UDC - 123.456.789

A NEW APPROACH FOR SPEAKER ADAPTATION WITH TTS MODELS BASED ON TRANSFORMER AND FEW-SHOT LEARNING METHOD

Gor P. Matevosyan
Russian-Armenian University
123 Hovsep Emin St., Yerevan 0051, RA
gor.matevosyan@student.rau.am
ORCID iD: 0000-0002-7466-5078
Republic of Armenia

Abstract

Deep learning methods are showing state-of-the-art results in many fields of machine learning and speech synthesis with multiple speaker voices is not an exception. Adapting the existing multispeaker text-to-speech (TTS) model to new speakers remains a challenging task. Current methods of synthesizing speech with the voice of unseen speakers (referring to the speaker recordings not included in the model's training data) are based on either transfer-learning methods or fine-tuning of existing multispeaker TTS model which requires a lot of data (30 minutes at least). In the current work a new approach has been developed to synthesize unseen speaker speech with only about 3 minutes of data based on the same TTS model and a few-shot learning method for speaker embeddings.

Key words: multispeaker text-to-speech model, few-shot learning, voice cloning, transformer.

Introduction

After decades of research, text-to-speech synthesis or the process of creating natural voice from text remains a difficult task. There are numerous TTS systems available today that can produce natural-sounding voices that are incredibly near to human ones. Unfortunately, many of these systems acquire the ability to synthesize text using just one voice. The purpose of this study is to develop a TTS system capable of efficiently generating natural speech for a diverse range of speakers who are not necessarily encountered during the training phase. The process by which these models are created is called Voice Cloning. It has various use-cases including recording podcasts without voice recordings or personalizing digital assistants such as Apple's Siri.

Conflict setting

There has been a growing interest in end-to-end solutions throughout time. Tacotron 2 [1] learned directly from text-audio pairs using WaveNet [2] as a vocoder to transform generated spectrograms by sequencing with attention model architecture [3] that encrypts text and decrypts spectrograms; however, it was limited to a single speaker. Later, Gibiansky et al. [4] presented a multi-speaker Tacotron variant to learn a latent speaker embedding for each training speaker from training data. Deep Voice 3 [5] developed a fully convolutional encoder-decoder architecture that supports thousands of speakers from [6] LibriSpeech data. Previously we have obtained a TTS model, which is able to generate spectrograms of multiple speakers as well. However, these systems are not able to generate speech with the voice of an unseen speaker.

Voiceloop [7] suggested an innovative architecture capable of producing speech from unheard voices during training. However, this method needs more tens (30 minutes or even more in some

cases) minutes of speech and transcripts. Later, transfer-learning-based methods were invented [8] to use only a few seconds of audio to generate speech with that voice. However, these methods do not have the desired Mean Opinion Score (MOS). They do not sound like professional recordings because they used a predefined network for speaker embedding prediction for which the TTS model is not adapted.

To overcome these problems, we introduce a new method to get embedding for unseen speakers having no more than 3 minutes of speech using the same TTS model which will later condition that embedding to generate speech. To achieve this, we have changed the concatenative speaker embedding condition method with a complex layer to make our model more sensitive for speakers. We freeze the whole model except the embedding layer to finetune with 3 minutes of data and get the desired speaker embedding. This approach has few advantages over others. First, it uses the same weights to generate speech for all speakers, even for unseen ones and we have just to store embeddings. Next, it does not require significant data to achieve desired results, and finally, the training time to adapt to new speakers does not make the model forget old ones.

Methods and Models

Text to speech is a one-to-many mapping problem as a single letter or phoneme can be pronounced differently in different contexts. This section reviews the methods based on transfer-learning [8]. It covers our model's core components, FastSpeech2, FastPitch and TransformerTTS and methods intuition for solving unseen speaker problems.

A Multi-speaker Text-to-speech Synthesis Approach Based on transfer learning system consists of three components: a *speaker-encoder* that generates a fixed-dimensional embedding vector from a few seconds of reference speech delivered by a target speaker; a *synthesizer* that generates a mel-spectrogram from an input text and an embedding vector; and a *neural vocoder* that infers time-domain waveforms from the synthesizer's mel-spectrograms. At inference time, the speaker encoder receives a brief reference utterance from the target speaker as input and creates an embedding vector based on its internal learnt speaker characteristics space. The synthesizer accepts a phoneme (or grapheme) sequence as input and creates a mel-spectrogram with the speaker encoder embedding vector as a condition. Finally, the vocoder takes the synthesizer's output and creates the speech waveform. Following results can be accomplished by training a neural network model on a text-independent speaker verification task to optimize the GE2E loss [10], such that embeddings of the same speaker's utterances have a high cosine similarity. In contrast, those of different speakers' utterances are widely separated in the embedding space. The main disadvantage of this method is that it uses a pre-trained network to extract embeddings from a single utterance, leading to a leak of data about emotions in embeddings because of GE2E loss. Also, at inference, there is no possibility to fix some embeddings if there are any problems. The overall architecture of the method is described in Fig. 1.

