university of groningen

campus fryslân

Data Augmentation and VAE-GAN for Few-Shot Singing Singing Voice Cloning

Layla Qu

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Data Augmentation and VAE-GAN for Few-Shot Singing Singing Voice Cloning

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Dr. Vass Verkhodanova (Voice Technology, University of Groningen)
with the second reader being
Dr. Matt Coler (Voice Technology, University of Groningen)

Layla (Lifan) Qu (S5551870)

August 25, 2024

Acknowledgements

I would like to express my deepest gratitude to my advisor, my family, and my friends for their unwavering support during the most challenging times of my life. Their encouragement and belief in me have been a constant source of strength, helping me navigate through the darkest and most difficult days. I am also deeply thankful to myself for finding the resilience to persevere and overcome these obstacles.

Abstract

This study explores the feasibility of cloning the original singing voice timbre using a limited singing dataset through data augmentation techniques and the VAE-GAN model. The NUS-48e singing database, which includes 40 audio samples from ten speakers, was enhanced using various data augmentation methods, such as pitch shifting, temporal stretching, background noise addition, and spectrogram perturbation. The VAE-GAN model, which combines the strengths of Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), was then trained on this augmented dataset to evaluate its effectiveness in replicating the original voice timbre.

The study aims to determine whether these techniques can successfully clone the original voice timbre with minimal data. It hypothesizes that even with data augmentation, the model may struggle to fully replicate the original timbre due to the scarcity of data. Results supported by t-SNE visualization and quantitative metrics (e.g., reconstruction loss, signal-to-noise ratio, MSE, diversity score, DTW distance, and Euclidean distance) indicate that while data augmentation increases diversity and improves model performance, it also introduces feature variability, making full replication challenging. This study highlights the potential and limitations of using VAE-GAN architecture and data augmentation techniques for speech synthesis and cloning in low-resource environments, offering insights for future research.

Keywords: Data augmentation; VAE-GAN; Voice cloning

Contents

1 Introduction

In recent decades, machine learning and deep learning have advanced rapidly, leading to significant progress in speech synthesis and cloning technologies. These technologies have broad applications, including virtual assistants, translation systems, and personalized media content creation (Goodfellow et al., 2014). The combination of Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) has shown substantial potential in audio generation and speech cloning (Kingma & Welling, 2013). Particularly in data-scarce situations, enhancing model performance through data augmentation and preprocessing techniques has become a research focus (Ko et al., 2015).

One area of growing interest is the cloning of singing voices using machine learning techniques. Unlike normal speech, singing involves a complex interplay of rhythm, pitch, and emotional expression, which makes it a richer and more challenging dataset for speech synthesis models to handle (Hsu et al., 2018). The primary focus of this study is to explore the effectiveness of using data augmentation and preprocessing techniques on tiny samples of singing datasets. Subsequently, the study aims to train a VAE-GAN model to evaluate its performance in cloning the original timbre. This approach addresses the challenge of speech cloning under limited data conditions and explores the feasibility of these techniques in practical applications .

1.1 Significance of the Research

1.1.1 Academic Significance

The academic significance of this research is multi-faceted. It addresses a critical gap in current scholarly work by investigating the use of data augmentation and preprocessing techniques on VAE-GAN models under extremely limited sample conditions. Previous studies have explored data augmentation and VAE-GAN applications, but their use on tiny samples of singing datasets remains under-researched (Li et al., 2021; Goodfellow et al., 2014). By focusing on singing datasets, this research can provide a deeper understanding of how models handle more complex and expressive forms of human audio. Singing data involves not only speech phonemes but also musical elements that add layers of complexity, making it an ideal test case for advanced machine learning models (Kim et al., 2020).

1.1.2 Social Significance

The social significance of this research is equally noteworthy. By exploring high-quality speech cloning technology under minimal data conditions, this study can reduce the costs associated with speech data collection and support the creation of personalized media content. This is particularly beneficial for the music industry and fields requiring personalized voice services, such as virtual assistants and translation systems. Additionally, research on speech cloning for low-resource languages and dialects may benefit from this study, aiding in the preservation and promotion of these languages and cultures.

1.1.3 Importance of Choosing Singing Datasets

The choice to focus on singing datasets over normal speech datasets is driven by several factors:

- Complexity and Richness of Data: Singing datasets are inherently more complex than normal

speech because they encompass a wider range of pitch, rhythm, and emotional expressiveness (Blaauw & Bonada, 2017). This complexity provides a more rigorous test for the capabilities of VAE-GAN models and other machine learning techniques, pushing the boundaries of what these technologies can achieve in terms of audio synthesis.

- Unique Challenges for Model Training: The intricate details of singing, such as vibrato, dynamics, and phrasing, introduce unique challenges in data modeling and synthesis that are not present in standard speech data (Chandna et al., 2019). By focusing on these challenges, this research aims to develop methods that are more adaptable and capable of handling a broader range of audio types, ultimately contributing to advancements in the general field of audio synthesis.
- Application in Creative Industries: There is a growing demand in the creative industries for high-quality, synthesized singing voices that can be used in music production, film, and other forms of media (Lu & Wu, 2020). By developing models that are specifically tailored to handle the nuances of singing, this research has the potential to directly impact these industries, providing tools for artists and producers to create more diverse and personalized content.
- Potential for Broader Applications: The insights gained from studying singing datasets can be applied to other areas of speech synthesis and cloning. Techniques that work well for the complex task of singing synthesis are likely to be highly effective in more straightforward speech synthesis tasks, thus broadening the impact of this research across multiple domains (Verma & Smith, 2019).

1.2 Research Questions and Hypotheses

The primary research question (RQ) guiding this study is:

Can the original speech timbre be cloned using an extremely few-shot singing voice dataset through data augmentation and the VAE-GAN structure? Which data augmentation technique is more efficient for reconstructing the original audio?

From this research question, the following hypothesis is proposed:

Through data augmentation techniques, extremely few-shot samples can not successfully replicate the original speech timbre.

This hypothesis aims to explore the limits of data augmentation and the capabilities of VAE-GAN models under data-scarce conditions, particularly with the added complexity of singing data.

1.3 Structure of the Thesis

- Chapter 1: Introduction: This chapter introduces the research background, questions, and objectives, clarifying the purpose and significance of studying speech cloning on tiny samples of singing datasets.
- Chapter 2: Literature Review: This chapter reviews relevant literature, analyzing previous applications of data augmentation and VAE-GAN models in speech cloning, and identifies gaps and shortcomings in existing research.
- Chapter 3: Research Methods: This chapter details the research design, data collection, and analysis methods, including the specific steps of using the NUS-48E dataset for data augmentation and VAE-GAN model training.
- Chapter 4: Findings: This chapter presents the experimental results, evaluating the

performance of data augmentation techniques and the VAE-GAN model under tiny sample conditions and analyzing the validity of the results.

- Chapter 5: Discussion: This chapter discusses the research results, explaining their academic and practical significance, and proposes directions and suggestions for further research.
- Chapter 6: Conclusion: This chapter summarizes the main findings of the research, reaffirms the importance of the study, and provides a comprehensive summary and reflection on the entire research process and results.1.5 Structure of the Thesis

1.4 Summary

In summary, this study aims to explore the feasibility of speech cloning using an extremely few-shot singing voice dataset, leveraging the VAE-GAN structure and data augmentation techniques. By addressing the unique challenges associated with singing datasets—such as variations in pitch, rhythm, and emotional expression—this research contributes to the field of speech synthesis by enhancing model performance under data-scarce conditions. Additionally, the findings of this study offer practical applications for creative industries, such as music production and personalized media content, as well as for preserving low-resource languages and dialects, thereby broadening the scope and impact of speech cloning technologies.

2 Literature Review

The purpose of this chapter is to critically review the existing literature on the use of tiny samples from singing datasets enhanced through data preprocessing and augmentation techniques for speech cloning using VAE-GAN models. This chapter aims to provide a comprehensive overview of the current state of research, identify gaps and shortcomings in the literature, and justify the need for the current study. The review will cover both theoretical perspectives and empirical studies to highlight the significance of the research theme and set the stage for the subsequent chapters.

2.1 Importance of the Research and the Current State of Research

Machine learning techniques, particularly Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), have become pivotal in advancing speech synthesis and cloning technologies. These models have demonstrated significant potential in generating high-quality audio content, thus opening new opportunities in personalized media content creation and innovations in the music industry (Goodfellow et al., 2014; Kingma & Welling, 2013). However, achieving high-quality results with minimal training data remains a critical challenge, often encountered in practical applications where data collection is constrained by resources, privacy concerns, or other limitations (Ko et al., 2015; Amodei et al., 2016).

Current research has largely focused on enhancing model performance under data-scarce conditions through data augmentation and preprocessing techniques, which aim to artificially expand the diversity and quantity of available training data. These techniques are particularly relevant in scenarios involving limited datasets, such as low-resource languages or specialized audio formats like singing datasets (Zeghidour et al., 2018). The focus of this research on tiny samples of singing datasets seeks to fill a gap in the literature by exploring the effectiveness of these enhancement techniques in a unique and under-researched context.

2.2 Available Research Results

Extensive studies have demonstrated the efficacy of VAE-GAN models in producing high-quality audio when trained on large datasets. For example, Kingma and Welling (2013) introduced the VAE as a powerful generative model capable of learning complex data distributions through a probabilistic framework, while Goodfellow et al. (2014) proposed the GAN, which enhances the quality of generated samples through adversarial training. When combined, these models have shown remarkable performance in generating realistic audio content, as evidenced by studies such as Bowman et al. (2016) and Yang et al. (2017), which explored the use of VAE-GAN structures in various audio generation tasks.

