Naturalness and the Utility of Synthetic Speech in Model Pre-training

Diptasree Debnath, Asad Ullah, Helard Becerra Martinez, Andrew Hines

School of Computer Science, University College Dublin, Ireland

{diptasree.debnath,asad.ullah}@ucdconnect.ie, {helard.becerra,andrew.hines}@ucd.ie

Abstract

Foundational models have advanced speech technology while introducing privacy concerns due to the sources and volume of pre-training data required. Synthetic speech could be an alternative as short utterances are indistinguishable from natural speech but limitations in prosody and tonal variation impact longer durations. We investigate if synthetic text-to-speech (TTS) systems have reached a point where it can substitute for natural speech in pre-training models for speech-based downstream tasks, e.g. phoneme recognition (PR). We also explore the degree to which these synthetic samples can be used when data augmentation is required. We pre-train three models using (i) natural speech; (ii) synthetic TTS cloned speech matched to the natural speakers; (iii) unmatched speech using standard voices provided by the state of the art VITS TTS system. They were fine-tuned for a PR task and results show TTS data does not currently contain the long term speech characteristics to replace natural speech in pre-training but has potential for low resource data augmentation.

Index Terms: synthetic speech, TTS quality, phoneme recognition

1. Introduction

The recent advancements in speech technology powered by foundational models are undeniable [1, 2]. However, these advancements raise concerns about privacy due to their reliance on vast amounts of real human speech data. In this context, the use of synthetic speech data steps in as a transformative solution, enabling the ethical, inclusive, and adaptable development of speech models. This approach also safeguards the user's privacy and guarantees adherence to user data usage principles. Synthetic speech generation has also made significant improvement thanks to deep learning models [3, 4, 5]. These models can produce realistic voices, blurring the line between human and machine, and they can ease the use of natural speech for many speech-based technological tools (e.g., voice assistant [6], machine translation [7]). However, several aspects still need to be explored to implement highly efficient speech-based tools. For instance, current evaluation methods for speech quality might not be keeping pace with the rapid advancements of synthetic speech generation, making it difficult to pinpoint areas for further improvement [8]. Also, even the most advanced systems can struggle with conveying emotion and natural intonation, especially when creating longer pieces of speech, leading to a feeling of artificiality that remains an open area of research [9].

As synthetic speech generation technology advances, concerns regarding cloning and identity theft are emerging [10]. With the ability to create highly realistic speech that mimics a specific person's voice, malicious actors could potentially exploit this technology for fraudulent purposes. Imagine a scenario where someone's voice is used to impersonate them on a phone call, enabling unauthorised transactions or spreading misinformation. This concern is particularly serious because unlike other forms of identity theft, voice cloning can be achieved without physical access to a person.

To address these aspects, we conducted a set of experiments training and analysing the performance of different speech-based models using both natural and synthetic speech. We rely on the LibriSpeech dataset [11] for our natural speech subset; meanwhile, two synthetic speech subsets were generated using a cutting-edge TTS system, the Variational Inference with adversarial learning for end-to-end Text-to-Speech (VITS) [3]. For the first synthetic speech set, the TTS system cloned voices matching those in the natural speech set (matched synthetic). As for the second synthetic speech set, we used standard voices available at the TTS system (unmatched synthetic). These speech datasets were used to train different models and analyse the key factors that influence the effectiveness of using synthetic speech for training speech-based models.

Three experiments were carried out for this study. The first experiment compares the performance of a Autoregressive Predictive Coding (APC) [12, 13] model pre-trained with natural and synthetic speech targeting a PR task. The second experiment investigates these pre-training datasets using a speech quality model, analysing their quality predictions distribution and the speaker characteristics (pitch, speaking rate, and intensity) of the datasets. Finally, a third experiment explores how TTS augmentation compares to 100% synthetic speech or natural speech augmented with noise/pitch perturbation/accented natural speech under a low resource scenario where a small amount of natural speech is available. This study will shed light on the current capabilities of synthetic speech for replacing natural speech in pre-training of speech models. It will also reveal crucial factors influencing model performance, allowing us to refine synthetic speech generation for optimal results.

2. Datasets

For this experiment we have generated two synthetic speech dataset equivalent to a natural speech dataset and used those datasets to pre-train a self supervised model with a downstream task of PR. We also used a pre-trained self-supervised model to predict quality of the audio samples used in pre-training to find out the correlation between the quality of pre-trained data and the downstream prediction task.