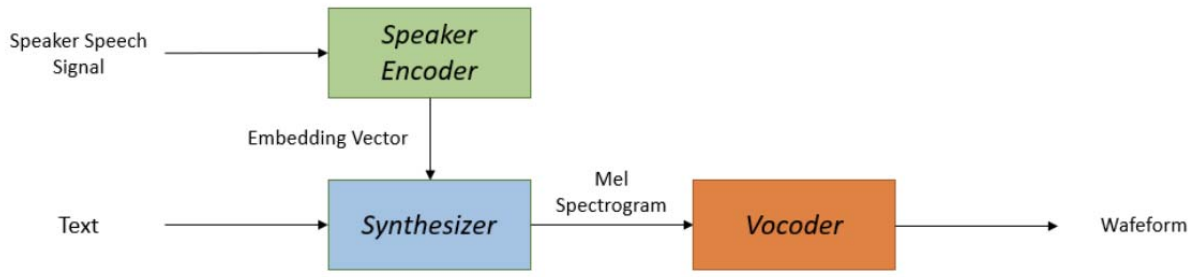

Fig. 1 High-level overview of the system

Our previous model has generated speeches with a high MOS score (4.173) for a large range of speakers using no more than 3 minutes of audio for each speaker. We achieved those results thanks to Transformer based architecture based on the Multi-Head Attention mechanism [11] and pitch and duration prediction during inference. The overall architecture of our previous model is described in Fig. 2. The following section shows how we use this model as a base, change the concatenative speaker embedding conditioning mechanism with a more complex layer, and adapt it to new speakers without forgetting old ones.

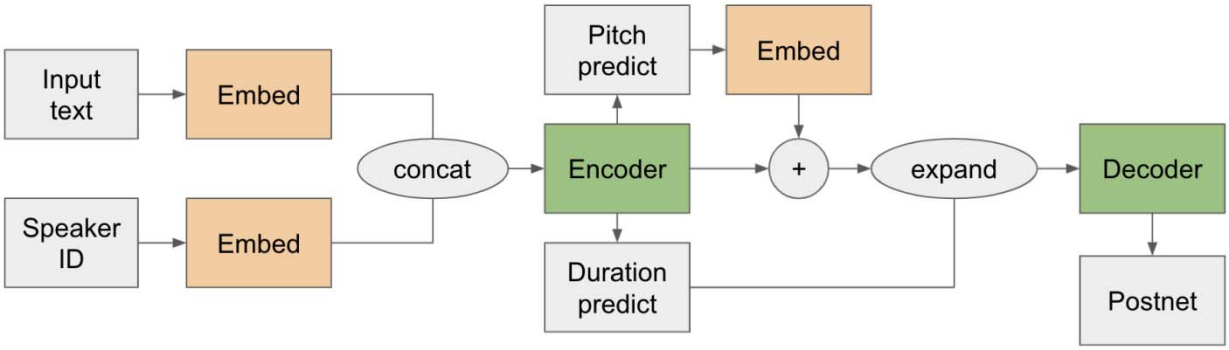

Fig. 2 The overall architecture of our previous TTS model

As a base, we use the model described in Fig. 2, a stack of Multi-Head Attention Transformer blocks. Both Encoder and Decoder consist of Self-Attention-Dense Blocks (SADB) followed by Self-Attention-Conv Blocks (SACB). This model can generate speech with more than 800 different voices with a MOS score higher than 4.1.

