Speech based evaluation system: A case study on speech synthesis

Adithya P, Bharad  
*Department of Computer Science and Engineering,  
Amrita School of Computing, Bengaluru*Amrita Vishwa Vidyapeetham, India  
[adithyaprasanna121@gmail.com](mailto:adithyaprasanna121@gmail.com), [ba.aaditya03@gmail.com](mailto:ba.aaditya03@gmail.com)

*Abstract*— This paper presents an extensive case study on exploring the use of speech synthesis in a speech-based evaluation system. The study delves into the viability of employing artificially generated speech for offering comments, feedback and delivering prompts within such a system. The case study primarily concentrates on the oral assessment application domain, illuminating the details involved in the design and implementation of the speech synthesis component. We assess the system's effectiveness through user evaluation, focusing on both objective metrics and subjective measures, aiming to highlight the potential impact in real-world scenarios.

Keywords— Speech Processing, Speech synthesis, speech-based evaluation, user evaluation.

# Introduction

The incorporation of speech synthesis technology into various applications has spurred extremely in the current time as it can improve on users experience and accessibility. One prospective research field is the integration of speech synthesis within speech-based evaluation tools. Such tools can have a significant function in many areas, among which are education, training and assessment. Speech synthesis based tools thus become capable of delivering the dynamic and customized feedback to users in a way that is interactive and natural. Although the advantages associated with the integration of speech synthesis are convincing, the effectiveness of speech synthesis within the speech-based evaluation tools remains largely unknown.

Building upon the foundation of artificial speech generation, this research endeavours to explore a specifically designed assessment system. Evaluated on both objective and subjective metrics through robust evaluation methods, the study explores leveraging state of the art AI methodologies allows us to better understand the impact speech synthesis has on the overall user engagement and learning outcomes. Our findings will highlight the possibilities and challenges associated with this integration, paving the way for the development of future generations of intelligent and engaging assessment tools

In the context of learning, training, and assessing, the study emphasize its ability to change established pedagogy by integrating a speech synthesis into our evaluation system. This aims to reveal new levels of user participation and learning efficiency.

[results, what has been done and achieved]

# RELATED WORKS

Speech synthesis holds significant importance in the realm of futuristic evaluation tools. This innovative process allows for the transformative integration of auditory experiences in assessments, introducing a novel dimension of feedback and learning efficacy. This significance lies not only in its diverse applications across domains like education, training, skill assessments but also in accessibility, engagement and personalization. This research adds a layer of benefit not only to users directly experiencing the novel assessment systems but also to educators and trainers and assessment developers.

Gao et al. [1] examine a pilot study about using personality traits to control the perceived personality of synthetic voices in human-robot interaction. Their work investigates the possibility of influencing personality traits, primarily extroversion and agreeableness, via a state-of-art DNN speech synthesis system. The researchers used an English native speaker voices dataset, carried out listening tests, and assessed the personality trait synthesis in the generated voices. They conclude that controlling the perceived personality may be possible, but obstacles and limits as necessitating the wider and more diversified data set are mentioned. Possible future work includes applying the synthetic voices in multi-turn conversational agent setups and investigating their effect in different cultures.

AI-based synthetic speech is investigated by researches to their everyday perception. By using Google’s Wavenet they examine its capability in the study of speech-in-noise perception in not only the younger but also the older adults. The study, Herrmaann et al. [2] shows that AI speech like human speech shows masking release, but older people have lower benefit from such effects. Though modern AI speech is potentially important for audiology and phonetics research, it is recognized with less frequency by older adults, is considered as more natural and is more difficult to tell apart between human and AI speech by them in comparison with the younger adult population. This exemplifies the possibility of AI-generated speech for researching speech processing in elderly people due to its natural sound.

