Cadre d'apprentissage conjoint d'un système de synthèse de la parole à partir du texte et de la conversion de voix dans contexte d'apprentissage profond

EXPRESSION - Proposition de stage de Master Recherche en Informatique

Octobre 2021

Mots-clés

Text-to-Speech synthesis, Voice conversion, Deep learning

Contexte général

La synthèse de la parole à partir du texte et la conversion de voix sont deux techniques de génération de la parole distincte. La synthèse de la parole (*en.* Text To Speech (TTS)) est un processus qui permet de générer la parole à partir d'une séquence de graphèmes ou de phonèmes. Quant à elle, la conversion de la voix permet de convertir de la parole produite par une voix source en une autre voix cible. Ces procédés trouvent leur application dans des domaines tel que l'Apprentissage des Langues Assisté par Ordinateur, par exmeple.

Cependant, ces deux processus partagent certaines briques notamment le vocodeur qui permet de générer de la parole à partir de caractéristiques acoustiques ou du spectrogramme. La qualité de ces deux technologies a été nettement améliorée grâce, notamment, à la disponibilité de bases de données massives, à la puissance des machines de calcul et à la mise en place du paradigme d'apprentissage profond (en. Deep Learning). En revanche, restituer ou contrôler l'expressivité, et plus généralement tenir compte des informations suprasegmentales, reste un défi majeur pour ces deux technologies.

Ce sujet de stage vise à mettre en place un cadre commun aux deux technologies. Nous visons un cadre d'apprentissage profond conjoint permettant de générer de la parole (voix cible) que cela soit à partir de la parole (voix source) ou du texte.

Travail demandé

Nous proposons de découper le déroulement de ce travail de stage en 3 phases.

— Phase 1 : Mise en place d'un environnement de travail qui consiste à fixer les baselines respectifs pour les deux technologies en se basant sur l'état de l'art.
Le *Blizzard challenge* pour la synthèse de la parole, et le *Voice conversion challenge* pour la conversion de la voix.

— Phase 2 : Proposer un modèle acoustique permettant de prédire des caractéristiques acoustiques (mel-spectrogramme) à partir de caractéristiques issues du texte ou de la parole.
— Phase 3 : Évaluation objective et subjective des deux modes (Synthèse et conversion)

Environnement technique

— Python, HTML/CSS, JavaScript
— pyTorch, ScikitLearn
— TorchAudio

Encadrement

Ce stage se déroulera à Lannion, dans les locaux de l'IRISA, au sein de l'ENSSAT école d'ingénieurs située en centre ville. Il sera ancadré par des membres de l'équipe EXPRESSION : Aghilas Sini, Pierre Alain, Damien Lolive et Arnaud Delhay-Lorrain.

Merci d'adresser vos messages à tous les contacts : aghilas.sini@irisa.fr, pierre.alain@enssat.fr, damien.lolive@irisa.fr, arnaud.delhay@irisa.fr

Références

[1] R. Prenger, R. Valle, and B. Catanzaro. Waveglow : A flow-based generative network for speech synthesis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621, 2019.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783, 2018.

[3] P. Taylor. *Text-to-speech synthesis*. Cambridge university press, 2009.

[4] G. Zhao, S. Ding, and R. Gutierrez-Osuna. Foreign accent conversion by synthesizing speech from phonetic posteriorgrams. In *INTERSPEECH*, pages 2843–2847, 2019.

[5] Y. Zhao, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda. Voice conversion challenge 2020 : Intra-lingual semi-parallel and cross-lingual voice conversion. *arXiv preprint arXiv :2008.12527*, 2020.