ITAcotron 2: Transfering English Speech Synthesis Architectures and Speech Features to Italian

Anna Favaro[1], **Licia Sbattella**[2], **Roberto Tedesco**[2] and **Vincenzo Scotti**[2]
[1]CIMeC, Università degli Studi di Trento
Corso Bettini 31, 38068, Rovereto (TN), Italy
[2]DEIB, Politecnico di Milano
Via Golgi 42, 20133, Milano (MI), Italy
anna.favaro@studenti.unitn.it licia.sbattella@polimi.it
roberto.tedesco@polimi.it vincenzo.scotti@polimi.it

Abstract

End-to-end deep learning models have pushed forward significantly many tasks of Natural Language Processing (NLP). However, most of these models are trained for languages providing many resources (such as English), and their behaviour is hardly studied in other languages due to resource shortage. To cope with these situations, it is common practice to employ *transfer learning*. With this work, we wanted to explore the cross-language transferability of a Text-to-Speech (TTS) architecture and the re-usability of the surrounding components that complete a speech synthesis pipeline. To do so, we fine-tuned an English version of the Tacotron 2 TTS, with speaker conditioning, to Italian (hence *ITAcotron 2*). The human evaluation –carried on 70 subjects– showed that the language adaptation was indeed successful.

1 Introduction

The development of Text-to-Speech (TTS) synthesis systems is one of the oldest problems in the Natural Language Processing (NLP) area and has a wide variety of applications (Jurafsky and Martin, 2009). Such systems are designed to output the waveform of a voice uttering the input text string. In the last years, the introduction of deep learning-based approaches, and in particular the end-to-end ones (Shen et al., 2018; Ping et al., 2018; Ren et al., 2019; Hsu et al., 2019), led to significant improvements.

Most of the evaluations carried out on these models are performed on languages with many available resources, like English. Thereby, it is hard to tell whether and how good these models and architectures are general across languages. With this work, we proposed to study how these models behave with less-resourced languages.

To evaluate the transferability of a TTS architecture to a different language, the effectiveness of training a new model starting from –and fine-tuning– another one, and to verify the effect on training convergence, we experimented with English and Italian languages. In particular, we started from the English TTS Tacotron 2 and fine-tuned its training on a collection of Italian corpora. Then, we extended the resulting model, with speaker conditioning; the result was an Italian TTS we named ITAcotron 2.

ITAcotron 2 was evaluated, through human assessment, on intelligibility and naturalness of the synthesised audio clips, as well as on speaker similarity between target and different voices. In the end, we obtained reasonably good results, in line with those of the original model.

We divide the rest of this paper into the following sections. In Section 2 we explain the problem and the available solutions. In Section 3 we present the corpora employed to train and test out the model. In Section 4 we explain the structure of the synthesis pipeline we are proposing and how we adapted it to Italian from English. In Section 5 we describe the experimental approach we followed to assess the model quality. In Section 6 we comment on the results of our model. In Section 7 we sum up our work and suggest possible future extensions.

2 Background

Modern, deep learning-based TTS pipelines are composed of two main blocks: a *spectrogram predictor* and a *vocoder* (Jurafsky and Martin, 2009). These components take care of, respectively, converting a string of characters to a (mel-scaled) spectral representation of the voice signal and converting the spectral representation to an actual waveform. Optionally, input text –apart from normalisation– undergoes phonemisation to present the input to the spectrogram predictor as a sequence of *phonemes* rather than *graphemes*.

Recent end-to-end solutions for spectrogam prediction are built with and *encoder-decoder* architecture (Wang et al., 2017; Shen et al., 2018; Ping et al., 2018; Ren et al., 2019). The encoder maps the input sequence to a hidden continuous space, and the decoder takes care of generating autoregressively the spectrogram from the hidden representation. To produce the alignment between encoder and decoder, an *attention mechanism* (Bahdanau et al., 2015) is introduced between these two blocks.

Among the available architectures for spectrogram prediction, *Tacotron* (Wang et al., 2017), and in particular its advanced version Tacotron 2 (Shen et al., 2018), seems to be the most flexible and re-usable.