2.1. Pre-training data

To comprehensively evaluate the impact of synthetic speech on PR model pre-training, we use three datasets:

Natural: Our baseline dataset consists of 100 hours of real human speech data from the LibriSpeech [11] corpus, specifically the Train-Clean-100 subset. **Matched Synthetic:** This dataset mirrors Train-Clean-100 speakers using cloned speaker synthetic speech generation to match the natural speakers. **Unmatched Synthetic**: Synthetic speaker using standard speaker voices from VITs system unmatched to Train-Clean-100 apart from gender balance.

We are not only investigating the effect of using synthetic speech, but also how the size of the pre-training data influences performance. To address this, each dataset (natural and synthetic) has been divided into three subsets – 25 hours, 50 hours, and 75 hours. These subsets preserve the original speech content and speaker distribution across all three datasets. Each 25-hour subset encompasses 62 speakers, with an even balance of 31 males and 31 females. This provides a consistent setup for analyzing the impact of both synthetic speech and dataset size on PR model performance.

For our experiment we are using LibriSpeech[1] [11] as our natural speech dataset. The LibriSpeech corpus consists of 1,000 hours of read English speech from LibriVox audiobooks for Automatic Speech Recognition (ASR) and related tasks. This dataset has gained widespread adoption as a benchmark in various research areas like ASR [14], speaker identification [15], and language modelling [16]. This freely available resource features 2,484 speakers reading diverse texts (fiction, non-fiction, poetry) from Project Gutenberg[2]. The speech underwent manual segmentation and annotation, resulting in time-aligned word labels with their corresponding transcript.

Our **Natural** dataset is divided into balanced subsets, and we are using the Train-Clean-100 subset to train the models for our experiment. The Train-Clean-100 subset contains 100.6 hours of Speech recordings with high audio quality and verified transcripts. This subset of LibriSpeech features recordings from 251 speakers. There is a balanced mix of genders, with 125 females and 126 males represented. While the average speaker contributes roughly 24 minutes of audio, there is some variation, e.g. recording range fro 5.44 – 25.25 mins.

To finetune the PR models, we used a 10-hour subset of speech data extracted from LibriSpeech's Train-Clean-360 subset. We then evaluated the models' performance on the Test-Clean subset of LibriSpeech. Both the fine-tuning data (10 hours) and the test data used phoneme labels from [17].

This experiment utilises a VITS model [3] trained on a vast and diverse dataset of roughly 1,200 speakers which includes 895 speakers from the LibriTTS corpus [18], which was specifically created from the LibriVox project for training TTS models. Notably, LibriTTS shares the same speakers as LibriSpeech.

The VITS model, generously provided by our industry partner Xperi, generates the synthetic audio by taking a speaker ID and a transcript as input. The generated audio mimics the speaking style of the specified speaker, which the model has encountered during training. This significantly simplifies the challenge of creating a balanced synthetic dataset that reflects the speaker and speech characteristics found in LibriSpeech.

We generated two synthetic datasets: **Matched Synthetic** containing the same speakers from the LibriSpeech Train-Clean-100 subset and **Unmatched Synthetic** with the same content but speakers replaced with random speakers from the Train-Clean-360 subset, maintaining speaker gender. As previously explored by [19], all LibriSpeech subsets share similar speaker and speech characteristic distributions. Therefore, both our synthetic datasets inherit this characteristic. The raw transcripts within LibriSpeech lacked punctuation, potentially affecting the generated speech's fluency and intonation. To address this, we employed a deep learning punctuator model [20] to preprocess the transcripts before feeding them into the TTS model. This step ensures high-quality synthetic speech generation.

3. Pretraining with TTS Speech

To address privacy concerns associated with using real human speech to train speech-based models, this experiment explores the possibility of replacing the natural speech with TTS generated synthetic speech in the pre-training of speech based models like APC model for PR. We investigate how this substitution impacts the performance of the model on downstream tasks. We have pre-trained the APC model with 25, 50, 75 and 100 hours of natural, matched synthetic and unmatched synthetic dataset. Following pre-training, these models are fine-tuned with 10 hours data from Train-Clean-360. Finally the performance of the models are evaluated on the Test-Clean subset.

