CATOTRON – A Neural Text-to-Speech System in Catalan

Baybars Külebi[1], *Alp Öktem*[1], *Alex Peiró-Lilja*[2], *Santiago Pascual*[3*], *Mireia Farrús*[2]

[1]Col·lectivaT
[2]Universitat Pompeu Fabra
[3]Universitat Politècnica de Catalunya

{bkulebi,alp}@collectivat.cat, {alex.peiro,mireia.farrus}@upf.edu, santi.pascual@upc.edu

Abstract

We present Catotron, a neural network-based open-source speech synthesis system in Catalan. Catotron consists of a sequence-to-sequence model trained with two small open-source datasets based on semi-spontaneous and read speech. We demonstrate how a neural TTS can be built for languages with limited resources using found-data optimization and cross-lingual transfer learning. We make the datasets, initial models and source code publicly available for both commercial and research purposes.

Index Terms: text-to-speech, sequence-to-sequence, Catalan.

1. Introduction

Nowadays, neural text-to-speech (TTS) systems have achieved levels of quality close to human speech . However, building such systems in low-resource languages remains still a challenge. To fill this gap, Col·lectivaT is one of the growing initiatives fostering open language resources and tools for linguistic and collaborative work and research, which apply to low-resourced languages such as Catalan, a Romance language spoken by about 9 million people. There exists work on speech technology on Catalan. However, they are either closed-sourced [1, 2] or based on older architectures [3, 4] which are outdated compared to state-of-the-art neural systems.

In this light, we present *Catotron*: a neural Catalan TTS system based on *Tacotron 2* architecture [5]. In our resource-limited setup, we optimized a multispeaker dataset *ParlamentParla*, which consists of non-read speeches from the sessions of the Catalan Parliament, and applied cross-lingual transfer learning to obtain best possible results from the only publicly available clean speech dataset. We made the source code, models and speaker adaptation scripts publicly available under the CC-BY 4.0 license[1].

2. Database

2.1. Sources

Traditionally, data used for building TTS systems are designed from scratch and recorded for the specific task with careful instructions. During the recent years, the utility of found data is being investigated [6, 7]. Catalan, as a minority language, does not have many resources that are easily accessible and speaker diarized at the same time. Hence for this work, we wanted to test the feasibility of the use of found data as well as specifically designed and recorded TTS speech corpora. We also wanted to see whether the developer community can take advantage of found data in developing applications of speech technology.

2.1.1. ParlamentParla

This a speech corpus based on the parliamentary sessions of the Catalan regional government (*Generalitat*), from June 2008 until July 2018. The audio files are scraped from the website of the Catalan parliament[2], and matched with their corresponding proceedings in PDF. The matched long audio files are then aligned to their corresponding text similarly to the segmentation of the Librispeech dataset [8]. To obtain a single-speaker corpus, we created a subset from the recordings of the deputy who has the most hours of recording, Artur Mas, the president of Generalitat between 2011 and 2015.

2.1.2. FestCat

FestCat corpus was part of a larger project to build the first statistical speech synthesis system, in Catalan and open source [4]. In addition to developing the front-end for Catalan with the Festival toolkit [9], it also entailed the design and recording of a speech corpus, with open licenses[3]. The corpus is based on voices of 10 people with their audio recordings totaling to approximately 28 hours. For training the models, we chose *Ona* (female) and *Pau* (male) speakers, who have the most amount of recording, reaching 10 hours each.

2.2. Preparation

We segmented the audio recordings into intervals of duration shorter than 7 seconds. This was to ensure convergence given memory restrictions of the neural network training. Segmentation boundaries were selected automatically considering both the pauses in the speech audio and punctuation. For the case of Festcat, this decreased the total duration of the recordings to 4 hours for Ona and 4 hours 16 minutes for Pau. For ParlamentParla, in addition to the orthographic segmentation, we have further filtered the recordings based on the quality of the transcriptions. Using a scoring function based on [8], we eliminated the segments with possible disfluencies and transcription errors. This effectively reduced the total amount of recordings of Artur Mas, from approximately 25 hours to 5 hours and 30 minutes. Finally, we normalized the silences longer than 500 ms to 500 ms exact, eliminated any silence at the beginning or the end of the segment and added 100 ms silence paddings to the beginning and end of the segments.

3. Architecture

3.1. Model description

Our complete TTS architecture consists of two components: 1) Tacotron 2 [5] to map input text to a mel-scale spectrogram

*S. Pascual is currently at Dolby Laboratories, Barcelona, Spain.

[1]https://creativecommons.org/licenses/by/4.0/

[2]https://www.parlament.cat/

[3]http://festcat.talp.cat/

representation, and 2) a neural vocoder to transform generated spectrograms into output waveforms. Tacotron 2 is a widely-used system in TTS research capable of producing an adequate synthesis quality with as little as 24 hours of data. Regarding the neural vocoder, we experimented with two alternatives: *WaveGlow* [10] and *MelGAN* [11]. *WaveGlow* offers fast parallelized inference operations and high generation quality, which were two important factors in our design. MelGAN was recently proposed as an alternative lightweight neural vocoder at the expense of some generation quality.

