Master research Internship

Internship report

French voice cloning and stylistic control with VALL-E

Domain: Artificial Intelligence - Sound

Author:
Yasser El Ayyachy

Supervisor:
David Guennec
Damien Lolive
Équipe Expression

ENSSAT
LANNION

Abstract: This internship report details the process of adapting the VALL-E neural codec language model for French voice cloning, focusing on overcoming the challenges posed by limited data resources in the French language. The work began with data collection and preparation, utilizing datasets such as Common Voice, which required careful preprocessing to ensure consistency and quality. The project was based on Bark, a GPT-style generative audio model, which —unlike VALL-E— does not use phonemes, making it less suited for languages like French with distinct phonetic characteristics. To address this, a key innovation was the implementation of a phoneme-based approach, significantly improving the naturalness and prosody of the synthesized speech. The training and optimization of the adapted model faced several challenges, including initial failures to generate coherent audio and delays in accessing powerful servers. Through iterative refinements, particularly in phonemization techniques and model architecture, these challenges were addressed, resulting in an improved voice cloning model capable of producing French speech. The report also delves into the evaluation and validation of the model, highlighting the impact of speaker bias and recording quality on the final output.

Contents

1 Introduction

The human voice is a powerful instrument of expression, intricately tied to our identity and interactions. The human voice serves as a complex, multifaceted instrument integral to both individual identity and social interactions. Functioning beyond basic communicative purposes, the voice acts as a medium conveying a spectrum of emotional states, cognitive processes, and elements of cultural background. Each voice possesses a distinctive combination of tonal qualities, inflection patterns, and rhythmic characteristics, facilitating the transmission of information that extends beyond mere verbal content. The vocal mechanism is capable of producing a wide array of sounds, ranging from subtle variations detectable in whispered speech to the pronounced acoustic features evident in loud vocalizations. This diversity in vocal expression plays a crucial role in the conveyance of a broad spectrum of human emotions and thoughts.

The human voice transmits information that extends beyond mere linguistic messages, encapsulating aspects of the speaker's identity. It functions as a medium facilitating interpersonal connections, overcoming geographical and cultural divides. Vocal expressions enable individuals to narrate personal experiences, disseminate knowledge, and articulate emotions. Variations in pitch, tone, and speed contribute to the complexity of human vocalization, reflecting the varied experiences and backgrounds of speakers. These vocal characteristics play a significant role in the communication process, serving as indicators of individual identity and emotional states.

The human voice also plays a crucial role in the formation and maintenance of personal relationships and societal bonds. Infants show an innate preference for their mother's voice, essential for early bonding. For instance, DeCasper and Fifer's study [1] revealed that newborns could recognize and preferentially produce their mother's voice, underlining its role in the initial stages of infant-mother bonding.

Speech synthesis technology, particularly neural Text-to-Speech (TTS), represents a major area in computational linguistics and acoustic engineering, involving interdisciplinary collaboration from computer science, linguistics, and psychoacoustics [2]. Key advancements include the development of neural TTS models like Deep Voice [3] and WaveNet [4], which have notably improved voice quality in terms of intelligibility and naturalness [5].

Speech synthesis has evolved significantly over the years, progressing from early text-to-speech (TTS) systems to contemporary neural network-based models. This evolution reflects advancements in both technology and our understanding of the linguistic and acoustic aspects of speech. In this internship report, we explore the development and application of modern speech synthesis techniques, specifically focusing on adapting these technologies to French voice cloning.

The report begins with an overview of the historical approaches that laid the foundation for early TTS systems. These early models, while limited by the computational resources of their time, were instrumental in shaping the field. As we examine the progression of TTS technology, we observe key developments that have led to modern systems capable of producing highly natural and expressive speech.

Central to this report is the adaptation of the VALL-E architecture, which has demonstrated significant capabilities in speech synthesis, including its proficiency in zero-shot learning and in-context voice cloning. While originally developed for English and Mandarin, adapting VALL-E to French presented unique challenges, primarily due to the scarcity of high-quality French speech datasets. The work undertaken during this internship aimed to address these challenges through data collection, phoneme-based model enhancements, and an exploration of training and optimization strategies.

The subsequent sections of the report discuss the development pipeline, from the preprocessing of datasets such as Common Voice to the fine-tuning of neural vocoder models like Bark. Emphasis is placed on the technical adaptations required to handle French-specific phonemes and the iterative process involved in optimizing model performance. Additionally, challenges related to training time and computational resources are highlighted, given the extensive time required for training on large datasets, both locally and on high-performance computing clusters.

Finally, the report presents an evaluation of the models, utilizing both objective and subjective metrics to assess the quality of the synthesized speech. The results are discussed in terms of speaker variability, dataset quality, and the impact of audio normalization and enhancement techniques. The report concludes with insights into the limitations encountered and potential directions for future work in French voice cloning.

2 Speech Synthesis

Speech synthesis, is the artificial production of human speech, transforming written text into spoken language. This complex process encompasses multiple stages, including text normalization, linguistic analysis, and acoustic synthesis. It draws upon the fields of phonetics and phonology to accurately model the variations in pitch, tone, and rhythm characteristics of natural speech. Modern advancements in speech synthesis often utilize artificial intelligence and machine learning, particularly deep learning, to enhance the naturalness and intelligibility of synthesized speech by training on extensive datasets of recorded human speech. Within this domain, voice cloning represents a specialized subset that focuses on creating digital replicas of specific individuals' voices. By capturing and reproducing unique vocal characteristics such as timbre, pitch, accent, and speaking style, voice cloning enables the generation of speech that closely mimics the original speaker. This involves sophisticated machine learning techniques, including neural networks, to analyze and model these vocal features from audio recordings. Speech synthesis, inclusive of voice cloning, finds applications in various fields such as virtual assistants, accessibility tools, personalized customer service, entertainment, and digital avatars. These technologies play a crucial role in enhancing human-computer interaction by enabling more natural and intuitive communication.

Understanding the history of speech synthesis, from its rudimentary beginnings to the sophisticated AI-driven methods of today, provides valuable context for appreciating the technological advancements and future potential of this field.

2.1 Historic approaches in TTS

This section presents an academic overview of the historic approaches in TTS and voice conversion, tracing the field's evolution from the seminal mechanical innovations of the 18^{th} century to the sophisticated digital technologies of today. It highlights key contributions and technological milestones that have shaped the development of speech synthesis.

The history of text-to-speech synthesis dates back to the 18^{th} century, with significant contributions from Christian Gottlieb Kratzenstein and Wolfgang von Kempelen. Kratzenstein, in the 1770s, developed mechanical resonators for vowel sound emulation, representing an early effort to replicate human vocal tract configurations [6]. Concurrently, von Kempelen advanced the field with a speaking machine that aimed to mimic the human vocal system, including the lungs, larynx, and vocal cords, using mechanical components. His work, particularly in the latter half of the 18^{th}

century, laid the foundation for future developments in speech synthesis [7].

2.2 Contemporary/recent advances

As the 20^{th} century progressed, articulatory synthesis emerged, involving complex modeling of the vocal tract with principles from differential calculus and fluid mechanics. This endeavor aimed to simulate vocal tract behavior under various physiological conditions to replicate human speech dynamics more accurately. Key figures in this domain include Coker and Fujimura (1966), who worked on the specification of the vocal tract area function, and Fant (1960), who contributed to the acoustic theory of speech production [8].

Linear Prediction Synthesis, developed in the 1960s, marked a novel approach in telecommunications by predicting speech samples based on previous ones using a mathematical model. This method was known for its bandwidth efficiency and clarity. Key contributors to Linear Prediction Synthesis include Atal and Schroeder (1968) [9], who explored predictive coding of speech, and Saito and Itakura (Jan 1967), who examined the statistical optimum recognition of speech spectral density [10].

In the domain of Concatenative Synthesis and Unit Selection, which predated Linear Prediction Synthesis, starting with pioneers like Sagisaka, Campbell, and Hunt in the late 1980s and 1990s worked on unit selection methods using phonetic units like diphones and triphones to improve speech synthesis quality and versatility [11].

HMM-based Statistical Parametric Speech Synthesis, though emerging as a more flexible approach in the late 1990s and early 2000s, faced challenges with naturalness compared to Unit Selection. Nonetheless, this method was notable for its adaptability and intelligibility in capturing human speech patterns [12].

Finally, the integration of parametric prediction with unit selection in Hybrid Parametric/Unit Selection Systems represented a synthesis of the strengths of both methods, leading to notable improvements in quality. This combination harnessed the adaptable, context-sensitive characteristics of parametric methods with the naturalness of unit selection techniques, culminating in more lifelike and versatile synthesized speech [13].

2.3 The rise of Neural models

The shift to contemporary TTS and voice cloning technologies, particularly since the mid-2010s, represents a significant evolution in the field. This period is marked by the transition from mechanical and rule-based systems to advanced digital technologies. The integration of computational methods, machine learning, and digital signal processing has notably transformed speech synthesis. Notably, technologies like WaveNet, introduced in 2016, and BigVGAN, developed in the early 2020s, have redefined benchmarks in naturalness and expressiveness of synthesized speech. These technologies set new standards by significantly enhancing the realism and fluidity of generated speech compared to previous systems.

In the realm of artificial intelligence and speech technology, TTS and voice cloning represent remarkable milestones. These technologies not only exemplify the synthesis of human-like speech from text but also the creation of unique, personalized voices through cloning. This essay delves into the intricate world of TTS and voice cloning, exploring their evolution, underlying principles, and the transformative impact they hold in various sectors.

TTS technology converts written text into spoken words, enabling machines to communicate with a human touch. This seemingly simple process is underpinned by complex algorithms and models that analyze and attempts to reproduce the nuances of human speech. Creating voices that are not only intelligible but also expressive and natural-sounding, enhances user experience in applications ranging from virtual assistants to accessibility tools for those with speech impairments.

Voice cloning, a specific application within TTS technology, involves creating a digital replica of a specific human voice, distinguishing itself from standard TTS by its ability to replicate individual vocal characteristics. Unlike TTS, which generally synthesizes speech from text using various pre-designed voice models, voice cloning specifically targets the replication of a unique voice's tone, inflection, and nuances using deep learning techniques, requiring samples of the original voice for accurate cloning [14]. Its applications range from personalizing digital assistants and providing speech capabilities to those who have lost their voices, to uses in entertainment. However, voice cloning also presents significant ethical challenges, particularly concerning consent and the potential for misuse, echoing concerns similar to those raised by deepfake technologies. These ethical considerations are crucial in the development and application of voice cloning technologies.

The intersection of TTS and voice cloning with fields such as machine learning, linguistics, and digital signal processing has led to rapid advancements. From rudimentary robotic voices to the sophisticated, emotive speech we encounter today, these technologies have undergone a dramatic transformation. They stand as testaments to human ingenuity, opening up new avenues for human-computer interaction and presenting both opportunities and challenges for the future.

2.3.1 Acoustic Models in Speech Synthesis

An acoustic model in speech synthesis is a component that translates textual information into acoustic features, such as phonetic and prosodic properties. These features represent the building blocks of speech, including aspects like pitch, duration, and timbre. The acoustic model's role is crucial in determining how speech sounds in terms of these acoustic characteristics. In the realm of modern TTS, significant advancements have been made in the development of acoustic models. The inception of this era can be traced back to the introduction of WaveNet. The introduction of WaveNet represents a significant leap in text-to-speech technology. WaveNet's architecture is a fully probabilistic and autoregressive model, processing audio data sample-by-sample, which is a key to its ability to generate highly realistic speech. This efficiency in handling high-resolution data is achieved without compromising on the quality of speech synthesis. Notably, WaveNet outperformed traditional TTS systems in terms of naturalness, as rated by human listeners. Its versatility extends beyond speech, showing promising results in music generation as well. The use of dilated convolutions in WaveNet is particularly noteworthy, as it allows the network to efficiently encapsulate information from extensive temporal windows, crucial for producing coherent speech. This breakthrough has not only set new standards in speech synthesis but also influenced subsequent research in the field, pushing the boundaries of what's achievable in TTS systems.

