Improving Speech Naturalness and Nuance using HiFiGAN-Hubert-Soft Vocoder: A Case Study of the Voicebox TTS model

Kulboboev Shukhrat
Department of Measurement and Information System
Budapest University of Technology and Economics
Budapest, Hungary
shukhrat.kulboboev@edu.bme.hu

Mohammad Salah Al-Radhi
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Budapest, Hungary
malradhi@tmit.bme.hu

***Abstract*—Text-to-speech (TTS) technology has significantly transformed human-machine interactions, facilitating seamless communication between humans and computers. However, achieving high-quality TTS remains a significant challenge, especially in synthesizing natural and nuanced speech. In this study, we investigate the potential of HiFiGAN-Hubert-Soft (HHS) vocoders to enhance the performance of TTS models, with a focus on integrating the HHS vocoder into the Voicebox TTS model—a versatile and scalable TTS system developed by Meta AI. Through both subjective (mean opinion score) and objective (speech similarity and visualization metric) evaluations, we illustrate that the HHS vocoder significantly enhances the naturalness and nuance of synthesized speech compared to the baseline HiFiGAN vocoder. This improvement is particularly pronounced in cases where pronunciation variations are minor or context-dependent. Our findings emphasize the potential of HHS vocoders in elevating TTS performance and laying the foundation for further advancements in TTS technology.**

Keywords— Generative AI model, Hubert-Soft vocoder, Speech synthesis, Text-to-Speech, Voicebox.

I. INTRODUCTION

In the rapidly evolving landscape of communication technologies, Text-to-Speech (TTS) stands out as a transformative force, reshaping the way we engage with information. Beyond its role in converting written text into spoken words, TTS plays a fundamental role in accessibility and inclusivity. This technology not only facilitates efficient message delivery but also addresses the diverse needs of individuals who may have different preferences or require alternative means of communication.

As we explore deeper into the area of speech-related fields, the applications of TTS become even more pronounced [1] [2]. Beyond its conventional use in articulating written content, TTS finds itself at the forefront of innovations in emotional conversion [3]. This requires the nuanced ability of the software to infuse spoken words with varying emotional tones, adding a layer of expressiveness to communication. Furthermore, TTS contributes significantly to the enhancement of single-channel communication, enabling clearer and more effective dissemination of information through diverse platforms [4].

In tandem with TTS, other speech-processing technologies are making substantial improvements. Emotional conversion techniques are not limited to TTS alone; they extend to broader applications in voice conversion, enabling the modification of the speaker's voice characteristics [3]. This breakthrough has implications for personalization and adaptation in communication, catering to individual preferences and needs. Moreover, the advancement of bandwidth extensions for narrowband speech highlights the commitment of speech-related technologies to overcome barriers and ensure high-quality communication experiences [5]. The ability to enhance the clarity and fidelity of speech in constrained bandwidth scenarios opens doors to improved connectivity, particularly in situations where conventional communication channels may face limitations.

II. RELATED WORK AND BACKGROUND

A. Generative speech models

Current speech generative models [7] are often task-specific and trained on different datasets. One common type of task is audio style conversion, which aims to modify specific speech characteristics while preserving others. Voice conversion, emotion conversion, and speech enhancement are examples of this category. Many of these models require supervised learning using data pairs that only differ in one attribute, such as emotion. However, obtaining such data is challenging, and annotating attributes like speaking style can be difficult. As a result, these models are often trained on small datasets and do not generalize well.

Another common task is controllable text-to-speech synthesis, which aims to synthesize speech in a target audio style given text [8]. While some styles, like voice, can be specified through labels or pre-trained embeddings [9], others, like prosody, are difficult to annotate or embed. Previous attempts to control these styles using residual embeddings have not been successful due to their limited capacity and oversimplified distribution modeling [10].

Infilling is another task to predict speech given context and optionally text guidance. Unlike explicit style embedding, infilling models predict speech coherent with the context [11]. This approach is similar to large language models (LLMs), which specify the task through context. While this is promising for building large-scale generalist models using minimal explicit supervision, most previous work with text guidance assumes a deterministic mapping from text and context to target, which is only realistic for very short segments [12]. As a result, models with these assumptions can typically only infill segments up to 1 second.

