Software Impacts 8 (2021) 100055

Contents lists available at ScienceDirect

Software Impacts

journal homepage: www.journals.elsevier.com/software-impacts

Original software publication

ICE-Talk 2: Interface for Controllable Expressive TTS with perceptual assessment tool

Noé Tits *, Kevin El Haddad, Thierry Dutoit

Numediart Institute, ISIA Lab, University of Mons, 31 Boulevard Dolez, 7000 Mons, Belgium

ARTICLE INFO

Keywords:
Controllable
Expressive
Speech synthesis
Interface
Perceptual experiment
Assessment
Evaluation

ABSTRACT

In this paper, we present open-source[1] tools that facilitates the use of controllable TTS systems in experiments, towards the democratization of TTS systems across domains. ICE-Talk is a 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. A tool to design a perceptual experiment is provided and consists of three steps: pre-synthesizing samples covering the 2D plot representing controllable dimensions, including this interface inside a template question, and integrate it in a Mechanical Turk system called turkle.

Code metadata

Current code version	v2
Permanent link to code/repository used for this code version	https://github.com/SoftwareImpacts/SIMPAC-2020-65
Permanent link to Reproducible Capsule	https://codeocean.com/capsule/6457578/tree/v1
Legal Code License	Apache 2.0
Code versioning system used	git
Software code languages, tools, and services used	python
Compilation requirements, operating environments & dependencies	
If available Link to developer documentation/manual	https://github.com/noetits/ICE-Talk/blob/master/README.md
Support email for questions	

1. Introduction and motivations

Speech Synthesis is an important component of Human–Robot Interaction. 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 [1].

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.

This interface allows for the control over the synthesis parameters of a DL-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. Indeed in speech synthesis, it is well known that objective measures of quality can sometimes be misleading because they do not always correlate well with the subjective perception. There exists tools to evaluate the naturalness of synthesized speech with subjective tests. However, the field of Controllable Expressive Speech synthesis needs experiments and protocols to assess the controllability of such systems.

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail address: noe.tits@alumni.umons.ac.be (N. Tits).

[1] https://github.com/noetits/ICE-Talk.

https://doi.org/10.1016/j.simpa.2021.100055
Received 20 December 2020; Accepted 22 December 2020

2. Related work

As of today, there are some open-source 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 [2,3]. 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.

ICE-Talk [4] provides a web interface capable of visualizing and exploring a space of voice expressiveness and synthesize corresponding expressive speech, it is a proof of concept based on [5]. However the controllable aspect of such system is difficult to assess and would need the design of perceptual experiments in which a user has to solve a task that would measure this controllability.

In this paper, we present an extended version of ICE-Talk that allows to study the controllability of a Controllable Expressive TTS. It is an integration of the interface inside a questionnaire template, making it possible to build perceptual experiments involving a user to interact with ICE-talk.

3. Description of ICE-Talk 2

3.1. System architecture

Fig. 1 depicts the different components of the system architecture. It sis 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 (see Fig. 2).

3.2. Controllable DL-based TTS

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

A TTS seq2seq model typically consists of an encoder–decoder structure. Text is encoded as a latent representation that is then decoded with an attention based decoder to predict a mel-spectrogram later inverted to an audio waveform.

In [5], to obtain a voice style representation for controllable expressiveness, a mel-spectrogram encoder is added. It consists of a stack of 1D convolutional layers, followed by an average pooling, to obtain and 8D encoding vector. 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.