Many works have been developed to introduce conditioning into Tacotron, obtaining a fine-grained control over different prosodic aspects. The *Global Style Token* (GST) approach enabled control over the speaking style in an unsupervised manner (Wang et al., 2018). Another controllable aspect is the speaker voice, introduced through additional *speaker-emebeddings* extracted through a speaker verification network (Jia et al., 2018). Finally Suni et al. (2020) proposed a methodology to control *prominence* and *boundaries* by automatically deriving prosodic tags to augment the input character sequence. It is also possible to combine multiple techniques into a single conditioned architecture, as shown by Skerry-Ryan et al. (2018).

Neural vocoders completed the deep learning TTS pipeline improving consistently the quality of synthesised voice (van den Oord et al., 2016; Kalchbrenner et al., 2018; Kumar et al., 2019; Yang et al., 2021). These vocoders substituted the Griffin-Lim algorithm (Griffin and Lim, 1983), which was characterised by artifacts and poor audio quality, especially if compared with newer neural approaches. These components, differently from the spectrogram predictors, do not strictly depend on the input language. Their primary role is to invert a spectral representation into the time domain; thus, they are thought to be *language-agnostic*.

As premised, the available models are primarily trained and evaluated on English corpora due to data availability. A general solution for data scarcity is to leverage a technique called *transfer learning* (Yosinski et al., 2014), which consists of re-using the hidden layers of a pre-trained deep neural network as inputs for a different task. For our work, we applied a variant of transfer learning called *fine-tuning*, where we used the pre-trained weights of the network as initialisation for the actual training on the new task (Yosinski et al., 2014).

3 Corpora

For the scope of this work, we considered three different corpora of Italian speech. All corpora are composed of read speech. We reported the main statistics about the corpora in Table 1. All clips were re-sampled at 22 050 Hz.

Mozilla Common Voice[1] (MCV) is a publicly available corpus of crowd-sourced audio recordings (Ardila et al., 2020). Contributors can either donate voice by reading a prompted sentences or validate clips by listening to others' recordings. The samples in this corpus have a sample rate of 48 000 Hz.

VoxForge[2] (VF) is a multilingual open-source speech database that includes audio clips collected from speaker volunteers. The samples in this corpus have a sample rate of 16 000 Hz.

Ortofonico (Ort.) is a subset of the CLIPS[3] corpus, a corpus of Italian speech collected for a project funded by the Italian Ministry of Education, University and Research. Audio recordings come from radio and television programs, map task dialogues, simulated conversations, and text excerpts read by professional speakers. The samples in this corpus subset have a sample rate of 22 050 Hz.

Apart from the three presented corpora, we used some clips from a private collection of audiobooks in the human evaluation step. We reported further details in Section 5.

4 ITAcotron 2 synthesis pipeline

The model we proposed and evaluated is called ITAcotron 2. It is an entire TTS pipeline, complete with speaker conditioning, based on Tacotron 2 (Shen et al., 2018; Jia et al., 2018). The pipeline is composed of a phonemiser, a speaker encoder (used for the conditioning step), a spectrogram predictor, and a neural vocoder. We reported a scheme of the pipeline in Figure 1.

The core part of the model we are presenting is the spectrogram predictor. We referred to the Tacotron 2 implementation and weights provided

[1] https://commonvoice.mozilla.org
[2] http://www.voxforge.org
[3] http://www.clips.unina.it

Table 1: Statistics on the considered corpora for the Italian fine tuning of the spectrogram predictor: Mozilla Common Voice (MCV), VoxForge (VF) and Ortofonico (Ort.).

