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

Possibly, start with a TTS lexicon with some definitions. Meanwhile, I'll make a list of TSS-specific words that may be worth explaining: coarticulation, context, rich context, spectral envelope, fundamental frequency

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
GENERIC FRAMEWORK OF SPSS:

Statistical parametric speech synthesis (SPSS) refers to a group of TTS synthesis methods where the relation between features computed on the input text and output acoustic features are modeled by a statistical model (called an acoustic model). A complete SPSS framework thus also includes a pipeline to extract features from the text to synthesize as well as a system able to reconstruct an audio waveform from the acoustic features produced by the acoustic model (such a system is called a vocoder). Unlike the acoustic model, these two parts of the framework may be entirely engineered and make no use of statistical methods. If it is possible to condition the acoustic model in such a way that the characteristics of the generated voice are modified, then the framework is a multispeaker TTS synthesis system could also be if conditioning on the vocoder. Rephrase this. .

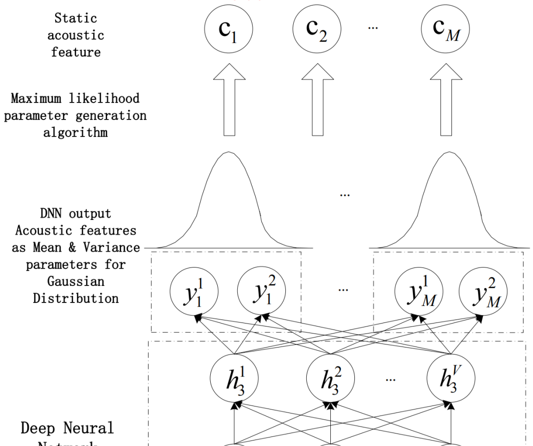
Processing of text into features (mainly linguistic contexts) - I haven't found a good source for this even though all the papers I cite use it... Maybe [5] section 2.3 and 3.1.

SOTA ON MULTISPEAKER TTS:

Previous state of the art in SPSS includes hidden Markov models (HMM) based speech synthesis. In this framework, 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 the distributions output by HMMs using maximum likelihood talk about MLPG too. See "Speech parameter generation algorithms for HMMbased speech synthesis" . They are then fed 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. Also, include the picture from [2]?

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

I'll take care of placing images correctly later...


Static acoustic feature
c_1
c_2
...
c_M
Maximum likelihood parameter generation algorithm
...
DNN output Acoustic features as Mean & Variance parameters for Gaussian Distribution
y_1^1
y_1^2
...
y_M^1
y_M^2
h_3^1
h_3^2
...
h_3^V
Deep Neural

contexts, and for a more powerful model?. They demonstrate improvements over the speech quality with a number of parameters similar to that of the HMM-based approach. Later researches corroborate these findings [5].

Also read [6]

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

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.