Following Wavenet, the introduction of Tacotron, marked a major step forward [15]. Tacotron simplified the speech synthesis process by directly mapping character sequences to spectrograms using a sequence-to-sequence model, streamlining the traditionally complex pipeline of TTS systems. Further refinement of WaveNet resulting in Tacotron2 [16] demonstrated how conditioning WaveNet on mel spectrograms could yield even more natural-sounding speech. The key innovation lies in conditioning the WaveNet model on mel spectrograms, which are a more efficient representation of audio compared to raw waveforms. This method simplifies the training process and

enhances the overall quality of the synthesized speech, marking an important step in making TTS more lifelike and realistic. This approach underscores the critical role of advanced neural network architectures in the continuous evolution of TTS technologies.

2.3.2 Advancements in Neural Vocoders

A vocoder in speech synthesis is a technology used to synthesize audible speech from acoustic features. It converts these features, which include aspects like the frequency spectrum of speech, into the final sound waveform.

In the domain of neural vocoders, recent advancements have greatly enhanced the quality and efficiency of speech synthesis. HiFi-GAN [17] marked a significant advancement in using Generative Adversarial Networks for high-fidelity and efficient speech synthesis. Its architecture includes a novel approach of multi-period discriminator which effectively captures periodic patterns in audio, crucial for realistic speech synthesis. This model demonstrates a higher mean opinion score (MOS) than existing models like WaveNet, indicating superior audio quality. Importantly, HiFi-GAN achieves this high-quality output with remarkable efficiency, synthesizing speech much faster than real-time on standard computing hardware. This advancement in TTS technology underscores the potential of GANs in achieving realistic and computationally efficient speech synthesis, marking a significant step towards more natural and accessible TTS systems. BigVGAN [18] is another key development, featuring a large-scale model with up to 112 million parameters. It has set new standards in terms of naturalness and clarity of generated audio. BigVGAN's innovative approach incorporates periodic activations and anti-aliased representations, which contribute to its exceptional performance across various speakers and recording conditions. The model's ability to handle challenging out-of-distribution scenarios with high fidelity marks it as a state-of-the-art development in TTS technology, pushing the boundaries of audio quality and adaptability in speech synthesis.

2.3.3 End-to-end models

Recent trends in TTS technology indicate a move towards integrating different model architectures to create more efficient and versatile systems. The integration of variational autoencoders and adversarial learning in TTS was explored in VITS(2021) [19]. The key innovation is the efficient end-to-end learning process, which results in more natural sounding audio compared to existing two-stage TTS models. The method employs variational inference augmented with normalizing flows and an adversarial training process, enhancing the generative modeling's expressive power. This approach is pivotal in addressing the one-to-many problem in TTS, where a text input can be spoken in multiple ways, with variations such as different pitches and rhythms. The paper's findings show that their method outperforms existing TTS systems, achieving a level of naturalness in speech synthesis that closely approaches ground truth. This development underlines the significance of combining advanced neural network techniques in creating more expressive and natural-sounding synthetic speech.

3 State of The Art in TTS/VC

3.1 Vall-E

In the specialized domain of voice cloning within TTS technology, the introduction of models such as VALL-E [20] and its extension, VALL-E X [21], heralds a new era of technological innovation. These models collectively represent a paradigm shift in the realm of TTS and cross-lingual speech synthesis. VALL-E, as a pioneering neural codec language model, reframes TTS as a conditional language modeling task, using discrete acoustic codes from neural audio codecs. This approach diverges significantly from traditional methods reliant on continuous signal regression, setting a new precedent in speech synthesis.

VALL-E's extensive training on a vast corpus of 60K hours of diverse English speech marks an unprecedented leap in zero-shot TTS capabilities. It demonstrates high proficiency in synthesizing highly natural and personalized speech, capturing the nuances of unseen speakers with only a minimal acoustic prompt. The extension of this model, VALL-E X, takes this innovation further into the domain of cross-lingual speech synthesis. VALL-E X not only retains the speaker's unique voice across languages but also addresses the challenges of foreign accent in cross-lingual settings, improving upon the capabilities of previous models.

This section aims to provide an integrated review of both VALL-E and VALL-E X, delving into their shared foundations and distinct advancements. We will explore their intricate architectures, evaluate their performance in respective domains, and highlight their superior capabilities in speaker similarity, speech naturalness, and cross-lingual adaptability. Additionally, we will critically assess the challenges and potential enhancements for these models, setting the stage for future developments in TTS technology. Through this comprehensive analysis, we aim to shed light on how VALL-E and VALL-E X are reshaping the landscape of speech synthesis, opening new avenues for natural and accessible communication technologies.

3.2 Overview

VALL-E represents an important shift compared to more traditional TTS approaches by framing speech synthesis as a conditional language modeling task. As can be seen in Figure 1 VALL-E creates specific acoustic tokens based on a 3-second recorded sample and a phoneme prompt. The recorded sample provides speaker information, while the prompt dictates the content. These tokens are then converted into the final speech waveform using a neural codec decoder. This innovative methodology enables VALL-E to produce personalized, high-quality speech and showcases its exceptional performance in zero-shot scenarios with unseen speakers.

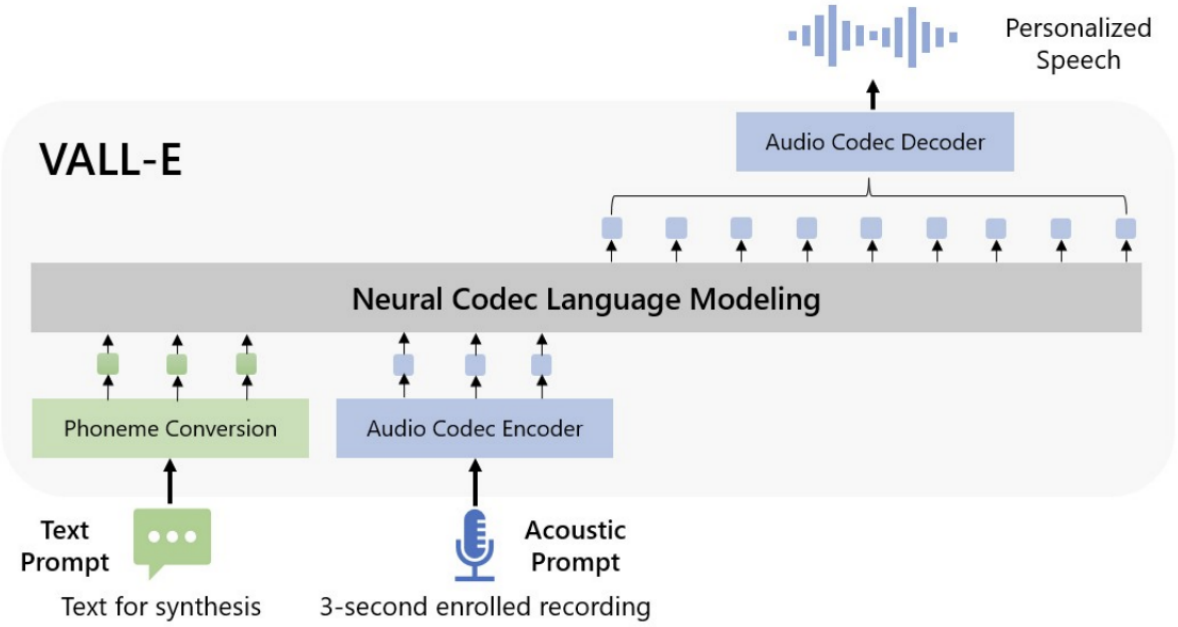

Figure 1: Overview of the VALL-E model architecture. Figure from the original Vall-E paper [20].

3.3 Encodec

As previously discussed, Vall-E's initial task is to generate tokens from audio recordings. To achieve this, Vall-E utilizes an architecture introduced by Meta AI in 2022 known as Encodec. As depicted in 2 , Encodec's architecture is composed of three primary components. First, an encoder processes the audio input and produces a latent space representation of the sample. This is followed by a quantization layer that compresses the latent representation into tokens using Vector Quantization, which Vall-E then employs. Finally, a decoder reconstructs the audio signal from these tokens.

3.3.1 Vector Quantization

The quantization step is critical and requires further clarification. Initially, the encoder produces a continuous latent representation, consisting of a set of vector representations. These vectors are then processed through a vector quantization layer.

Vector quantization serves to compress data, reduce noise, extract features, and offer an efficient representation of audio waveforms. This is particularly important given the high information density of audio data, which is typically represented by encoded vectors with high dimensionality.

Therefore, quantization is necessary to reduce the dimensionality of these encoded vectors, facilitating data compression. Vector quantization accomplishes this by transforming the continuous latent representation into a discrete one, based on a set of codebooks.

A codebook is a small collection of vectors that represent the range and variety of the encoded vectors. It functions similarly to a dictionary of sound patterns. Rather than storing every detail of a sound, the process identifies the closest pattern in the codebook and saves the corresponding reference.

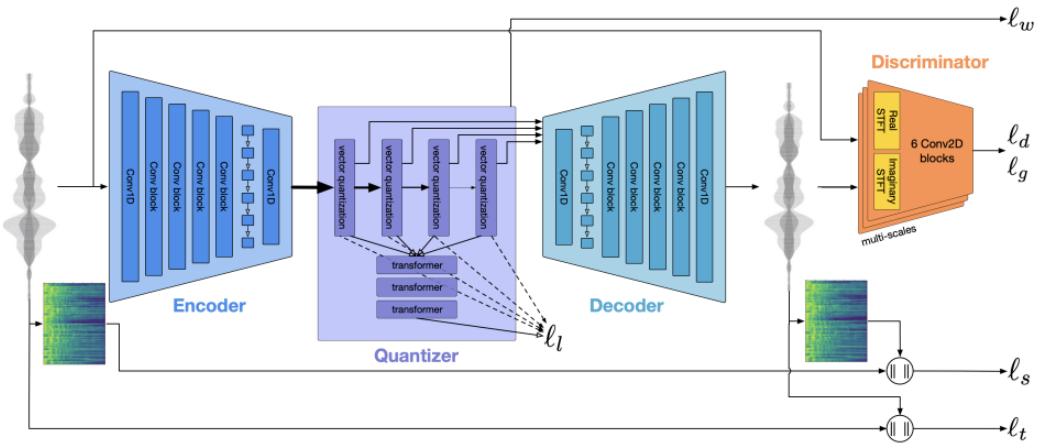

Figure 2: EnCodec : an encoder decoder codec architecture which is trained with reconstruction (ℓ_f and ℓ_t) as well as adversarial losses (ℓ_g for the generator and ℓ_d for the discriminator). The residual vector quantization commitment loss (ℓ_w) applies only to the encoder. Optionally, we train a small Transformer language model for entropy coding over the quantized units with ℓ_l, which reduces bandwidth even further. Figure from the original Encodec paper [22].

This process of mapping audio data to the nearest vector (or codeword) in the codebook allows for a compact representation of the audio data. For example, 16 codebooks might be used for a 12kbps-bandwidth audio waveform. The discrete quantized representation can be converted back into a vector by summing the corresponding codebook entries, a step performed just before the data is passed to the decoder as can be seen in Figure 3.

3.3.2 Residual Vector Quantization

A single quantizer layer would require an extremely large codebook to accurately represent compressed data, with the codebook size growing exponentially as the data's dimensionality increases. Encodec implements residual vector quantization (RVQ) which adresses this challenge by offering a more efficient approach to quantization through a series of codebooks applied in a cascading manner. The core concept behind RVQ is to achieve a progressively finer approximation of high-dimensional vectors by breaking down the quantization process into multiple stages as illustrated in Figure 4. Initially, the primary codebook performs a first-level quantization of the input vector. The residuals, which are the differences between the original data vectors and their quantized counterparts, are then further quantized using a secondary codebook.

This process is repeated across several stages, with each subsequent codebook focusing on quantizing the residuals left over from the previous stage. By decomposing the quantization task into these smaller, sequential steps, RVQ effectively captures the intricacies of the data with high precision while significantly reducing the computational demands typically associated with quantizing high-dimensional vectors using a single, large codebook.