We have changed the concatenative method of applying speaker embedding with a more complex layer. Let us suppose that we have phoneme embeddings $E_s = E_{symbol}(x) \in R^{n \times d1}$ and speaker embedding $E_{sp} = E_{speaker}(s) \in R^{n \times d2}$, where n is the length of the input sequence (the length of phonemes of graphemes), x is an input sequence, E_{symbol} is the symbol encoder, and $E_{speaker}$ is a speaker encoder, $d1$ is the dimension of phoneme embedding and $d2$ is the dimension of speaker embedding. We changed the concatenative method $E = concatenate(E_{symbol}, E_{speaker}, axis = -1) \in R^{n \times d1+d2}$ with a more complex layer by adding a liner layer with bias. The final equation of merging symbol and speaker embeddings is below.

$$E_{concat} = concatenate(E_{symbol}, E_{speaker}, asix = -1) \in R^{n \times d1+d2}$$
$$E = W * E_{concat} + b \quad (1)$$

where W and b are trainable parameters. The intuition is that each speaker embedding dimension relates to each dimension of symbol embedding, which gives merged contextual embedding of symbols and speakers.

We found that by picking a random vector from speaker embedding space or merging few speaker embeddings with sum, mean or max pooling, the model generates speech with voices of unseen speakers. So we came to the idea that for every new unseen voice, we can find an embedding. Suppose we have a pre-trained model described in Figure 1 and data of a new unseen speaker. Our goal is to find an embedding with which the model will generate speech with an unseen speaker voice. To achieve this, we first have to find an initialization of that embedding e and, next, finetune that

embedding. First, we use two approaches for initialization - random initialization and picking one from the training set with a similar voice. Experiments show that convergence is faster in the second case, but the first case does not require any human interaction. In both cases, the results are very similar. Next, we change the $E_{speaker}$ layer in the model with our initialized embedding e. We freeze the whole model except e and train few hundred steps on the new data. Next, we pick the resulted embedding. In this way, the model can find such embedding with which it can generate speech of unseen voice.

To train the model using a new data, we employ the following loss previously introduced by us. It is a weighted sum of L1 loss for pitch, duration, and target mel-spectrogram and categorical cross-entropy for the speaker (as a weight of this loss, we use 0). L1 loss and categorical cross-entropy loss functions have these formulas:

$$L1(y,\hat{y}) = \sum_{i=1}^{n}|y_i - \hat{y}_i|. \quad (2)$$

Where y is the actual value, $\hat{y}$ is the predicted value, n is the vector size of vectors y and $\hat{y}$. With weights w_s, w_y, w_p, the total loss of the model will be:

$$L = w_s * CCE(s,\hat{s}) + w_y * L1(y,\hat{y}) + w_p * L1(p,\hat{p}) \rightarrow min \quad (3)$$

Research results

To evaluate the results, we use the Mean Opinion Score (MOS). We finetune the model for our internal speakers and it performs slightly better than other solutions as shown in Tab. 1.

Table 1

Mean Opinion Score (MOS) for speech naturalness with 95% confidence intervals

Method	MOS Score
Tacotron2 + GST - Zero-shot	2.67 ± 0.10
Expressive Neural Voice Cloning - Zero-Shot	3.56 ± 0.09
Expressive Neural Voice Cloning - Adaptation Whole	3.75 ± 0.09
Expressive Neural Voice Cloning - Adaptation Decoder	3.61 ± 0.09
Proposed model	3.83 ± 0.12

Overall, our solution works better and requires a small amount of data for unknown speakers which is a huge advantage compared to the existing solutions.

Conclusion

The challenge of speech synthesis for multiple voices is a rapidly developing area using deep learning methods. In this paper we have explored a new model of multispeaker text-to-speech generation which uses the fraction of the data required compared to existing methods (3 mins VS 30 mins) for speaker recordings which have not been previously used in the training data.

References

1. Shen J., et al. Natural TTS Synthesis by Conditioning WaveNet on mel-spectrogram predictions (2018) //2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).- p. 4779-4783.
2. van den Oord A., et al. Wavenet: A generative model for raw audio (2016) //arXiv preprint arXiv: 1609.03499
3. Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate (2015) //CoRR, vol. abs/1409.0473, 2015.
4. Gibiansky A., Arik S., Diamos G., Miller J., Peng K., Ping W., Raiman J., Zhou Y. Deep voice 2: Multi-speaker neural text-to-speech (2017) //Advances in Neural Information Processing Systems 30.- p. 2962-2970
5. Ping W., Peng K., Gibiansky A., Arik S., Kannan A., Narang Sh., Raiman J., Miller J. Deep voice 3: 2000-speaker neural text-to-speech (2018) //International Conference on Learning Representations, 2018.
6. Panayotov V., Chen G., Povey D., Khudanpur S. Librispeech: An asr corpus based on public domain audiobooks (2015) //2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).- p. 5206-5210.
7. Taigman Y., Wolf L., Polyak A., Nachmani E. Voiceloop: Voice fitting and synthesis via a phonological loop (2018) //International Conference on Learning Representations, 2018.
8. Ruggiero G., et al. Voice Cloning: a Multi-Speaker Text-to-Speech Synthesis Approach based on Transfer Learning (2021) //arXiv preprint arXiv:2102.05630
9. Nachmani E., Polyak A., Taigman Y., Wolf L. Fitting new speakers based on a short untranscribed sample (2018) //CoRR, vol. abs/1802.06984
10. Wan L., Shan Wang Q., Papir A., Lopez Moreno I., Generalized end-to-end loss for speaker verification (2018) //2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).- p. 4879-4883.
11. Vaswani, A., et al. Attention is all you need (2017) //Advances in neural information processing systems 30.- p. 5998-6008.