Corpus	Time (h)			Clips			Speakers		
	Train	Validation	Test	Train	Validation	Test	Train	Validation	Test
MCV	79.07	26.45	26.42	50 322	16 774	16 775	5151	3719	3743
VF	13.62	1.74	1.75	7176	913	918	903	584	597
Ort.	2.94	0.36	0.32	1436	164	159	20	20	20

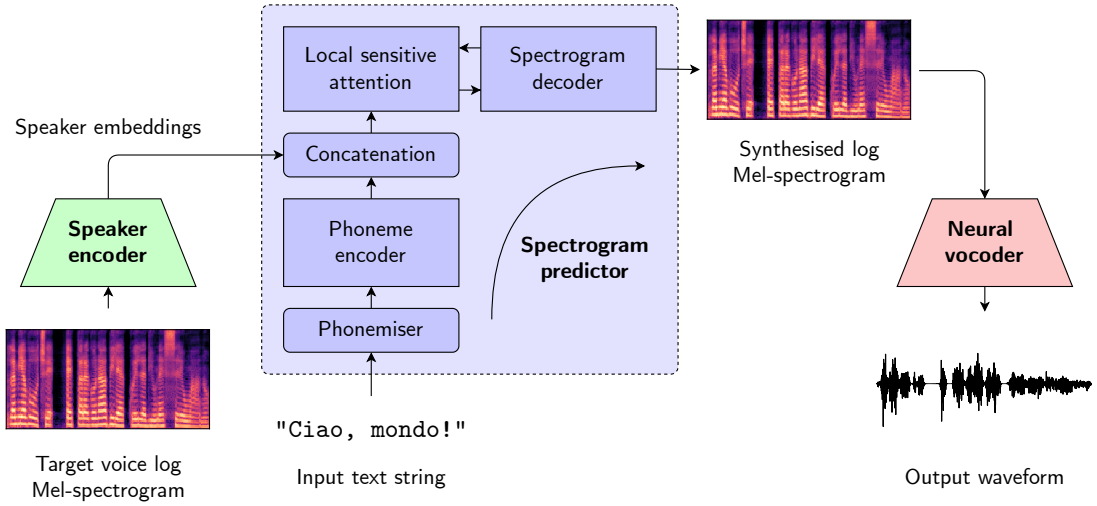

Figure 1: ITAcotron 2 synthesis pipeline.

by Mozilla[4] (Gölge, 2020). The model uses a *phoneme encoder* to represent the input sequence to utter, and an *autoregressive decoder* to generate the target spectrogram; an intermediate attention mechanism provides the input-output alignment. With respect to the original implementation, we only extended the employed phonemiser[5] to accommodate Italian's accented vowels as additional input characters. Code and pre-trained weights for speaker encoder and vocoder came from the Tacotron 2 same source.

We divided the fine-tuning process of the spectrogram predictor into two steps. In this way, we iteratively improved the output quality.

The former used only the data coming from the MCV corpus, which constituted the majority of the available data. Due to the low quality of the input audio recordings, we leveraged this step mostly to drive the network's weights towards the target language. The noisy and sometimes poorly uttered clips of this corpus resulted in an awful quality of the synthesised clips, which sometimes were impossible to understand. This fine tuning was performed for 52 271 update steps (identified trough validation) on mini-batches containing 64 clips each (Other hyper-parameters were left unchanged from the reference implementation).

The latter fine-tuning leveraged both VF and Ort. corpora. Audio clips in these corpora had a noticeable higher quality than those of MCV in terms of audio cleaning and speaker articulation. As a result, the outputs of this final stage had significantly less background noise, and the content was highly intelligible. We performed this second fine-tuning for 42 366 update steps (identified trough validation) on mini-batches containing 42 clips each (Other hyper-parameters were left unchanged from the reference implementation).

To achieve speaker conditioning, we concatenated the encoder representation of the spectrogram predictor with a *speaker embedding*. These embeddings are extracted from a speaker verification model (Chung et al., 2020), similar to that of the reference work by Jia et al. (2018). For the

[4]Repository link: https://github.com/mozilla/TTS, reference commit link: https://github.com/mozilla/TTS/tree/2136433

[5]https://pypi.org/project/phonemizer/

vocoder, instead, we adopted the more recent *Full-Band MelGAN* (FB-MelGAN) vocoder (Yang et al., 2021).