It is important to note that Vall-E only employs the first two components of a pretrained Encodec model for token generation.

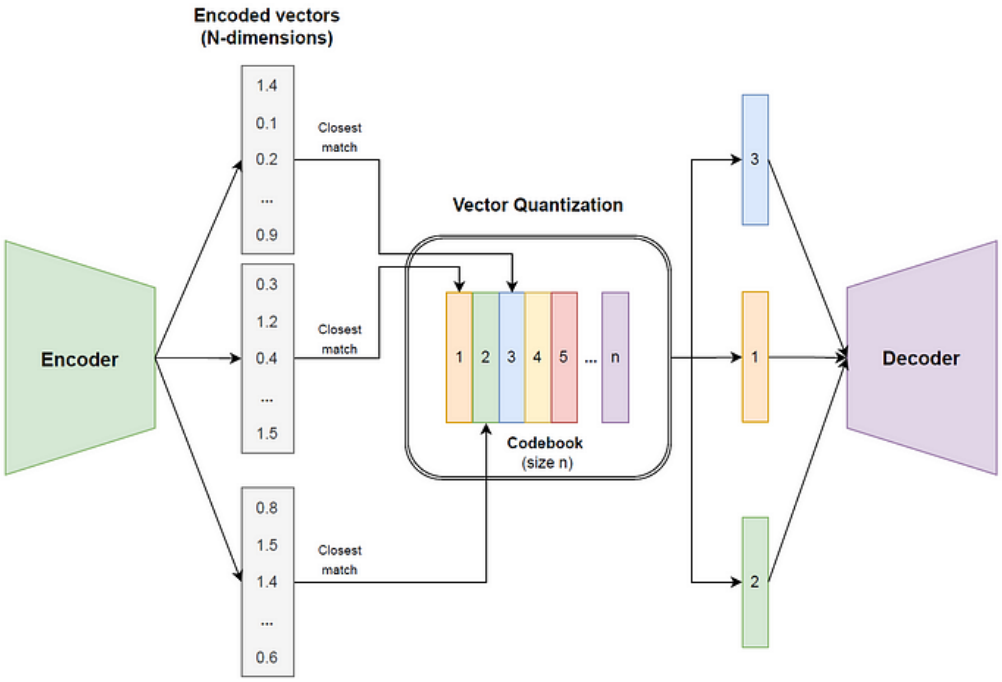

Figure 3: Illustration of the vector quantization process where a continuous latent vector is mapped to a discrete codeword in the codebook.

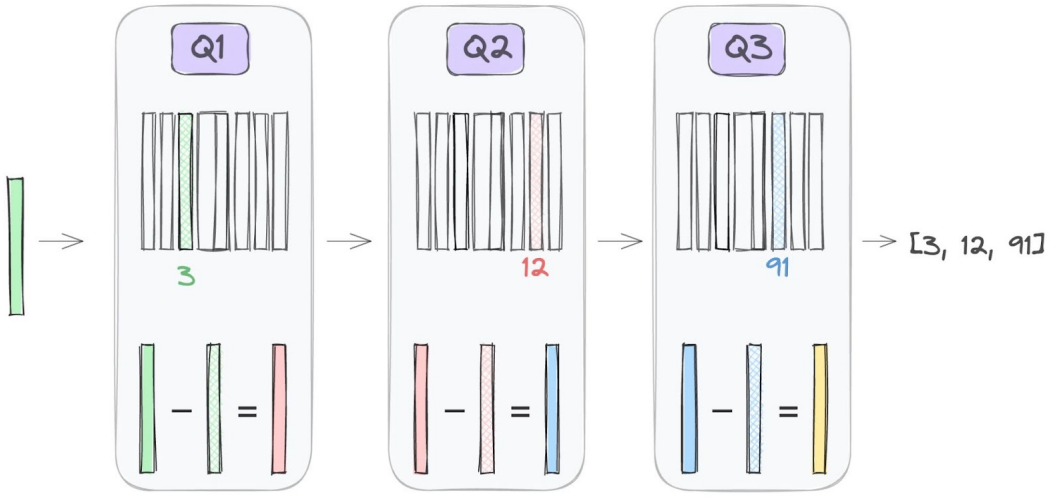

Figure 4: Illustration of Residual Vector Quantization (RVQ) process.

3.4 Model Design: Architecture and Training

The architecture of VALL-E integrates autoregressive (AR) and non-autoregressive (NAR) models. As shown in Figure 5, the AR component is designed to generate tokens from the first quantizer, focusing on capturing broader acoustic properties like the speaker's identity. In contrast, the NAR

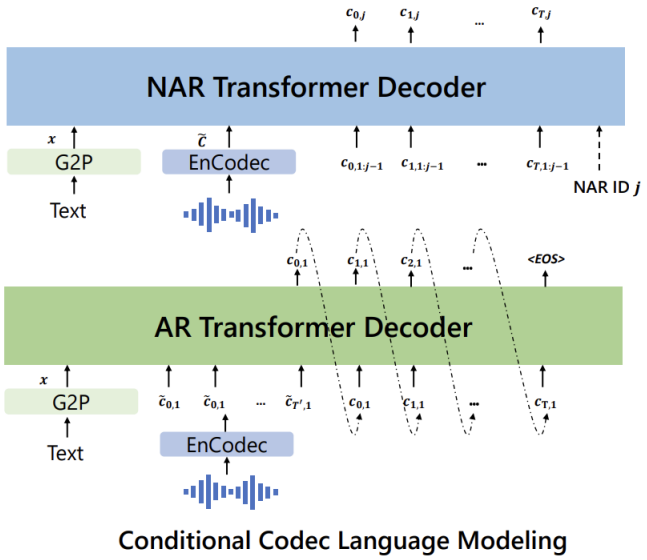

Figure 5: The structure of the conditional codec language modeling, which is built in a hierarchical manner. Figure from the original Vall-E paper [20].

model attends to finer acoustic details as captured by subsequent quantizers. This design balances speech quality with inference speed and accommodates various speaking styles and lengths. The foundation of VALL-E's success lies in its training on the LibriLight [23] corpus, an extensive dataset with 60K hours of English speech from over 7,000 distinct speakers. This corpus provides an array of speakers and prosodies, ensuring effective generalization across various speech patterns. The data's volume, vastly exceeding that of existing TTS systems, significantly enhances VALL-E's robustness and versatility, allowing it to adapt to a wide spectrum of vocal attributes.

3.5 Capabilities: In-Context Learning and Zero-Shot Performance

A standout feature of VALL-E is its in-context learning capability, reminiscent of large text-based language models. This attribute allows VALL-E to synthesize speech for previously unseen speakers without requiring fine-tuning, maintaining high fidelity in speaker similarity and speech naturalness. The model's prowess in zero-shot scenarios is further evidenced by its ability to retain the acoustic environment and the speaker's emotional nuances as inferred from input prompts.Moreover, VALL-E's proficiency extends to the nuanced realm of emotional TTS, a subset of speech synthesis focused on generating speech with specific emotional undertones. Remarkably, VALL-E demonstrates the capability to maintain the emotional tone of the input prompt in its output, effectively transferring the emotional context from the source to the synthesized speech. This feature is exemplified by tests using acoustic prompts from the EmoV-DB [24] dataset, which contains speech samples across five distinct emotions, and VALL-E successfully retains these emotions in the synthesized output without any specific fine-tuning for emotional speech synthesis.

3.6 Cross-Lingual Extensions: VALL-E X

In addition to its remarkable capabilities in monolingual TTS, the advent of VALL-E X marks an extension into the realm of cross-lingual speech synthesis. VALL-E X adeptly synthesizes speech in foreign languages, maintaining the essence of the original speaker's voice and emotional tone. This innovation is pivotal for applications in multilingual contexts and broadens the scope of VALL-E's usability in sophisticated speech-to-speech translation tasks, thereby addressing a broader spectrum of linguistic diversity.

3.7 Comprehensive Performance Evaluation

Both VALL-E and VALL-E X have been subjected to rigorous performance evaluations across varied datasets. They have demonstrated superior performance over existing zero-shot TTS systems, particularly in terms of speaker similarity and speech naturalness. VALL-E's capability to achieve a CMOS score comparable to human speech recordings demonstrates its advanced synthesis capabilities. These evaluations highlight the model's proficiency in generating speech that is natural, diverse, and closely aligned with the specific characteristics of individual speakers.

3.8 Implications and Prospects for VALL-E in Text-to-Speech Synthesis

The introduction of VALL-E and its cross-lingual counterpart, VALL-E X, represents a highlight of TTS technology. Their innovative approaches, extensive training datasets, and advanced architectural frameworks have set new benchmarks in speech synthesis. As these models evolve, they promise a future where nuanced, natural, and diverse speech synthesis is accessible across languages and speaker profiles, marking a significant stride towards universal

4 Data

Although the VALL-E architecture offers several advantages, such as its zero-shot TTS capabilities, ability to generate highly natural and personalized speech, and effective handling of speaker nuances with minimal input, the development of a french speaking version is confronted with a landscape of challenges, central to which is the scarcity and variability of quality in available French speech data. This section of the research, integral to my internship, aims to address these challenges, acknowledging that while neural speech synthesis technologies have made significant strides, their effectiveness is heavily contingent on the quality and quantity of the training data.

The evolution of speech synthesis, especially end-to-end systems, has increasingly relied on large quantities of high-quality, dedicated speech data. However, this presents a unique challenge in the context of the French language. The majority of benchmark studies and advancements in speech synthesis have been based on extensive English language datasets like LJSpeech, VCTK, LibriTTS, and ARCTIC, which are predominantly composed of dedicated data specifically created for speech synthesis tasks. In stark contrast, dedicated French speech datasets are limited in volume and scope, often featuring a single speaker or lacking the diversity seen in English datasets [25].

Furthermore, while initiatives using variable quality and non-dedicated data for speech synthesis exist in English, such as the Librispeech corpus, this aspect of speech synthesis is largely unexplored in French. This gap in research and development poses a significant hurdle for adapting VALL-E to French, as we must consider how neural speech synthesis systems perform when trained on data

that is of inferior quality compared to the standards of the field and, crucially, not specifically dedicated to speech synthesis tasks [25].

4.1 Data Acquisition and Quality Constraints

One of the most significant challenges in advancing neural speech synthesis technology, such as VALL-E for French, lies in the acquisition of ample and high-quality training data. This challenge is exacerbated for languages with less global prevalence compared to English. For French, the scarcity of comprehensive and diverse datasets significantly hampers the development of robust and versatile speech synthesis models. While there are notable French datasets like FrenchSiwis [26] and SynPaFlex [27], they are limited in terms of size and diversity, often featuring voices from a narrow range of speakers.

The scarcity of comprehensive French datasets for multi-speaker speech synthesis presents a notable contrast to the abundance of English language resources. While the situation for mono-speaker speech synthesis in French is more favorable, especially for datasets ranging from 1 to 10 hours, this disparity still highlights the need for more inclusive and representative data in speech synthesis models. It underscores the importance of harvesting as much data as possible to enhance the development of robust and versatile speech synthesis technologies for French, aiming to bridge the gap in resource availability between languages.

The research by Guennec et al. [28] offers a promising direction for addressing the challenge of data scarcity in French speech synthesis, particularly in multi-speaker contexts. Their study emphasizes the nuanced process of speaker selection in training datasets, especially when data resources are limited. By focusing on selecting speakers who closely resemble the target voice in characteristics like timbre, pitch, and speaking style, their approach significantly impacts the quality and effectiveness of voice cloning. This method presents a viable solution for enhancing the development of robust and diverse speech synthesis models for French, leveraging the available data to its fullest potential.

This strategic selection enhances the quality and authenticity of the synthesized voice, particularly in scenarios where data is scarce. The key lies in the diversity and representativeness of the chosen voices within the dataset. A well-curated dataset, incorporating a range of dialects, accents, and speech patterns, can lead to more versatile and inclusive speech synthesis models. This approach not only addresses the challenge of limited data but also ensures a broader representation, crucial for developing speech synthesis technologies that can cater to various user needs.