Voicebox is a text-guided infilling model that overcomes these limitations by leveraging the conditional neural field (CNF) model [13], which can parameterize any distribution. This allows Voicebox to infill speech of any length and can be trained on in-the-wild datasets with rich variation.

B. Voicebox TTS model

Voicebox is a versatile text-conditioned speech generative model that can handle various tasks, including speech synthesis, voice conversion [14], and speech enhancement [15]. Unlike previous models that rely on speech style labels, Voicebox does not require explicit audio style information. Instead, it learns to infer the speech style from the surrounding speech and text transcript. This approach makes Voicebox more scalable and data-efficient, as it can be trained on larger datasets without requiring specialized labels. Voicebox's ability to handle various speech-generative tasks makes it a powerful tool for various applications, including speech synthesis, voice cloning, and audio editing.

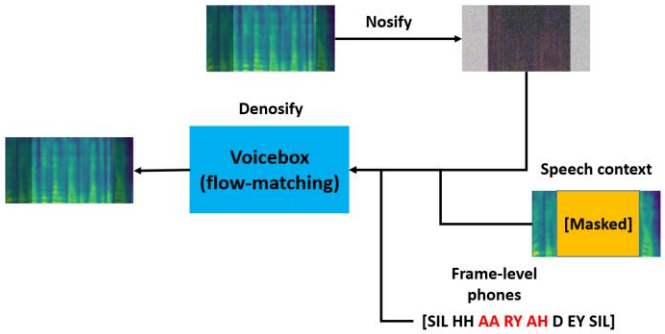

Figure 1: The Voicebox model acquires knowledge of the inverse transformation process, enabling it to convert noise back into audio by considering both textual information and speech context through the application of a flow-matching objective. From a noisy distribution (top right in the image), it progressively builds the masked part of the sequence by denoising it using the provided text and the unmasked parts of the audio sequence. As a result, the Voicebox model gives accurate and clear results the same as the original audio at the end.

Voicebox employs a non-autoregressive (NAR) continuous normalizing flow (CNF) model, which differs from traditional autoregressive models' ability to process context information from the past and the future. This approach enables Voicebox to generate more natural and coherent speech, as it can consider the broader context of the conversation. The CNF model is trained using a recently proposed method called flow-matching (see Fig.1) [16], which allows for efficient and scalable training of CNFs using a simple vector field regression loss. This approach makes Voicebox a powerful tool for speech synthesis and other speech-related tasks.

C. Generative adversarial networks(GAN) based vocoders

Text-to-speech (TTS) technology has been the subject of extensive research and development in recent years, driven by the increasing demand for natural and human-like speech synthesis in various applications, including virtual assistants, chatbots, and educational software. TTS systems have traditionally employed concatenative and parametric vocoders [17], but these approaches have limitations in generating natural-sounding speech, especially in reproducing subtle pronunciation variations, context-dependent speech patterns, and emotional expressions.

In recent years, generative adversarial networks (GANs) have emerged as a powerful tool for TTS, potentially generating more natural and nuanced speech [18]. One of the most successful GAN-based TTS approaches is HiFiGAN, which has demonstrated impressive results in synthesizing high-quality speech with various accents and speaking styles [19].

However, HiFiGAN still faces challenges in capturing the nuances of human speech, particularly in producing subtle pronunciation variations and context-dependent speech patterns [20]. To address these limitations, researchers have explored the use of HiFiGAN Hubert Soft (HHS) vocoders [21], which integrate HiFiGAN with Hubert-Soft, a neural vocoder that utilizes a hierarchical representation of spectral features.

Our study aims to investigate further the potential of HHS vocoders in elevating TTS performance by integrating them into the Voicebox TTS model—a versatile and scalable TTS system developed by Meta AI. We will evaluate the HHS vocoder's performance using subjective and objective measures and compare its performance to the baseline HiFiGAN vocoder. Our findings will contribute to understanding the capabilities of HHS vocoders in improving TTS quality and will lay the foundation for further advancements in TTS technology.

