Interspeech 2018
2-6 September 2018, Hyderabad

Speech synthesis in the wild

Ganesh Sivaraman, Parav Nagarsheth, Elie Khoury

Pindrop, Atlanta USA

{gsivaraman,pnagarsheth,ekhoury}@pindrop.com

Abstract

Speech synthesis has wide range of applications in modern artificial intelligence technologies. Most state-of-the-art speech synthesis systems usually require high quality recordings of large amounts of speech data of the target speaker. We focus on low-budget speech synthesis. Our software deals with methods to perform statistical parametric speech synthesis using unlabeled and mixed quality speech data sourced from the internet. An average voice model trained using DNN is adapted to a target speaker using different speaker adaptation strategies. Preprocessing methods like speech enhancement, diarization and segmentation are applied to the sourced data. Utterance selection based on Mean cepstral distortion and forced alignment confidence are applied to prune the noisy and mis-aligned data. The mixed quality data thus pre-processed is then used to adapt the average voice model and duration models to the target speaker.

The software to be demonstrated automates the whole procedure from preprocessing to synthesis. The software will be demonstrated by performing live synthesis using audio sourced from Youtube.

Index Terms: speech synthesis, utterance selection, speaker adaptation, speech enhancement

1. Introduction

Speech synthesis has many applications in several technologies involving human computer interaction. With the wide popularity of artificial intelligence technologies, there is a growing demand for personalized AI systems that talk in personalized voices. Current state-of-the-art speech synthesis systems require hours of labeled clean speech data and several hours of model training in order to synthesize a good voice. With the availability of large amount of speech data for public personalities on social media platforms, they can be utilized for generating synthetic voices. However, online (or in-the-wild) speech data is typically conversational, noisy, and devoid of any ground truth transcriptions. In this demo we perform low-budget speech synthesis using mixed quality limited training data available online with no ground truth transcriptions.

Our approach to speech synthesis is statistical parametric speech synthesis (SPSS) using DNN acoustic models [1] [2]. We map linguistic features containing phone, syllable and duration information to acoustic Mel cepstral features and frame wise pitch (F0) frequencies using DNNs, The acoustic features are then used to generate synthetic speech using the WORLD vocoder [3]. We train DNN acoustic models to map linguistic features containing phoneme, syllable and duration information to acoustic features on a large number of speakers. We call this an Average Voice Model (AVM). We explain the AVM training procedure and the feature normalizations in section 2.3. We then adapt the AVM on the target speaker's speech to synthesize the voice for the target speaker [4]. Since the speech data and their corresponding transcriptions for the target speaker are noisy we explore utterance selection techniques using Mean Cepstral Distortion (MCD) [5] and Automatic Speech Recognition (ASR) confidence scores. The utterance selection methods are outlined in section 3.4.

The demo will demonstrate the ability of performing English speech synthesis using limited amount of speech found online without any groundtruth transcription for popular personalities of interest.

2. Speech synthesis system

In this paper we use a standard DNN based SPSS architecture. Figure 1 shows the overall block diagram of the speech synthesis system. Note that this is the architecture of the system used to synthesize a test utterance. The details of the training, and adaptation of the DNN AVMs and duration models are provided in the upcoming subsections.

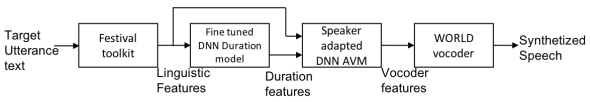

Figure 1: *Block diagram of the speech synthesis system.*

2.1. Linguistic feature extraction

The Festival toolkit [6] is used to extract the linguistic features. The linguistic features encode information regarding the phone identity, syllable stress, position of the phoneme within the syllable, quinphone context, adjacent syllable context, and syllable stress features of adjacent syllables.

2.2. Duration model

The duration feature for every linguistic feature is represented as the time durations of the 5 HMM states that model the current phoneme.

The duration model generates HMM state level durations for every linguistic feature during speech synthesis. A 6-hidden layer DNN with tanh activation functions is trained on the VCTK dataset training set [7] to map the linguistic feature representations to the duration features. This is the baseline average duration model which we later adapt to a target speaker.

2.3. Average Voice Model

The Average Voice Model (AVM) maps the frame level linguistic phonetic features to vocoder features which consist of 60 dimensional Mel cepstral coefficients, band aperiodicities for 5 bands spanning the spectral range, and log F0. These features are normalized and appended with deltas and double deltas to form the output feature vector for the AVM. A 6 layer

feedforward DNN is trained to learn the mapping from frame-wise linguistic features to the vocoder features. The DNN thus trained on the VCTK training set formed our AVM.