In the context of neural speech synthesis for languages with limited data resources like French, the emphasis is shifting towards effective dataset curation as a means to compensate for data scarcity. Good dataset curation, focusing on quality and relevance of the data, can significantly enhance model performance, even with smaller datasets. During the internship, this approach will be a key area of investigation, alongside efforts to expand the datasets themselves. By concentrating on both refining the curation process and augmenting the volume of available data, the goal is to advance speech synthesis capabilities in underrepresented languages.

4.2 The Issue of Data Representativeness and Bias

Addressing the challenge of data representativeness and bias in neural speech synthesis, particularly for French, involves a multi-faceted approach to dataset curation. Representativeness influences the inclusiveness and accuracy of speech synthesis models, necessitating datasets that encompass

a broad spectrum of dialects, accents, and speech patterns. Dialects in French, which vary in pronunciation, vocabulary, and grammar, include Parisian French, Quebec French, and African French dialects. To mitigate biases and ensure inclusiveness, it's crucial to include linguistic variations across different age groups, genders, and socio-economic backgrounds, without strictly adhering to their proportional representation in the population. [29].

Moreover, the impact of dataset bias on the fairness and effectiveness of machine learning models, including those used in speech synthesis, has been extensively studied. One significant observation is that the fairness of a model can be influenced by how representative the training data is of the diverse linguistic characteristics it aims to capture. For instance, using synthetic data in training can affect the model's bias and fairness.

Different synthetic data generation methods can result in varying levels of bias in models. For instance, PATE-GAN, which focuses on differential privacy, tends to amplify bias compared to methods like CTGAN and CopulaGAN. This amplification in PATE-GAN occurs due to the increased noise introduced for privacy protection, which can distort statistical relationships and elevate correlations between sensitive and non-sensitive attributes. Conversely, CTGAN and CopulaGAN, by generating features with reduced correlation, demonstrate a lesser degree of bias. This variation highlights the critical need to consider the impact of data generation techniques on bias in machine learning models. [30].

Incorporating these insights into the development of neural speech synthesis models for French or other languages with diverse dialects and accents can lead to more representative, fair, and inclusive outcomes. This approach not only addresses the technical challenges of speech synthesis but also aligns with the broader goals of creating equitable and accessible AI technologies.

5 Evaluation Methodology

The evaluation of synthesized speech, a crucial step in model development, presents a complex challenge. The subjective nature of human perception of speech makes it difficult to establish a universally applicable evaluation standard. While subjective methods, such as the Mean Opinion Score (MOS), provide valuable insights into human perception, they are inherently limited by their reliance on individual judgments, which can vary widely. For instance the paper by Kirkland et al. [31] underscores the variability and potential impact of different testing approaches on the perceived quality of synthesized speech. Regrettably, in the current state of the field, subjective testing remains the principal method for assessing speech quality, as there is a lack of objective measures that consistently correlate well with human perception. This reality points to an ongoing need for research and development in this critical area of speech synthesis technology.

The selection of speakers in speech synthesis, particularly those who exhibit similarities to the target voice in aspects such as timbre, pitch, and speaking style, presents a promising research direction. This approach can potentially enhance the quality of synthesized speech and may contribute to the validity and reliability of evaluation methods in text-to-speech systems. However, it's important to note that this is an emerging area of study and not an absolute requirement across the speech synthesis community. Further research and exploration are needed to fully understand its impact and efficacy. The study's findings indicate the necessity for rigorous and transparent evaluation methods in TTS research, especially when dealing with languages that have limited data resources. This aligns with the broader need for diverse, high-quality datasets to foster advancements in speech synthesis technology. Conversely, objective evaluation methods, relying on

quantifiable metrics like Mel-Cepstral Distortion (MCD) or Signal Noise Ratio (SNR), provide consistency but may fail to capture nuances in human perception. This dichotomy necessitates a balanced approach, combining both subjective and objective methods, to accurately evaluate the performance of neural speech synthesis models. Notably, models like MOSNet [32] or WV-MOS [33], designed as automated substitutes for human testing, bear inherent biases. These models, typically trained on English or Japanese data, may show skewed performance, especially when evaluating audio samples that deviate from the typical 3-5 second duration commonly used in MOS tests. This dichotomy between subjective and objective evaluation underscores the need for a balanced approach in accurately assessing neural speech synthesis models.

6 Internship Objectives and Methodology

As we move into the practical phase of this research, the internship aimed to apply advanced voice cloning and style management techniques specifically to the French language, building upon the VALL-E architecture previously discussed. The project centered on adapting the VALL-E model, originally designed for languages like English and Mandarin with abundant and varied datasets, to French—a language with relatively limited data resources.

The first phase of the project involved the replication of the VALL-E model's architecture for English, which served as a baseline for subsequent adaptations. This required an in-depth understanding of the architectural frameworks that underpinned VALL-E, as well as the preparation of a suitable French corpus. Given the model's dependency on large and diverse datasets, the data collection and preparation process was critical and involved the careful selection of high-quality audio data from various sources, as outlined in the following sections.

Once the dataset was curated, the focus shifted to model development and adaptation. The challenge here was not just to replicate the English model but to modify it in a way that accommodated the linguistic and stylistic nuances of French. This involved a significant amount of experimentation with phonemization techniques, as French pronunciation differs considerably from English. The phonemization process was essential for enabling the model to effectively handle the phonetic subtleties of French.

Another major hurdle was managing the high computational cost associated with training large language models. The project involved optimizing training procedures and experimenting with different architectures to reduce the time and resources required while still achieving high-quality voice synthesis. As part of this, Bark [34], a GPT-style generative audio model that is similar to Vall-E while being more forgiving to train, was used as a base for fine-tuning. Special attention was thus given to the unique aspects of the Bark architecture, which was leveraged for its advanced capabilities in generating expressive and natural-sounding speech.

The final stages of the project involved rigorous evaluation and validation of the adapted model's performance. Various objective and subjective evaluation methods were employed, including the use of MOS (Mean Opinion Score) tests to assess the naturalness and intelligibility of the synthesized speech. The model's performance was also compared against the original English version to gauge the effectiveness of the adaptation process.

7 Data Collection and Preparation

The training and fine-tuning of speech synthesis models like Vall-E necessitate the use of extensive, diverse, and high-quality datasets. In the original Vall-E paper, the authors utilized the "Libri-Light" dataset for English language training. In contrast, our reproduction of Vall-E used the "LibriTTS" dataset for English training, and a multilingual version with an architecture similar to that of Vall-E, called Bark was fine-tuned using the Common Voice dataset. Below, we provide a detailed comparison and analysis of these datasets to highlight their suitability and distinct features.

7.1 The Libri-Light Dataset

The Libri-Light dataset is an extensive corpus designed to benchmark automatic speech recognition (ASR) systems with limited or no labeled data. This dataset, derived from the LibriVox project, contains over 60,000 hours of English speech. Its large size and the fact that it includes both clean and noisy audio make it ideal for unsupervised and semi-supervised learning approaches. The dataset is divided into several subsets to facilitate research under varying conditions of data availability. Ref. Table 1:

- Unlabeled Speech Training Set: This is the largest portion of the dataset, consisting of over 57,706 hours of speech across 9,860 books, amounting to 219,041 files. The subset is further divided into smaller partitions:
  - The "unlab-60k" subset contains 57,706 hours of data from 7,439 speakers.
  - The "unlab-6k" subset offers a reduced version with 5,770 hours and 1,742 speakers. The smallest subset, "unlab-600," includes 577 hours from 489 speakers.
- Limited Resource Training Set: This subset provides labeled data but in limited quantities, supporting experiments in low-resource ASR:
  - The "train-10h" set includes 10 hours of speech, evenly split between male and female speakers.
  - The "train-1h" and "train-10m" subsets offer progressively smaller amounts of data, with 1 hour and 10 minutes of labeled speech, respectively.
  - Dev & Test Sets from LibriSpeech: For evaluation purposes, Libri-Light includes development and test sets derived from the LibriSpeech corpus. These sets include both "clean" and "other" (Noisy) conditions, with approximately 5.4 hours of data in each condition and contributions from 30 to 40 speakers.
  - Unaligned Text Training Set: Additionally, the dataset includes an unaligned text training set derived from the LibriSpeech language model (librispeech-LM) corpus, which contains 800 million tokens and a vocabulary of 200,000 words.

The primary advantage of Libri-Light lies in its scale and diversity, providing a rich resource for training robust models capable of generalizing across various speaking styles and acoustic conditions. However, it's important to note that while Libri-Light is extensive, its primary focus is on ASR rather than TTS (Text-to-Speech), which might limit its direct applicability to speech synthesis tasks without significant preprocessing and adaptation.

Table 1: Libri-Light Dataset Statistics. Directly from the LibriLight paper [23]

Subset	Hours	Books	Files	Per-Speaker Hours	Total Speakers
Unlabeled Speech Training Set					
unlab-60k	57706.4	9860	219041	7.84	7439
unlab-6k	5770.7	1106	21327	3.31	1742
unlab-600	577.2	202	2588	1.18	489
Subset	**Hours**	**Per-Speaker Minutes**	**Female Speakers**	**Male Speakers**	
Limited Resource Training Set					
train-10h	10	25	12	12	24
train-1h	1	2.5	12	12	24
train-10m*	10min	2.5	2	2	4
Dev & Test Sets from LibriSpeech					
dev-clean	5.4	8	20	20	40
dev-other	5.3	10	16	17	33
test-clean	5.4	8	20	20	40
test-other	5.1	10	17	16	33
Subset	**Tokens**	**Vocabulary Size**			
Unaligned Text Training Set					
librispeech-LM (in-domain)	800M	200K			

7.2 The LibriTTS Dataset

The LibriTTS dataset [35], derived from LibriSpeech, is specifically designed for Text-to-Speech (TTS) applications and consists of approximately 585 hours of English speech data at a 24kHz sampling rate. Unlike Libri-Light, LibriTTS includes precise alignment between text and speech, making it highly suitable for TTS training. The dataset is categorized into train-clean, train-other, dev-clean, dev-other, and test-clean subsets, with clean versions containing high-quality, noise-free recordings, and other versions including more challenging data with varied levels of noise and speaker accents.

The creation of the LibriTTS dataset involved an extensive pipeline to ensure the quality and usability of the data for TTS tasks. The process began with aligning long-form audio recordings with their corresponding texts and segmenting them into sentence-level units. To maintain high-quality alignments, utterances with mismatches between the audio and text—due to errors like inaccuracies in the text, reader-induced modifications, or text normalization issues—were systematically excluded. Text preprocessing was carried out by splitting book-level texts into paragraphs and sentences, followed by normalizing non-standard words using a weighted finite state transducer-based text normalizer.

In handling the unprocessed audio and text materials, the second phase involved extracting multi-paragraph text that matched each chapter-level audio file. This was achieved by running automatic speech recognition (ASR) on the chapter audio and matching the transcription with the corresponding book text. The critical alignment between the audio and text was conducted using a modified version of YouTube's "auto-sync" feature, which employs a bidirectional LSTM-based acoustic model and a miniature tri-gram language model to force-align the transcript to the audio.

Post-processing steps were meticulously implemented to refine the final dataset. This involved filtering out sentences with potential errors, normalizing audio polarity, removing excessive silences, and calculating signal-to-noise ratios (SNR) to exclude low-quality recordings. As a result of these stringent filtering criteria, the final corpus contained a significantly smaller volume of data than the original LibriSpeech corpus—approximately 60% less—primarily due to strict alignment and SNR-based filtering requirements.

The carefully curated LibriTTS dataset stands out for its high-quality, well-annotated speech data, making it a valuable resource for TTS model training. However, its smaller size compared to Libri-Light may limit its ability to capture the full diversity of English speech, which could

necessitate the use of supplementary data to enhance the robustness of TTS models trained on this corpus.

7.3 The Common Voice Dataset

The Common Voice dataset [36], developed by Mozilla, is a massively multilingual corpus aimed at improving the accessibility and inclusivity of speech recognition technology. For our purposes, the French subset of this dataset was utilized to fine-tune the Bark model. Common Voice collects speech data through crowd-sourcing, with speakers from various linguistic backgrounds contributing recordings. The dataset for French alone contains over 1,000 hours of validated speech data, offering a broad range of accents, dialects, and recording conditions.

