Master research Internship

Bibliographic 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 bibliographical study presents a comprehensive exploration of VALL-E, an advanced neural codec language model, and its application in French voice cloning and stylistic control. It encompasses a detailed review of the evolution of text-to-speech (TTS) and voice conversion technologies, from historical approaches to the latest advancements in neural vocoders and end-to-end models. The study highlights VALL-E's innovative zero-shot capabilities and in-context learning, particularly its proficiency in maintaining speaker identity, emotional nuances, and acoustic environments across languages. Critical to this analysis is an examination of specific challenges in French voice synthesis, including the intricacies of data acquisition, the complexity of balancing subjective and objective measures in evaluation methodologies, the implications of data representativeness and inherent biases, and the exploration of future directions in French voice synthesis technology. This review sets a solid groundwork to more effectively address these challenges during the internship.

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 considerably from early TTS systems to current neural network-based models. This evolution reflects continuous technological progress and a deeper understanding of linguistic and acoustic elements. This review explores the latest advancements in TTS and voice conversion technologies, tracing the transition from initial approaches to the modern techniques that are redefining the field. It highlights how historical methodologies have laid the groundwork for today's innovations in generating lifelike and natural-sounding synthetic speech. In section 2, we start with an overview of the foundational approaches that laid the groundwork for early TTS systems. This section delves into the initial developments in speech synthesis, marked by their relative simplicity and the constraints of computational capabilities at the time. Providing this historical context is essential for understanding the substantial advancements that occurred in the field in the subsequent years. As we move through the timeline of TTS technology, we witness a series of evolutionary steps, each marked by groundbreaking innovations. From the initial text-to-speech algorithms constrained by the technology of their time to the advent of deep learning models that revolutionized the field, this narrative showcases the relentless pursuit of more natural, versatile, and expressive forms of synthetic speech.

At the heart of this narrative is an in-depth analysis of VALL-E, a pioneering model representing the pinnacle of modern TTS technology. VALL-E's novel approach to speech synthesis epitomizes

the remarkable advancements in the field, offering levels of naturalness and expressiveness in synthetic speech that were once thought unattainable. Through the lens of VALL-E's development and capabilities, we gain insights into the cutting-edge techniques that are pushing the boundaries of what's possible in speech synthesis. For this reason the VALL-E model will be our main focus in the third section. As the primary subject of this internship, VALL-E represents a significant leap in human-computer interaction capabilities within the TTS domain. The analysis delves into the innovative aspects of VALL-E, including its advanced machine learning algorithms, natural language processing techniques, and the remarkable improvement it brings to the realism of synthetic speech. By concentrating on VALL-E, the review not only explores its current technological achievements but also discusses its potential implications and future applications in the broader context of voice synthesis technology.

The fourth and final section of the study specifically addresses the significant challenges posed by data scarcity in French voice cloning within TTS technology. The scarcity of comprehensive and diverse French language datasets is a primary obstacle, hindering the development of robust and accurate voice cloning models. This limitation becomes particularly evident when compared to languages like English, which benefit from an abundance of available speech data. We explore how this data scarcity impacts the quality and versatility of synthesized French voices, including the difficulties in capturing regional accents and informal speech patterns. The review also discusses ongoing efforts and potential strategies to mitigate these challenges, such as data augmentation and cross-lingual model training, to advance the field of French voice cloning.

2 Text-to-Speech & Voice Conversion

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.

2.1 Historic approaches in TTS

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].

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].

2.2 Current approaches

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 voice 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.

2.2.1 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.

2.2.2 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.

2.2.3 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 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.

3.1 Overview of VALL-E

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.