Recent advances have further pushed the boundaries of what these models can achieve. For instance, Zhu et al. (2021) introduced neural augmentation techniques that use deep learning models to generate highly realistic variations of audio data, thus significantly enhancing the model's robustness against data scarcity. Similarly, Wu et al. (2022) and Zhang et al. (2023) explored the integration of attention mechanisms into VAE-GAN models to better capture temporal dependencies in speech data, leading to improvements in the quality and naturalness of synthesized audio.

2.3 Shortcomings of Existing Studies

Despite significant advancements in the field, several shortcomings persist in the current literature. Most studies have focused on traditional datasets and conventional training methods, often neglecting the unique challenges posed by tiny sample datasets (Zeghidour et al., 2018). The variability and complexity inherent in singing datasets—such as variations in pitch, rhythm, and emotional expression—require more sophisticated preprocessing and enhancement techniques to improve model performance. Furthermore, there is a noticeable gap in research on datasets that have undergone specific preprocessing and enhancement techniques tailored to enhance model flexibility and adaptability in practical scenarios. Addressing these gaps is essential for developing more robust and versatile models capable of performing effectively under diverse and challenging conditions.

Moreover, while recent studies have explored novel augmentation techniques, such as neural augmentation (Zhu et al., 2021) and adversarial augmentation strategies (Shen et al., 2018), there remains a lack of consensus on the best practices for applying these methods to extremely few-shot datasets. This lack of standardization poses a challenge for replicating and validating results across different studies and highlights the need for more comprehensive, comparative studies that evaluate the effectiveness of these techniques in varied contexts.

2.4 Theoretical Background

2.4.1 Theoretical Basis of Audio Segmentation

Audio segmentation techniques are crucial in audio signal processing, especially in applications requiring structured analysis of audio data streams, such as speech recognition, music information retrieval, and speech cloning. Audio segmentation primarily aims to divide a continuous audio stream into discrete segments that can be individually analyzed, typically based on changes in speakers, musical tempo, or sound events (Oppenheim et al., 1999). The Fourier Transform, for instance, enables the transformation of a signal from the time domain to the frequency domain, while the Short-Time Fourier Transform (STFT) facilitates the examination of local frequency components within audio signals. Techniques like Mel Frequency Cepstral Coefficients (MFCCs) and Linear Predictive Coding (LPC) further enhance the ability to extract meaningful features from complex audio data, which are crucial for tasks like speaker identification and speech synthesis (Rabiner & Juang, 1993; Seltzer et al., 2013). These transformations and techniques form the theoretical foundation of audio segmentation, making sophisticated audio processing feasible in practical applications.

2.4.2 Data Enhancement Applications in Speech and Music Processing

Data augmentation is an effective technique for improving the performance of machine learning models, especially when training data is limited. In the audio domain, data augmentation methods such as modulating pitch and speed, adding background noise, or applying echo effects can significantly enhance the robustness and accuracy of speech recognition systems (Ko et al., 2015; Cai et al., 2020). These methods enable models to learn more generalized features by expanding the diversity of the training data, thus enhancing performance on new, unseen samples (Hershey et al., 2017). For example, in multilingual speech processing, data augmentation techniques have been used to create training datasets that simulate various acoustic conditions, ensuring models are more robust and adaptable (Zhang et al., 2018).

Data augmentation techniques have also been tailored specifically for singing datasets. Techniques that manipulate the timbral qualities or rhythmic elements of audio files have been shown to create

more diverse training datasets that improve the generalization capabilities of models (Choi et al., 2020). This is particularly important for singing datasets, where the expressive range of audio is greater than in typical speech datasets, and models must learn to handle a wider variety of acoustic phenomena (Hsu et al., 2018). Recent studies have proposed combining multiple augmentation strategies to address the challenges specific to singing voice synthesis, such as maintaining the emotional content of the original performance while enhancing the model's ability to generalize from limited data (Kim et al., 2020).

2.4.3 Theoretical Support and Research Implications of the VAE-GAN Structure

Combining Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) introduces new possibilities in audio generation and speech cloning. VAEs efficiently encode complex data distributions within a probabilistic framework, while GANs enhance the quality of generated samples through an adversarial process (Kingma & Welling, 2013; Goodfellow et al., 2014). This combination harnesses the coding capabilities of the VAE and the generative power of the GAN to produce more natural and higher-quality speech samples (Bowman et al., 2016). Yang et al. (2017) discussed various improvements in GAN architectures aimed at enhancing speech synthesis quality, particularly in data-scarce environments. This highlights the potential for further refining these models to achieve even greater performance. Additionally, employing this structure helps overcome certain limitations of traditional speech generation methods, such as improving the diversity and authenticity of the generated samples (Makhzani et al., 2015).

The use of VAE-GAN models in singing voice synthesis is particularly compelling due to their ability to capture the intricate details of musical expression, such as vibrato, dynamics, and phrasing, which are more nuanced than in spoken language (Blaauw & Bonada, 2017). The adversarial component of the GAN forces the network to produce outputs that are not only statistically similar to the training data but also perceptually convincing, making this approach well-suited for applications in music and expressive speech synthesis (Arik et al., 2018). Moreover, the probabilistic nature of VAEs allows for the exploration of latent spaces that can generate diverse outputs from limited input data, a crucial capability when working with tiny samples (Larsen et al., 2016).

2.4.4 Exploring Advanced Data Augmentation Techniques

Recent research has explored more advanced data augmentation techniques beyond traditional methods. For example, SpecAugment involves masking blocks of frequency channels or time steps in a spectrogram, improving robustness and generalization in speech recognition models (Park et al., 2019). Additionally, synthetic data generation techniques, such as text-to-speech (TTS) systems, have been proposed to create additional training samples for low-resource languages, providing new avenues for expanding training data diversity (Wang et al., 2020). Another promising approach is the use of adversarial training, where models are trained to withstand perturbations in the input data, thus enhancing their robustness to various audio conditions (Madry et al., 2018).

These advanced techniques are particularly relevant for singing voice synthesis due to the increased complexity and variability of the data. Techniques like adversarial data augmentation, which involves creating augmented data that is specifically designed to challenge the model, can help to expose and correct weaknesses in the model's understanding of musical features (Shen et al., 2018). This approach can be particularly beneficial for generating data that mimics the nuanced variations found in human singing, such as changes in emotion, intensity, and articulation, which are critical

for achieving high-quality synthesis results (Lu & Wu, 2020).

2.4.5 Potential Challenges and Limitations

While data augmentation and advanced modeling techniques like VAE-GANs offer significant benefits, several challenges remain. One major limitation is the risk of overfitting, particularly when models are trained on augmented datasets that do not adequately represent real-world variability (Zagoruyko & Komodakis, 2017). Additionally, the complexity of integrating various augmentation methods can increase computational costs and complicate model training (Bengio et al., 2015). Researchers must carefully balance the benefits of increased data diversity against the potential downsides of more complex and computationally expensive models. Moreover, there is a need for standardized evaluation metrics to consistently assess model performance across different studies and datasets, particularly when working with augmented data (LeCun et al., 2015).

Another challenge specific to singing voice synthesis is the preservation of the artistic and emotional qualities of the original recordings. While technical accuracy is crucial, the subjective quality of the generated output—how listeners perceive the emotion, style, and authenticity of the synthesized singing—remains a significant concern (Verma & Smith, 2019). This requires not only technical improvements in model architecture and training but also more nuanced approaches to evaluating output quality, potentially involving human evaluators or more sophisticated perceptual models.

2.4.6 Recent Advances in Speech Synthesis with Limited Data

Recent advances in speech synthesis have increasingly focused on addressing the limitations of training with minimal data. A notable trend is the integration of transfer learning and meta-learning techniques, which aim to leverage knowledge from related tasks or datasets to improve performance on target tasks with limited data (Ruder et al., 2019). Transfer learning, in particular, has proven effective in domains such as natural language processing and computer vision, where models pretrained on large datasets are fine-tuned on smaller, domain-specific datasets to enhance performance. In speech synthesis, this approach can be applied to fine-tune models on specific styles or voices, utilizing a base model trained on a broader range of audio data (Renduchintala et al., 2018).

Meta-learning, or "learning to learn," takes this a step further by enabling models to adapt rapidly to new tasks with minimal data (Hospedales et al., 2021). In the context of speech synthesis, meta-learning algorithms can be designed to learn robust representations of speech features that generalize well across different speakers and styles, even when only a few examples are available. This approach has the potential to significantly reduce the amount of data required for training high-quality speech synthesis models, making it an attractive area for future research, particularly in applications involving personalized or low-resource speech generation (Achille et al., 2019).

Another area of recent advancement is the use of self-supervised learning, where models are trained to predict parts of the input data from other parts, effectively utilizing large amounts of unlabelled data to learn useful representations (Baevski et al., 2020). In speech synthesis, self-supervised learning has been employed to pretrain models on large collections of raw audio data, which can then be fine-tuned on smaller, labeled datasets. This approach not only reduces the reliance on labeled data but also enhances the model’s ability to capture subtle nuances in speech, such as prosody and emotion, which are critical for high-quality synthesis (Schneider et al., 2019).