3.2. Training

3.2.1. Tacotron 2

Given the relatively small amount of data that we had for each speaker, we decided to apply transfer learning on out-of-the-box English model provided in the NVIDIA Tacotron 2 repository[4] trained with the *LJSpeech dataset*[5]. For training our models, we used the scripts provided in the repository, using a batch size of 64 samples, a learning rate of 0.001 and a dropout rate of 0.3. Convergence was relatively fast; for all the data sets approximately between 35-40 epochs. We also implemented an additional linguistic *front-end* to process Catalan ortography.

3.2.2. Neural vocoder

As a first approach, we used WaveGlow with NVIDIA's out-of-the-box English model trained on the LJSpeech dataset. Nonetheless, WaveGlow did not operate properly across male speakers as the pre-trained model was trained with a female voice. Secondly, we experimented with MelGAN using the original training parameters published in the MelGAN paper and available in the official repository[6]. We modified the architecture slightly to match the feature extraction to match Tacotron 2, and also adapted to the FestCat training set. We noticed a faster and better convergence by applying a two-timescale update rule (TTUR) [12], which makes the discriminator in a GAN learn quicker than the generator (from $\times 2$ to $\times 4$). This feeds better features via back-propagation to the generator on what it needs in order to generate more realistic outcomes. In this scenario we set the learning rate for the generator to $lr_G = 1e^{-4}$, and the discriminator's to $lr_D = 3e^{-4}$.

4. Open access

Synthesis samples of all three speakers can be listened in our project blog. Moreover, we release the following models, source code and speaker adaptation scripts for developers and researchers who would like to build applications and contribute to the project:

Project blog with links to models (Ona, Pau, Waveglow, MelGAN) and samples: http://collectivat.cat/blog/2019-12-05-speech-synthesis-dl/

Demo: http://catotron.collectivat.cat/

Catotron GPU: http://github.com/CollectivaT-dev/catotron

Catotron CPU: http://github.com/CollectivaT-dev/catotron-cpu

IPython notebooks for inference and speaker adaptation: http://github.com/CollectivaT-dev/TallersParla

[4]https://github.com/NVIDIA/tacotron2

[5]https://keithito.com/LJ-Speech-Dataset/

[6]https://github.com/descriptinc/melgan-neurips

5. Conclusions

We have presented Catotron, a neural text-to-speech system in Catalan, completely open-source, based on state-of-the-art neural speech synthesis techniques and open licensed corpora. Catotron takes a step forward to develop state-of-the-art speech synthesis systems for low-resourced and minority languages.

6. Acknowledgements

This work was subsidised by the Catalan Department of Culture. A part of the funding comes from the financing administered by the inheritance board of the Generalitat de Catalunya. The last author has been funded by the Agencia Estatal de Investigación (AEI), Ministerio de Ciencia, Innovación y Universidades and the Fondo Social Europeo (FSE) under grant RYC-2015-17239 (AEI/FSE, UE). Part of this work has been carried out using an NVIDIA GPU Titan Xp generously provided by NVIDIA Company. This research was also partially supported by the project TEC2015-69266-P (MINECO/FEDER, UE).

7. References

[1] H. Schulz, M. Ruiz, and J. A. R. Fonollosa, "TECNOPARLA - Speech technologies for Catalan and its application to speech-to-speech translation," *Procesamiento del Lenguaje Natural*, vol. 41, pp. 319–320, 2008.

[2] J. B. Mariño, J. Padrell, A. Moreno, and C. Nadeu, "Monolingual and bilingual Spanish-Catalan speech recognizers developed from speechdat databases," in *Workshop on Developing Language Resources for Minority Languages: Reusability and Strategic Priorities, LREC*, Athens, Greece, 2000, pp. 57–61.

[3] B. Külebi and A. Öktem, "Building an open source automatic speech recognition system for Catalan," in *IberSPEECH*, Barcelona, Spain, 2018, pp. 25–29.

[4] A. Bonafonte, J. Adell, I. Esquerra, S. Gallego, A. Moreno, and J. Pérez, "Corpus and voices for Catalan speech synthesis." in *LREC*, Marrakech, Morocco, 2008.

[5] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," *CoRR*, vol. abs/1712.05884, 2017.

[6] A. Stan, O. Watts, Y. Mamiya, M. Giurgiu, R. A. Clark, J. Yamagishi, and S. King, "Tundra: a multilingual corpus of found data for tts research created with light supervision." in *INTERSPEECH*, 2013, pp. 2331–35.

[7] E. L. Cooper, "Text-to-speech synthesis using found data for low-resource languages," Ph.D. dissertation, Columbia University, 2019.

[8] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *ICASSP*, South Brisbane, Queensland, Australia, 2015, pp. 5206–10.

[9] A. Black, P. Taylor, R. Caley, and R. Clark, "The Festival speech synthesis system," 1998.

[10] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," *CoRR*, vol. abs/1811.00002, 2018.

[11] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," in *NeurIPS*, Vancouver, Canada, 2019, pp. 14 881–92.

[12] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *NeurIPS*, Long Beach, CA, USA, 2017, pp. 6626–37.