Université de Liège

Faculté des Sciences Appliquées

Automatic Multispeaker 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
Processing of text into features
Present the generic framework of SPSS since it doesn't change much over the years?
SOTA ON MULTISPEAKER TTS:

Previous state of the art is it really sota when concatenative often leads to more natural results? to review in TTS include hidden Markov models (HMM) based speech synthesis, which is a statistical parametric speech synthesis (SPSS) method. HMMs learn a distribution over mel-frequency cepstral coefficients (MFCC) with energy, their delta and delta-delta coefficients [1]. These speech parameters are derived from their distribution by maximum likelihood before going through a vocoder such as MLSA [3]. The input text to generate is processed into a sequence of linguistic contexts. The HMM parameters to use for speech generation are distributed conditionally to these contexts. Indeed, contexts are clustered with decision trees and an HMM is learned for each cluster [8] (effectively partitioning the training set). It is possible to modify the voice generated by conditioning on a speaker or tuning these parameters with adaptation or interpolation techniques (e.g. [7] elaborate a bit on these techniques?), making HMM-based speech synthesis a multispeaker TTS system. Compare with concatenative? see [9] and https://ieeexplore.ieee.org/document/541110

Improvements to this framework were later brought by feed-forward and recurrent deep neural networks (DNN and RNN respectively) as a result of progress in both hardware and software. [9] proposes to replace entirely the decision tree-clustered HMMs in favor of a DNN. They argue for better data efficiency as the training set is no longer fragmented in different clusters of contexts. They demonstrate improvement over the speech quality with a similar number of parameters as in the HMM-based approach.

Several authors propose to replace the decision trees by a DNN, arguing for better data efficiency and for more representational power of complex dependencies [2, 4, 5, 6]. Most demonstrate improved speech quality for a similar number of parameters [2, 5, 6].

Entirely read [2] if missing infos about HMMs + DNNs

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

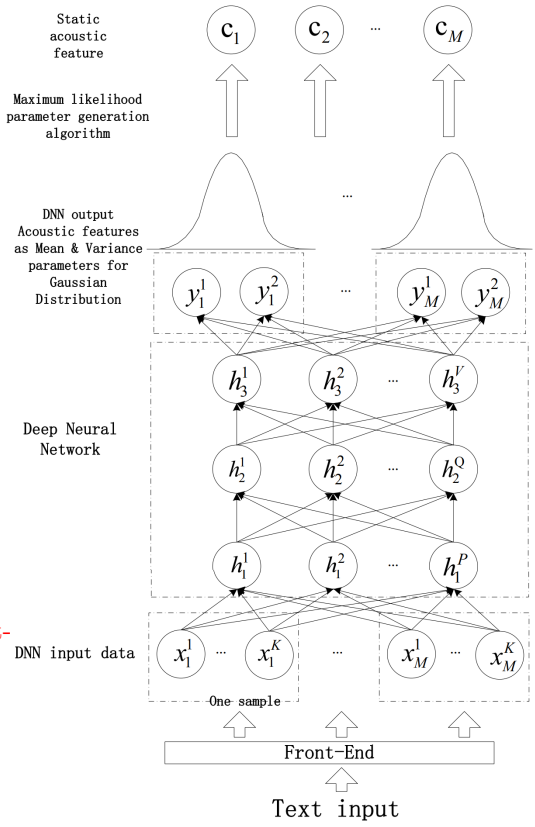

Figure 1: caption

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

References

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

[2] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda. The effect of neural networks in statistical parametric speech synthesis. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4455–4459, April 2015.

[3] 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.

[4] Heng Lu, Simon King, and Oliver Watts. Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis.

[5] Y. Qian, Y. Fan, W. Hu, and F. K. Soong. On the training aspects of deep neural network (dnn) for parametric tts synthesis. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 3829–3833, May 2014.

[6] Xiang Yin, Ming Lei, Zhiliang Hong, Frank K. Soong, Lei He, Zhen-Hua Ling, and Li-Rong Dai. Modeling dct parameterized f0 trajectory at intonation phrase level with dnn or decision tree. In INTERSPEECH, 2014.

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

[8] 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.

[9] H. Zen, 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.