D. A Comparison of discrete and soft speech units for improved voice conversion

Discrete Content Encoder: The discrete content encoder encompasses feature extraction and k-means clustering processes to convert an audio utterance into a sequence of discrete speech units (refer to Fig. 2b). The feature extraction step can employ various techniques, spanning from low-level descriptors to self-supervised models such as HuBERT. In the clustering step, similar features are grouped into a dictionary of discrete speech units. Previous research indicates that clustering features derived from large self-supervised models enhances the quality of speech units and contribute to improved voice conversion. Ultimately, the discrete content encoder transforms an input utterance into a sequence of discrete speech units represented as $<d_1, ..., d_T>$.

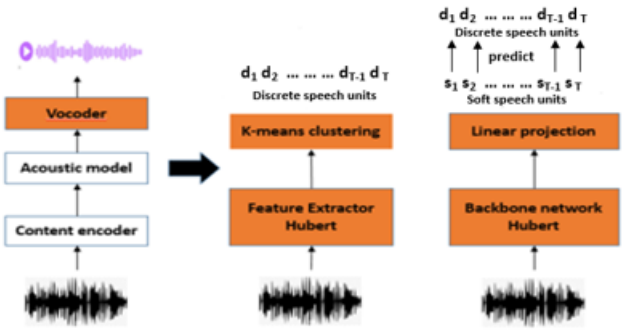

Figure 2: a) Vocoder system. b) The discrete content encoder clusters audio features to produce a sequence of discrete speech units. c) The soft content encoder is trained to predict the discrete units. The acoustic model transforms the discrete/soft speech units into a target spectrogram. The vocoder converts the spectrogram into an audio waveform.

Soft Content Encoder: Directly utilizing the output of the feature extractor for soft speech units may seem intuitive, but prior research indicates that these representations retain significant speaker-related information, rendering them unsuitable for voice conversion (as later confirmed in our experiments). Instead, we choose to train the soft content encoder to predict a distribution over a discrete unit.

The training process for the soft content encoder is illustrated in Figure 2c. When presented with an input utterance, the first step involves extracting a sequence of discrete speech units $< d_1, \dots, d_T >$ as reference labels. Next, the utterance undergoes processing through a backbone network (HuBERT). The output from HuBERT is then projected through a linear layer to generate a sequence of soft speech units $< s_1, \dots, s_T >$. Each soft unit serves as a probability distribution over the discrete units in the dictionary:

$$p(d_t = i \mid \mathbf{s}_t) = \frac{\exp(\text{sim}(\mathbf{s}_t, \mathbf{e}_i)/\tau)}{\sum_{k=1}^{K} \exp(\text{sim}(\mathbf{s}_t, \mathbf{e}_k)/\tau)},$$

Each soft unit is represented by an embedding vector e_i, which is learned during training. The cosine similarity between the soft and discrete units is computed using sim(·, ·), and the temperature parameter τ controls the degree of concentration of the probability distribution. The average cross-entropy between the distributions and discrete targets $< d_1, \dots, d_T >$ is minimized to update the encoder, including the backbone network. When it comes to testing time, the soft content encoder transforms the input audio into a sequence of soft speech units $< s_1, \dots, s_T >$, which is then forwarded to the acoustic model for further processing.

III. Problem Statement

In the realm of Text-to-Speech (TTS), continuous advancements have led to the development of diverse systems adapted for various purposes over an extended duration. A prominent challenge within this domain is the tendency of synthesized voices to show a robotic, inexpressive, and non-authentic quality. To address this issue, the newly developed Voicebox [1] TTS model is implemented with the HiFiGAN-Hubert-Soft vocoder [6]. This strategic choice is aimed at achieving a higher quality of synthesized speech that aligns with the requirements for a genuinely human-like voice. The utilization of the HiFiGAN-Hubert-Soft vocoder represents a deliberate approach to advance the authenticity and expressiveness of the synthesized voice, contributing to the ongoing efforts to overcome the limitations associated with traditional TTS systems. This endeavor holds significance in the context of human-robot interaction and other applications where the fidelity and naturalness of synthesized speech are crucial factors.