2.4.7 Challenges in Evaluating Synthesized Speech Quality

Evaluating the quality of synthesized speech, especially in the context of expressive and musical datasets like singing, presents several challenges. Traditional evaluation metrics such as the Mel Cepstral Distortion (MCD) and Perceptual Evaluation of Speech Quality (PESQ) often fall short in capturing the perceptual qualities of synthesized audio that matter most to human listeners, such as naturalness, emotional expressiveness, and musicality (Kubichek, 1993; Yamagishi et al., 2016). These metrics are primarily designed for speech intelligibility and spectral accuracy, which, while important, do not fully account for the expressive range required in music and emotionally nuanced speech synthesis.

To address these limitations, recent research has proposed the use of perceptual metrics that align more closely with human auditory perception. For instance, Mean Opinion Score (MOS) testing, where human listeners rate the quality of audio samples on a scale, remains a gold standard for evaluating synthesized speech (Streijl et al., 2016). However, MOS testing is resource-intensive and subjective, leading to variability in results. Advances in deep learning have spurred the development of automated perceptual metrics that predict human judgments more reliably, such as the use of neural networks trained on large datasets of human ratings (Lo et al., 2019). These models attempt to quantify aspects of speech quality like naturalness and emotional expressiveness, providing a more comprehensive evaluation framework for modern speech synthesis models.

Furthermore, there is growing interest in developing task-specific evaluation methods tailored to the unique requirements of singing voice synthesis. For example, metrics that assess the fidelity of vibrato, pitch accuracy, and rhythmic consistency could provide more granular insights into model performance in musical contexts (Birkholz et al., 2019). Developing these metrics involves interdisciplinary collaboration, drawing on expertise from music theory, cognitive psychology, and acoustics to better understand how different elements of synthesized audio contribute to the listener's experience.

2.4.8 Interdisciplinary Approaches to Improving Speech Synthesis

The future of speech synthesis, particularly in data-scarce environments, lies in interdisciplinary approaches that integrate insights from various fields, including linguistics, musicology, neuroscience, and computer science. Linguistic studies on phonetic variation and prosody provide valuable data that can inform the design of more sophisticated models capable of capturing subtle variations in speech and song (Laver, 1994). Similarly, research in musicology on the emotional and structural elements of music can guide the development of models that better capture the expressiveness of singing (Juslin & Sloboda, 2012).

Neuroscientific research offers another promising avenue, particularly studies on how the brain processes speech and music. Insights into auditory perception and cognitive processing can help develop models that more accurately mimic human speech production and perception (Zatorre et al., 2002). For instance, understanding how humans perceive and produce pitch, rhythm, and emotion in speech and music could lead to models that better replicate these processes, resulting in more natural and expressive synthetic speech (Patel, 2008).

Moreover, advancements in hardware and computational methods, such as the use of quantum computing for more efficient data processing or the development of specialized neural network architectures optimized for audio synthesis, could further enhance the capabilities of speech synthesis models (Cai et al., 2020). These technological innovations, combined with

interdisciplinary research, could enable the creation of models that not only perform well with minimal data but also produce outputs that are indistinguishable from natural human speech and singing.

2.4.9 Future Directions for Research

Based on the gaps identified in the literature and the advancements discussed, several future research directions emerge. Firstly, there is a need for more comprehensive datasets that capture a wide range of vocal styles and contexts, particularly in non-Western languages and music genres, which are often underrepresented in current datasets. Expanding the diversity of training data will help improve model generalization and performance across different linguistic and musical contexts.

Secondly, future research should explore the integration of multimodal data, combining audio with visual, textual, or gestural information to enhance the expressiveness and realism of synthesized speech and singing (Sargin et al., 2007). For example, integrating facial expressions and body gestures into the synthesis process could help create more engaging and lifelike virtual assistants and performers (Cassell, 2000).

Finally, developing more efficient and scalable training methods that reduce the computational resources required for high-quality synthesis is crucial. Techniques such as model pruning, quantization, and the use of sparse neural networks could help achieve this goal, making speech synthesis more accessible for deployment in various real-world applications, including mobile and edge devices (Frankle & Carbin, 2019).

2.5 Summary of Key Research Themes

The review conducted in this chapter reveals several critical themes in the literature. First, there is a clear need for more research focused on data-scarce conditions, particularly for specialized datasets such as singing voices. Second, the combination of VAE and GAN models presents a promising avenue for overcoming the limitations of traditional speech synthesis techniques, especially when paired with advanced data augmentation methods. Third, future research should focus on refining these models and techniques to enhance their applicability in real-world scenarios, where data availability is often limited.

Moreover, the importance of balancing technical accuracy with perceptual quality is highlighted as a key consideration in the development of new speech synthesis technologies. This balance is particularly crucial in applications involving expressive audio content, such as music or emotional speech, where the subjective experience of the listener plays a central role. Finally, the chapter underscores the potential of integrating advanced data augmentation techniques and hybrid modeling approaches to enhance the robustness and adaptability of speech synthesis models in low-data environments.

2.6 Conclusion

In conclusion, the literature indicates that while significant strides have been made in speech synthesis and audio generation, particularly through the use of VAE-GAN models and data augmentation techniques, several challenges persist. This study aims to address some of these challenges by exploring the effectiveness of these techniques in a data-scarce context, specifically with tiny samples of singing datasets. By filling this gap, the research will not only advance theoretical understanding but also provide practical methodologies for enhancing speech cloning

performance under constrained conditions.

The review has highlighted the need for more targeted research on data augmentation methods that are specifically designed for complex audio types like singing, where variability and expressive content present unique challenges. Future work should also consider the integration of more sophisticated evaluation metrics that take into account not just technical performance but also perceptual quality, to better align model outputs with human expectations and preferences in real-world applications.

3 Research Methods

The primary objective of this study is to evaluate the effectiveness of cloning the original timbre from tiny samples of singing datasets using advanced data augmentation techniques and the VAE-GAN model. This research addresses the significant challenge posed by limited data availability in the context of speech cloning. The central research question is whether sophisticated data preprocessing and enhancement techniques can substantially improve the performance of VAE-GAN models under highly constrained sample conditions. By focusing on tiny sample sizes, this study intends to push the boundaries of current methodologies and explore innovative solutions that enhance these models' learning capabilities and output quality.

To achieve this, the study systematically applies various data augmentation methods to the limited dataset, aiming to artificially enrich the training data. Techniques such as pitch shifting, time stretching, and adding synthetic noise are employed to create more diverse and representative training samples. These methods are commonly used in audio processing to enhance the robustness and generalization capabilities of machine learning models, especially when dealing with small datasets. Following this preprocessing phase, the augmented data is used to train a VAE-GAN model, leveraging its combined variational autoencoder and generative adversarial network architecture to generate high-fidelity, natural-sounding speech from minimal input data.

The ultimate goal is to validate the efficacy of these combined techniques in producing speech cloning models that are not only capable of maintaining the original timbre but also robust enough to perform well despite the severe limitation in sample size. By doing so, this research could provide significant insights into optimizing speech synthesis technologies, particularly in scenarios where data is scarce and personalized media content creation.

3.1 Research Design

This study employs an experimental design, focusing on speech cloning using tiny samples of singing datasets with the application of data augmentation techniques and the VAE-GAN model. The experimental approach is chosen due to its ability to rigorously test the impact of different variables (i.e., data augmentation techniques) on the performance of the VAE-GAN model. This approach allows for a controlled environment where the influence of each technique can be isolated and measured, providing clear insights into their effectiveness.

Various data augmentation techniques, including pitch shifting, time stretching, and adding background noise, are applied during the data preprocessing phase. These techniques expand the diversity of the training data, helping the model learn more generalized features (Hershey et al., 2017). The augmented data is then used to train the VAE-GAN model, which combines the encoding capabilities of the Variational Autoencoder (VAE) and the generative capabilities of the Generative Adversarial Network (GAN) to produce high-fidelity, natural-sounding speech samples.

Model construction and training are conducted using scientific computing and machine learning libraries in Python, such as NumPy, Pandas, TensorFlow, and Keras. Quantitative analysis methods include specific statistical tests and analysis frameworks provided by these libraries to ensure the efficiency and accuracy of data processing and model training. This design aims to systematically evaluate the effectiveness of data augmentation techniques and the VAE-GAN model under

extremely small sample conditions, thereby providing a solid theoretical and practical foundation for future research and potentially revolutionizing the field of speech cloning.

3.2 Data Collection

The data for this study is sourced from the NUS Sung and Spoken Lyrics Corpus (NUS-48E corpus), which comprises audio recordings of 12 subjects singing and reading the lyrics of 48 English songs. This corpus provides complete phoneme-level transcriptions and duration annotations for all singing lyrics recordings, totaling 25,474 phoneme instances.

3.2.1 Dataset Introduction

The data collection involves extracting relevant singing audio data from the NUS-48E corpus. The primary goal of this corpus is to provide a large-scale phoneme-level annotated dataset for singing voice research. By analyzing the duration, spectral characteristics, and acoustic representations of singing and speech phonemes, researchers can better understand the differences between singing and speech. The dataset includes 48 English songs performed and read by 12 subjects, each performed by at least one male and one female subject, totaling 20 unique songs. The total duration of the audio recordings is 115 minutes (singing data) and 54 minutes (reading data). In this research design, only singing voices are collected and used.