The strength of the Common Voice dataset lies in its diversity and inclusiveness. The dataset covers a wide variety of speakers and recording conditions, making it particularly valuable for building models that need to perform well across different dialects and accents. However, due to the crowd-sourced nature of the data, the audio quality can be inconsistent, with some recordings containing background noise or other artifacts. This variability can be both a challenge and an advantage, depending on the robustness required of the final model.

Each of these datasets presents unique strengths that cater to different aspects of speech synthesis and recognition. Libri-Light provides an extensive, albeit ASR-focused, corpus suitable for large-scale unsupervised learning. LibriTTS, on the other hand, offers high-quality, well-annotated data specifically tailored for TTS applications, albeit on a smaller scale. Finally, Common Voice provides a diverse and inclusive dataset with multilingual support, crucial for building models that generalize across different linguistic and cultural contexts, despite potential inconsistencies in audio quality.

In the context of our work, the selection of LibriTTS for English and Common Voice for French was guided by the specific needs of the TTS and multilingual capabilities we aimed to develop. LibriTTS's alignment of text and speech ensured effective English training for Vall-E, while the diversity of Common Voice enabled Bark to handle the rich variety of French speech. These choices underscore the importance of dataset characteristics in shaping the performance and generalization capabilities of speech synthesis models.

8 Model Development/Adaptation

In our project, we sought to adapt and utilize Vall-E for both English and French language applications. Initially, we recreated Vall-E for English by leveraging an unofficial adaptation of the model, trained on the LibriTTS dataset. This dataset, with its high-quality text-to-speech alignment, allowed us to effectively replicate Vall-E's capabilities for English speech synthesis. However, when attempting to extend this approach to the French language using the Common Voice dataset, several significant challenges arose.

8.1 Phonemization

One of the primary issues encountered during the adaptation of Vall-E for French was related to phonemization—the process of converting text into phonemes, which are the distinct units of sound in a language. Vall-E's architecture, which is optimized for languages like English with well-established phoneme datasets, struggled with the phonetic intricacies of French. The Common Voice

dataset, while rich in diversity and volume, presented challenges in this regard, as French phonemization required more complex and nuanced handling. The model's existing phoneme processing mechanisms were not easily transferable to French, leading to inaccuracies in speech synthesis.

8.2 High Training Time

Another significant challenge was the extensive training time required by Vall-E. Training the model on the Common Voice dataset on our local machine, equipped with two Nvidia RTX 4090 GPUs (each with 24GB VRAM)took approximately one and a half weeks, assuming no interruptions due to system maintenance or other issues. This long training time posed logistical challenges, particularly given the iterative nature of model development, where frequent adjustments and retraining are often necessary to fine-tune the model's performance. On the Jean-Zay supercluster, which I was only able to access nearly three months into the internship, the training time was reduced to over two days, utilizing five Nvidia A100 GPUs, each with 80GB VRAM. These time constraints significantly impacted the speed of experimentation and model optimization throughout the project.

Due to these difficulties, we decided to explore an alternative approach for French speech synthesis by using Bark, a model developed by Suno.ai. Bark shares some similarities with Vall-E and other models in the field, as it employs GPT-style transformers to generate audio. However, Bark differentiates itself by not relying on phonemes. Instead, it uses high-level semantic tokens derived directly from the initial text prompt, allowing for a broader generalization beyond just speech synthesis.

8.3 Architecture of Bark

Bark's architecture involves multiple sub-models, each playing a critical role in the generation of audio from text:

- Encodec : Similar to Vall-E, Bark uses Encodec to compress audio data into a smaller, more manageable form, which is then used for further processing.
- Semantic Model: The semantic model in Bark is a GPT-2-like causal autoregressive transformer that processes the initial text prompt. This model, consisting of 80 million parameters, converts the text into semantic tokens that capture the essence of the input. Unlike Vall-E, which relies on phonemes, Bark's use of semantic tokens allows it to handle a wider range of inputs, including non-speech sounds and music.
- Coarse Acoustics Model: Following the semantic model, Bark uses a coarse acoustics model, also GPT-2-like, to predict the coarse acoustic tokens that represent the general structure of the audio waveform. This model processes the semantic tokens and outputs tokens corresponding to the first two audio Encodec codebooks, which capture the broader characteristics of the sound.
- Fine Acoustics Model: The final component is the fine acoustics model, a non-causal GPT-like autoencoder. This model refines the audio by predicting the tokens for the remaining six codebooks, which capture the finer details of the sound. Together, the coarse and fine models allow Bark to generate a detailed and realistic audio waveform from the initial semantic tokens.

8.4 Advantages and Limitations of Bark

Bark's architecture offers several advantages over traditional phoneme-based models like Vall-E, particularly in its ability to handle diverse input types and its flexibility in generating creative outputs. However, these strengths come with trade-offs:

- High Variance in Output: Due to its generative nature, Bark can produce outputs with significant variance. The multiple GPT-style components introduce a level of randomness that can lead to outputs with varying quality and coherence. This variance can be problematic, especially when consistent and high-quality speech synthesis is required.
- Output Duration Limit: Bark is optimized for generating outputs with a maximum duration of around 14 seconds. This limitation makes it unsuitable for processing long-form text inputs directly, restricting its use in applications that require extended audio outputs.

9 Preprocessing, Training and Optimization

9.1 Data Preprocessing

Before fine-tuning the Bark models, we performed a crucial preprocessing step to ensure that the audio data was in a consistent format compatible with the original Bark model's requirements. All audio data was converted to a standardized format of 24kHz sampling rate, mono channel, and 16-bit depth. This decision was driven by two primary factors: first, the need to align with the format used in the pre-trained Bark models to maintain compatibility and ensure effective fine-tuning; and second, guidance from the Vall-E paper, which employed similar preprocessing settings. By adhering to these established standards, we aimed to optimize the model's ability to learn from the data and to replicate the success observed in prior research.

Additionally, as will be discussed in a subsequent section, a volume normalization preprocessing step was introduced as an experimental measure to further enhance the training process. Initially, the average volume of the training data was approximately -24 dB. To standardize this across all audio samples, the data was normalized to 6 dB, a level typically used for clear studio-quality audio. This adjustment aimed to provide a more consistent input for the model, potentially improving the clarity and effectiveness of the synthesized speech during fine-tuning.

9.2 Training Procedure

In this study, the Bark models were trained using a typical data subdivision approach, commonly employed in machine learning tasks, where the dataset was divided into three parts: 70% for training, 10% for validation, and 20% for testing. This partitioning strategy ensures that the models are exposed to a sufficient amount of data for learning while retaining a separate validation set for tuning and a distinct test set for final evaluation.

The training procedure was designed to optimize the performance of the semantic, coarse, and fine models previously mentioned. The base models were acquired from Hugging Face [34], then fine-tuned over five epochs, with the AdamW optimizer configured with a learning rate of 1e-5. To prevent overfitting, a weight decay of 0.01 was applied, and a gradient clipping strategy with a maximum norm of 1.0 was employed.

In addition to the standard data, we also trained our models on volume-normalized data and enhanced volume-normalized data. This was done to assess the impact of different preprocessing techniques, particularly in handling variations in audio volume that are common in real-world datasets.

The model architecture included several advanced features aimed at enhancing training efficiency and performance. Mixed precision training using bfloat16 was utilized, leveraging the capabilities of modern GPUs to accelerate computations while conserving memory. Additionally, a LoRA (Low-Rank Adaptation) mechanism was integrated into the model to allow efficient fine-tuning of pre-trained parameters, particularly beneficial given the model's large scale. The LoRA module was configured with a dimension of 64, a scaling factor of 1, and a dropout rate of 0.1.

To optimize the learning process further, a linear learning rate scheduler was implemented, with 60 warmup steps to gradually ramp up the learning rate, stabilizing the early stages of training. The overall training steps were determined dynamically, based on the size of the training dataset and the specified number of epochs, ensuring that the model was trained for a sufficient number of iterations.

During training, the model's performance was continuously monitored on the validation set. The validation process involved computing the cross-entropy loss between the predicted and actual semantic tokens, allowing for fine-tuning of the model and early detection of potential overfitting.

The combination of these strategies—advanced optimization techniques, careful monitoring, and a robust data subdivision—ensured that the model was trained effectively, with attention to both computational efficiency and generalization to unseen data. This approach aligns with best practices in the field, leveraging both traditional methods and cutting-edge innovations to achieve optimal model performance.

However, despite the effectiveness of the training process and the achievement of low loss values, the initial fine-tuning attempts were unsuccessful. The fine-tuned Bark model consistently failed to generate coherent speech, often producing either empty audio clips or noise. This issue prompted an investigation into the underlying causes, where it became apparent that phonemization was a significant factor.

Bark, unlike some other models, does not incorporate phonemes directly into its architecture. Given the differences between French and English pronunciation, this lack of phonemization was identified as a potential hindrance to the model's ability to accurately generate French speech. Various strategies were tested to incorporate phonemes into the text model, including concatenating the latent representation of phonemes with the existing latent text representation. However, these methods did not yield the desired improvements.

An alternative approach was then implemented, drawing on insights from the XPhoneBERT paper[37]. This approach involved using a BERT-like sequential latent representation derived directly from phonemes, while excluding the original text data. The reasoning was that phonetic representation inherently encapsulates both semantic and phonetic information, which could be particularly advantageous for speech synthesis in French. By employing the pre-trained model from XPhoneBERT, the fine-tuned Bark model demonstrated notable improvements in generating speech, particularly in terms of naturalness and prosody.

This adjustment underscores the importance of accounting for linguistic characteristics in multilingual text-to-speech models and illustrates the effectiveness of phoneme-based approaches in addressing language-specific challenges.

10 Evaluation and Validation

In this section, we present a comprehensive evaluation of our fine tuned Bark model. The evaluation is divided into two primary components: objective metrics and subjective assessments. The objective evaluation involves quantifiable measures that provide insights into the performance of the models based on specific criteria such as accuracy, intelligibility, and quality of the generated speech. These metrics are crucial for understanding how well the model performs in controlled, measurable aspects and are often used to compare against baseline models or other state-of-the-art approaches.

However, TTS systems must not only perform well according to objective standards but also meet the subjective expectations of listeners, the standards of which may differ considerably among each other. Therefore, we complement our objective analysis with subjective evaluations, most notably the Mean Opinion Score (MOS) test. The MOS test is a widely-used method to assess the naturalness and overall quality of synthesized speech by having human listeners rate the audio samples. This combination of objective and subjective evaluations provides a holistic view of the model's performance, ensuring that the synthesized speech is both technically sound and pleasing to the end-users.

In the following subsections, we will delve into the specific methodologies used for each type of evaluation, the results obtained, and the implications these results have for the overall performance of the model.

10.1 Objective Measures

In this subsection, we focus on the objective quality measurement of our model's performance, employing a series of established metrics to quantitatively assess the quality of the synthesized speech. Our evaluation methodology involves a comparative analysis across multiple configurations of our synthesized data. Specifically, we compare the ground truth data with synthesized data generated by Bark under different conditions. First, the model was fine-tuned on non-normalized data, followed by fine-tuning on volume-normalized data. Additionally, we explored whether applying audio restoration techniques to both the training data and/or the ground truth data could further improve the model's performance.

To ensure a thorough evaluation, we utilized four key objective metrics all based on the MOS(Mean Opinion Score) metric:

- **WV-MOS** (Waveform Mean Opinion Score): An objective speech quality measure based on direct MOS score prediction by a fine-tuned wave2vec2.0 model [38].
- **UTMOS** (Universal Text-to-Speech MOS): A metric based on ensemble learning, combining strong and weak learners. The strong learners are obtained by fine-tuning self-supervised learning (SSL) models, while the weak learners use non-neural machine learning methods to predict scores from SSL features [39].
- **DNSMOS** (Deep Noise Suppression MOS): An objective perceptual speech quality metric-that provides MOS estimates and is specifically designed to evaluate Deep Noise Suppression (DNS) methods [40].
- **SIGMOS** (Signal MOS): A newly developed objective metric that assesses full-band audio, introduced in the ICASSP 2024 Speech Signal Improvement Challenge [41].

