results, consistent with its stochastic nature. While some configurations achieved high naturalness scores, the method often favored characteristics from one parent voice over the other. This suggests that while the genetic approach has the potential for creating unique voices, it may require further refinement to consistently produce balanced blends.

PCA demonstrated the most consistent performance across different numbers of components, with higher dimensionality leading to results closer to LERP. The lower Euclidean distances in PCA results indicate closer proximity to original voices, but slightly lower MOS scores suggest a trade-off between consistency and perceived quality.

A key finding of this study is the apparent inverse relationship between voice quality/naturalness and novelty. Methods producing more novel voices often scored lower on quality and naturalness MOS, highlighting the challenge of creating voices that are both unique and high-quality.

These findings have significant implications for voice blending technology. While complex methods like the genetic algorithm can produce interesting and potentially more unique voices, simpler methods like LERP remain competitive in terms of perceived quality and naturalness. The consistency of PCA results suggests its potential as a reliable method for applications requiring stable voice blending outcomes.

Limitations of this study include the focus on blending only two parent voices and the use of fixed blend ratios for LERP and SLERP. Future research should explore blending multiple voices, varying blend ratios, and investigating hybrid approaches, such as combining PCA with genetic algorithm fine-tuning.

In conclusion, this research contributes to the ongoing development of ethical and creative voice synthesis technologies. By offering alternatives to direct voice cloning, these voice blending techniques address associated legal and ethical concerns while opening new avenues for applications in entertainment, assistive technologies, and personalized AI interfaces. As the field of voice synthesis continues to evolve, balancing innovation with ethical considerations will remain crucial in shaping the future of this technology.

6. REFERENCES

Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Moreno, I. L., & Wu, Y. (2019, January 2). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. arXiv.org. https://arxiv.org/abs/1806.04558

Kaneko, T., & Kameoka, H. (2017, December 20). Parallel-data-free voice conversion using cycle-consistent adversarial networks. arXiv.org. https://arxiv.org/abs/1711.11293

Kim, J., Kong, J., & Son, J. (2021, June 11). Conditional variational Autoencoder with adversarial learning for end-to-end text-to-speech. arXiv.org. https://arxiv.org/abs/2106.06103

Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016, September 19). WaveNet: A generative model for raw audio. arXiv.org. https://arxiv.org/abs/1609.03499

Qian, K., Zhang, Y., Chang, S., Cox, D., & Hasegawa-Johnson, M. (2021, March 13). Unsupervised speech decomposition via triple information bottleneck. arXiv.org-https://arxiv.org/abs/2004.11284

Shen, J., Pang, R.., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R.., Saurous, R. A., Agiomyrgiannakis, Y., & Wu, Y. (2018, February 16). Natural TTS synthesis by conditioning WaveNet on Mel Spectrogram predictions. arXiv.org. https://arxiv.org/abs/1712.05884

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y., & Saurous, R. A. (2018, March 23). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. arXiv.org. https://arxiv.org/abs/1803.09017