Titre : Optimisation de l'évaluation humaine pour la qualité de la parole synthétique : apprentissage actif pour l'adaptation

Mots clefs : Apprentissage actif, évaluation de la qualité de voix synthétique, modélisation de l'évaluation subjective, sélection de données d'entraînement pour adaptation au domaine

Contexte :

L'objectif principal des systèmes de synthèse vocale (TTS), de conversion vocale et de traduction parole-parole est de synthétiser ou de générer un signal vocal de haute qualité. En règle générale, la qualité de la parole synthétique est évaluée subjectivement par des auditeurs humains. Ce test d'écoute vise à comparer le degré de similitude avec la parole humaine plutôt qu'avec la parole d'une machine. Le principal défi de l'évaluation de la qualité de la parole synthétique réside dans la recherche d'un équilibre entre le coût et la fiabilité de l'évaluation. Alors que le coût d'un test d'écoute humaine est élevé, une évaluation automatique de la qualité peut, elle, s'avérer moins fiable. De plus, la définition de la qualité peut varier [7]. La qualité de la sortie TTS peut être décrite selon divers aspects tels que l'intelligibilité, le naturel, l'expressivité et la présence de bruit. Les différences fines entre deux signaux ne peuvent pas être évaluées avec précision grâce aux évaluations du score d'opinion moyen (MOS) [1]. De plus, l'évolution des systèmes TTS a modifié la nature de l'évaluation de la qualité. Des améliorations significatives de la qualité de la parole synthétique ont été réalisées au cours de la dernière décennie [2] et, alors que dans le passé, l'accent était mis sur l'intelligibilité dans la synthèse vocale, c'est désormais l'expressivité de la parole synthétique qui est mise en avant. Des efforts récents vers l'évaluation automatique de la parole synthétisée [4] ont démontré le succès des métriques objectives lorsque le domaine, le langage et le système sont limités. En plus de l'évolution de la qualité TTS au fil du temps, des études telles que [10] et [8] ont souligné la nécessité de nouvelles collectes de données et d'annotations pour l'adaptation du domaine et du langage.

Objectifs :

L'objectif principal de cette thèse est de proposer une approche d'apprentissage actif [9] minimisant l'intervention humaine pour une tâche subjective telle que l'évaluation automatique de la qualité de la parole synthétique. Le cœur de ce projet est un modèle objectif prédisant la qualité de la parole synthétique, qui nécessite des échantillons diversifiés et optimisant l'apprentissage. Nous souhaitons améliorer la précision de la prédiction de qualité de la voix synthétique ou permettre l'adaptation des prédicteurs de qualité à de nouveaux domaines et à de nouvelles générations de systèmes. Il est essentiel d'aborder différents aspects de la qualité, comme les exigences spécifiques aux domaines ou les variations linguistiques, en permettant l'acquisition de nouvelles données ou l'utilisation de modèles de recyclage en mettant un accent particulier sur des ensembles d'échantillons ciblés. L'objectif principal est de collecter et d'interroger efficacement les données afin de minimiser les lacunes d'informations, garantissant ainsi un ensemble de données complet maximisant l'amélioration des performances. Les deux principales adaptations qui seront étudiées dans ce projet sont le langage (adapter un prédicteur de qualité formé à une nouvelle langue) et la synthèse vocale expressive (adapter un prédicteur de naturel formé à un prédicteur de qualité de parole expressive). Cette adaptation pourrait potentiellement s'étendre à différents auditeurs et types de systèmes, par ex. systèmes avec différents modèles acoustiques ou vocodeurs. Dans ce contexte, la collecte de données (synthèse de nouveaux échantillons) est rentable, ce qui permet de se concentrer uniquement sur l'optimisation des requêtes pour identifier les échantillons les plus informatifs.
Dans un objectif secondaire, nous nous concentrerons sur la modélisation des désaccords des auditeurs en matière d'évaluation de la qualité. Cet objectif vise à aborder les différentes perspectives sur la perception de la qualité TTS. De plus, cet objectif visera une prédiction personnalisée de la qualité du TTS basée sur les définitions individuelles de la qualité des auditeurs. Par conséquent, l'analyse de scripts complexes peut révéler les défis restants dans le domaine de la synthèse vocale.

References:

[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.

Laboratoire hôte : LIUM

Localisation : Le Mans, France

Encadrants: Anthony LARCHER, Meysam SHAMSI

Profil du/de la candidat.e : Candidat.e motivé.e par l'intelligence artificielle possédant un diplôme de Master en informatique, traitement des données, traitement du signal, analyse de la parole et du langage ou dans un autre domaine en lien avec le sujet de thèse.

Instructions pour candidater : Envoyer un CV + lettre de motivation accompagnés des notes de master à : meysam.shamsi@univ-lemans.fr et anthony.larcher@univ-lemans.fr