The APC model is based on a 3-layer Long Short-Term Memory (LSTM) [21] architecture, each with hidden layer of 512 units. The APC model ingests a sequence of mel-spectrogram features as input, representing the spectral content of the speech signal. The model then processes these features and outputs a 512-dimensional vector representation for each input frame. During training, the model aims to predict the mel-spectrogram features of future frames, specifically a 3-step prediction horizon. The model's prediction accuracy is evaluated using Mean Squared Error (MSE) [22] loss between predicted and actual future frames. To prepare the model for the downstream PR task, a feed-forward linear layer is added on top of the final LSTM layer's output. This linear layer transforms the 512-dimensional representation into a probability distribution over the total number of phonemes in the target language. A log softmax function is applied to convert the linear layer's output into probabilities suitable for classification. Finally, the cross-entropy loss function is used to measure the difference between the predicted phoneme probabilities and the ground truth phoneme labels. Other hyper-parameter settings are consistent with [12].

3.1. Natural vs. synthetic speech pretraining

Figure 1 shows the phoneme classification accuracy as percentage against the pre-training dataset size in hours for the three pre-training dataset. The natural, matched synthetic and unmatched synthetic datasets are represented by blue, magenta and green color respectively. A significant performance gap can be seen between models pre-trained with natural and synthetic data. Regardless of the pre-training data size (25, 50, 75, or 100 hours), models trained on natural data consistently outperform those trained on synthetic data. Interestingly, a minor difference exists between the performance of models pre-trained with matched and unmatched speaker synthetic datasets. After 50 hours of pre-training or more, the unmatched speaker model shows a slight edge over the matched speaker model. Notably, achieving performance comparable to models trained with natural data requires considerably more synthetic data – approximately 50 hours or more. For instance,

[1] https://www.openslr.org/12

[2] https://www.gutenberg.org/

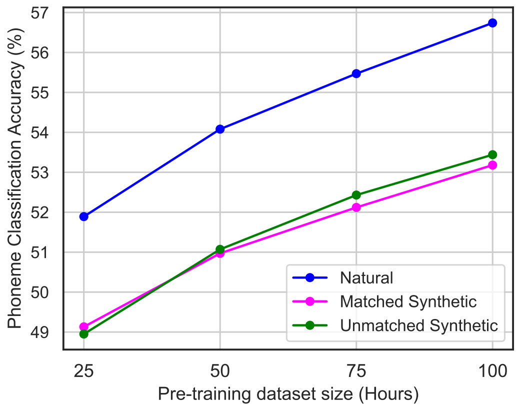

Figure 1: *Performance of APC model for natural (blue) and synthetic (green and magenta) pre-training data of different size (25, 50, 75 and 100 hours)*

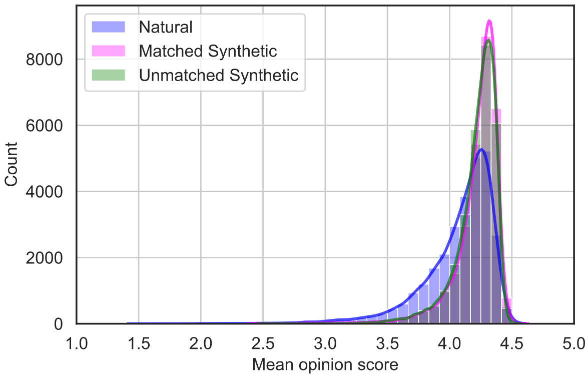

Figure 2: *MOS score distribution for the pre-training datasets: natural (blue), matched speaker synthetic (magenta), and unmatched speaker synthetic dataset (green)*

to reach the accuracy achieved by a model pre-trained with 25 hours of natural data (around 52%), we would require 75 hours of synthetic data. Similarly, to match the performance of a model pre-trained with 50 hours of natural data, we would need over 100 hours of synthetic data.

3.2. Exploring TTS quality and speech characteristics

To gain insights into the significant performance gap observed between models pre-trained with natural and synthetic speech, we have analysed the speech quality and various speech characteristics of these datasets. This analysis aims to identify potential shortcomings in synthetic speech compared to natural speech and understand the root causes of the performance difference. Descriptions of the models used to estimate speech quality and the algorithms for analyzing speech characteristics can be found in sections 3.2 and 3.2, respectively.