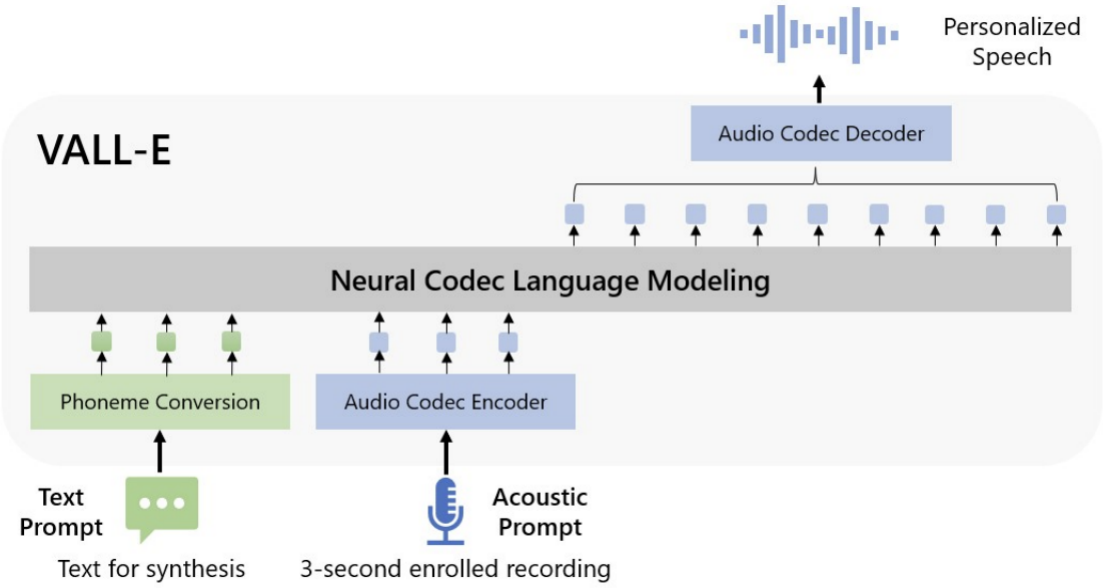

Figure 1: Overview of the VALL-E model architecture. Figure from the original Vall-E paper [20].

3.2 Model Design: Architecture and Training

The architecture of VALL-E integrates autoregressive (AR) and non-autoregressive (NAR) models. As shown in Figure 2, 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 model attends to finer acoustic details as captured by subsequent quantizers. This design balances

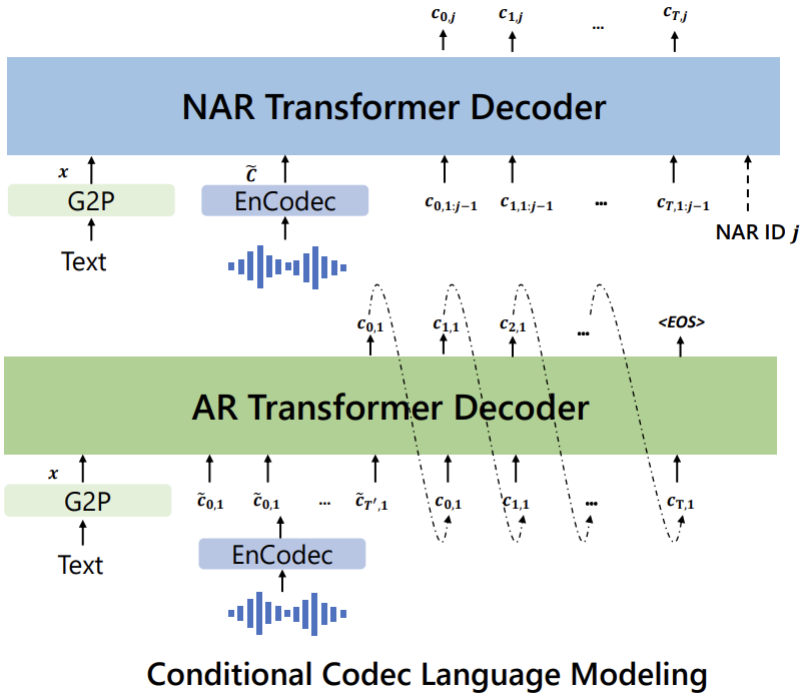

Figure 2: The structure of the conditional codec language modeling, which is built in a hierarchical manner. Figure from the original Vall-E paper [20].

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 [22] 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.

3.3 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 [23] 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.

3.4 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.

3.5 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. Notably, VALL-E's ability to achieve a CMOS score comparable to human speech recordings is proof to its advanced synthesis capabilities. These evaluations underscore the models' proficiency in generating speech that is not only natural and diverse but also finely attuned to the specific characteristics of individual speakers.

3.6 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 applicability in TTS technology.

4 Challenges

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 [24].

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 [24].

This section aims to explore these challenges in detail, investigating how the scarcity and variability of quality in French speech data impact the process of cloning and synthesizing French voices using VALL-E. The focus will be on examining the difficulties in producing quality French speech synthesis comparable to the state of the art under these constraints, and exploring potential solutions to overcome these challenges. This exploration not only addresses a critical gap in the field of speech synthesis but also sets the stage for innovative approaches in adapting advanced speech synthesis technologies like VALL-E to new languages and linguistic environments.

4.1 Data Acquisition and Quality Constraints

One of the most significant challenges in advancing neural voice 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 voice synthesis models. While there are notable French datasets like FrenchSiwis [25] and SynPaFlex [26], 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 voice 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 voice synthesis models. It underscores the importance of harvesting as much data as possible to enhance the development of robust and versatile voice synthesis technologies for French, aiming to bridge the gap in resource availability between languages. The research by Guennec et al. [27] offers a promising direction for addressing the challenge of data scarcity in French voice 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 voice 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 voice synthesis models. This approach not only addresses the challenge of limited data but also ensures a broader representation, crucial for developing voice synthesis technologies that can cater to various user needs. In the context of neural voice 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 voice synthesis capabilities in underrepresented languages.

