Title: Optimizing Human Intervention for Synthetic Speech Quality Evaluation: Active Learning for Adaptability

Keywords: Active Learning, Synthetic Speech Quality Evaluation, Subjective Quality Modeling, Training Set Design for Domain Adaptation

Context:

The primary objective of Text-to-Speech (TTS), speech conversion and speech to speech translation system is to synthesize or generate a high-quality speech signal. Typically, the quality of synthetic speech is subjectively evaluated by human listeners. This listening test aims to assess the degree of similarity to human speech rather than machine-like speech. The main challenge in assessing synthetic speech quality lies in finding a balance between the cost and reliability of evaluation. When the cost of conducting a human listening test is high, an automatic quality evaluation may be less reliable. Additionally, the definition of quality can be varied in different perspectives [7]. The quality of TTS output can be described in terms of various aspects such as intelligibility, naturalness, expressiveness, and the presence of noise. Furthermore, fine differences between two signals cannot be precisely tracked through Mean Opinion Score (MOS) ratings [1]. Moreover, the evolution of TTS systems has altered the nature of quality evaluation. Significant improvements in synthetic speech quality have been made over the last decade [2]. And while in the past, the emphasis was on intelligibility in speech synthesis, today, the focus is more on the expressiveness of synthetic speech. Recent efforts toward the automatic evaluation of synthesized speech [4] have demonstrated the success of objective metrics when the domain, language, and system are limited. In addition to the evolution of TTS quality over time, studies such as [10] and [8] have emphasized the need for new data collection and annotation for domain and language adaptation.

Objective:

The main objective of this thesis is to propose an active learning approach [9], where human intervention should be minimum, for a subjective task such as automatic evaluation of synthetic speech quality. The core of this framework would be an objective model as synthetic quality predictors, which require a diverse and efficient training samples. It is desired to efficiently improve the precision of synthetic quality prediction or adapt the synthetic quality predictors to new domains and new generation of systems. It is essential to address different aspects of quality, domain-specific requirements, and linguistic variation through the acquisition of new data or retraining models with a specific emphasis on targeted sample sets.

The main goal is to efficiently collect and query data to minimize information gaps, ensuring a comprehensive dataset for adaptation in order to maximize the performance improvement. The main adaptations that will be investigated in this project are language (adapting a trained quality predictor to a new language) and expressive speech synthesis (adapting a trained naturalness predictor to an expressive speech quality predictor). This adaptation could potentially extend to different listeners and system types, e.g. systems with different acoustic models or vocoder. In this context, the data collection (synthesizing new samples) is cost-effective, which allows focusing on only query optimization to identify the most informative samples. In a secondary objective, we will focus on modeling listeners' disagreements in quality evaluation. This objective aims to address the diverse perspectives on the perception of TTS quality. Furthermore, this objective will work towards personalized quality prediction for TTS based on listeners' individual definitions of quality. Consequently, analysing challenging scripts can reveal remaining challenges in the Text-to-Speech field.

Reference:

[1] Joshua Camp et al. “MOS vs. AB: Evaluating Text-to-Speech Systems Reliably Using Clustered Standard Errors”. In: Interspeech. 2023, pp. 1090–1094.
[2] Erica Cooper and Junichi Yamagishi. “How do Voices from Past Speech Synthesis Challenges Compare Today?” In: Proc. 11th ISCA Speech Synthesis Workshop (SSW 11). 2021, pp. 183–188. doi: 10.21437/SSW.2021-32.
[3] Erica Cooper et al. “Generalization ability of MOS prediction networks”. In: ICASSP. IEEE. 2022, pp. 8442–8446.
[4] Wen Chin Huang et al. “The VoiceMOS Challenge 2022”. In: Interspeech. 2022, pp. 4536–4540. doi: 10.21437/Interspeech.2022-970.
[5] Georgia Maniati et al. “SOMOS: The Samsung Open MOS Dataset for the Evaluation of Neural Text-to-Speech Synthesis”. In: Interspeech. 2022, pp. 2388–2392. doi: 10.21437/Interspeech.2022-10922.
[6] Felix Saget et al. “LIUM-TTS entry for Blizzard 2023”. In: Blizzard Challenge Workshop. 2023. doi: hal.science/hal-04188761.
[7] Fritz Seebauer et al. “Re-examining the quality dimensions of synthetic speech”. In: Proc. 12th ISCA Speech Synthesis Workshop (SSW2023). 2023, pp. 34–40. doi: 10.21437/SSW.2023-6.
[8] Thibault Sellam et al. “SQuId: Measuring speech naturalness in many languages”. In: ICASSP. IEEE. 2023, pp. 1–5.
[9] Burr Settles. “Active learning literature survey”. In: (2009).
[10] Wei-Cheng Tseng, Wei-Tsung Kao, and Hung-yi Lee. “DDOS: A MOS Prediction Framework utilizing Domain Adaptive Pre-training and Distribution of Opinion Scores”. In: Interspeech. 2022, pp. 4541–4545.

Host laboratory : LIUM

Location : Le Mans, France

Supervisors : Anthony Larcher, Meysam Shamsi

Applicant profile : Candidate motivated by Artificial Intelligence, with Master's degree in Computer Science, Signal processing, Speech analysis or related fields

Instructions for Application: Send **CV + letter/message of motivation + master’s note** to : meysam.shamsi@univ-lemans.fr and anthony.larcher@univ-lemans.fr