ԽՈՍՆԱԿԻ ՁԱՅՆԻ ՀԱՐՄԱՐԵՑՄԱՆ ՆՈՐ ՄՈՏԵՑՈՒՄ ՏՐԱՆՍՖՈՐՄԵՐՆԵՐԻ ՎՐԱ ՀԻՄՆՎԱԾ TTS ՄՈԴԵԼԻ ԵՎ ՄԻ ՔԱՆԻ ՓՈՐՁՈՎ ՈՒՍՈՒՑՄԱՆ ՄԵԹՈԴԻ ԿԻՐԱՌՄԱՄԲ

Գ.Պ. Մաթևոսյան
Հայ-Ռուսական համալսարան

Խորը ուսուցման մեթոդները ցույց են տալիս բարձր մակարդակի արդյունքներ մեքենայական ուսուցման բազմաթիվ ոլորտներում, և բազմակի «խոսնակներով» խոսքի սինթեզը բացառություն չէ: Գոյություն ունեցող բազմախոսնակ տեքստից խոսք (TTS, text to speech) մոդելը նոր խոսնակներին հարմարեցնելը մնում է դժվարին խնդիր: Անտեսանելի խոսնակների (անտեսանելին վերաբերում է խոսնակի ձայնագրություններին, որոնք ներառված չեն մոդելի ուսուցման տվյալներում) ձայնով խոսքի սինթեզման ներկայիս մեթոդները հիմնված են կամ փոխանցման ուսուցման (transfer-learning) մեթոդների կամ գոյություն ունեցող բազմախոսնակ TTS մոդելի ճշգրիտ կարգավորման վրա (fine-tuning), որը պահանջում է բազմաթիվ տվյալներ (առնվազն 30 րոպե): Տվյալ աշխատանքում նոր մոտեցում է մշակվել անտեսանելի խոսնակի խոսքի սինթեզման համար՝ ընդամենը մոտ 3 րոպե տևողությամբ տվյալներով, հիմնված նույն TTS մոդելի և մի քանի փորձով ուսուցման (few-shot learning) մեթոդի վրա:

Բանալի բառեր. բազմախոսնակ տեքստից խոսք մոդել, մի քանի փորձով ուսուցում, ձայնի կլոնավորում, տրանսֆորմեր:

НОВЫЙ ПОДХОД ДЛЯ АДАПТАЦИИ ГОЛОСА СПИКЕРА С ИСПОЛЬЗОВАНИЕМ TTS-МОДЕЛЕЙ НА ОСНОВЕ ТРАНСФОРМЕРА И МЕТОДА ОБУЧЕНИЯ С НЕСКОЛЬКИМИ ПОПЫТКАМИ

Г.П. Матевосян
Российско-Армянский университет

Методы глубокого обучения показывают высокие результаты во многих областях машинного обучения, и синтез речи с голосами нескольких спикеров не является исключением. Адаптация существующей мультиспикерной модели преобразования текста в речь (TTS, text to speech) к новым спикерам остается сложной задачей. Современные методы синтеза речи с голосом невидимого спикера (невидимый относится к записям спикера, не включенным в данные обучения модели) основаны либо на методах трансфер-обучения, либо на тонкой настройке (fine-tuning) существующей мультиспикерной модели TTS, которая требует большого количества данных (минимум 30 минут). В настоящей работе был разработан новый подход к синтезу речи невидимого спикера, с данными всего за 3 минуты, на основе той же модели TTS и метода обучения с несколькими попытками.

Ключевые слова: мультиспикерная модель текста в речь, обучение с несколькими попытками, клонирование голоса, трансформер.

Submitted on 05.10.2021.
Sent for review on 06.10.2021.
Guaranteed for printing on 11.11.2021.