IV. Methodology

To assess the efficacy of the modified Voicebox in generating high-quality speech from text, we conducted a comprehensive two-stage investigation. In the initial phase, we precisely trained the Voicebox model from the ground up for a duration of four days, closely monitoring parameter changes throughout the process. Subsequently, upon completing the training, we transitioned to the inference stage, where speech was generated using a Mel-spectrogram. This section presents a detailed exposition of every part of our methodology, encompassing data preparation, model structure, training procedures, and the inference process, each explained and illustrated for clarity and transparency.

A. Data

We trained the only English model using the LJSpeech dataset [22], a public-domain speech dataset consisting of 13,100 short audio clips of a single female speaker reading passages from 7 non-fiction books. A transcription is provided for each clip. Each clip is accompanied by a corresponding transcription, with durations between 1 to 10 seconds and a combined length of around 24 hours. To phonemically annotate and align the transcript, we employed the Montreal Forced Aligner (MFA) [23], utilizing its phone set derived from a modified version of the international phonetic alphabet (IPA). Additionally, word position 10 postfixes were introduced. The audio content was represented as an 80-dimensional log Mel spectrogram, and we employed a HiFi-GAN Discrete vocoder, trained on a 60,000-hour English speech dataset, to generate the waveform.

B. Model

The Voicebox [1] audio model employs a transformer [24] featuring convolutional positional embedding [25]. This model consists of 24 layers, 16 attention heads, and dimensions of 1024/4096 for the embedding/feed-forward network (FFN). With a total of 63.7 million parameters, skip connections are incorporated to link symmetric layers, following a UNet

[1] https://github.com/lucidrains/voicebox-pytorch

architecture style (e.g., connecting the first layer to the last layer, the second layer to the second-to-last layer, and so forth).

C. Training

The Voicebox model was trained from scratch with 100 epochs using NVIDIA TITAN Xp GPU and CUDA Version: 11.4. For training efficiency, audio length was capped at 1,600 frames and chunked randomly if the length exceeded this threshold. The Adam [26] optimizer is used with a peak learning rate of le-4, linearly warmed up for 5K steps, and linearly decays over the rest of the training. For audio models, we clipped the gradient norm to 0.2 for training stability. The audio sequence was masked with p_{drop} = 0.3/0.2. We constantly checked changes in epoch loss and learning rate during the training. We started the learning rate with 1.18e-5 and epoch loss 0.788 after the first epoch training.

D. Inference

The inference stage involved generating speech from text using the trained Voicebox model, the HiFiGan Hubert Discrete (HHD) vocoder, and the HiFiGan Hubert Soft (HHS) vocoder. Given a text input, the model first converts the text into phonemes. The phonemes are then represented as an 80-dimensional log Mel spectrogram. The spectrogram was then passed through the Voicebox audio model, which generates a waveform. The waveform was then processed by the HiFiGAN Discrete vocoder to produce the final synthesized speech. Then, we replaced the HHD vocoder with the HHS vocoder to check the speech quality for comparison.

V. EVALUATION RESULTS

This section conducts a comparative analysis of two vocoders, HHS and HHD, with a focus on evaluating speech quality and naturalness. The results are presented through both subjective and objective perspectives. Subjective assessments are from the preferences and perceptions of test participants who, on a scale from 1 to 5, rated speech samples—where 5 signifies excellence and 1 denotes poor quality. On the other hand, objective metrics draw from the quantifiable outcomes of the training process, including loss and accuracy metrics. This dual approach offers a comprehensive understanding of the vocoders' performance across perceptual and quantitative dimensions.

HHS vocoders exhibit the potential to enhance the authenticity and smoothness of Text-to-Speech (TTS), especially in generating nuanced pronunciation variations and speech patterns dependent on context. The HHS vocoder demonstrated notable superiority over the baseline HiFiGAN Discrete vocoder in terms of mean opinion score (MOS), audio similarity, and visualization metrics. However, HHD vocoder inference speed is higher at approximately 0.39 sec than HHS vocoder inference speed at about 1 sec. It means that HHD vocoder generates wave from melspectogram almost 3 times faster than its counterpart.

