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, linguistic context, rich context, spectral envelope, fundamental frequency, supra-segmental

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

2.1 Statistical parametric speech synthesis

Statistical parametric speech synthesis (SPSS) refers to a group of data-driven TTS synthesis methods that emerged in the late 1990s. In SPSS, the relation between features computed on the input text and output acoustic features are modeled by a statistical generative model (called the 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 use of no statistical methods. If it is possible to condition parts of the framework in such a way that the characteristics of the generated voice are modified, then the framework is a multispeaker TTS synthesis system.

The processing of text into features can be nearly inexistent as it can be very extensive. Pronunciation is an intricate process that depends on a wide range of linguistic contexts. Providing those greatly reduces the extent of the task to be learned by the acoustic model, but may require complex natural language processing (NLP) techniques or accuracy trade-offs, especially for rare or unknown words. Linguistic contexts are retrieved on different levels: utterance, phoneme, syllable, word and phrase. For each of those elements, their neighbouring elements of the same level are usually considered, as well as the elements lower in the hierarchy it comprises. For example, a given frame will contain a word, the two previous words, the two following words and the syllables contained in all those words. The position of each element with regard to its parent element can be included (e.g. fifth word in a sentence), as well as grammatical information such as part of speech. For syllables, the lexical stress and accent can be predicted by a statistical model.

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 [4] section 2.3 and 3.1. or [5].

Talk about evaluation metrics (mainly MOS)?

2.2 Multispeaker TTS state of the art

Previous state of the art in SPSS includes hidden Markov models (HMM) based speech synthesis [5]. The speech generation pipeline is laid out in figure 1. In this framework, the acoustic model is a set of HMMs trained to produce a distribution over mel-frequency cepstral coefficients (MFCC) with energy, their delta and delta-delta coefficients.

These speech parameters are derived from the distributions output by HMMs using maximum likelihood talk about MLPG too. See "Speech parameter generation algorithms for HMM-based 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 multi-speaker TTS system. Compare with concatenative see [9] and https://ieeexplore.ieee.org/document/541110.

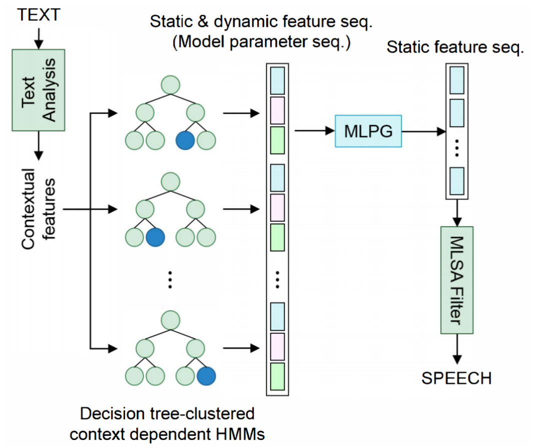

Figure 1: The general HMM-based TTS synthesis approach.

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, 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. Their best model is a DNN with 4 layers of 256 units using a sigmoid activation function. Subjects assessing the quality of the generated audio samples report that the DNN-based models produces speech that sounds less muffled than that of the HMM-based models. Later researches corroborate these findings [4]. [2] additionally studies the effect of replacing MLPG with another DNN. The combinations of HMM/DNN and MLPG/DNN give rise to four possible frameworks, the novel ones being HMM+DNN and DNN+DNN[1], while HMM+MLPG and DNN+MLPG are respectively the frameworks described in [5] and [9]. Each DNN they used is 3 layers deep with 1024 units using a sigmoid activation function. The MOS of each combination is reported in figure **??**. How much do we care about their results? Should I make this part shorter? How about a table at the end that groups all MOS

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?

[1] Note that since the two networks are consecutive in the framework, they can be considered as a single network.

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] 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.

[5] Yoshihiko; Toda Tomoki; Zen Heiga; Yamagishi Junichi; Oura Keiichiro Tokuda, Keiichi; Nankaku. Speech synthesis based on hidden markov models. Proceedings of the IEEE, 101, 05 2013.

[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.