Sujet de stage de master recherche

Génération de parole avec un contrôle de l'accent : Étude de la robustesse de modèles de plongement (embedding) d'accent

Contacts : philippe.martin@irisa.fr, vincent.barreaud@irisa.fr

Localisation : Le stage s'effectuera au sein de l'équipe Expression de l'IRISA, dans les locaux situés à Lannion.

Mots clefs : Apprentissage profond ; Synthèse de la parole ; Classification d'accent ; Évaluation automatique ; Traitement automatique des langues ;

1 Objet du stage

Ce stage a pour but de compléter le travail de thèse sur le contrôle d'un système de synthèse vocale (TTS) à partir de caractéristiques d'accent [4]. Son objectif est de concevoir des expériences pour détecter la présence, dans les modèles d'accents [7] [8] [2] [3], de caractérisations d'accents établies par la linguistique. Cette étude peut s'apparenter à une tâche de démêlage d'embeddings, relativement à l'environnement d'enregistrement, à la physionomie du locuteur, à son rythme d'élocution, à sa langue natale, etc.

2 Contexte et motivation

Cette thèse s'inscrit dans le cadre du projet EVA (Explicite Voice Attribute) de l'ANR (Agence Nationale de Recherche). L'objectif de ce projet est de déchiffrer les codes des voix humaines par l'apprentissage de représentations explicites et structurées des attributs de la voix (la voix est-elle "grave", "rauque", "soufflée" ? Le locuteur a-t-il un accent ? Est-il "jeune" ou "vieux" ? ...). Le principal cas d'usage est l'anonymisation de la voix : afin de permettre des enregistrements vocaux conformes au RGPD, les systèmes de conversion vocale pourraient être configurés pour supprimer les attributs fortement associés à l'identité d'un locuteur.

Pour cette thèse, l'attribut choisi est complexe, puisqu'il s'agit de l'accent de locuteurs s'exprimant en anglais (l'anglais pouvant être la langue maternelle ou non). Or, en linguistique, l'accent en anglais se caractérise par des aspects segmentaux (comme la distance entre les voyelles) et non segmentaux (le rythme, la présence d'accents toniques). D'autre part, les modèles d'embeddings d'accent, du fait de leur architecture (réseaux de neurones convolutionnels ou de type transformer), ne capturent pas explicitement ces caractéristiques. Il est donc difficile de garantir que les caractéristiques recherchées soient bien représentées dans l'espace d'embeddings et, par extension, si le système de TTS est bien piloté par les bonnes caractéristiques [4][6].

C'est pourquoi nous envisageons d'établir un protocole d'évaluation des différents modèles d'embeddings d'accent sur une variété de corpus [5] [1] par rapport à ces caractéristiques et, en premier lieu, le rythme.

3 Missions

Pour répondre aux objectifs de ce stage la/le candidat(e) devra :

- Prendre en main le processus de création d'embeddings d'accents
- Identifier les métriques qui prendront compte des caractéristiques choisies (en premier lieu le rythme).
- Adapter les modèles d'embeddings d'accents choisis au processus
- Développer des scripts (python) pour le calcul des métriques
- Rédiger un document récapitulatif des expériences.
  Nous souhaiterions que cela aboutisse sur un article scientifique rédigé par la/le candidat(e).
- Possiblement, proposer une fonction de perte qui permette en prendre en compte les caractéristiques absentes des embeddings.

4 Compétences requises

- Connaissance du langage Python : *La/le candidat(e) devra être à l'aise avec le Python car c'est le language que nous utilisons très majoritairement.*
- Notion en apprentissage profond : *La/le candidat(e) sera amené(e) à utiliser et à modifier légèrement le comportement de modèle d'apprentissage profond.*
- Notion en système Linux : *La/le candidat(e) sera amené à utiliser des machines de calcul distantes utilisant différentes variantes d'Ubuntu.*
- Rigueur : *La/le candidat(e) réalisera de nombreux tests et mesures. Les résultats devront être reproductibles, documentés et clairs.*
- Curiosité : *C'est un incontournable pour faire de la recherche.*

Références

[1] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber. Common voice : A massively-multilingual speech corpus. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France, May 2020. European Language Resources Association.

[2] N. Bafna and M. Wiesner. LID models are actually accent classifiers : Implications and solutions for LID on accented speech. In *Interspeech 2025*, pages 1488–1492. ISCA.

[3] P. Foley, M. Wiesner, B. Odoom, L. P. Garcia Perera, K. Murray, and P. Koehn. Where are you from ? geolocating speech and applications to language identification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, pages 5114–5126. Association for Computational Linguistics.

[4] L. Ma, Y. Zhang, X. Zhu, Y. Lei, Z. Ning, P. Zhu, and L. Xie. Accent-VITS : Accent transfer for end-to-end TTS. In *Man-Machine Speech Communication - 18th National Conference, NCMMSC 2023, Proceedings*, pages 203–214. Springer Science and Business Media Deutschland GmbH.

[5] X. Shi, F. Yu, Y. Lu, Y. Liang, Q. Feng, D. Wang, Y. Qian, and L. Xie. The accented english speech recognition challenge 2020 : Open datasets, tracks, baselines, results and methods. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6918–6922. IEEE.

[6] W. Wang, Y. Song, and S. Jha. USAT : A universal speaker-adaptive text-to-speech approach. 32 :2590–2604.

[7] J. Zhong, K. Richmond, Z. Su, and S. Sun. AccentBox : Towards high-fidelity zero-shot accent generation. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025. ISSN : 2379-190X.

[8] J. Zuluaga-Gomez, S. Ahmed, D. Visockas, and C. Subakan. CommonAccent : Exploring Large Acoustic Pretrained Models for Accent Classification Based on Common Voice. In *Interspeech 2023*, pages 5291–5295, 2023.