By comparing the scores obtained under these different conditions, we aim to gain insights into the effects of normalization and audio restoration on the overall quality of synthesized speech. This comparative analysis helps us determine the most effective preprocessing steps for enhancing the performance of French voice cloning models according to objective standards.

In the following sections, we will present the results obtained from this evaluation, providing a detailed analysis of each metric and discussing the implications of our findings.

Before diving into the results of the synthesized speech, it is crucial to examine the characteristics of the original data to establish a baseline for comparison.

10.2 Original Data Analysis

The scatter plots in Figure 6 showcase the distribution of the WV-MOS and DNSMOS scores for the original dataset. The analysis of these plots reveals several key insights:

- WV-MOS Distribution: nlike DNSMOS, the WV-MOS scores are more concentrated between the 2 to 4 range as can be seen in Figure 6. This concentration suggests that, for most of the original samples, the perceived naturalness of the waveforms was moderate to good. However, there are some negative outliers in the WV-MOS scores. These negative values likely indicate samples cluttered with external noise, a phenomenon that makes sense given the highly varied quality and clarity of the Common Voice data. In these cases, having personally listened to the audio clips, the WV-MOS model doesn't seem perform correctly, resulting in these negative outlier values.

- DNSMOS and SIGMOS Variability: In contrast, the DNSMOS and SIGMOS scores are more varied throughout the entire spectrum. This variability is expected, as both metrics are indicators of the clarity of the speech. The broad range in DNSMOS scores, from 1.0 to 4.5, reflects the presence of both clear and noisy samples within the dataset. Similarly, SIGMOS, which focuses on the signal integrity and intelligibility, shows a wide distribution, further confirming the diverse quality of the original audio data.

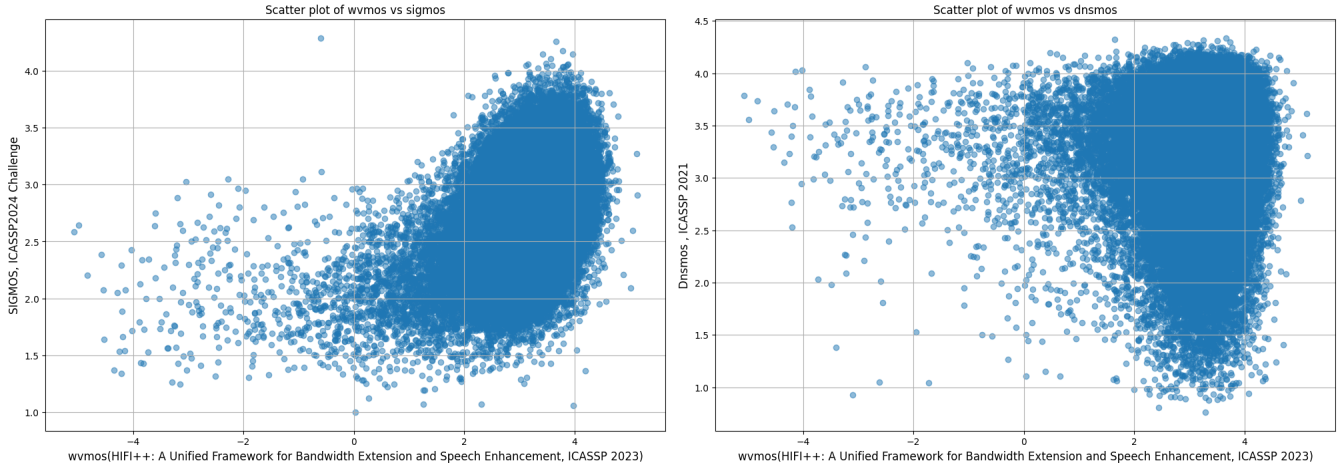

Figure 6: Scatter plots of SIGMOS vs WV-MOS and DNSMOS vs WV-MOS scores on the French CommonVoice dataset.

These initial observations set the stage for understanding how different preprocessing techniques, such as normalization and audio restoration, affect the synthesized speech's quality. By establishing this baseline, we can better appreciate the improvements or drawbacks introduced by our preprocessing strategies.

10.3 Comparison of Original and Synthesized Data

In this evaluation, three categories of data are compared: Ground Truth Data, which refers to the original recordings used for evaluation; Synthesized Data, which is the output generated by the model trained without any volume normalization; and $Synthesized_N$ Data, which represents the synthesized audio generated by the model trained on volume-normalized data.

As shown in Table 2, the ground truth data generally scores better across all metrics compared to the synthesized data. The DNSMOS score for the original data averages 3.241, indicating a relatively high level of clarity and noise suppression. This is higher than the 2.839 average DNSMOS for the synthesized data, which suggests that the synthesized speech, in its unaltered form, struggles with noise, likely due to the inherent variability and noise present in the training data. This finding is consistent with expectations; synthesizing speech from data with varied quality and clarity—like the Common Voice dataset—can introduce noise, affecting the DNSMOS score.

The same trend is observed in the SIGMOS and UTMOS scores. The original data shows higher averages (2.758 for SIGMOS and 2.504 for UTMOS) than the synthesized data (2.222 and 1.778, respectively). These differences indicate that the synthesized speech is less intelligible and has a lower overall perceived quality. This decline in performance can be attributed to the challenges of generating high-quality, natural-sounding speech from diverse and potentially noisy training data.

The WV-MOS, which focuses on the naturalness of the waveform, also reflects this trend. The original data has a WV-MOS of 3.136, while the synthesized data scores lower at 2.710. Notably, the WV-MOS has a higher standard deviation (0.747) compared to the other metrics, suggesting greater variability in the perceived naturalness of the synthesized speech. This variability could be due to the complex nature of the original dataset, which may contain instances of poor-quality speech that the model struggles to synthesize accurately.

10.4 Impact of Normalization

When the synthesized data ($Synthesized_N$) is generated from a model trained on volume-normalized data, there is a notable improvement across all metrics as illustrated in Table 2. The DNSMOS score for $Synthesized_N$ is 3.226, nearly matching the score of the original data. This indicates that volume normalization during training significantly reduces noise in the synthesized speech, leading to a clearer output that is almost as good as the original recordings.

Similarly, the SIGMOS for $Synthesized_N$ improves to 2.529 from 2.222 in the non-normalized synthesized data. Although this is still slightly lower than the original data's SIGMOS, the improvement suggests that normalization helps in preserving the signal quality, making the synthesized speech more intelligible. The UTMOS also shows a recovery, improving to 2.247 from 1.778, which reflects an overall enhancement in the perceived quality of the speech output.

The WV-MOS for $Synthesized_N$ is 2.987, which is closer to the original data's score of 3.136. This improvement indicates that normalization contributes to generating more natural-sounding waveforms, although the synthesized speech still doesn't fully match the original data's naturalness. The slightly lower standard deviation (0.647) for $Synthesized_N$ in WV-MOS suggests that

normalization may also reduce variability in the naturalness of the synthesized output.

	DNSMOS	**SIGMOS**	**UTMOS**	**WV-MOS**
Ground Truth	3.241	2.758	2.504	3.136
	(0.506)	(0.392)	(0.566)	(0.694)
Synthesized	2.839	2.222	1.778	2.71
	(0.536)	(0.365)	(0.464)	(0.747)
Synthesized$_N*$	3.226	2.529	2.247	2.987
	(0.505)	(0.445)	(0.494)	(0.647)

Table 2: This table provides a comparison of the average scores and standard deviations for four key metrics: DNSMOS, SIGMOS, UTMOS, and WV-MOS, across the two different training data conditions. $*$: this data refers to inference audio generated by the Bark model, which was trained on volume-normalized data. The values in parentheses represent the standard deviation for each metric, providing an indication of the variability in the scores.

10.5 Mel Cepstral Distortion

While the previously discussed metrics, such as WV-MOS and DNSMOS, provide objective measures of perceived speech quality based on MOS scales, it is also useful to evaluate the spectral differences between the synthesized and ground truth audio. For this purpose, we turn to the Mel Cepstral Distortion (MCD) metric, which quantifies the difference between the spectral properties of two audio signals and is often employed in speech synthesis and voice conversion tasks to measure the similarity between the generated and the ground truth speech.

Table 3 presents the MCD results for both the synthesized audio and the synthesized audio generated with volume normalization, offering a complementary view of the similarity between the generated speech and the ground truth.

For the standard synthesized data, the average Mel Cepstral Distortion (MCD) is 6.559, calculated from 319,017 frames. In contrast, when examining the synthesized data with a normalization preprocessing step, the average MCD significantly improves to 5.536, based on 173,226 frames. MCD values in this range suggest that the synthesized speech might be perceived as fairly close the target in terms of spectral quality 3. Moreover, this reduction in MCD reflects the positive impact of volume normalization during training, resulting in synthesized speech that is spectrally closer to the original. The notable improvement —over 18%— highlights that normalization not only enhances noise suppression and signal clarity, as observed in earlier metrics, but also effectively reduces the spectral distortion in the speech signal.

	Average MCD	**Number of frames**
Synthesized	6.559	319017
Synthesized$_N*$	5.536	173226

Table 3: The table provides a comparison of Mel Cepstral Distortion scores, across two different training data conditions.

10.6 Discussion

The results from Table 2 indicate that volume normalization during training ($Synthesized_N$) has a positive impact on the quality of synthesized speech. The MCD results in Table 3 further emphasize that impact. The normalized model produces synthesized speech with significantly less spectral distortion, aligning more closely with the original recordings. Across all objective metrics, normalized synthesized data shows substantial improvements compared to non-normalized synthesized data. However, despite these improvements, the synthesized audio based on the model trained on normalized data still does not entirely reach the quality of the original recordings, suggesting that while normalization during training is beneficial, it is not a complete solution. Further refinements in preprocessing, model architecture, or training strategies may be necessary to fully close the gap between synthesized and original speech quality.

10.7 Subjective Evaluation - MOS Test

To assess the perceived quality of the speech generated by our models, a MOS test was also conducted. This test focused on three categories of audio samples: natural speech, synthesized speech, and synthesized speech with normalization preprocessing. A total of 900 samples were included in the test, with 300 randomly selected samples from each category. The samples were chosen without filtering for specific speakers or audio quality, ensuring that the evaluation reflected the overall performance of the models across a diverse set of data.

10.8 Test Design and Sampling Strategy

The MOS test was distributed internally among the team members that are experts in the field. Given the limited sample size and the internal distribution of the test, there is a concern that the sampling might not naturally balance across the three classes due to the law of large numbers not applying as strongly in this context. To address this, the test was designed with an adaptive sampling strategy that adjusted the probability of sampling from each category based on the distribution of samples already seen by each participant.

Initially, the test employed a uniform distribution, assigning equal probabilities to each of the three categories. As the test progressed, the sampling strategy was dynamically adjusted to favor categories that had been under-sampled, ensuring that each category received adequate representation in the final results. This approach helped mitigate any potential bias that could arise from unequal sampling and ensured that the MOS scores provided a fair assessment of each category.

The sampling adjustments were based on the number of samples previously viewed by each participant. If a category was under-represented in a participant's history, the probability of selecting samples from that category was increased. Conversely, categories that had been more frequently sampled were less likely to be selected in subsequent rounds. This adaptive strategy was designed to achieve a more balanced distribution of evaluations across the three categories, thereby enhancing the reliability of the MOS results.

10.9 MOS Test Results

The results of the Mean Opinion Score (MOS) test, as shown in Table 4, provide a direct comparison of the perceived quality of three types of speech data: the original (ground truth) recordings,

	MOS	Standard Deviation
Ground Truth	4.350427	1.011283
Synthesized	2.432099	1.068327
Synthesized$_N$*	2.456897	0.963619

Table 4: The table provides the Mean Opinion Score test results across two different training data conditions.

synthesized speech without normalization, and synthesized speech with normalization preprocessing.