Notice that while we fine-tuned the spectrogram synthesis network, we did not apply the same process to the speaker embedding and neural vocoder networks. We did so because we wanted to observe the zero-shot behaviour of these networks in the new language. In this way, we could assess whether the two models are language-agnostic.

5 Evaluation approach

Similarly to Jia et al. (2018), we divided the evaluation process of the fine-tuned model into two listening tasks:

- evaluation of *Intelligibility and Naturalness* (I&N) of the speaker-conditioned synthesised samples;
- evaluation of *Speaker Similarity* (SS) of the speaker-conditioned synthesised samples.

For both tasks we asked subjects to rate different aspects in a 1 to 5 scale, with 0.5 increments (ITU-T Recommendation, 1999), of the various stimuli (i.e. audio clips). We divided the 70 participants into 20 experimental groups for both listening tasks. We prompted participants of each group with the same stimuli.

In the I&N tasks, we assigned each group with 4 clip pairs, for a total of 160 clips among all groups. Each clip pair was composed of a real clip (ground truth) coming from one of the corpora (including an additional private corpus of audio-books) and a synthetic clip generated in the voice of the ground truth, but with different speech content (i.e. the same voice uttered a different sentence). At this step, we asked subjects to rate the intelligibility and naturalness of each clip separately. Clips were presented in a random order (to avoid biases) and were rated right after listening.

In the SS tasks, we assigned each group with with 16 clips split into 4 subsets, for a total of 160 clips among all groups. We divided the SS task into three further sub-tasks. Each subset was composed of a synthetic clip and three real clips. Subjects compared the synthetic clip to each of the other three real clips:

1. real clip containing an utterance in the voice of the same speaker of the synthetic one (*same speaker* comparison sub-task);
2. real clip containing an utterance in the voice of a different speaker having the same gender of the speaker of the synthetic one (*same gender* comparison sub-task);
3. real clip containing an utterance in the voice of a different speaker having different gender of the speaker of the synthetic one (*different gender* comparison sub-task).

At this step, we asked subjects to rate how similar the synthetic voice was to the one we paired it with (knowing that the fixed clip was synthetic and the other three real). Real clips were presented in a random order (to avoid biases), and subjects rated the similarity after listening to a synthetic-real pair.

6 Results

Table 2: Results of the listening tasks. MOS values are reported as *average* $\pm$ *standard deviation*.

Task	Sub-task	Model	MOS
I&N	Intelligibility	ITAcotron 2	4.15 ± 0.78
		Ground truth	4.43 ± 0.74
	Naturalness	ITAcotron 2	3.32 ± 0.97
		Ground truth	4.28 ± 0.86
SS	Same speaker	ITAcotron 2	3.45 ± 1.07
	Same gender	ITAcotron 2	2.78 ± 1.01
	Different gender	ITAcotron 2	1.99 ± 1.08

We reported the Mean Opinion Score (MOS) of each task in Table 2. The overall scores were satisfying and reflected the intentions and the expectations underlying this research.

Concerning the I&N evaluation, the first thing that jumps to the eye is the high intelligibility score, very close to real clips. This high score provides clear evidence of how easy it was to understand the linguistic content of the synthetic clips. The naturalness score is lower than that of intelligibility, meaning that it is still possible to distinguish between real and fake clips.

Concerning the SS evaluation, instead, the thing that jumps to the eye is the progressive drop in the MOS value. This reduction is precisely the expected behaviour: changing the speaker should lead to lower similarity, especially when the two speakers have different gender. The value obtained for the same speaker sub-task seems promising. The reduction in speaker similarity observed in different speaker sub-task showed that the synthetic

clips' voice is distinguishable from those of the same gender. The further drop observed in different speaker similarity evaluations underlined that the network learned to separate even better these aspects, as we expected considered the general difference in pitch ranges between the two genders (Leung et al., 2018).

The figures we obtained are quite similar to those obtained by Jia et al. (2018) on similar tasks for English. However, we choose not to report a direct comparison against the work mentioned above as it focuses on English and the tasks are not perfectly comparable with ours. Nevertheless, obtaining scores that are similar to the ones provided by that work, is a hint that our approach seems sound.