Speech quality: Speech quality for the pre-training datasets was evaluated using a self-supervised wav2vec 2.0 architecture[3] [23] fine-tuned for Mean Opinion Score (MOS) prediction [24]. The open-source baseline, detailed in the paper [25], refines the pre-trained model by adding an output layer that aggregates features and trains it using an L1 loss function. Convergence is ensured by monitoring the MSE between predicted and target MOS. Training ceases if there is no improvement in MSE for 20 consecutive epochs. The data for fine-tuning came from main track BVCC dataset [26] of the VoiceMOS challenge. The dataset includes corresponding MOS ratings collected through a unified listening test. We utilised the standard training and development splits provided by the challenge to create our training and validation sets. This approach ensures that unseen synthesis systems, speakers, texts, and listeners are held out in the development set, while maintaining similar overall rating distributions across both sets.

Speech characteristics: This paper focuses on analysing three key speech characteristics: pitch, intensity, and rate. Pitch refers to the perceived highness or lowness of a sound, and in speech, it's primarily determined by vocal fold vibration frequency [27]. To estimate pitch, we employed the CREPE [28] algorithm, available on PyPI and implemented using TensorFlow, provides a predict function for pitch estimation. We set the `step_size` hyper-parameter to 30 to set the pitch estimation interval to 30 ms. Additionally, we enabled Viterbi smoothing for the pitch curve by setting the `viterbi` argument to True. The default model size, "full," was used as recommended in [28]. The predict function returns three lists: timestamps, predicted fundamental frequency in Hz, and voicing confidence i.e. confidence in the presence of a pitch. We only consider pitches with a confidence score exceeding 75%.

Speech intensity refers to the perceived loudness or strength of the sound. In this context, it reflects the amount of acoustic energy produced by the speaker and is typically measured in decibels (dB). We opted to estimate loudness using Loudness Units relative to Full Scale (LUFS) [29] and it was performed using the `pyloudnorm` [30] library.

Speech rate, also known as speaking rate or tempo, refers to the speed at which a person speaks. It is typically measured in Words Per Minute (WPM) by calculating the duration of speech and dividing it by the number of words spoken. We obtained word counts from the available transcripts in LibriSpeech dataset [11] and measured audio duration using `librosa.get_duration` after removing silence at the beginning and end with `librosa.effects.split`. This approach calculates WPM for each speech sample.

Figure 2 compares the MOS distribution of the three datasets. As in Figure 1, the `natural`, `matched synthetic` and `unmatched synthetic` datasets are represented by blue, magenta and green color respectively. The MOS distribution reveals a potential mismatch between perceived and actual speech quality. While the density plots for the synthetic datasets appear narrower and taller compared to natural speech, with most scores concentrated between 3.5 and 4.5, this suggests a higher number of synthetic samples receiving high quality ratings. However, this observation contradicts the performance of the APC model, which shows a significant gap between models trained on natural and synthetic data. This discrepancy implies that the speech quality model might not be capturing the nuances that are crucial for the PR task. In other words, the model might be giving high MOS scores to synthetic speech samples that lack the natural qualities needed for optimal phoneme recognition.

Figure 3 compares the speech characteristics between natural and synthetic datasets. Despite using cloned versions of

[3] https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec

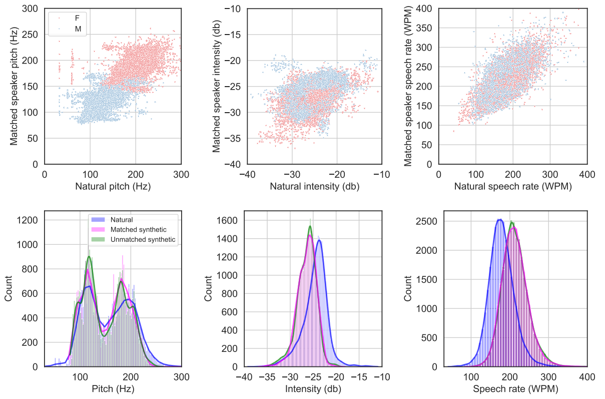

Figure 3: *Top: Speech characteristics scatter plot for natural dataset vs matched speaker dataset where dots color indicate female (red) and male (blue). Bottom: Speech characteristics (pitch, intensity and speech rate) distributions for natural (blue) matched synthetic (magenta), unmatched synthetic (green) pre-training datasets.*