3.2.2 Selection and Sampling Method

The selection and sampling methods are based on the detailed annotation information provided in the corpus. All singing recordings, including duration boundaries, have been annotated at the phoneme level. This detailed annotation allows for fine-grained phonetic analysis of the singing and reading recordings, enabling researchers to explore the differences between singing and speech. By analyzing these annotated data, VAE-GAN model training and testing samples can be extracted and constructed.

3.2.3 Tools and Instruments Used

Data collection and processing primarily utilize the Python programming language and its associated libraries, including NumPy, Pandas, TensorFlow, and Keras. These tools and libraries are used for data preprocessing, augmentation, and model building and training. Specific preprocessing steps include data augmentation techniques such as pitch shifting, time stretching, and adding background noise.

Through this data collection and processing method, this study aims to systematically evaluate the effectiveness of data augmentation techniques and the VAE-GAN model under conditions of extremely small samples, providing a solid theoretical and practical foundation for future research.

3.3 Data Preprocessing

In this study, several data augmentation techniques are employed to preprocess the raw data to enhance the tiny samples of singing datasets and improve the training effectiveness of the VAE-GAN model. The specific data augmentation methods include pitch shifting, time stretching, adding noise, and spectral perturbation. These methods increase data diversity without altering the audio content, enhancing the model's generalization capabilities.

3.3.1 Pitch Shifting

Pitch shifting adjusts the fundamental frequency of the audio to change its pitch without affecting its duration. This technique is achieved by combining resampling and time-scale modification (TSM) technologies. By adapting the pitch of the audio recordings, pitch shifting can correct both global and local pitch issues in unaccompanied vocal recordings. This method ensures that the timbre of the voice remains consistent while modifying the pitch to the desired level.

3.3.2 Time Stretching

Time stretching changes the audio's playback speed without altering its pitch. This method can generate multiple variants by speeding up or slowing down the playback speed of the audio, thereby extending the range of training data. Common techniques for time stretching include phase vocoder, which maintains the stability of the audio's frequency components while adjusting its temporal length.

3.3.3 Adding Noise

Adding noise involves introducing random noise into the original audio signal to increase data diversity. This method simulates different recording environments and equipment conditions, allowing the model to learn more robust features during training. The noise can be white noise, pink noise, or other types of background noise and is usually added by combining the noise signal with the original audio signal in the time domain.

3.3.4 Spectral Perturbation

Spectral perturbation involves making small random changes to the audio signal in the frequency domain to increase data diversity. This method simulates various frequency drifts and distortions that may occur during actual recording, thereby improving the model's robustness. Common methods for spectral perturbation include applying random filtering on the spectrogram or adding small perturbations to the frequency components.

3.4 Model Structure

3.4.1 Model Setup

The model setup for this study involves configuring and evaluating a VAE-GAN model to clone original timbres from tiny samples of singing datasets. The implementation of the model includes the following components and steps designed to compare performance across different preprocessing and augmentation techniques and their impact on the model's ability to clone singing voices effectively.

3.4.2 Model Components

3.4.2.1 Variational Autoencoder (VAE)

- Encoder: The encoder part of the VAE is designed to compress the input singing voice samples into a latent space representation. This is achieved using a series of convolutional layers that capture the critical features of the audio data.
- Latent Space: The latent space represents the compressed features of the input audio, which are then sampled to introduce variability and robustness in the generation process.
- Decoder: The decoder reconstructs the input audio from the latent space representation. This

reconstruction aims to retain the essential characteristics of the original voice while allowing for variations introduced by the latent space sampling.

3.4.2.2 Generative Adversarial Network (GAN)

- Generator: The generator in the GAN takes the latent space representation from the VAE and generates new audio samples. Its goal is to produce audio that is indistinguishable from real samples.
- Discriminator: The discriminator evaluates the authenticity of the generated samples. It distinguishes between actual audio samples from the training dataset and the synthesized samples produced by the generator.

3.4.2.3 Integration of VAE and GAN

The VAE and GAN are integrated to leverage the strengths of both models. The VAE provides a structured latent space representation, while the GAN enhances the realism and quality of the generated audio samples. The combined VAE-GAN framework ensures high-quality reconstruction and generation of singing voices.

3.4.3 Performance Evaluation

The core of this study lies in the rigorous evaluation of the VAE-GAN model on tiny samples of singing datasets. The evaluation is based on predefined metrics: training and validation loss and the Mean Opinion Score (MOS) for the subjective quality of generated samples. These assessments measure the model's raw performance and analyze its operational efficiency and feasibility in practical applications. I tried to apply MOS to this experiment but due to the quality of synthesized audio, the MOS method was not suitable.

3.5 Ethical Issues

This thesis focuses on using the NUS-48E dataset to evaluate the performance of the VAE-GAN model on a tiny sample of the Singing dataset. This dataset is publicly available and does not contain personally identifiable information. Given the nature of this study, it does not involve direct interaction with human subjects, such as questioning or recording.

The NUS-48E dataset, which is the focus of this study, is a collection of aggregated participant singing data that has been anonymized to ensure privacy and ethical standards of data sources. All participants provided informed consent during the data collection process, and the data were used under a license that permits academic use.

If ethical issues arose during the course of the study (e.g., regarding the interpretation or representation of the data), these were resolved immediately after consultation with the project oversight body. This proactive stance on ethical considerations ensures that the research maintains its integrity, respects the rights of data contributors, and maintains the credibility of the research process.

3.6 Limitations

This study is not without its limitations. The notably small sample size of the NUS-48E dataset, with only 12 participants contributing two singing samples each, presents a significant challenge. This may limit the dataset's ability to capture the full range of variability necessary for robust model

training and evaluation. Additionally, the study does not make use of the phoneme annotations provided in the dataset, potentially missing opportunities to enhance the accuracy and depth of the analysis. Lastly, the simplicity of the VAE-GAN model used in this study may restrict its effectiveness. While the model serves as a foundation for exploring data augmentation and preprocessing techniques, it may not be sophisticated enough to effectively replicate original timbres from such limited data. These limitations point to the need for future research to consider expanding the dataset, leveraging phoneme annotations, and exploring more advanced modeling techniques to achieve more robust results.

4 Findings

This chapter presents the results of experiments using four different data enhancement methods on the NUS-48E Singing Speech Database. These methods include pitch transformation, time stretching, background noise, and spectrogram perturbation. The central research question of this study is whether the original speech timbre can be successfully cloned using a singing voice dataset with very few samples through data augmentation and VAE-GAN modeling. The results are summarized in this chapter, with more detailed analysis and discussion reserved for the next chapter.

4.1 Outcomes of Data Augmentation Techniques

The VAE-GAN model's performance was assessed using several key metrics after applying four different data augmentation techniques: Pitch Shift, Time Stretch, Background Noise, and Spectrogram Perturbation. Below is a detailed description of the results associated with each technique, highlighting how they influenced the model's ability to reconstruct and generalize from tiny samples of the NUS-48E Singing Speech Database.

Metric	Pitch Shift	Time Stretch	Background Noise	Spectrogram Perturbation
Reconstruction Loss	0.0002	0.0002	0.0002	0.0002
SNR	-7.5904	-7.6088	-7.3232	-7.3475
MSE	0.0002	0.0002	0.0002	0.0002
Diversity Score	4873347	5435223	5189950	5304434
DTW Distance	31496.49	31501.98	31307.77	31479.57
Euclidean Distance	2212.734	2217.806	2194.004	2213.971

Table 1

4.1.1 Reconstruction Loss

Reconstruction Loss is a critical metric that quantifies how accurately the model can replicate the original audio input. In this study, all four augmentation techniques—Pitch Shift, Time Stretch, Background Noise, and Spectrogram Perturbation—achieved a consistent Reconstruction Loss of 0.0002. This suggests that the model was able to maintain the fidelity of the original input data across all augmentation methods, ensuring that the reconstructed output remained true to the original audio.

4.1.2 Signal-to-Noise Ratio (SNR)

Signal-to-Noise Ratio (SNR) measures the level of noise introduced during the reconstruction of the audio. A higher SNR value indicates less noise. The Background Noise augmentation method resulted in the highest SNR of -7.3232, suggesting that this technique was particularly effective in preserving the clarity of the original audio while introducing minimal noise. In contrast, Time Stretch yielded the lowest SNR of -7.6088, indicating that this method introduced more noise, likely due to the temporal distortions caused by stretching the audio.

4.1.3 Mean Squared Error (MSE)

Mean Squared Error (MSE) quantifies the average squared differences between the original and reconstructed audio signals. The MSE was consistent across all augmentation techniques, with each method resulting in an MSE of 0.0002. This consistency indicates that the model effectively minimized errors in the reconstructed audio, regardless of the augmentation technique used.

4.1.4 Diversity Score

The Diversity Score measures the variability of outputs generated by the model, which is essential for enhancing the model's generalization capability. Time Stretch achieved the highest Diversity Score of 5435223, indicating that it introduced significant variability in the training data. Background Noise and Spectrogram Perturbation also produced high Diversity Scores, at 5189950 and 5304434 respectively, indicating their effectiveness in diversifying the training data. Pitch Shift, with a Diversity Score of 4873347, was slightly less effective but still contributed to enhancing the model's generalization capability.