7 Conclusion

This paper showed the approach we followed in our work to adapt a speech synthesis pipeline from English to Italian. The procedure is language-agnostic; however, the spectrogram prediction network requires fine-tuning data in the target language. To show how some pipeline components can be used out-of-the-box (i.e. without language adaptation), we also introduced a speaker embedding network (to achieve speaker conditioning) and a neural vocoder. Opinion scores from a human evaluation session showed that the adaptation was successful in terms of intelligibility and naturalness. Concerning speaker conditioning, the result was not as sharp as for the first evaluation, yet we obtained a satisfying similarity score, matching that of the reference model.

In future work, to derive speaker discriminative representations, we could refine the speaker encoder on Italian multi-speaker speech data. In doing so, we will assess the impact of employing a network refined on a target language for deriving descriptive features for speakers of that language. Finally, since ITAcotron 2 is not completely able to isolate the speaker voiceprint from the prosody of the reference audio, we suggest conditioning its generative performance on independent auxiliary representations as in Skerry-Ryan et al. (2018) and Wang et al. (2018). For instance, one intended to capture the speaker's accent and one the speaker's voiceprint.

Acknowledgments

This work was partially supported by the European Union's Horizon 2020 project *WorkingAge* (grant agreement No. 826232).

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4218–4222. European Language Resources Association.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Joon Son Chung, Jaesung Huh, Seongkyu Mun, Minjae Lee, Hee-Soo Heo, Soyeon Choe, Chiheon Ham, Sunghwan Jung, Bong-Jin Lee, and Icksang Han. 2020. In defence of metric learning for speaker recognition. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 2977–2981. ISCA.

Daniel W. Griffin and Jae S. Lim. 1983. Signal estimation from modified short-time fourier transform. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '83, Boston, Massachusetts, USA, April 14-16, 1983*, pages 804–807. IEEE.

Eren Gölge. 2020. Solving attention problems of tts models with double decoder consistency.

Po-chun Hsu, Chun-hsuan Wang, Andy T. Liu, and Hung-yi Lee. 2019. Towards robust neural vocoding for speech generation: A survey. *CoRR*, abs/1912.02461.

ITU-T Recommendation. 1999. *P.910: Subjective video quality assessment methods for multimedia applications*.

Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez-Moreno, and Yonghui Wu. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 4485–4495.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., USA.

Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aäron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. 2018. Efficient neural audio synthesis. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2415–2424. PMLR.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C. Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14881–14892.

Yeptain Leung, Jennifer Oates, and Siew Pang Chan. 2018. Voice, articulation, and prosody contribute to listener perceptions of speaker gender: A systematic review and meta-analysis. *Journal of Speech, Language and Hearing Research (Online)*, 61(2):266–297.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. In *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016*, page 125. ISCA.

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Ömer Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2018. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3165–3174.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ-Skerrv Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, pages 4779–4783. IEEE.

R. J. Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, and Rif A. Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4700–4709. PMLR.

Antti Suni, Sofoklis Kakouros, Martti Vainio, and Juraj Simko. 2020. Prosodic prominence and boundaries in sequence-to-sequence speech synthesis. *CoRR*, abs/2006.15967.

Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards end-to-end speech synthesis. In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pages 4006–4010. ISCA.

Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A. Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5167–5176. PMLR.

Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, and Lei Xie. 2021. Multi-band melgan: Faster waveform generation for high-quality text-to-speech. In *IEEE Spoken Language Technology Workshop, SLT 2021, Shenzhen, China, January 19-22, 2021*, pages 492–498. IEEE.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3320–3328.

A Code base and model weights

The source code developed during this project is available at the following link: https://github.com/vincenzo-scotti/ITAcotron_2. Inside the repository we also provide the links to download the weights of the fine-tuned model ITAcotron 2, for Italian speech synthesis. We remind that the original source code we forked, and the weights of the speaker encoder and neural vocoder, were taken from the reference open source project developed by Mozilla[4].