The original recordings, labeled as "Ground Truth," achieved an average MOS of 4.350 with a standard deviation of 1.011. At first glance, this high MOS score might seem inconsistent with the mixed quality of Common Voice data, which is often characterized by background noise and varying recording conditions. However, the relatively large standard deviation indicates a wide range of participant opinions, suggesting that testers did not uniformly agree on the quality of the audio.

This discrepancy could stem from the fact that different testers may prioritize different aspects of speech quality. For instance, some participants may focus primarily on whether the voice sounds human and intelligible, potentially overlooking background noise or recording artifacts. Others may weigh factors like noise more heavily in their evaluations, leading to lower scores. The presence of outliers in the ratings—likely caused by these differences in perception—explains why the average MOS is higher than expected, as the score reflects a blend of differing priorities among the testers.

In contrast, the synthesized speech without any normalization preprocessing received a significantly lower average MOS of 2.432, with a standard deviation of 1.068. This score highlights the challenges faced by the model in generating speech that matches the naturalness and quality of the original recordings. The higher standard deviation indicates greater variability in the perceived quality of these synthesized samples, suggesting that while some synthesized outputs may be closer in quality to the original, others fall short, leading to a broader range of participant responses.

For the synthesized speech generated by the model trained with normalization preprocessing ($Synthesized_N$), the average MOS was 2.457, only slightly higher than the score for the non-normalized synthesized speech. However, the lower standard deviation of 0.964 suggests that normalization contributed to a more consistent quality across the samples. While normalization seems to have stabilized the perceived quality, the marginal improvement in the average MOS indicates that it did not significantly enhance the naturalness or overall quality of the synthesized speech.

Interestingly, this small gain in subjective evaluation contrasts with the more noticeable improvements in objective metrics like MCD, where normalization led to a clearer reduction in spectral distortion. This disparity may point to potential biases in the way testers evaluated the samples. There are several possible explanations for this contradiction:

- **Focus on Perceptual Naturalness**: Testers may have prioritized naturalness in their evaluations—i.e., whether the speech sounded "human"—over other factors like background noise or gain inconsistencies. This could explain why normalization, which reduces technical artifacts such as volume inconsistency, had less of an impact on their subjective ratings.

- **Neglect of Background Artifacts**: Since the MOS test explicitly asked participants to assess the quality of speech, testers might have overlooked issues like background noise or

gain variability in both natural and synthesized samples. This would especially affect the ratings for the ground truth data, as some participants may not have penalized noise or other recording issues as heavily as expected.

- **Combination of Perception Biases**: The testers' focus on speech quality in terms of naturalness, coupled with a tendency to overlook non-speech-related artifacts, could have led to the observed disconnect between subjective MOS results and objective evaluations. This suggests that while normalization helped address technical issues, the testers may have been more influenced by broader perceptual characteristics, leading to only minor improvements in their subjective assessments.

Overall, the results highlight that while normalization contributes to more consistent quality in synthesized speech, it may not sufficiently address the broader perceptual factors that drive subjective evaluations. Further refinements in synthesis techniques, possibly addressing prosody, clarity, and other perceptually salient aspects, are likely required to achieve significant gains in perceived speech quality.

10.10 Variability in MOS Scores and the Impact of Audio Quality

The variability in the MOS (Mean Opinion Score) values, as illustrated in the Figure 7, highlights the significant fluctuations in perceived audio quality across different samples within each category—natural, synthesized, and synthesized with normalization preprocessing. This variability is in every class, where the MOS scores span the entire spectrum from 1.0 to 5.0.

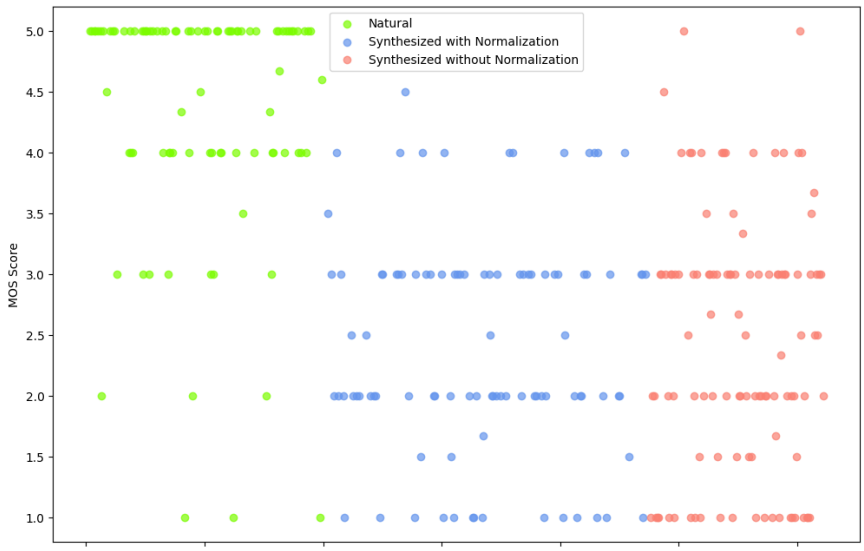

Figure 7: Distribution of the MOS Across Natural, Synthesized, and Synthesized with Normalization Audio Samples.

Upon closer examination, the synthesized audio samples, regardless of whether they were produced using a model trained with normalized data or not, exhibit a strong correlation between the quality of the ground truth recordings and their respective MOS scores. Specifically, the synthesized samples that received the lowest MOS scores often correspond to ground truth audio that suffered

from poor quality due to external noise, low bitrate, or other recording imperfections. Conversely, the highest scoring synthesized samples were generally based on ground truth recordings of superior quality, with minimal noise and clear speech.

This observation underscores a critical challenge in voice: the fidelity of the output is not solely dependent on the synthesis model itself but is also heavily influenced by the quality of the input data and training data alike. When the ground truth audio is of high quality, the synthesized output is more likely to be perceived as natural and clear, leading to higher MOS scores. However, when the training data and/or input data is marred by noise or other issues, even sophisticated models struggle to produce high-quality output, resulting in lower MOS scores.

This variability points to a potential area for enhancement—specifically, the application of audio enhancement or denoising techniques during the training and prior to the synthesis process as well. By improving the quality of the input data, it may be possible to achieve more consistent and higher-quality synthesized speech, as evidenced by the improved MOS scores for samples based on high-quality ground truth audio.

11 Audio Denoising and Enhancement

The observed variability in the MOS scores, particularly among synthesized audio samples, can be traced back to the varying quality of the ground truth recordings used both in training and testing. This variability underscores the necessity for robust preprocessing techniques, such as denoising and audio enhancement, to mitigate the impact of poor-quality input data. To address these issues, we implemented a denoising and enhancement pipeline based on the methodology used by Resemble-enhance [42], which has shown effectiveness in improving speech quality under challenging conditions.

11.1 Denoising and Enhancement Pipeline

The Denoiser module is designed to separate speech from unwanted background noise within an audio sample. This separation is achieved using a UNet-based model, a type of neural network architecture commonly used in tasks requiring precise localization and segmentation. The model operates on the complex spectrogram of the noisy audio, predicting a magnitude mask and a phase rotation. The magnitude mask selectively attenuates the noise while preserving the speech signal, and the phase rotation aligns the phase information of the spectrogram, resulting in cleaner audio output. This approach is consistent with the methodology described in the AudioSep [43] framework. Following denoising, the Enhancer module further refines the audio by addressing any remaining distortions and extending the bandwidth of the speech signal, thereby improving its perceptual quality. The enhancement process involves a latent conditional flow matching (CFM) model, which functions in two distinct stages:

- The first stage employs an autoencoder that compresses a clean Mel spectrogram into a compact latent representation. This representation is then decoded and converted back into a waveform using a vocoder.
  - **GAN-Based Vocoder Losses**: To train this autoencoder-vocoder system, Generative Adversarial Network (GAN)-based vocoder losses were used. These include multi-resolution Short-Time Fourier Transform (STFT) losses, which are calculated by com-

paring the STFT of the generated and target audio at different time and frequency resolutions.

  - **Discriminator Losses**: Additionally, the model is trained with discriminator losses, derived from a discriminator network that tries to distinguish between real and generated audio.

- After training the autoencoder in Stage 1, its parameters are frozen, and the focus shifts to training the CFM model. This model is conditioned on a blend of Mel spectrograms derived from both the noisy and denoised versions of the audio. This blended Mel spectrogram is used to predict the latent representation of the clean speech.

  - The **CFM model** is based on a non-causal WaveNet architecture, which processes audio in a sequential manner but without assuming a fixed direction of time, allowing it to better model the nuances of speech signals.
  - The training of the CFM model employs the I-CFM (Implicit Conditional Flow Matching) objective. This objective guides the model to transform an initial point in the latent space, which is a blend of the noisy latent representation and Gaussian noise, into a point that closely matches the distribution of clean speech latents. This transformation ensures that the enhanced audio is as close as possible to the clean reference, even if the original input was degraded.

This denoising and enhancement pipeline was applied to both a randomly selected set of audio clips from 100 different speakers and the lowest scoring clips from the MOS test discussed in the previous section. The results, presented in subsequent tables, demonstrate the pipeline's effectiveness in improving the quality of audio recordings, particularly in instances where the original recordings were of poor quality.

11.2 Experiments

Before presenting the experimental results, an important question must be addressed: Could there be a speaker bias affecting the outcomes, given the observed variability in the quality of synthesized audio? In other words, are certain speakers' voices inherently better suited for voice cloning? The answer to this question remains uncertain. The variability in results -while it can be- is not necessarily due to factors like gender bias in training data or accent bias (such as differences between speakers from France and Quebec) as it is likely further influenced by the crowd-sourced nature of the Common Voice dataset, where each speaker uses their own device to record their voice, resulting in significant variability in audio quality. As such, it is difficult to determine whether certain voices are inherently better suited for cloning, or if the differences stem from the recording conditions.

This discrepancy is highlighted in Figure 8, which illustrates the variability in objective metric scores for each of the 100 randomly selected speakers. As the figure shows, there is a considerable range in the performance of the objective metrics across different speakers, suggesting that the quality of the initial recordings has a notable impact on the subsequent results.

Table 5 presents the results for the randomly selected set of 100 speakers' audio clips, showing the objective metrics for ground truth, denoised, and enhanced audio samples.

For the randomly selected audio clips, the ground truth scores establish a moderate quality baseline, with a DNSMOS of 3.342. After denoising, there was a slight decrease in DNSMOS to

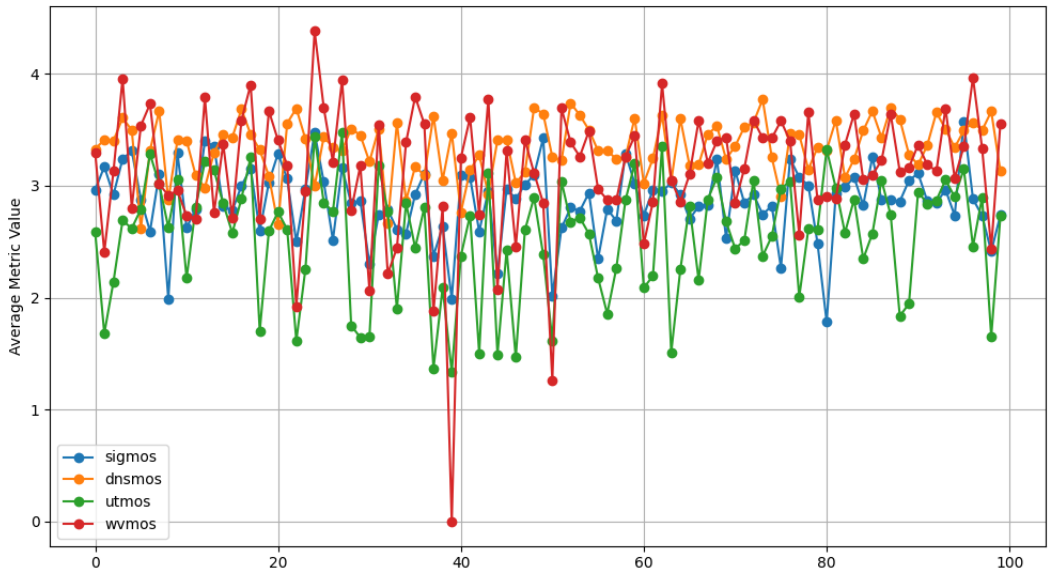

Figure 8: Variability of Objective Metric Scores Across 100 Selected Speakers.