A. Subjective metric

In the evaluation of MOS, we conducted a listening test among 10 graduate students to compare the performance of various vocoders integrated into our system. The subjective evaluation was conducted in a classroom environment. Participants were seated in a quiet room and headphones were provided for the test. The audio was played on the computer. The conditions were set to minimize external distractions, creating an optimal environment for focused listening. Participants were tasked with evaluating converted speeches by listening to the original voice from the complete dataset, the HHS vocoder voice, and the HHD vocoder voice. Quality ratings were assigned using the Absolute Category Rating (ACR) scale, signifying the absolute category rating, where a score of 5 denotes excellence representing speech that is virtually indistinguishable from natural human speech. and 0 reflects poor quality. The ACR scale ratings obtained from the participants were subjected to statistical analysis, revealing significant trends in perceived speech quality. The mean ACR score for our TTS model was 4.32 in an original voice, 3.84 in an HHS vocoder, and 3.6 in an HHD vocoder. As depicted in Figure 3 and the obtained analysis result, our findings indicate that the HHS vocoder outperformed, achieving high results close to the original source, followed by the HHD vocoder.

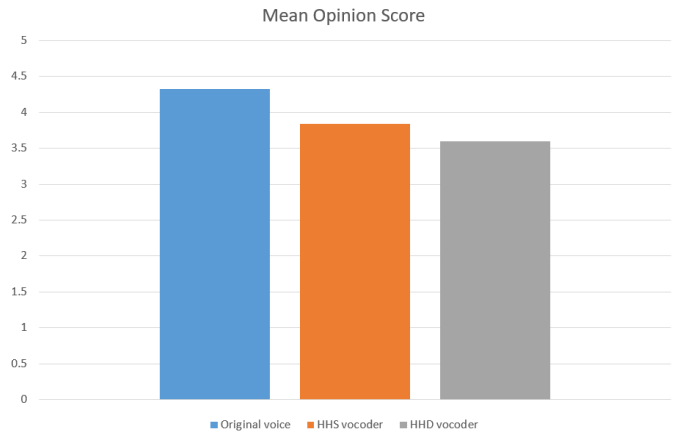

Figure 3: Spectral results of two vocoders and original voices.

B. Objective metric

Visualization Metric: In the assessment of objective results, we studied the performance of two vocoders, HHS and HHD (see Fig. 4). The primary objective was the perfect integration of the HHS vocoder into the Voicebox TTS model. Originally implemented post-training for inference, our subsequent decision favored the HHS vocoder to enhance overall performance. The HHS vocoder has distinct advantages, producing more natural-sounding speech while minimizing error introduction. Moreover, the performance of HHS proves superior in most cases when evaluated based on spectrograms. Despite its computational demands, the HHS vocoder showcases its prowess, offering enhanced fidelity in speech synthesis. It's essential to highlight that the HHD vocoder also exhibited commendable performance, with some deviations compared to the original voice in select instances. The trade-offs between the HHS and HHD vocoders underscore the nuanced considerations in selecting the optimal model for our TTS application.

