gipsa-lab Cognitive Robotics, Interactive Systems, & Speech Processing (CRISSP) team	Proposition de Master 2/PFE « Exploring style embeddings of an expressive text-to-speech system »

General context

Within the THERADIA project (see https://www.theradia.fr/), we are developing a conversational agent for computer-assisted therapies. The avatar will be driven by text augmented with expressive tags generated by an emotion-aware dialog system.

The problem

We already have two implementations of a French end-to-end text-to-speech system with neutral reading style (Tacotron2 [Shen et al. 2018] and FastSpeech2 [Ren et al. 2020]).

These systems have a common structure: a text encoder converts input characters into contextual embeddings. A second module aligns the text with audio frames. The audio decoder converts these aligned contextual embeddings into a spectrogram, that is further converted into speech by a neural vocoder (wavenet, waveglow, hifigan, etc). Both systems have been trained using more than 100h from audiobooks read by 5 speakers. As for data-driven approaches, the overall performance is quite high and synthetic speech natural-sounding (see http://www.gipsa-lab.fr/~martin.lenglet/segmentation_impact).

Expressive TTS is becoming a hot topic in speech technology nowadays: a so-called style encoder is added to the neutral TTS and tries to disentangle emotional dimensions from linguistic prosody that is computed from the raw text. This module is trained to generate a latent space with few dimensions from example spectrograms as input, and coordinates in this latent space are used to bias the output of the text encoder of the neutral TTS. In particular, the Google speech group [Wang et al., 2018] introduced the Global Style Tokens (GST) model that projects the expressive latent space on controllable vectors, as an attempt to disentangle emotional dimensions. Nevertheless, the interpretability of these dimensions remains unclear.

We already developed tools to explore latent spaces, in the particular case of the contextual embeddings built by the text encoders. We have demonstrated that letter-to-sound mapping (including prediction of liaisons and positioning of pauses) is performed quite accurately; and duration, melody, spectral tilt of phonemes, etc. can also be retrieved from the embeddings. [Lenglet et al. JEP 2022, Lenglet et al. Interspeech 2022]

The main objective of this work is to similarly explore the embeddings computed by the style encoder in order to gain interpretability and compare different solutions.

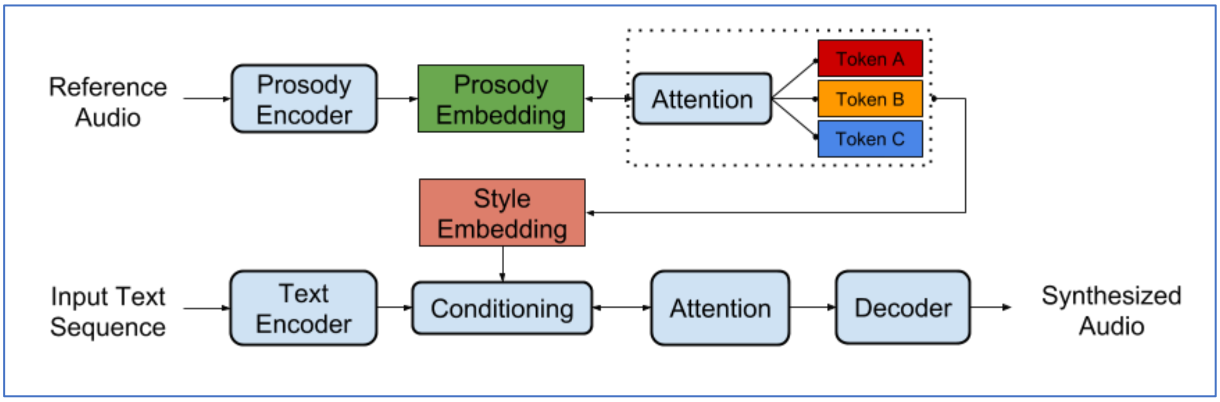

Figure 1. Style embeddings computed by a prosody encoder that conditions the text encoder of an end-to-end TTS.

The work

We already collected 10 hours of expressive utterances: one of our speakers was instructed to utter the same text with 11 different attitudes (suppliante, réconfortante, pensive, incrédule, étonnée, désolée, déterminée, etc).
The proposed internship will consist in:

- Implementing and grafting a GST component onto our two neutral TTS
- Exploring style embeddings using standard data analysis tools (t-SNE, PCA, MDS, LDA, etc.)
- Proposing methods to evaluate the controllability and navigability of the latent space

Training allowance

The work is financed by the project THERADIA. The by-law monthly allowance is around 580€.

Contact

Gérard BAILLY, DR CNRS, gerard.bailly@gipsa-lab.fr 04 76 57 47 11
Olivier PERROTIN, CR CNRS, olivier.perrotin@gipsa-lab.fr
Martin LENGLET, PhD, martin.lenglet@gipsa-lab

Refs:

1. Shen J. et al. (2018). Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In Proc. ICASSP (pp. 4779-4783). IEEE
2. Wang, Y., Stanton, D., Zhang, Y., Ryan, R. S., Battenberg, E., Shor, J., & Saurous, R. A. (2018, July). Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In International Conference on Machine Learning (pp. 5180-5189). PMLR.
3. Kwon, O., Jang, I., Ahn, C., & Kang, H. G. (2019, June). Emotional speech synthesis based on style embedded Tacotron2 framework. In 2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC) (pp. 1-4). IEEE.
4. Lenglet, M., Perrotin, O., & Bailly, G. (2022, June). Modélisation de la Parole avec Tacotron2: Analyse acoustique et phonétique des plongements de caractère. In JEP 2022-34e Journées d'Études sur la Parole.
5. Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. (2020). Fastspeech 2: Fast and high-quality end-to-end text to speech. arXiv preprint arXiv:2006.04558.
6. Lenglet, M., Perrotin, O., & Bailly, G. (2022, September). Speaking Rate Control of end-to-end TTS Models by Direct Manipulation of the Encoder's Output Embeddings. In Interspeech 2022 (pp. 11-15). ISCA.
7. Shin, Y., Lee, Y., Jo, S., Hwang, Y., Kim, T. (2022) Text-driven Emotional Style Control and Cross-speaker Style Transfer in Neural TTS. Proc. Interspeech 2022, 2313-2317, doi: 10.21437/Interspeech.2022-10131