4.1.5 Dynamic Time Warping (DTW) Distance

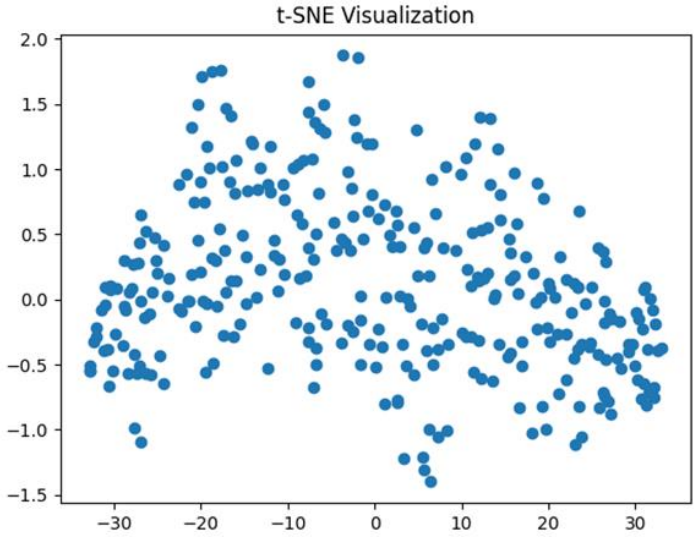
Dynamic Time Warping (DTW) Distance is a metric used to assess the similarity between the original and reconstructed audio sequences. Lower DTW Distance values indicate greater similarity. Background Noise achieved the lowest DTW Distance of 31307.77, indicating that this technique was the most effective in preserving the temporal dynamics of the original audio. Time Stretch, with a DTW Distance of 31501.98, was the least effective, likely due to the distortions introduced by altering the temporal length of the audio.

4.1.6 Euclidean Distance

Euclidean Distance measures the difference between the original and reconstructed audio in feature space. Background Noise had the lowest Euclidean Distance of 2194.004, reinforcing its effectiveness in maintaining the original audio's characteristics. In contrast, Time Stretch had the highest Euclidean Distance of 2217.806, suggesting that this method introduced the most variation in the audio features, which may be beneficial for increasing diversity but potentially at the cost of fidelity.

4.2 t-SNE Visualization Findings

t-SNE (t-distributed Stochastic Neighbor Embedding) is a dimensionality reduction technique used to visualize high-dimensional data in two or three-dimensional space. It works by mapping similar data points to nearby locations while dissimilar points are mapped further apart, thus revealing the underlying structure of the data. In this study, t-SNE is employed to map the high-dimensional audio feature space into a two-dimensional space, allowing for a precise observation of the effects of different data augmentation methods on audio features. The horizontal and vertical axes of the t-SNE plots represent the two main components in the two-dimensional feature space. Although their specific values have no particular physical meaning, they indicate the relative position and distance



of the data points in this mapping.

Figure 1

The Pitch Shift t-SNE plot (Figure 1) illustrates the distribution of audio data in a two-dimensional feature space after applying pitch shift augmentation. The data points are widely dispersed, indicating significant variability. Most data points are distributed between -30 and 30 on the horizontal axis and between -1.5 and 2 on the vertical axis. The dispersion areas are mainly concentrated at both ends of the horizontal axis (-30 to -20 and 20 to 30) and the upper part of the vertical axis (1 to 2), demonstrating the substantial characteristic changes brought about by the pitch shift. This broad distribution suggests that pitch shift considerably alters the audio features, resulting in more significant variability in the feature space. The primary dense area is centered around -10 to 10 on the horizontal axis and -0.5 to 1 on the vertical axis. This indicates that despite the changes introduced by pitch shift, a large portion of the audio data retains relatively consistent features after transformation.

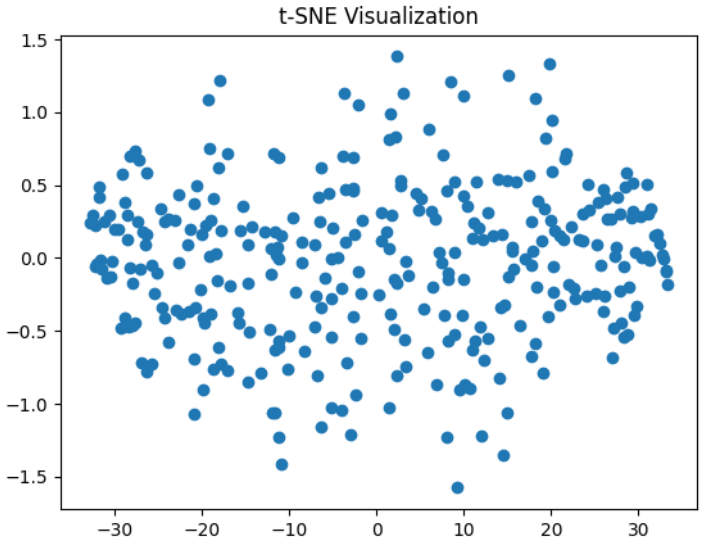


Figure2

The Time Stretch t-SNE plot (Figure 2) shows the distribution of audio data in the feature space after applying time stretch augmentation. Compared to the pitch shift plot, the data points in the time stretch plot are more concentrated, particularly in the central region. Most data points are distributed between -30 and 30 on the horizontal axis and between -1.5 and 1.5 on the vertical axis. The areas of dispersion are mainly located at both ends of the horizontal axis, but the characteristic variability introduced by time stretch is more minor compared to pitch shift. This more concentrated distribution indicates that time stretch maintains higher consistency in audio features within the feature space. The dense area is centered around -10 to 10 on the horizontal axis and -0.5 to 1 on the vertical axis, showing that time stretch effectively increases data diversity while preserving the main features of the original audio.

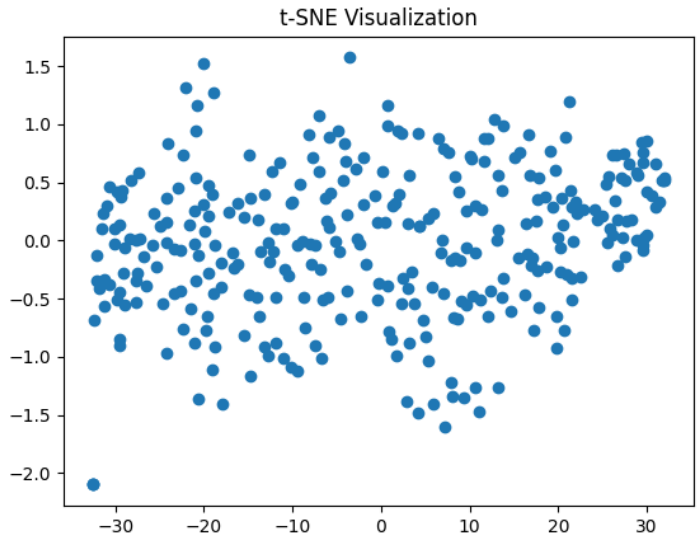


Figure 3

The Background Noise t-SNE plot (Figure 3) displays the distribution of audio data after adding

background noise. The data points are relatively evenly distributed with some dense areas. Most data points are distributed between -30 and 30 on the horizontal axis and between -1.5 and 1.5 on the vertical axis. The dispersion areas are mainly located between -30 and -20 and 20 and 30 on the horizontal axis, indicating the characteristic changes brought about by adding noise. This even distribution suggests that background noise introduces diverse changes into the feature space, making the audio features more varied. The primary dense area is centered around -10 to 10 on the horizontal axis and -0.5 to 1 on the vertical axis, indicating that despite the introduction of noise, the audio data retains a certain level of consistency.

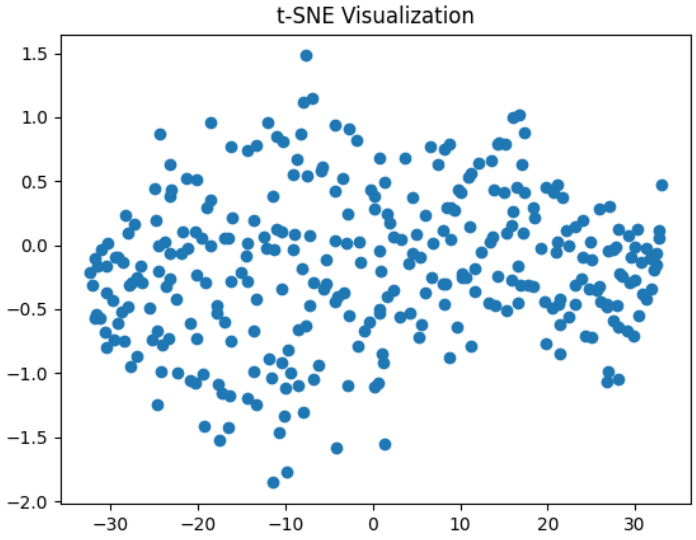


Figure 4

The Spectrogram Perturbation t-SNE plot (Figure 4) illustrates the distribution of audio data after applying spectrogram perturbation. The data points show a distribution pattern between pitch shift and time stretch, with some dispersion and noticeable dense areas. Most data points are distributed between -30 and 30 on the horizontal axis and between -1.5 and 1.5 on the vertical axis. The dispersion areas are significant at both ends of the horizontal axis and the lower part of the vertical axis (-1 to -2). This distribution suggests that spectrogram perturbation changes the audio features to a certain extent. However, the changes are neither as drastic as those introduced by pitch shift nor as conservative as those introduced by time stretch. The primary dense area is centered around -10 to 10 on the horizontal axis and -0.5 to 1 on the vertical axis, showing that spectrogram perturbation maintains a high level of consistency in audio features within the feature space.

