ICE-Talk: an Interface for a Controllable Expressive Talking Machine

Noé Tits [iD]*, Kevin El Haddad* [iD]*, Thierry Dutoit*

Numediart Institute, University of Mons

{noe.tits, kevin.elhaddad, thierry.dutoit}@umons.ac.be

Abstract

ICE-Talk is an open source[1] web-based GUI that allows the use of a TTS system with controllable parameters via a text field and a clickable 2D plot. It enables the study of latent spaces for controllable TTS. Moreover it is implemented as a module that can be used as part of a Human-Agent interaction.

Index Terms: expressive, controllable, speech synthesis, human-computer interaction, deep learning, web interface

1. Introduction and Motivations

Speech Synthesis is an important component of Human-Robot Interaction [1]. However as of today, expressiveness in speech generated by Text-to-Speech (TTS) systems is under-explored in such interactions. The reason is the difficulty of accessing the variables controlling speech expressiveness in a deep learning-based TTS system [2].

To tackle this problem we propose a tool allowing the control of these variables through a graphical interface, thus contributing to the democratization of the use of Deep Learning (DL)-based TTS systems in Human-Agent interaction (HAI) applications. The goal of this tool is to be generic enough to be connected in any HAI application.

This interface allows for the control over the synthesis parameters of a deep learning-based model through its latent space directly and intuitively in a graphical way. It therefore allows the implementation of several interesting applications and experiments such as listening tests for the evaluation of such systems thanks to easy prototyping of experiments. An example of experiment is available. This tool also enables the possibility of studying the impact of expressive synthesized speech in Human-Robot interaction.

2. Related Work

As of today, there are some web interfaces allowing the use of DL TTS models[2]. They allow to write text, that is sent to the model and get the synthesized speech as an audio object that one can listen. The text is therefore the only control variable that we can access.

Recently, an interface[3] allowing to give an audio for TTS with speaker characteristics was developed based on the research of Tacotron team [3, 4]. It allows to select a reference audio file and synthesize speech from text imitating the voice of the reference. It is however not possible to interact with a latent space representing acoustic variability.

In this paper we provide a web interface capable of visualizing and exploring a space of voice expressiveness and synthesize corresponding expressive speech.

---

[1]https://github.com/noetits/ICE-Talk
[2]https://github.com/keithito/tacotron
[3]https://github.com/CorentinJ/Real-Time-Voice-Cloning

3. Description of ICE-Talk

3.1. System architecture

Figure 1 depicts the different components of the system architecture. It is constituted of a DL unsupervised TTS model trained on an expressive dataset (see Section 3.2).

To make the model available as a web service and communicate information of text, audio and style between the web interface and the TTS model, the Falcon Web framework[4] is used.

Falcon allows to bridge the gap between a python code and a web interface, allowing the use of Deep Learning frameworks through a web application.
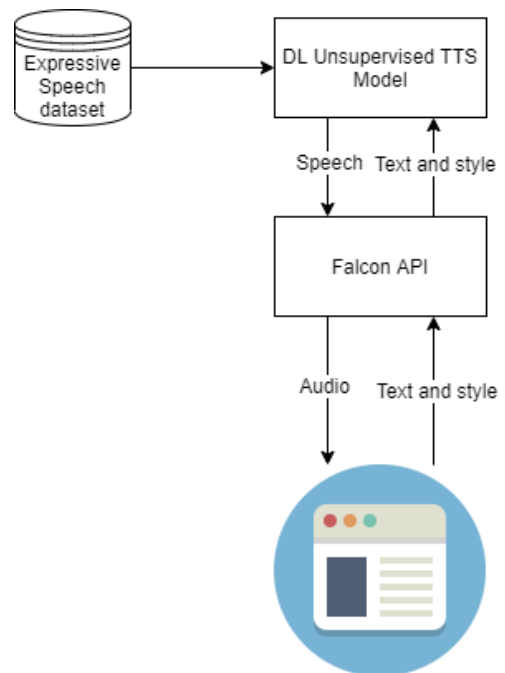

Figure 1: *System architecture*

3.2. DL Unsupervised Model

As a use case, we use a modified version of *Deep Convolutional Text-to-Speech* (DCTTS) [5], a state-of-the art Deep-Learning Sequence-to-Sequence (seq2seq) model with an output controllable through a Latent Space designed to represent variations in voice style as described in [6].

A TTS seq2seq model is typically consists of an encoder-decoder structure. Text is encoded as a latent representation

---

[4]https://falcon.readthedocs.io/en/stable/

that is then decoded with an attention based decoder to predict a mel-spectrogram later inverted to an audio waveform.