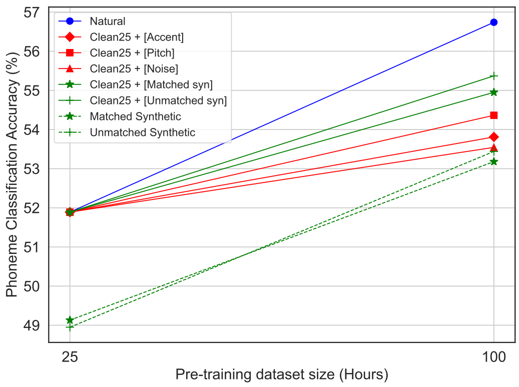

Figure 4: *Performance of different augmented data with and without natural data seed (25 hours) in pre-training*

the speakers from the natural dataset in the matched synthetic dataset, the natural speech exhibits a wider range in both pitch and intensity. Speech rate, however, appears similar but offset across between natural and synthetic datasets, indicating a lower average speaking rate compared to synthetic speech. Interestingly, the matched-speaker data has a high correlation with natural speech for pitch, yet the natural speech distribution remains broader. This suggests the synthetic data may lack natural intonation and prosody, leading to less long-term variation in pitch. Similarly, the intensity distribution for matched synthetic data is narrower despite a lower correlation with natural speech. This might indicate limitations in replicating natural emphasis variations that occurs due to prosody and intonation. The distribution plots further highlight the similarity in speech characteristics between the matched and unmatched synthetic datasets. This aligns with their comparable performance observed in the APC model. The reduced variability in synthetic speech characteristics, particularly pitch and intensity, could potentially contribute to the performance gap between models trained on natural and synthetic data.

4. Data augmentation with synthetic speech

Pre-training with synthetic speech yielded lower performance compared to natural speech. Here we investigate the use of TTS synthetic speech for data augmentation, like for low-resource scenarios. Ullah et al. [31] explore various augmentation strategies for limited pre-training data scenarios. They consider a 25-hour subset (`Clean25`) from `Train-Clean-100` as a low-resource scenario representing the maximum natural data available. They then applied speech modifications like pitch modification, accent augmentation, and noise addition to the `Clean25` data, creating 75 additional hours. This augmented data was combined with the clean data to form new datasets (`Clean25 + [Pitch]`, `Clean25 + [Accent]`, `Clean25 + [Noise]`). These datasets were used to pre-train APC models, followed by evaluation on the same downstream modeling task employed in our experiment. To evaluate TTS speech augmentation, we created two 100-hour datasets: `Clean25 + [Matched synthetic]` and `Clean25 + [Unmatched synthetic]`. These datasets combined the `Clean25` baseline data with 75 hours of matched-speaker and unmatched-speaker synthetic speech data, respectively. We then used these datasets to pre-train the APC model, and the results are presented in Figure 4 plotting phoneme classification accuracy for different pre-training scenarios. We compare models pre-trained with 25 and 100 hours of data, using natural speech (blue), synthetic speech (dashed green), data augmentation with synthetic speech (solid green), and other augmentation strategies (red). Pre-training with only synthetic speech (dashed green) results in the lowest performance. Interestingly, data augmentation using synthetic speech (solid green) outperforms all other augmentation methods (red). This improvement might be attributed to the introduction of novel content through synthetic speech augmentation. This additional variety potentially enriches the phoneme dictionary learned by the model, leading to better performance.

5. Discussion and Conclusions

Despite recent advancements in TTS models, we show synthetic speech currently lacks the naturalness needed to fully replace natural speech for pre-training speech-based models. Even using cloned TTS voices fail to achieve comparable performance on a downstream PR task, likely due to reduced variability in pitch and intensity when compared to natural speech. Additionally, speech quality metrics do not capture naturalness deficiencies beyond single utterance lengths. Nevertheless, synthetic speech shows promise as a data augmentation strategy to improve performance for low-resource languages. Future work should explore the use of diverse downstream tasks to evaluate TTS systems for long-term speech quality aspects like prosody and naturalness in extended speech segments, moving beyond single-utterance quality assessments. Additionally, merging augmentation methods holds promise. Combining perturbations that introduce signal variation with TTS could leverage the strengths of both approaches. Perturbations offer diverse signal variations, while TTS can add variety through a wider range of phone and word variety in generated utterance content. Finally, incorporating different TTS systems into the training data could be beneficial. This might introduce more natural speaker variability and potentially improve model performance.