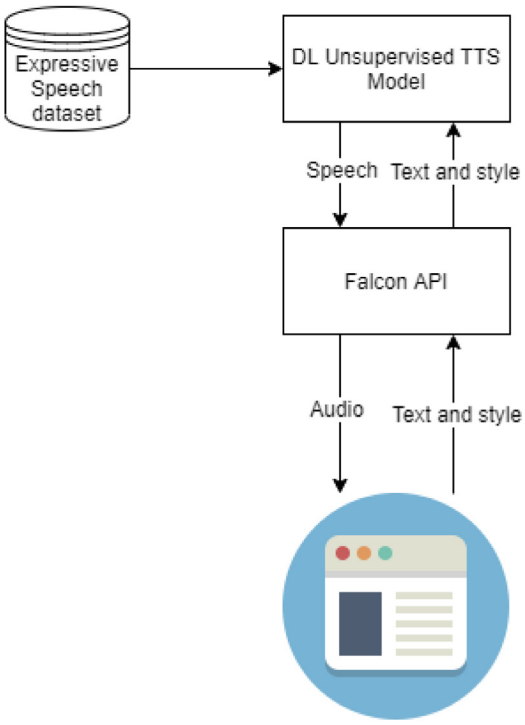

Fig. 1. System architecture.

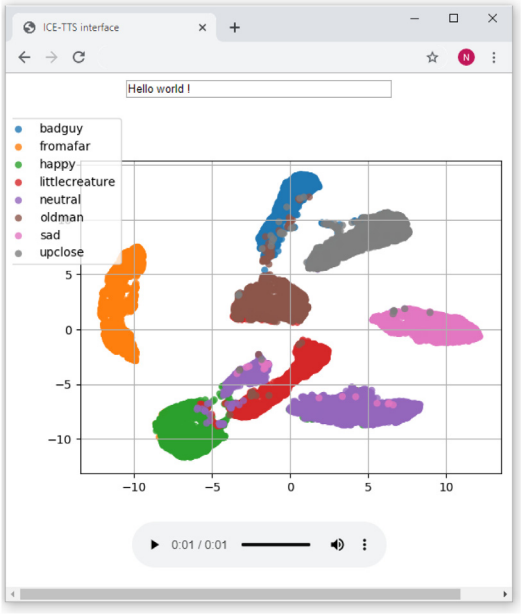

Fig. 2. ICE-Talk web interface. It is constituted of a text field, an image representing a latent space of vocal variability contained in the dataset, and an audio player to listen to the synthesized utterance.

[3] https://github.com/CorentinJ/Real-Time-Voice-Cloning.
[4] https://falcon.readthedocs.io/en/stable/.