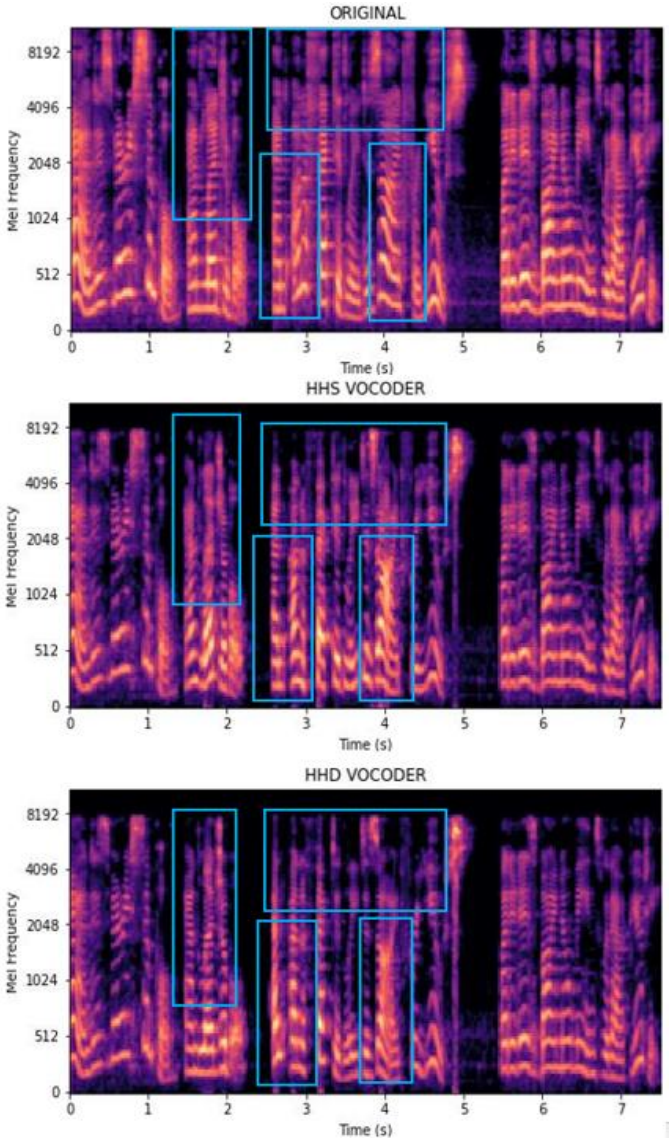

Figure 4: Mel-spectrograms extracted from synthesized speech samples.

Audio Similarity Metric: The assessment of audio similarity metrics involved the utilization of specialized algorithms integrated into dedicated Blue2Digital [27] software tools. The software utilizes an algorithm that produces an estimate of the audio signal's frequency content to compose the magnitude of the function that calculates the windowed discrete-time Fourier transform for the given audio input. The algorithm includes computing cross-correlation in the spatial and frequency domain rather than comparing audio files directly exhibiting superior performance. This software compares the similarity of two audios and we compared two audios taken from different vocoders with original audio. According to the data presented in the figure, the HHS vocoder exhibited superior performance, averaging 17.43% (refer to Fig. 5) similarity, compared to the HHD vocoder, which averaged 16.90%. This indicates that the HHS vocoder consistently outperformed the HHD vocoder in generating audio signals more closely resembling the original ones.

It is crucial to acknowledge that the audio similarity method faced some technical limitations related to audio synchronization, impacting the achieved result percentages. Despite efforts to synchronize two sample audios as closely as possible, the results did not meet our initial expectations. Nevertheless, it is noteworthy that, even under these constraints, the HHS vocoder consistently outperformed the HHD vocoder in this particular evaluation.

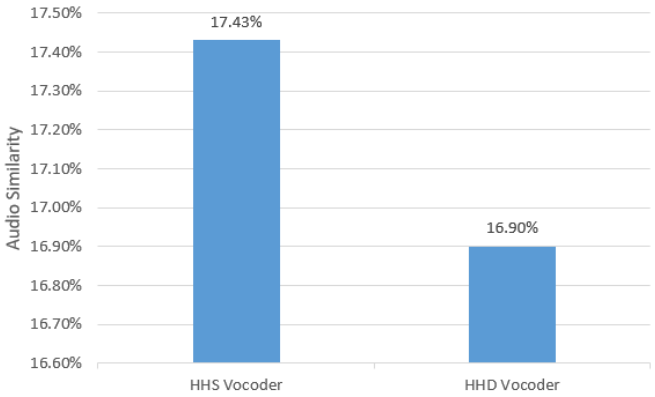

Figure 5: Audio similarity metrics table defined by the Blue2Digital software.

