Université de Liège

Faculté des Sciences Appliquées

Automatic Voice Cloning Across Languages

Author:
Corentin Jemine

Supervisor:
Prof. Gilles Louppe

Academic year 2018 - 2019

Graduation studies conducted for obtaining the Master's degree in Data Science by Corentin Jemine

1 Abstract

To do when I'll have a good overview of the project. Try to answer:

- What is the goal of the application? What are its requirements, what is the setting, what kind of data are we going to use it on?
- What is zero-shot voice cloning? How does it fit in here (difference between an online and offline approach)?
- What are the particularities of our implementation (both model and datasets), what are its upsides and downsides (for example: requires huge datasets but fast inference)?
- What did we ultimately achieve? How good are our results?

2 Introduction

Concise presentation of the problem
Note that layers will be explained in an upcoming section
Preprocessing of text into phonemes?
SOTA ON MULTISPEAKER TTS:
First SPSS methods [2 - 20] of https://arxiv.org/pdf/1606.06061.pdf

Previous state of the art in TTS include hidden Markov models (HMM) based speech synthesis, which is a statistical parametric speech synthesis (SPSS) method. HMMs are trained to synthesize mel-frequency cepstral coefficients (MFCC) with energy, their delta and delta-delta coefficients [1]. The result is passed through a vocoder[1] such as MLSA [2]. The spectral parameters, pitch parameters and state durations of the model are conditioned on the phonemes context such that different contexts are clustered by a decision tree and a distribution is learned for each cluster [4]. It is thus possible to modify the voice generated by tuning the parameters with adaptation or interpolation techniques (e.g. [3]), effectively making HMM-based speech synthesis a multispeaker TTS system. Compare with concatenative [5] ?

[5] proposed to model
Wavenet:

Breakthrough in TTS with raw waveform gen https://deepmind.com/blog/wavenet-generative-model-raw-audio/ ?? Dilated causal convolutions Condition on a speaker identity

Tacotron
Deep voice (1, 2, 3 + few samples), Tacotron 2
SV2TTS
Extensions?

[1] Specifically in TTS, some authors define a vocoder as a voice encoder that retrieves speech parameters to be used in synthesis. The more common definition however, is that of a function that generates a raw audio waveform from temporal features such as MFFC. This is the one we will use. Review this

References

[1] Kallirroi Georgila. Speech Synthesis: State of the Art and Challenges for the Future, page 257–272. Cambridge University Press, 2017.

[2] S. Imai. Cepstral analysis synthesis on the mel frequency scale. In ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 8, pages 93–96, April 1983.

[3] Takayoshi Yoshimura, Takashi Masuko, Keiichi Tokuda, Takao Kobayashi, and Tadashi Kitamura. Speaker interpolation in hmm-based speech synthesis system. In EUROSPEECH, 1997.

[4] Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. In EUROSPEECH, 1999.

[5] H. Ze, A. Senior, and M. Schuster. Statistical parametric speech synthesis using deep neural networks. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 7962–7966, May 2013.