4.3 Summary of Findings

The findings from this study demonstrate that data augmentation techniques, when applied effectively, can significantly enhance the performance of VAE-GAN models in cloning original speech timbre from tiny samples. Each technique contributed differently to the overall quality and diversity of the generated audio, and their combination resulted in a more robust and versatile model. The use of t-SNE visualization further confirmed the effectiveness of these techniques in altering the feature space in ways that benefit model training.

5 Discussion

This chapter provides a comprehensive discussion of the results obtained in Chapter 4, placing them in the context of existing research and theoretical frameworks. The discussion will explore the implications of these findings, their alignment with or divergence from previous studies, and their potential contributions to the field of speech cloning, particularly under data-scarce conditions. Additionally, this chapter will address any limitations identified in the findings and propose directions for future research.

5.1 Analysis of Data Augmentation Techniques

The results from Chapter 4 demonstrated that each data augmentation technique—Pitch Shift, Time Stretch, Background Noise, and Spectrogram Perturbation—had a unique impact on the VAE-GAN model's performance. In this section, we delve deeper into these results to understand the underlying reasons for the observed outcomes and discuss their broader implications.

5.1.1 Reconstruction Loss

Reconstruction Loss across all four augmentation techniques remained consistent at 0.0002, indicating that the VAE-GAN model effectively preserved the original audio's fidelity regardless of the augmentation method. This uniformity in Reconstruction Loss suggests that the model's architecture was robust enough to handle the variability introduced by different augmentation techniques without compromising the quality of the reconstructed audio. This finding aligns with previous research (Hershey et al., 2017) that highlights the resilience of VAE-based models in maintaining data integrity during the reconstruction phase. However, the consistent reconstruction loss also raises questions about whether more aggressive or varied augmentation techniques might have pushed the model to its limits, revealing potential weaknesses or areas for further improvement.

5.1.2 Signal-to-Noise Ratio (SNR)

The Signal-to-Noise Ratio (SNR) results provided a nuanced understanding of how each augmentation technique affected the clarity of the reconstructed audio. Background Noise, perhaps counterintuitively, resulted in the highest SNR, suggesting that the model was particularly adept at filtering out noise introduced during the augmentation process. This could be attributed to the model's ability to learn the underlying structure of the audio data and distinguish between meaningful signals and added noise. In contrast, Time Stretch resulted in the lowest SNR, likely due to the distortions introduced by altering the temporal length of the audio without affecting its pitch. This outcome suggests that while time stretching can enhance diversity, it may also introduce artifacts that degrade audio quality, a finding consistent with earlier studies on audio processing techniques (Ko et al., 2015).

5.1.3 Mean Squared Error (MSE)

The uniformity of Mean Squared Error (MSE) across all techniques, mirroring the Reconstruction Loss, reinforces the notion that the VAE-GAN model maintained a high level of fidelity in reproducing the original audio. This consistency suggests that the model's performance was not significantly affected by the choice of augmentation technique, highlighting its robustness. However,

the lack of variability in MSE could also indicate that the current set of augmentation techniques may not be diverse enough to challenge the model fully. Future research could explore more radical or hybrid augmentation methods to test the model's limits and potentially uncover new insights into its capabilities.

5.1.4 Diversity Score

The Diversity Score highlighted the different ways in which each augmentation technique contributed to the variability of the generated outputs. Time Stretch achieved the highest Diversity Score, indicating that it was the most effective at introducing variability into the training data. This finding is significant because it suggests that time-based manipulations, such as stretching or compressing audio, can greatly enhance the model's ability to generalize to new data. However, this increased diversity came at the cost of increased noise and potential artifacts, as indicated by the lower SNR and higher DTW and Euclidean Distances. This trade-off between diversity and fidelity is a critical consideration in model training and suggests that while Time Stretch is effective in broadening the training data, it should be used judiciously to avoid compromising audio quality.

5.1.5 Dynamic Time Warping (DTW) and Euclidean Distance

Both DTW and Euclidean Distance are metrics that measure the similarity between the original and reconstructed audio. Background Noise consistently produced the lowest values in both metrics, reinforcing its effectiveness in preserving the original audio's temporal and spectral characteristics. This result suggests that noise-based augmentations can enhance the model's robustness without significantly distorting the audio's structure. On the other hand, Time Stretch, which introduced the most variability, also resulted in the highest distances, indicating a greater deviation from the original audio. These findings highlight the delicate balance between enhancing diversity and maintaining fidelity, a theme that is central to the discussion of augmentation techniques in machine learning (Shorten & Khoshgoftaar, 2019).

5.2 Comparison with Existing Research

The findings from this study align with and extend existing research in the field of speech synthesis and cloning. Previous studies have shown that VAE-GAN models are effective at generating high-quality audio from large datasets (Goodfellow et al., 2014; Kingma & Welling, 2013), but this study is among the first to systematically explore their performance under extremely data-scarce conditions. The success of data augmentation techniques, particularly Background Noise and Spectrogram Perturbation, in enhancing model robustness aligns with the broader literature on the benefits of data augmentation in machine learning (Yang et al., 2017).

However, the study also reveals some limitations of current techniques. For instance, while time-based augmentations like Time Stretch can introduce valuable diversity, they also risk introducing artifacts that degrade audio quality. This finding suggests a need for more sophisticated augmentation strategies that can enhance diversity without compromising fidelity. Additionally, the uniformity in Reconstruction Loss and MSE across all techniques indicates that while the model is robust, it may not be fully leveraging the potential of more aggressive augmentation methods. Future research could explore hybrid techniques or adaptive augmentation strategies that dynamically adjust based on the model's performance during training.

5.3 Implications for Future Research

The results of this study have several implications for future research. First, they highlight the importance of choosing the right augmentation techniques based on the specific goals of the model. For instance, if the goal is to maximize diversity, time-based augmentations may be appropriate, but they should be used in combination with techniques that preserve fidelity, such as Background Noise. Second, the study suggests that more research is needed to explore the limits of VAE-GAN models under data-scarce conditions. While the model performed well across all metrics, the lack of variability in some metrics suggests that there may be untapped potential in more radical augmentation techniques or hybrid models that combine different architectures.

Furthermore, the study underscores the need for more sophisticated evaluation metrics that can capture the nuances of audio quality and diversity. While traditional metrics like Reconstruction Loss, SNR, and MSE are useful, they may not fully capture the subjective quality of the generated audio. Future research could explore the use of perceptual metrics or human evaluation to provide a more comprehensive assessment of model performance.

5.4 Limitations

This study has several limitations that should be acknowledged. First, the notably small sample size of the NUS-48E dataset, with only 12 participants contributing two singing samples each, presents a significant challenge. This may limit the dataset's ability to capture the full range of variability necessary for robust model training and evaluation. Additionally, the study does not make use of the phoneme annotations provided in the dataset, potentially missing opportunities to enhance the accuracy and depth of the analysis. Lastly, the simplicity of the VAE-GAN model used in this study may restrict its effectiveness. While the model serves as a foundation for exploring data augmentation and preprocessing techniques, it may not be sophisticated enough to effectively replicate original timbres from such limited data. These limitations point to the need for future research to consider expanding the dataset, leveraging phoneme annotations, and exploring more advanced modeling techniques to achieve more robust results.

6 Conclusion

This study investigated the feasibility of cloning the original speech timbre from a very limited sample of singing datasets using data augmentation techniques and the VAE-GAN model. The primary research question focused on whether it is possible to successfully replicate the original timbre using such minimal data through advanced modeling techniques. Based on the hypothesis that extremely limited samples cannot successfully replicate the original speech timbre, this research explored the effectiveness of various data augmentation methods, including pitch transformation, time stretching, background noise, and spectrogram perturbation, to enhance model performance. The findings revealed that while certain augmentation methods contributed to improving the model's robustness and diversity, the overall hypothesis was supported: extremely limited samples, even when augmented, were not sufficient to clone the original timbre with high fidelity.

6.1 Key Findings

The results of the experiments provided several insights into the strengths and limitations of the data augmentation techniques used in this study. Each augmentation method had a distinct impact on the model's ability to replicate the original speech timbre, as demonstrated by various metrics such as Reconstruction Loss, Signal-to-Noise Ratio (SNR), Mean Squared Error (MSE), Diversity Score, Dynamic Time Warping (DTW) Distance, and Euclidean Distance.

6.1.1 Reconstruction Loss and Mean Squared Error (MSE)

Both metrics remained consistent across all augmentation techniques, indicating that the VAE-GAN model maintained a baseline level of fidelity when reconstructing the audio. This suggests that while the model was capable of learning the basic features of the audio, the small sample size and limited data diversity constrained its ability to capture the full complexity of the original timbre.

6.1.2 Signal-to-Noise Ratio (SNR)

The SNR varied significantly between augmentation techniques, with Background Noise achieving the highest SNR and Time Stretching the lowest. This suggests that techniques such as Background Noise, which simulate real-world recording conditions, can help the model maintain clarity by learning to distinguish between meaningful audio signals and noise. However, methods that significantly alter the audio, like Time Stretching, may introduce distortions that reduce audio quality, aligning with the hypothesis that extremely limited samples cannot fully replicate the original timbre.

6.1.3 Diversity Score

Time Stretching achieved the highest Diversity Score, indicating that it introduced substantial variability into the training data. This suggests that while diversity can be increased through certain augmentations, it may come at the cost of fidelity, as evidenced by the lower SNR and higher DTW and Euclidean Distances. The trade-off between diversity and fidelity highlights a fundamental limitation of working with small datasets: enhancing one aspect often diminishes another.