In Figure 6, we depict the evolution of loss and learning rate across epochs during the training process. The learning rate index fluctuated in the initial 40 epochs before stabilizing at 1.00E-05. The epoch loss demonstrated a significant reduction from 0.788 to 0.097, with minimal variation after 20 epochs. This linear progression signifies the efficient and successful implementation of the training process, with a considerable decrease in epoch loss. Additionally, the total estimated model parameter size is about 254.962 Mb. The Voicebox TTS model demonstrated efficient mel spectrogram generation, completing the process in an average time of 1.89 seconds during the inference process.

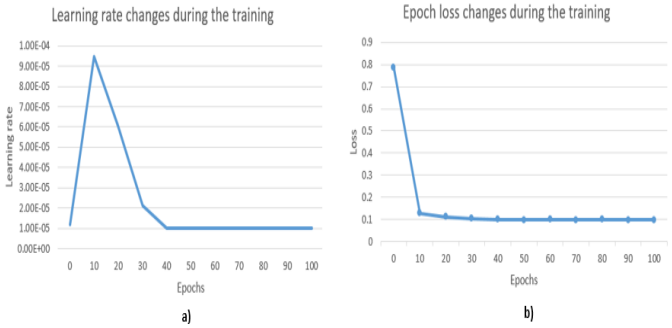

Figure 6: a) Dynamic changes in learning rate; b) Evolution of epoch loss.

VI. CONCLUSION AND DISCUSSION

The paper achieved a successful enhancement of the Voicebox text-to-speech (TTS) model by substituting the Hifigan-Hubert-Discrete (HHD) vocoder with the Hifigan-Hubert-Soft (HHS) vocoder. The HHS vocoder demonstrated superior effectiveness in generating precise and nuanced speech, particularly in

scenarios where pronunciation variations are subtle or context-dependent. This is attributed to the Voicebox model's emphasis on acoustic parameters over linguistic parameters. This improvement was substantiated through a comprehensive evaluation utilizing both subjective and objective (Audio Similarity and Visualization) metrics. The findings of this study hold substantial implications for the advancement of high-quality TTS systems. The successful integration of the HHS vocoder into the Voicebox TTS model showcases the potential of discrete vocoders to enhance the accuracy and nuance of synthesized speech. This breakthrough sets the stage for further strides in TTS technology, promising more realistic and engaging speech interactions between humans and machines.

Our future endeavors include the implementation of a universal vocoder to strengthen our model's performance with enhanced robustness and noise resistance. This forward-looking approach aims to contribute to the ongoing evolution of TTS technology, ensuring its adaptability and efficacy in diverse real-world scenarios.

ACKNOWLEDGMENT

The research was supported by the European Lighthouse project to Manifest Trustworthy and Green (ENFIELD), funded by the HORIZON Research and Innovation Programme under grant agreement No. 101120657. The research was partially supported by the EU project RRF-2.3.1-21-2022-00004 within the AI National Laboratory and by the National Research, and the Development and Innovation Office of Hungary (FK 142163 grant). The Titan Xp GPU used was donated by NVIDIA. We also extend our thanks to all the student participants who took part in the listening test.

REFERENCE

[1] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar, W. Hsu, "Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale", *International Conference on Neural Information Processing Systems (NeurIPS)*, pp.1-32, 2023.

[2] M.S. Al-Radhi, T.G. Csapó, G. Németh, "Continuous vocoder applied in deep neural network based voice conversion", *Multimedia Tools and Applications*, 78, pp. 33549-33572, 2019.

[3] Y. Xue, Y. Hamada, M. Akagi, "Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space", *Speech Communication*, 102, pp. 54-67, 2018.

[4] A. Li, C. Zheng, G. Yu, J. Cai and X. Li, "Filtering and Refining: A Collaborative-Style Framework for Single-Channel Speech Enhancement," *in IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2156-2172, 2022.

[5] P. Jax and P. Vary, "An upper bound on the quality of artificial bandwidth extension of narrowband speech signals", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, FL, USA, pp. I-237-I-240, 2002.