Ning and Co [3] focus on the latest results in the field of speech synthesis, particularly from deep learning and end-to-end approaches. These advancements have greatly transformed many applications in the field of human oriented speech interaction and conversational AI. Their paper commences by presenting traditional speech synthesis methods, focusing on the critical role of acoustic modelling in statistical parametric speech synthesis (SPSS) systems. It then gives a high-level summary of the advances in deep learning-based speech synthesis, where mainly end-to-end models have achieved compelling results. The paper deals with issues associated with deep learning in speech synthesis and provides avenues for addressing this challenge.

The authors Zhen-Hua et al [4] introduce HiNet, a neural vocoder that reconstructs speech waveforms from acoustic features with hierarchical prediction of amplitude and phase spectra. Different from existing neural vocoders, HiNet has an Amplitude Spectrum Predictor (ASP) and a Phase Spectrum Predictor (PSP). ASP, a simple DNN model, can predict log amplitude spectra which are then given to PSP for the purpose of phase recovery. The PSP solves the problem of phase warping by merging a neural source-filter waveform generator with a phase extractor. GANs are integrated into ASP and PSP. The cumulative outputs facilitate phoneme superposition by short-time Fourier synthesis. HiNet overperforms traditional vocoders while matching the performance of a 16-bit WaveRNN vocoder. Critically, the efficiency of HiNet is also maintained even when dealing with simplified models, thus cutting down the GPU processing time from 0.34 seconds to 0.19 seconds for generating the waveforms of a one-second 16 kHz speech without noticeable quality loss.

Yi Lei et al. [5] describe MsEmoTTS, a new multi-scale emotional speech synthesizer model aimed at achieving expressive synthetic speech in various applications. Unlike previous methods that concentrated on average style representation, MsEmoTTS incorporates three modules (GM, UM, LM) in an attention-based sequence-to-sequence model for the global emotion category, utterance-level emotion variation, and syllable-level emotion strength, respectively. Such approach allows flexible synthetic voice for emotional speech, by transfer of emotions from another reference voice, prediction of emotion from input text, and manual control of emotional strength. Evaluation on an emotional speech corpus of Chinese shows the effectiveness of MsEmoTTS in emotion transfer, text-based emotion prediction, and flexible emotion control, outperforming audio-based and text-based methods.

Wei Wang et al. [6] offer a novel speech synthesis framework accounts the problem of low-resource languages in Automatic Speech Recognition (ASR). The framework concentrates on the application of hidden units of self-supervised learning (SSL) from Hidden Unit BERT (HuBERT) as universal phonetic units, leveraging these SSL phonetic units to connect speech and text in diverse languages. Through data splicing, speech fragments from high-resource languages are smoothly connected to synthesized speech, which guarantees acoustic consistency with text of low-resource languages. In order to increase practicality, a confidence-based sampling technique is introduced which dramatically speeds up the ASR model convergence and improves overall performance. The experiments conducted on the Common-Voice dataset shows that the proposed model achieved superior relative gains in language-specific ASR, thereby proving the scalability of the framework when we incorporated a larger unsupervised speech corpus for synthesizing speech fragments in data splicing.

Automatic public speaking training has seen advancements in visual dashboards and emotion recognition, Wynn et al., [7] While some dashboards solely visualize speech parameters, others integrate emotion recognition. Studies on emotion in speech inform this exploration, as do investigations into the role of pitch. Studies that combine these elements specifically within the chosen language and context has further strengthened research and highlighted their own unique contributions.

The authors Takaaki Saeki et al., [8] present an incremental text-to-speech (TTS) method which is aimed at high-quality and low-latency synthesis through utilization of a pseudo lookahead generated with a pre-trained language model, namely GPT2. On the contrary, such an approach is unlike the other classical TTS models that concentrate on short linguistic pieces of text, at times only consisting of several words. To alleviate the problem of striking a balance between naturalness and synthesis latency, the proposed method predicts lookahead using GPT2 which gives a scope to estimate contextual information without waiting on the actual look-ahead. In addition to that, the authors put forward a language model-based fine-tuning approach to improve the performance of the pseudo lookahead. The experimental evaluations on LJSpeech dataset show that the proposed method achieves better synthesized speech qualities and latency than the existing approaches.