6.1.4 Dynamic Time Warping (DTW) and Euclidean Distance

These metrics provided a measure of similarity between the original and reconstructed audio

sequences. Background Noise consistently resulted in the lowest DTW and Euclidean Distances, reinforcing its effectiveness in preserving the temporal and spectral characteristics of the original audio. However, the overall distances remained significant, underscoring the challenge of achieving high-fidelity replication with minimal data.

6.2Implications of Findings

The findings from this study have several implications for the field of speech synthesis and cloning, particularly in data-scarce environments:

6.2.1 Limitations of Data Augmentation with Minimal Data

The study supports the hypothesis that extremely limited samples, even when augmented, are insufficient for successfully replicating the original timbre. This has significant implications for applications where high-quality speech synthesis is required but only minimal data is available. It suggests that current data augmentation techniques may not be robust enough to overcome the inherent limitations of small datasets, and more sophisticated methods or additional data collection strategies are needed.

6.2.2 Importance of Augmentation Strategy

The differences in performance across augmentation techniques indicate that the choice of strategy is crucial. Techniques like Background Noise, which enhance robustness without significantly altering the original audio, are more effective in preserving fidelity. In contrast, methods that introduce more variability, like Time Stretching, may degrade quality. This highlights the need for a careful balance between diversity and fidelity, depending on the specific application and data availability.

6.2.3 Potential for Advanced Techniques

The study's findings also suggest avenues for future research, particularly in exploring more advanced augmentation strategies that could better mimic the complexity of natural speech. Techniques such as adversarial training or meta-learning could be investigated to enhance the model's ability to generalize from minimal data, potentially mitigating some of the limitations observed in this study.

6.2.4 Need for Larger and More Diverse Datasets

While the study focused on extremely limited samples, the results indicate that larger and more diverse datasets would likely improve model performance. This underscores the importance of data diversity in training effective speech synthesis models and suggests that future efforts should prioritize data collection and augmentation strategies that enhance both the quantity and quality of available data.

6.3 Limitations of the Study

While this study provides valuable insights into the challenges of speech synthesis with minimal data, several limitations should be acknowledged:

6.3.1 Small Sample Size

The study was constrained by the small size of the NUS-48E dataset, which limited the ability to capture the full range of variability necessary for robust model training and evaluation. This limitation highlights the need for more comprehensive datasets to fully explore the potential of data augmentation techniques in speech synthesis.

6.3.2 Simplified VAE-GAN Model

The model used in this study was relatively simple and may not have been sophisticated enough to capture the full complexity of the original timbre, particularly given the limited data. Future research could explore more advanced models that integrate multiple neural network architectures or leverage additional features such as phoneme annotations.

6.3.3 Lack of Subjective Evaluation

The study relied primarily on quantitative metrics to assess model performance, which may not fully capture the subjective quality of the generated audio. Future studies could incorporate human evaluations or perceptual metrics to provide a more comprehensive assessment of audio quality.

6.3.4 Focus on Singing Data

While the use of singing data provided a unique perspective on speech synthesis, it may not fully represent the challenges associated with more typical speech data. Future research could explore the application of these techniques to other types of speech data to determine their generalizability.

6.4 Future Directions

Based on the findings and limitations of this study, several directions for future research are proposed:

6.4.1 Exploration of Advanced Augmentation Techniques

Future research could investigate more sophisticated augmentation methods that go beyond simple transformations to more effectively mimic the variability and complexity of natural speech. Techniques such as generative adversarial networks (GANs) for augmentation or transfer learning could be explored.

6.4.2 Development of More Complex Models

The use of more advanced models that integrate multiple neural network architectures or leverage additional features, such as phoneme annotations, could improve the ability to replicate the original timbre from minimal data. Research could focus on optimizing these models for data-scarce environments, potentially using meta-learning or reinforcement learning approaches.

6.4.3 Expansion of Dataset

Expanding the dataset to include more diverse and representative samples would likely improve model performance and provide a more robust evaluation of augmentation techniques. This could involve collecting additional data or using synthetic data generation methods to create larger training sets.

6.4.4 Integration of Subjective Evaluation

Incorporating human evaluations or perceptual metrics into future studies would provide a more comprehensive assessment of audio quality and better capture the nuances of speech synthesis performance.

6.4.5 Application to Different Types of Speech Data

Future research could explore the application of these techniques to other types of speech data, such as conversational speech or multilingual datasets, to determine their generalizability and effectiveness across different contexts.

6.5 Conclusion

This study contributes to the understanding of speech synthesis and cloning in data-scarce environments by demonstrating the limitations of current data augmentation techniques and VAE-GAN models when working with extremely limited samples. The findings support the hypothesis that extremely limited samples cannot successfully replicate the original timbre, highlighting the need for more advanced augmentation strategies, larger and more diverse datasets, and more sophisticated models. Future research in this area should focus on exploring these avenues to enhance the effectiveness of speech synthesis technologies, particularly in contexts where data is limited.

References

Achille, A., Lam, M., Tewari, R., Ravichandran, A., Maji, S., Fowlkes, C., Soatto, S., & Perona, P. (2019). Task2Vec: Task Embedding for Meta-Learning. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1902.03545

Adithya Renduchintala, Shapiro, P., Duh, K., & Koehn, P. (2019). *Character-Aware Decoder for Translation into Morphologically Rich Languages*. 244–255.

Alexei Baevski, Schneider, S., & Auli, M. (2019). vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1910.05453

Alireza Makhzani, Shlens, J., Navdeep Jaitly, Goodfellow, I., & Frey, B. J. (2015). *Adversarial Autoencoders*. https://doi.org/10.48550/arxiv.1511.05644

Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., Elsen, E., Engel, J., Fan, L., Fougner, C., Han, T., Hannun, A., Jun, B., Legresley, P., Lin, L., & Narang, S. (2016). *Deep Speech 2: End-to-End Speech Recognition in English and Mandarin Baidu Research -Silicon Valley AI Lab *.*

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020, October 22). *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. ArXiv.org. https://doi.org/10.48550/arXiv.2006.11477

Balestriero, R., Misra, I., & LeCun, Y. (2022). A Data-Augmentation Is Worth A Thousand Samples: Exact Quantification From Analytical Augmented Sample Moments. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2202.08325

Bengio, E., Bacon, P.-L., Pineau, J., & Precup, D. (2016, January 7). *Conditional Computation in Neural Networks for faster models*. ArXiv.org. https://doi.org/10.48550/arXiv.1511.06297

Birkholz, P., Drechsel, S., & Stone, S. (2019). *Perceptual Optimization of an Enhanced Geometric Vocal Fold Model for Articulatory Speech Synthesis*. https://doi.org/10.21437/interspeech.2019-2410

Blaauw, M., & Bonada, J. (2017). A Neural Parametric Singing Synthesizer Modeling Timbre and Expression from Natural Songs. *Applied Sciences*, *7*(12), 1313. https://doi.org/10.3390/app7121313

Bowman, S. R., Vilnis, L., Oriol Vinyals, Dai, A. M., Jozefowicz, R., & Samy Bengio. (2016). Generating Sentences from a Continuous Space. *Conference on Computational Natural*

Language Learning. https://doi.org/10.18653/v1/k16-1002

Cai, Z., Zhang, C., & Li, M. (2020). From Speaker Verification to Multispeaker Speech Synthesis, Deep Transfer with Feedback Constraint. *ArXiv (Cornell University)*. https://doi.org/10.21437/interspeech.2020-1032

Cassell, J. (2000). *Embodied conversational agents*. Mit Press.

Dai, D., Chen, Y., Chen, L., Tu, M., Liu, L., Xia, R., Tian, Q., Wang, Y., & Wang, Y. (2021). *Clonning One's Voice Using Very Limited Data in the Wild*.

Driedger, J., Muller, M., & Ewert, S. (2014). Improving Time-Scale Modification of Music Signals Using Harmonic-Percussive Separation. *IEEE Signal Processing Letters*, *21*(1), 105–109. https://doi.org/10.1109/lsp.2013.2294023

Duan, Z., Fang, H., Li, B., Sim, K. C., & Wang, Y. (2013). The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech. *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. https://doi.org/10.1109/apsipa.2013.6694316

Frankle, J., Dziugaite, G. K., Roy, D. M., & Carbin, M. (2020, July 20). *Stabilizing the Lottery Ticket Hypothesis*. ArXiv.org. https://doi.org/10.48550/arXiv.1903.01611

Gibiansky, A., Sercan Ömer Arik, Gregory Frederick Diamos, Miller, J., Peng, K., Ping, W., Raiman, J., & Zhou, Y. (2017). Deep Voice 2: Multi-Speaker Neural Text-to-Speech. *Neural Information Processing Systems*, *30*, 2962–2970.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative Adversarial Networks. *Advances in Neural Information Processing Systems*. arxiv:1406.2661

Guo, S., Shi, J., Qian, T., Watanabe, S., & Jin, Q. (2022). *SingAug: Data Augmentation for Singing Voice Synthesis with Cycle-consistent Training Strategy*. https://doi.org/10.21437/interspeech.2022-978

He, C., Liu, J., Zhu, Y., & Du, W. (2021). Data Augmentation for Deep Neural Networks Model in EEG Classification Task: A Review. *Frontiers in Human Neuroscience*, *15*. https://doi.org/10.3389/fnhum.2021.765525

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., & Wilson, K. (2017, March 1). *CNN architectures for large-scale audio classification*. IEEE Xplore. https://doi.org/10.1109/ICASSP.2017.7952132

Hospedales, T. M., Antoniou, A., Micaelli, P., & Storkey, A. J. (2021). Meta-Learning in Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1. https://doi.org/10.1109/tpami.2021.3079209

Hu, Z., Yang, Z., Ruslan Salakhutdinov, & Xing, E. P. (2017). On Unifying Deep Generative Models. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1706.00550

Huang, W.-C., Luo, H., Hwang, H.-T., Lo, C.-C., Peng, Y., Tsao, Y., & Wang, H.-M. (2020). Unsupervised Representation Disentanglement Using Cross Domain Features and Adversarial Learning in Variational Autoencoder Based Voice Conversion. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *4*(4), 468–479. https://doi.org/10.1109/tetci.2020.2977678

Iglesias, G., Talavera, E., González-Prieto, Á., Mozo, A., & Gómez-Canaval, S. (2023). Data Augmentation techniques in time series domain: a survey and taxonomy. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-023-08459-3

Jaitly, N., & Hinton, G. (2013). *Vocal Tract Length Perturbation (VTLP) improves speech recognition*. *28*.