4.2 Evaluation Methodology: Balancing Subjectivity and Objectivity

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. [28] 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 voice synthesis technology. The selection of speakers in voice 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 voice 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 voice 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 voice synthesis models. Notably, models like MOSNet [29] or WV-MOS [30], 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 voice synthesis models.

4.3 The Issue of Data Representativeness and Bias

Addressing the challenge of data representativeness and bias in neural voice synthesis, particularly for French, involves a multi-faceted approach to dataset curation. Representativeness influences the inclusiveness and accuracy of voice 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. [31].

Moreover, the impact of dataset bias on the fairness and effectiveness of machine learning models, including those used in voice 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. [32].

Incorporating these insights into the development of neural voice 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 voice synthesis but also aligns with the broader goals of creating equitable and accessible AI technologies.

4.4 Future Directions in Voice Synthesis for French

To address the future directions in voice synthesis for French in the context of the internship, a comprehensive approach is necessary. This involves include expanding the size and diversity of French voice datasets and developing methodologies for bias reduction and representativeness enhancement in training datasets. This aligns with the internship's objective of adapting VALL-E to French, necessitating substantial data quantities and high-quality datasets. The internship will entail rigorous work in dataset building and bias hunting, crucial for optimizing the quality and authenticity of vocal cloning. Additionally, adapting the model to the unique characteristics of the French language and potentially integrating specific stylistic data will be key areas of focus. This hands-on approach in dataset preparation, implementation, and evaluation of a French version of VALL-E aligns with current research trends, emphasizing deep learning-based techniques to improve synthetic speech's naturalness and accuracy.

5 Conclusion

In summary, the evolution of TTS and voice cloning technologies has been a journey marked by groundbreaking innovation, adapting to ever-changing linguistic and technological landscapes. From the initial mechanical resonators of the 18^{th} century to the current neural models, each step has been pivotal in shaping the field's trajectory. The advancements in computational methods, machine learning, and digital signal processing have revolutionized the quality of synthesized speech, surpassing previous benchmarks in naturalness and expressiveness.

The historical progression of speech synthesis technologies, from mechanical to digital, highlights a relentless pursuit of more realistic and human-like speech synthesis. Early efforts by pioneers like Kratzenstein and von Kempelen set the foundational principles, which were later refined through articulatory synthesis and linear prediction techniques. The emergence of concatenative synthesis and unit selection, followed by HMM-based statistical parametric speech synthesis, demonstrated the field's gradual shift towards more adaptable and nuanced approaches. The integration of these methods in hybrid parametric/unit selection systems paved the way for a synthesis of naturalness and versatility, significantly enhancing the lifelike quality of synthesized speech.

The contemporary era, marked by the advent of WaveNet and BigVGAN, represents a significant

leap in the field. These technologies have not only improved the realism of speech synthesis but also broadened its applications, ranging from virtual assistants to accessibility tools for speech-impaired individuals. Furthermore, the development of voice cloning technologies like VALL-E and its cross-lingual extension, VALL-E X, signifies a major advancement in creating personalized and adaptable speech synthesis across languages. These innovations underscore the growing potential of TTS systems to cater to a wide array of needs and preferences, bridging communication gaps and enhancing user experiences.

Despite these achievements, the field faces considerable challenges, especially in extending these technologies to languages with limited available data. The development of French-specific models, for instance, highlights the complexities involved in collecting and utilizing high-quality, diverse speech datasets. Addressing these challenges requires dedicated efforts towards inclusive data collection, representing a broader range of dialects, accents, and speech patterns. This focus on inclusivity is crucial for developing TTS technologies that are representative and accessible to diverse global populations.

Moreover, the ethical dimensions of voice cloning, particularly in terms of consent and the risk of misuse, present critical considerations for the future of these technologies. As the line between synthesized and human speech continues to blur, it becomes imperative to establish robust ethical frameworks and guidelines to prevent misuse and ensure responsible development and application of these technologies.

In conclusion, the field of TTS and voice cloning is navigating the delicate balance between technological advancement and ethical responsibility. As we progress, it's imperative to leverage these innovations for inclusive, ethical, and human-centric communication solutions. The transformative potential in human-computer interaction is vast, foreseeing a future where synthesized speech matches the richness and diversity of natural human speech. The internship's trajectory aligns seamlessly with this narrative, charting a course through the replication of VALL-E's English model and its subsequent adaptation to the French linguistic environment. This process entails a methodical evaluation of existing implementations, a rigorous development of a French-specific training corpus, and the fine-tuning of VALL-E to embrace the nuances of the French language. The final phase of benchmarking this French version against established models, underscoring a commitment to advancing voice synthesis technology and linguistic inclusivity.

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] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, *A survey on neural speech synthesis*, 2021.

[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] 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.

[23] 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].

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

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