	DNSMOS	**SIGMOS**	**UTMOS**	**WV-MOS**
Ground Truth	3.342	2.86	2.54	3.12
Denoised	3.33	2.77	2.55	3.31
Enhanced	**3.666**	**2.98**	**2.68**	**3.39**

Table 5: This table presents the results of objective metrics following enhancement and denoising processes applied to 100 randomly selected speakers' audio clips.

3.33, while other metrics like SIGMOS and UTMOS remained relatively stable, indicating that the denoising process effectively reduced noise without significantly affecting the speech quality. The enhanced audio, however, showed notable improvements across all metrics, particularly in DNSMOS (3.666) and WV-MOS (3.39), demonstrating the enhancement stage's effectiveness in improving perceptual quality and signal clarity.

The second experiment, summarized in Table 6, focused on the 100 lowest scoring audio clips from the MOS test, representing the most challenging cases with poor initial quality.

	DNSMOS	**SIGMOS**	**UTMOS**	**WV-MOS**
Ground Truth	3.05	2.54	2.29	2.84
Denoised	3.08	2.52	2.32	3.03
Enhanced	3.47	2.83	2.56	3.29

Table 6: This table shows the results of objective metrics following enhancement and denoising for the lowest scoring audio clips from the MOS test.

The ground truth audio in this set started with lower scores across all metrics compared to the random sample, with a DNSMOS of 3.05. After denoising, there was a slight improvement in

DNSMOS (3.08), though SIGMOS saw a marginal decline, indicating that while noise was reduced, there was a slight degradation in the perceived speech quality. The enhancement stage led to significant improvements in all metrics, with DNSMOS increasing to 3.47 and SIGMOS to 2.83, suggesting that the enhancement process can substantially improve both perceptual quality and overall usability, even in cases of poor initial audio quality.

11.3 Discussion

The results from both experiments underscore the pipeline's effectiveness in enhancing audio quality. The enhancement stage consistently provided substantial gains over denoising alone. The slight reduction in some metrics during denoising, especially in the most challenging cases, highlights a trade-off between noise reduction and signal integrity. However, the subsequent enhancement phase successfully mitigated this, leading to improved scores across all objective metrics.

These findings emphasize the value of a dual-stage audio processing approach, where denoising prepares the audio by removing noise, and enhancement refines the signal to achieve higher perceptual quality. This indicates that such an approach can be particularly beneficial when handling audio with variable quality, as evidenced by the results from both the randomly selected and poorly performing clips from the MOS tests.

12 Conclusion

During the internship, I explored and advanced the application of voice cloning and style management techniques, specifically within the context of the French language. The project began with the challenge of adapting the VALL-E architecture, originally developed with extensive resources in English and Mandarin, to French—a language with comparatively limited data resources in neural codec voice cloning.

To address this, a significant portion of the project was dedicated to the collection and preparation of diverse and high-quality French-language datasets. This involved curating data from sources such as Libri-Light, LibriTTS, and Common Voice, each presenting unique challenges in terms of data quality, consistency, and suitability for the task at hand. Through preprocessing, including standardization of audio formats and normalization of volume levels, these datasets were optimized for model training.

The adaptation of the model required careful consideration of linguistic nuances. Phonemization emerged as a critical component, particularly due to the absence of phonetic information in the original Bark model. Through experimentation, it was found that utilizing a BERT-like sequential latent representation directly from phonemes, rather than text, significantly improved the model's ability to generate natural and prosodically accurate speech. This approach was guided by recent advancements in phoneme-based models, such as those discussed in the XPhoneBERT [37], and highlighted the importance of phonetic information in capturing the details of French pronunciation.

Training and optimization posed further challenges, particularly in achieving a balance between computational efficiency and model generalization. Despite initial difficulties, including models that refused to generate coherent audio, iterative refinements and strategic modifications led to substantial improvements in model performance. The resulting voice cloning model demonstrated notable advancements in terms of naturalness and prosody, marking a significant step forward in the adaptation of text-to-speech technologies for French.

Evaluation and validation were key aspects of this project, ensuring that the model's outputs were not only accurate but also met the subjective and objective standards expected in high-quality speech synthesis. The variability in the quality of synthesized audio raised important questions regarding speaker bias, which were addressed through detailed analysis and experiments involving denoising and enhancement pipelines. The findings underscored the influence of the original recording quality on the final synthesized output, highlighting areas for future research and improvement.

In conclusion, this internship tackled important challenges in adapting voice cloning models for the French language. The project showed that it's possible to use phoneme-based representations effectively for text-to-speech in a language with fewer resources.

References

[1] A. J. DeCasper and W. P. Fifer, "Of human bonding: Newborns prefer their mothers' voices," *Science*, vol. 208, no. 4448, pp. 1174–1176, 1980. DOI: 10.1126/science.7375928.

[2] P. Taylor, *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[3] S. Ö. Arık, M. Chrzanowski, A. Coates, *et al.*, "Deep voice: Real-time neural text-to-speech," in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., ser. Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, pp. 195–204. [Online]. Available: https://proceedings.mlr.press/v70/arik17a.html.

[4] A. oord, S. Dieleman, H. Zen, *et al.*, "Wavenet: A generative model for raw audio," Google DeepMind, London, UK, 2016.

[5] Z. Mu, X. Yang, and Y. Dong, "Review of end-to-end speech synthesis technology based on deep learning," *ArXiv*, 2021.

[6] J. J. Ohala, "Christian gottlieb kratzenstein: Pioneer in speech synthesis," in *International Congress of Phonetic Sciences*, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:29706004.

[7] T. H. Tarnóczy, "The Speaking Machine of Wolfgang von Kempelen," *The Journal of the Acoustical Society of America*, vol. 21, no. 4 Supplement, pp. 461–461, 2005.

[8] P. Palo, "A review of articulatory speech synthesis," Ph.D. dissertation, 2006.

[9] B. Atal and M. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 3, pp. 247–254, 1979. DOI: 10.1109/TASSP.1979.1163237.

[10] B. Atal, "The history of linear prediction," *Signal Processing Magazine, IEEE*, vol. 23, pp. 154–161, 2006. DOI: 10.1109/MSP.2006.1598091.

[11] D. Guennec, "Study of unit selection text-to-speech synthesis algorithms," Ph.D. dissertation, 2016.

[12] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from hmm using dynamic features," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1995, 660–663 vol.1. DOI: 10.1109/ICASSP.1995.479684.

[13] J. Tao, L. Xin, and P. Yin, "Realistic visual speech synthesis based on hybrid concatenation method," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, pp. 469–477, 2009. DOI: 10.1109/TASL.2008.2011538.

[14] P. Neekhara, S. Hussain, S. Dubnov, F. Koushanfar, and J. McAuley, "Expressive neural voice cloning," in *Proceedings of The 13th Asian Conference on Machine Learning*, V. N. Balasubramanian and I. Tsang, Eds., ser. Proceedings of Machine Learning Research, vol. 157, PMLR, 2021, pp. 252–267. [Online]. Available: https://proceedings.mlr.press/v157/neekhara21a.html.

[15] Y. Wang, R. Skerry-Ryan, D. Stanton, *et al.*, "Tacotron: Towards end-to-end speech synthesis," Interspeech, 2017.

[16] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," IEEE, 2018.

[17] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," NeurIPS, 2020.

[18] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigVGAN: A universal neural vocoder with large-scale training," ICLR, 2023.

[19] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," ICML, 2021.

[20] C. Wang, S. Chen, Y. Wu, *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," 2023.

[21] Z. Zhang, L. Zhou, C. Wang, *et al.*, "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," 2023.

[22] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[23] J. Kahn, M. Rivière, W. Zheng, *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7669–7673. DOI: 10.1109/ICASSP40776.2020.9052942.

[24] A. Adigwe, N. Tits, K. E. Haddad, S. Ostadabbas, and T. Dutoit, *The emotional voices database: Towards controlling the emotion dimension in voice generation systems*, 2018. arXiv: 1806.09514 [cs.CL].

[25] A. Sini, L. Wadoux, A. Perquin, *et al.*, "Techniques de synthèse vocale neuronale à l'épreuve des données d'apprentissage non dédiées : Les livres audio amateurs en français," TAL, 2022.

[26] P.-E. Honnet, A. Lazaridis, P. Garner, and J. Yamagishi, "The siwis french speech synthesis database – design and recording of a high quality french database for speech synthesis," Tech. Rep., 2017.

[27] A. Sini, D. Lolive, G. Vidal, M. Tahon, and É. Delais-Roussarie, "SynPaFlex-corpus: An expressive French audiobooks corpus dedicated to expressive speech synthesis.," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, N. Calzolari, K. Choukri, C. Cieri, *et al.*, Eds., Miyazaki, Japan: European Language Resources Association (ELRA), 2018. [Online]. Available: https://aclanthology.org/L18-1677.

[28] D. Guennec, L. Wadoux, A. Sini, N. Barbot, and D. Lolive, "Voice cloning: Training speaker selection with limited multi-speaker corpus," in *Proc. 12th ISCA Speech Synthesis Workshop (SSW2023)*, 2023, pp. 170–176. DOI: 10.21437/SSW.2023-27.

[29] L. H. Clemmensen and R. D. Kjærsgaard, *Data representativity for machine learning and ai systems*, 2023. arXiv: 2203.04706 [stat.ML].

[30] A. Gupta, D. Bhatt, and A. Pandey, *Transitioning from real to synthetic data: Quantifying the bias in model*, 2021. arXiv: 2105.04144 [cs.LG].

[31] A. Kirkland, S. Mehta, H. Lameris, G. E. Henter, E. Sz'ekely, and J. Gustafson, "Stuck in the mos pit: A critical analysis of mos test methodology in tts evaluation," in *Proc. 12th ISCA Speech Synthesis Workshop (SSW2023)*, 2023.

[32] C.-C. Lo, S.-W. Fu, W.-C. Huang, *et al.*, "MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion," in *Proc. Interspeech 2019*, 2019, pp. 1541–1545. DOI: 10.21437/Interspeech.2019-2003.

[33] S. Ogun, V. Colotte, and E. Vincent, "Can we use common voice to train a multi-speaker tts system?" In *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 900–905. DOI: 10.1109/SLT54892.2023.10022766.

[34] *Bark hugging face web page*, https://huggingface.co/suno/bark, Accessed: 2024-09-04.

[35] H. Zen, V. Dang, R. Clark, *et al.*, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *Proc. Interspeech 2019*, 2019, pp. 1526–1530. DOI: 10.21437/Interspefech.2019-2441.

[36] R. Ardila, M. Branson, K. Davis, *et al.*, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[37] L. T. Nguyen, T. Pham, and D. Q. Nguyen, "XPhoneBERT: A Pre-trained Multilingual Model for Phoneme Representations for Text-to-Speech," in *Proceedings of the 24th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2023.

[38] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi++: A unified framework for bandwidth extension and speech enhancement," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10097255.

[39] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "Utmos: Utokyo-sarulab system for voicemos challenge 2022," in *Interspeech 2022*, 2022, pp. 4521–4525. DOI: 10.21437/Interspeech.2022-439.

[40] C. K. A. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6493–6497. DOI: 10.1109/ICASSP39728.2021.9414878.

[41] R. Cutler, N.-C. Ristea, A. Saabas, B. Naderi, S. Braun, and S. Branets, *Icassp 2024 speech signal improvement challenge*, Sep. 2023.

[42] *Resemble enhance*, https://huggingface.co/ResembleAI/resemble-enhance, Accessed: 2024-09-04.

[43] X. Liu, H. Liu, Q. Kong, *et al.*, "Separate what you describe: Language-queried audio source separation," in *Proc. Interspeech*, 2022, pp. 1801–1805.