2.4. Vocoder

The WORLD vocoder [3] is used to generate the speech waveforms from vocoder features estimated by the speaker adapted DNN AVM. The vocoder uses the estimated magnitude spectra with a Pitch Synchronous Overlap Add (PSOLA) method to generate the speech waveforms.

3. Proposed Methods to deal with data from the wild

The data found from Youtube is typically noisy, contains many pauses and incomplete sentence structures that are usual in conversational speech. There are also no ground truth transcriptions and obtaining the human transcriptions would be time consuming and very expensive. We employ a combination of techniques to generate labels, sanitize the mixed quality data and select good utterances for performing speech synthesis in the wild. Figure 2 shows the block diagram of the methods used to perform utterance selection and model adaptation for data from the wild.

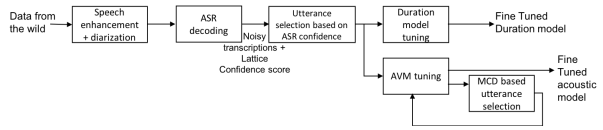

Figure 2: *Block diagram of the methods used to adapt the AVM and the duration model for the data from the wild.*

3.1. Speech Enhancement

Most often utterances downloaded from the internet contain background noise and music in them which are detrimental to speech synthesis. We perform a Wiener Filtering based speech enhancement on the downloaded audio file before further processing for synthesis.

3.2. Speaker diarization and segmentation

Most speech resources for celebrities online are from interviews and talk shows where typically two or more speakers (including the target subject) are speaking. We performed voice activity based segmentation and speaker diarization to cluster speech segments into different speakers and manually selected the cluster of belonging to the target speaker. Thus audio from any interview or talk show with predominantly speech content could be used to extract data for speech synthesis.

3.3. Transcribing using ASR

The audio downloaded from Youtube were transcribed using a conversational ASR system trained on the Fisher corpus. The ASR model was a DNN acoustic model trained on the Fisher corpus with i-vector based speaker adaptation which achieved a 23.7% word error rate (WER) on the highly challenging Fisher corpus development set.

3.4. Utterance selection methods

Since the duration models and AVMs are adapted to the test speakers, it was essential to prune the data based on the the quality of the utterance and the transcriptions. We perform Weiner filtering based speech enhancement as a preproecessing step, and Mean Cepstral Distortion (MCD) and ASR confidence measure based utterance selection for pruning the data for acoustic and duration model adaptation.

3.4.1. ASR confidence score

We computed the the difference between the total costs of the best and the second-best paths in the ASR decoding lattice and used it as a rough confidence measure of the ASR decoding. For every target speaker we sorted the adaptation set utterances based on the confidence measure and kept the top 80% of the utterances.

3.4.2. MCD based utterance selection

The MCD-based utterance selection is done in an iterative manner: after every Fine Tuning of the Speaker Acoustic Model, the utterances with high MCD (higher than a fixed threshold that was set empirically to 7.0) are removed from the training data.

4. Conclusion

Our software puts together all the components into one automated script that is capable of synthesizing a voice from a given Youtube video link. The software retrieves the chosen video from online, performs diarization, segmentation, speech enhancement, transcription, utterance selection to prepare the data for model training. The software then adapts the AVM and the duration models to the target speaker using the pre-processed data to create the synthetic voice. We will demonstrate the working of this automated software with live on the spot synthesis using audio sourced from the internet.

5. References

[1] Y. Qian, Y. Fan, W. Hu, and F. K. Soong, "On the training aspects of Deep Neural Network (DNN) for parametric TTS synthesis," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2014.

[2] H. Zen, "Acoustic modelling in statistical parametric speech synthesis - from HMM to LSTM-RNN," *Mlslp*, 2015.

[3] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, nov 2016.

[4] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for DNN-based speech synthesis," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2015.

[5] P. Baljekar and A. W. Black, "Utterance Selection Techniques for TTS Systems Using Found Speech," in *9th ISCA Speech Synthesis Workshop*, sep 2016, pp. 184–189.

[6] P. Taylor, A. W. Black, and R. Caley, "The Architecture of the Festival Speech Synthesis System," in *International Conference on Spoken Language Processing*. International Speech Communication Association, 1998.

[7] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit," University of Edinburgh. The Centre for Speech Technology Research (CSTR)., Edinburgh, Tech. Rep., 2016.