To obtain a voice style representation in [6], a mel-spectrogram encoder is added. It consists of a stack of 1D convolutional layers, followed by an average pooling. This operation ensures to obtain time-invariant information. It can thus contain information about statistics of prosody such as pitch average, average speaking rate, but not a pitch evolution.

3.3. Web Interface

The interface contains a 2D representation of a latent space which is an internal representation of the data distribution by the network. This 2D representation is obtained via a dimensionality reduction applied to the highly dimensional latent space of the system. The interface also contains a text box for the system's input and an audio player for the system's output.

The latent space represents the distribution of some controlling parameters (the expressiveness for instance) of the output speech, and is obtained after training.

By writing a text and clicking on a point on the 2D space, an audio signal is generated with the parameters values corresponding to the point clicked on.

The web interface is implemented in HTML5 and javascript to use the service.

There are several possibilities for dimensional reduction : UMAP, PCA or t-SNE. The click of the mouse is detected using javascript in pixels coordinates and mapped to the reduced data space.

Then Nearest Neighbour regression is used to compute the 2D data point, and a lookup table gives the corresponding 8D point of the latent space. The text and the 8D vector are fed to the model that generates the sentence and save it into a wav file. The audio wav file is then served and played as an HTML5 audio object.

4. Conclusions and Future Works

We presented an innovative interface for Controllable Expressive TTS called ICE-Talk that show a proof of concept of research results previously presented at Interspeech. It is open source and ready to be used with available pre-trained models.

As future works, other models available such as a multi-speaker TTS[5] could be integrated to be able to generate speech from different speakers, not based on references but on the latent space describing speaker characteristics.

5. Acknowledgements

6. References

[1] N. Tits, K. El Haddad, and T. Dutoit, "The Theory behind Controllable Expressive Speech Synthesis: A Cross-Disciplinary Approach," in *Human-Computer Interaction*. IntechOpen, 2019. [Online]. Available: http://dx.doi.org/10.5772/intechopen.89849

[2] N. Tits, "A Methodology for Controlling the Emotional Expressiveness in Synthetic Speech - a Deep Learning approach," in *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2019, pp. 1–5.

[3] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno *et al.*, "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis," *arXiv preprint arXiv:1806.04558*, 2018.

[4] C. Jemine *et al.*, "Master thesis: Automatic multispeaker voice cloning," 2019.

[5] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4784–4788.

[6] N. Tits, F. Wang, K. E. Haddad, V. Pagel, and T. Dutoit, "Visualization and Interpretation of Latent Spaces for Controlling Expressive Speech Synthesis through Audio Analysis," in *Proc. Interspeech 2019*, 2019, pp. 4475–4479. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1426
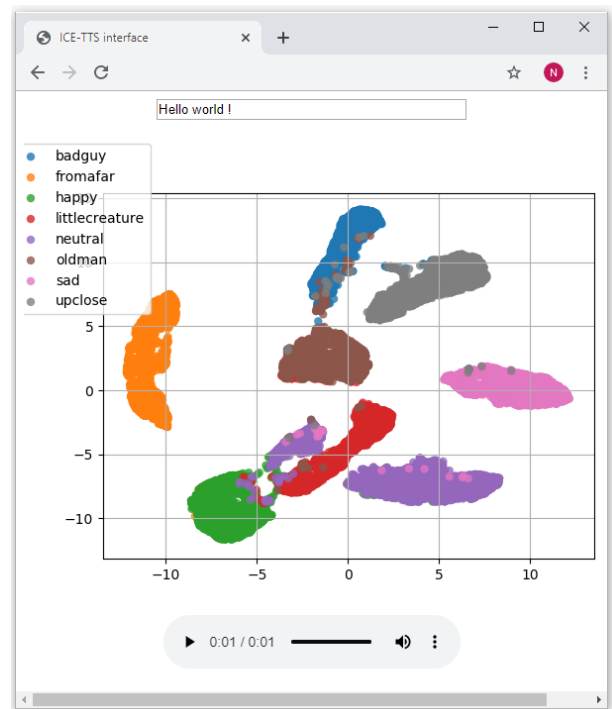

Figure 2: *ICE-Talk web interface. It is constituted of a text field, an image representing a latent space of vocal variability contained in the dataset, and an audio player to listen to the synthesized utterance.*

---

[5] https://github.com/CorentinJ/Real-Time-Voice-Cloning