[6] W.N. Hsu, B. Bolte, Y. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed, "Hubert: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units", *IEEE/ACM Transactions on Audio, Speech and Language Processing*, pp 3451–3460, 2021.

[7] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. J. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. "Language models are few-shot learners", *International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 1-25, 2020.

[8] X. Zhou, Z. Zhou and X. Shi, "FCH-TTS: Fast, Controllable and High-quality Non-Autoregressive Text-to-Speech Synthesis", *International Joint Conference on Neural Networks (IJCNN)*, Padua, Italy, 2022.

[9] M.S. Al-Radhi, T.G. Csapó, G. Németh, "Nonparallel Expressive TTS for Unseen Target Speaker using Style-Controlled Adaptive Layer and Optimized Pitch Embedding", *IEEE International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 176-181, 2023.

[10] E. Battenberg, S. Mariooryad, D. Stanton, R.J. Skerry-Ryan, M. Shannon, D. Kao, T. Bagby, "Effective Use of Variational Embedding Capacity in Expressive End-to-End Speech Synthesis", *International Conference on Learning Representations (ICLR)*, 2019.

[11] H. Bai, R. Zheng, J. Chen, X. Li, M. Ma, and L. Huang. "A3T: Alignment-aware acoustic and text pretraining for speech synthesis and editing". *In International Conference on Machine Learning*, 2022.

[12] M. Sharifi, M. Tagliasacchi. "SpeechPainter: Text-conditioned speech inpainting". *In Interspeech*, pp. 431- 435, Incheon, Korea, 2022.

[13] T.M. Nguyen, A. Garg, R.G. Baraniuk, A. Anandkumar, "InfoCNF: An Efficient Conditional Continuous Normalizing Flow with Adaptive Solvers", *International Conference on Learning Representations (ICLR)*, pp. 1- 17, 2019.

[14] K.T. Kaneko, K. Tanaka, N. Hojo, "StarGAN-VC: non-parallel many-to-many voice conversion using star generative adversarial networks", *IEEE Spoken Language Technology Workshop*, 2018.

[15] D.G. Synnaeve, Y. Adi, "Real-time speech enhancement in the waveform domain", i*n Interspeech*, pp. 3291- 3295, Shanghai, China, 2020.

[16] Y. Lipman, R. Chen, H. Ben-Hamu, M. Nickel, M. Le, "Flow Matching for Generative Modeling", *International Conference on Learning Representations (ICLR)*, pp.1- 28, 2023.

[17] M.S. Al-Radhi, T.G. Csapó, G. Németh, "A continuous vocoder using the sinusoidal model for statistical parametric speech synthesis", *Speech and Computer: 20th International Conference, SPECOM*, Leipzig, Germany, 2018.

[18] A. Wali, Z. Alamgir, S. Karim, A. Fawaz, M.B. Ali, M. Adan, M. Mujtaba, "Generative adversarial networks for speech processing: A review", *Computer Speech & Language*, 72, 2022.

[19] J. Kong, J. Kim, J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis", *Advances in Neural Information Processing Systems*, pp. 17022-17033, 2022.

[20] Z. Xu, S. Zhang, X. Wang, J. Zhang, W. Wei, L. He, S. Zhao, "MuLanTTS: The Microsoft Speech Synthesis System for Blizzard Challenge", *18th Blizzard Challenge Workshop*, Grenoble, France, 2023.

[21] B. Niekerk, M. Carbonneau, J. Zadi, M. Baas, H. Seute, H. Kamper, "A Comparison of Discrete and Soft Speech Units for Improved Voice Conversion", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6562-6566, 2022.

[22] K. Ito, "The LJ speech dataset", https://keithito.com/LJ-Speech-Dataset/, 2017.

[23] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, M. Sonderegger, "Montreal Forced Aligner: trainable text-speech alignment using Kaldi" , *in Proc. Interspeech*, pp. 498-502, 2017.

[24] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations", *Advances in neural information processing systems*, 2020.

[25] V. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need", *Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA, 2017.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", *International Conference on Learning Representations (ICLR)*, 2014.

[27] https://blue2digital.com/apps/compare-audios.html