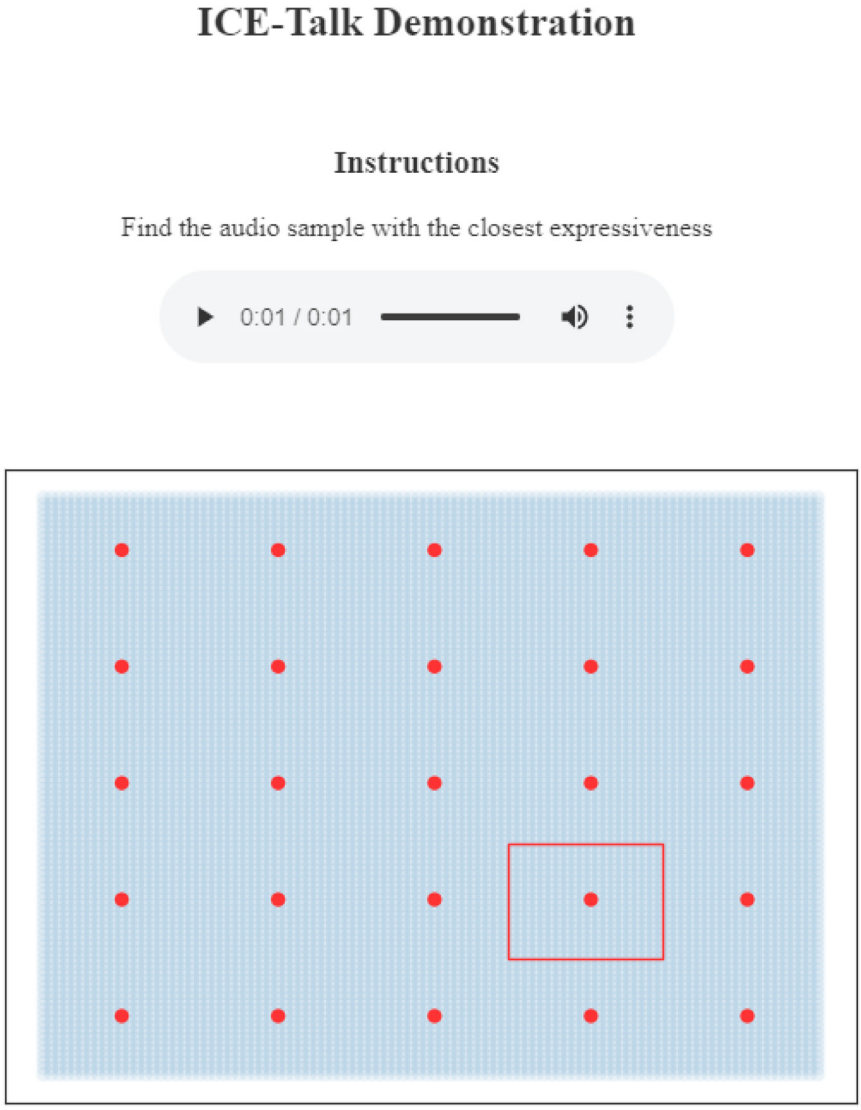

Fig. 3. Task example.

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.

3.4. Perceptual assessment tool

To study the controllability of a Controllable Expressive TTS system, we provide a tool to design perceptual tests.

First, we provide a python script to generate a set of predefined sentences covering the 2D latent space by discretizing it in a set of points. An interface can then be used to play the pre-synthesized samples corresponding to the different regions of the space. A demonstration is available from the github repository.[5]

We provide a template of question, depicted in Fig. 3, that includes this interface, and that can be integrated inside turkle,[6] an open-source web server equivalent to Amazon's Mechanical Turk that one can host on a server or run on a local computer. This template shows an reference audio that a user should find by exploring the 2D space by clicking in it.

It is then possible to ask participants to use the 2D interface to produce the same expressiveness as in a given reference. We assume that if a participant is able to locate in the space the expressiveness corresponding to the reference, it means he is able to use this interface to find the expressiveness he has in mind.

[5] https://github.com/noetits/ICE-Talk.
[6] https://github.com/hltcoe/turkle.

4. Conclusions and future works

We presented an extension of ICE-Talk, a proof of concept of research results of [5], which is a tool that allows to build perceptual experiments involving a user to interact with ICE-Talk.

This tool will enable the study and assessment of the controllability of Controllable Expressive TTS systems and how participants behave and feel with the system.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Noé Tits is funded through a FRIA (https://app.dimensions.ai/details/grant/grant.8952517) grant (Fonds pour la Formation à la Recherche dans l'Industrie et l'Agriculture, Belgium).

References

[1] 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, http://dx.doi.org/10.1109/ACIIW.2019.8925241.

[2] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I.L. Moreno, Y. Wu, et al., Transfer learning from speaker verification to multispeaker text-to-speech synthesis, in: Advances in Neural Information Processing Systems, 2018, pp. 4480–4490.

[3] C. Jemine, et al., Automatic Multispeaker Voice Cloning (Master thesis), Université de Liège, Liège, Belgique, 2019.

[4] N. Tits, K.E. Haddad, T. Dutoit, ICE-Talk: an interface for a controllable expressive talking machine, in: Proc. Interspeech 2020, 2020, pp. 482–483.

[5] N. Tits, F. Wang, K.E. Haddad, V. Pagel, T. Dutoit, Visualization and interpretation of latent spaces for controlling expressive speech synthesis through audio analysis, in: Proc. Interspeech 2019, 2019, pp. 4475–4479, http://dx.doi.org/10.21437/Interspeech.2019-1426.

[6] H. Tachibana, K. Uenoyama, 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.