6. Acknowledgements

This research was conducted with the financial support of Xperi Inc. and Science Foundation Ireland to the Insight Centre for Data Analytics (12/RC/2289_P2). For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. The authors would also like to thank ISCA and the organising committees of past Interspeech conferences for their help and for kindly providing the previous version of this template.

7. References

[1] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi *et al.*, "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 28 492–28 518.

[3] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning (ICML)*. JMLR-JOURNAL MACHINE LEARNING RESEARCH, 2021.

[4] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He *et al.*, "Naturalspeech: End-to-end text-to-speech synthesis with human-level quality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[5] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=piLPYqxtWuA

[6] D. O'shaughnessy, "Interacting with computers by voice: automatic speech recognition and synthesis," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1272–1305, 2003.

[7] K. Hashimoto, J. Yamagishi, W. Byrne, S. King, and K. Tokuda, "An analysis of machine translation and speech synthesis in speech-to-speech translation system," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5108–5111.

[8] A. Peiró-Lilja, G. Cámbara, M. Farrús, and J. Luque, "Naturalness and intelligibility monitoring for text-to-speech evaluation," *Proc. Speech Prosody 2022*, pp. 445–449, 2022.

[9] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[10] N. Amezaga and J. Hajek, "Availability of voice deepfake technology and its impact for good and evil," in *Proceedings of the 23rd Annual Conference on Information Technology Education*, 2022, pp. 23–28.

[11] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[12] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An Unsupervised Autoregressive Model for Speech Representation Learning," in *Proc. Interspeech 2019*, 2019, pp. 146–150.

[13] Y.-A. Chung and J. Glass, "Generative pre-training for speech with autoregressive predictive coding," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3497–3501.

[14] A. Baevski and A. Mohamed, "Effectiveness of self-supervised pre-training for asr," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7694–7698.

[15] Q.-B. Hong, C.-H. Wu, H.-M. Wang, and C.-L. Huang, "Combining deep embeddings of acoustic and articulatory features for speaker identification," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7589–7593.

[16] Z. Yue, F. Xiong, H. Christensen, and J. Barker, "Exploring appropriate acoustic and language modelling choices for continuous dysarthric speech recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6094–6098.

[17] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6419–6423.

[18] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *Proc. Interspeech 2019*, 2019, pp. 1526–1530.

[19] D. Debnath, H. Becerra Martinez, and A. Hines, "Well said: An analysis of the speech characteristics in the librispeech corpus," in *2023 34th Irish Signals and Systems Conference (ISSC)*, 2023, pp. 1–7.

[20] O. Guhr, A.-K. Schumann, F. Bahrmann, and H. J. Böhme, "Fullstop: Multilingual deep models for punctuation prediction," June 2021. [Online]. Available: http://ceur-ws.org/Vol-2957/sepp_paper4.pdf

[21] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[22] *Mean Squared Error*. New York, NY: Springer New York, 2008, pp. 337–339. [Online]. Available: https://doi.org/10.1007/978-0-387-32833-1_251

[23] 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*, vol. 33, pp. 12 449–12 460, 2020.

[24] W. C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS Challenge 2022," in *Proc. Interspeech 2022*, 2022, pp. 4536–4540.

[25] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of mos prediction networks," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8442–8446.

[26] E. Cooper and J. Yamagishi, "How do voices from past speech synthesis challenges compare today?" in *11th ISCA Speech Synthesis Workshop (SSW 11)*. ISCA, 2021.

[27] C. J. Plack and A. J. Oxenham, *Overview: The Present and Future of Pitch*. New York, NY: Springer New York, 2005, pp. 1–6. [Online]. Available: https://doi.org/10.1007/0-387-28958-5_1

[28] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 161–165.

[29] International Telecommunication Union (ITU), "Rec. ITU-R BS.1770, Algorithms to measure audio programme loudness and true-peak audio level." 2006.

[30] C. J. Steinmetz and J. D. Reiss, "pyloudnorm: A simple yet flexible loudness meter in python," in *150th AES Convention*, 2021.

[31] A. Ullah, A. Ragano, and A. Hines, "Reduce, reuse, recycle: Is perturbed data better than other language augmentation for low resource self-supervised speech models," *arXiv preprint arXiv:2309.12763*, 2023.