The authors Jordi Adelle et al. [9] describe a system to include filled pauses in the generation of conversational speech synthesis, giving prominence to the relevance of getting beyond the read-style synthesis so as to achieve more natural and conversational voices. They maintain that multimedia application styles like film dubbing, robotics, dialog systems, and multilingual broadcasting necessitate varied speech styles beyond traditional reading. The system applies machine learning methods in particular a language modelling combined with decision trees to predict and insert filled pauses in text. The algorithm is trained on an extemporaneous Spanish corpus collected within the LC-STAR project, concentrating on conversational interactions about tourist information. The authors underwent objective evaluations yielding up to 96% of precision in filled pause insertion and perceptual evaluation to show that filled pauses improve the naturalness and human-like qualities of synthetic speech. The research deals with the rhythmic characteristics of filled pauses in speech synthesis as well, pointing out a noticeable pace difference in filled pauses compared to any other elements of the speech.

# METHODOLOGY

# RESULTS AND ANALYSIS

##### CONCLUSION AND FUTURE SCOPE

##### References

1. Gao, Shilin, Matthew P. Aylett, David A. Braude, and Catherine Lai. ‘Synthesising Personality with Neural Speech Synthesis’. In Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue, edited by Svetlana Stoyanchev, Shafiq Joty, David Schlangen, Ondrej Dusek, Casey Kennington, and Malihe Alikhani, 393–99. Prague, Czechia: Association for Computational Linguistics, 2023.
2. Herrmann, B. The perception of artificial-intelligence (AI) based synthesized speech in younger and older adults. Int J Speech Technol 26, 395–415 (2023).
3. Ning Y, He S, Wu Z, Xing C, Zhang L-J. A Review of Deep Learning Based Speech Synthesis. Applied Sciences. 2019; 9(19):4050.
4. Y. Ai and Z. -H. Ling, "A Neural Vocoder With Hierarchical Generation of Amplitude and Phase Spectra for Statistical Parametric Speech Synthesis," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 839-851, 2020, doi: 10.1109/TASLP.2020.2970241
5. Y. Lei, S. Yang, X. Wang and L. Xie, "MsEmoTTS: Multi-Scale Emotion Transfer, Prediction, and Control for Emotional Speech Synthesis," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 853-864, 2022, doi: 10.1109/TASLP.2022.3145293.
6. W. Wang and Y. Qian, "Universal Cross-Lingual Data Generation for Low Resource ASR," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 32, pp. 973-983, 2024, doi: 10.1109/TASLP.2023.3345150.
7. Wynn, A.T., Wang, J., Umezawa, K., Cristea, A.I. (2022). An AI-Based Feedback Visualisation System for Speech Training. In: Rodrigo, M.M., Matsuda, N., Cristea, A.I., Dimitrova, V. (eds) Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners’ and Doctoral Consortium. AIED 2022. Lecture Notes in Computer Science, vol 13356. Springer, Cham.
8. T. Saeki, S. Takamichi and H. Saruwatari, "Incremental Text-to-Speech Synthesis Using Pseudo Lookahead With Large Pretrained Language Model," in IEEE Signal Processing Letters, vol. 28, pp. 857-861, 2021, doi: 10.1109/LSP.2021.3073869.
9. Adell, J., Bonafonte, A., Escudero, D. (2007). Filled Pauses in Speech Synthesis: Towards Conversational Speech. In: Matoušek, V., Mautner, P. (eds) Text, Speech and Dialogue. TSD 2007. Lecture Notes in Computer Science(), vol 4629. Springer, Berlin, Heidelberg.
10. M. Zhang, X. Zhou, Z. Wu and H. Li, "Towards Zero-Shot Multi-Speaker Multi-Accent Text-to-Speech Synthesis," in IEEE Signal Processing Letters, vol. 30, pp. 947-951, 2023, doi: 10.1109/LSP.2023.3292740.