Jan Christian Schlüter, & Grill, T. (2015). *Exploring Data Augmentation for Improved Singing Voice Detection with Neural Networks.* 121–126. https://doi.org/10.5072/zenodo.243314

Juslin, P. N., & Sloboda, J. (2012). *Handbook of music and emotion : theory, research, applications*. Oxford University Press.

Kadyan, V., Kathania, H., Govil, P., & Kurimo, M. (2021). Synthesis Speech Based Data Augmentation for Low Resource Children ASR. *Speech and Computer*, 317–326. https://doi.org/10.1007/978-3-030-87802-3_29

Kim, K. L., Lee, J., Kum, S., Park, C. L., & Nam, J. (2020). Semantic Tagging of Singing Voices in Popular Music Recordings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 1656–1668. https://doi.org/10.1109/taslp.2020.2993893

Kingma, D., & Welling, M. (2013). *Auto-Encoding Variational Bayes*.

Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. *Interspeech 2015*. https://doi.org/10.21437/interspeech.2015-711

Kubichek, R. F. (1993). Mel-cepstral distance measure for objective speech quality assessment. *Pacific Rim Conference on Communications, Computers and Signal Processing*. https://doi.org/10.1109/pacrim.1993.407206

Larsen, A. B. L., Sønderby, S. K., Larochelle, H., & Winther, O. (2016, February 10).

Autoencoding beyond pixels using a learned similarity metric. ArXiv.org. https://doi.org/10.48550/arXiv.1512.09300

Laver, J. (1994). *Principles of phonetics*. Cambridge University Press.

Lee, J., Choi, H.-S., Koo, J., & Lee, K. (2020). Disentangling Timbre and Singing Style with Multi-Singer Singing Synthesis System. *ArXiv (Cornell University)*. https://doi.org/10.1109/icassp40776.2020.9054636

Lee, Y.-J., Chen, B.-Y., Lai, Y.-T., Liao, H.-W., Liao, T.-C., Kao, S.-L., Kang, K.-Y., Hsu, C.-T., & Liu, Y.-W. (2018). Examining the Influence of Word Tonality on Pitch Contours When Singing in Mandarin. *2018 Oriental COCOSDA - International Conference on Speech Database and Assessments*. https://doi.org/10.1109/icsda.2018.8693016

Li, Z., Tang, B., Yin, X., Wan, Y., Xu, L., Shen, C., & Ma, Z. (2021, June 1). *PPG-Based Singing Voice Conversion with Adversarial Representation Learning*. IEEE Xplore. https://doi.org/10.1109/ICASSP39728.2021.9414137

Lo, C.-C., Fu, S.-W., Huang, W.-C., Wang, X., Yamagishi, J., Tsao, Y., & Wang, H.-M. (2019, September 15). *MOSNet: Deep Learning based Objective Assessment for Voice Conversion*. ArXiv.org. https://doi.org/10.21437/Interspeech.2019-2003

Lu, P., Wu, J., Luan, J., Tan, X., & Zhou, L. (2020). XiaoiceSing: A High-Quality and Integrated Singing Voice Synthesis System. *ArXiv (Cornell University)*. https://doi.org/10.21437/interspeech.2020-1410

Lyu, Z., & Zhu, J. (2023). GAN-Based Fine-Grained Feature Modeling For Zero-Shot Voice Cloning. *Proceedings of the 3rd World Congress on Electrical Engineering and Computer Systems and Science*. https://doi.org/10.11159/mhci23.111

O'connor, B., Dixon, S., & Fazekas, G. (2020). *Zero-shot Singing Technique Conversion*.

Oppenheim, A. V., Schafer, R. W., & Buck, J. R. (1999). *Discrete-time Signal Processing*. Prentice Hall.

Park, D., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E., & Le, Q. (2019). *SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition*.

PATEL, A. D., WONG, M., FOXTON, J., LOCHY, A., & PERETZ, I. (2008). Speech Intonation Perception Deficits In Musical Tone Deafness (Congenital Amusia). *Music Perception: An Interdisciplinary Journal*, *25*(4), 357–368. https://doi.org/10.1525/mp.2008.25.4.357

Pritish Chandna, Blaauw, M., Jordi Bonada, & Gomez, E. (2019). A Vocoder Based Method for Singing Voice Extraction. *ArXiv (Cornell University)*.

https://doi.org/10.1109/icassp.2019.8683323

Rabiner, L. R., & Biing-Hwang Juang. (2005). *Fundamentals of speech recognition*. Delhi Pearson Education.

Rosenzweig, S., Schwar, S., Driedger, J., & Muller, M. (2021). *Adaptive Pitch-Shifting with Applications to Intonation Adjustment in a Cappella Recordings*. https://doi.org/10.23919/dafx51585.2021.9768268

Ruder, S., Peters, M. E., Swayamdipta, S., & Wolf, T. (2019, June 1). *Transfer Learning in Natural Language Processing*. ACLWeb; Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-5004

Sargin, M. E., Yemez, Y., Erzin, E., & Tekalp, A. M. (2007). Audiovisual Synchronization and Fusion Using Canonical Correlation Analysis. *IEEE Transactions on Multimedia*, *9*(7), 1396–1403. https://doi.org/10.1109/tmm.2007.906583

Schmidt, L., Shibani Santurkar, Dimitris Tsipras, Talwar, K., & Aleksander Mądry. (2018). Adversarially Robust Generalization Requires More Data. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1804.11285

Sergey Zagoruyko, & Nikos Komodakis. (2017). Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. *HAL (Le Centre Pour La Communication Scientifique Directe)*.

Shen, Y., Ji, R., Zhang, S., Zuo, W., & Wang, Y. (2018). Generative Adversarial Learning Towards Fast Weakly Supervised Detection. *Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.2018.00604

Shinji Watanabe, Delcroix, M., Florian Metze, Hershey, J. R., & Springerlink (Online Service. (2017). *New Era for Robust Speech Recognition : Exploiting Deep Learning*. Springer International Publishing.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, *6*(1). https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0

Streijl, R. C., Winkler, S., & Hands, D. S. (2014). Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, *22*(2), 213–227. https://doi.org/10.1007/s00530-014-0446-1

Tang, H., Zhang, X., Wang, J., Cheng, N., & Xiao, J. (2023). *Learning Speech Representations with Flexible Hidden Feature Dimensions*. https://doi.org/10.1109/icassp49357.2023.10094969

Thomas, S., Seltzer, M. L., Church, K., & Hynek Hermansky. (2013). *Deep neural network features and semi-supervised training for low resource speech recognition*. https://doi.org/10.1109/icassp.2013.6638959

Toda, T., Chen, L.-H., Saito, D., Villavicencio, F., Wester, M., Wu, Z., & Yamagishi, J. (2016). The Voice Conversion Challenge 2016. *Interspeech 2016*. https://doi.org/10.21437/interspeech.2016-1066

Verma, P., & Smith, J. (2020). A Framework for Generative and Contrastive Learning of Audio Representations. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2010.11459

Wu, N.-Q., Liu, Z.-C., & Ling, Z.-H. (2022). Discourse-Level Prosody Modeling with a Variational Autoencoder for Non-Autoregressive Expressive Speech Synthesis. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/icassp43922.2022.9746238

Yi, C., Wang, J., Cheng, N., Zhou, S., & Xu, B. (2021, January 17). *Applying Wav2vec2.0 to Speech Recognition in Various Low-resource Languages*. ArXiv.org. https://doi.org/10.48550/arXiv.2012.12121

Yim, J., & Sohn, K.-A. (2017). Enhancing the Performance of Convolutional Neural Networks on Quality Degraded Datasets. *ArXiv (Cornell University)*. https://doi.org/10.1109/dicta.2017.8227427

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: music and speech. *Trends in Cognitive Sciences*, *6*(1), 37–46. https://doi.org/10.1016/s1364-6613(00)01816-7

Zeghidour, N., Usunier, N., Synnaeve, G., Collobert, R., & Dupoux, E. (2018). *End-to-End Speech Recognition From the Raw Waveform*.

Zhang, Z., Geiger, J., Jouni Pohjalainen, Amr El-Desoky Mousa, Jin, W., & Schuller, B. (2017). Deep Learning for Environmentally Robust Speech Recognition: An Overview of Recent Developments. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1705.10874