How to make someone speak a language that they don't know.

Anonymous Author(s)
Affiliation
Address
email

Abstract

We present a simple idea that allows to record a speaker in a given language and synthesize their voice in other languages that they may not even know. These techniques open a wide range of potential applications such as cross-language communication, language learning or automatic video dubbing. We call this general problem *multi-language speaker-conditioned speech synthesis* and we present a simple but strong baseline for it.

Our model architecture is similar to the encoder-decoder Char2Wav model [Sotelo et al., 2017] or Tacotron [Shen et al., 2017]. The main difference is that, instead of conditioning on characters or phonemes that are specific to a given language, we condition on a shared phonetic representation that is universal to all languages [Meier, 2016]. This cross-language phonetic representation of text allows to synthesize speech in any language while preserving the vocal characteristics of the original speaker. Furthermore, we show that fine-tuning the weights of our model allows us to extend our results to speakers outside of the training dataset.

1 Introduction

The goal of this work is to create a text-to-speech system able to generate audio in multiple languages for any given speaker. We further impose two requirements. First, the model should be able to copy the voice of an out-of-dataset speaker given only very limited data. Second, the model should be able to generate audio in any language, even when trained on a single-language speaker.

Such a system, paired with a word translation system, would enable anyone to speak in any language. It could be used by travelers to help them communicate in foreign countries or by movie producers to dub movies while keeping the original voices of their actors.

Our approach is to build a model able to generate speech in multiple languages. The model is trained with multiple speakers to let the model be aware of the variations between speakers and also to disentangle speech content from speaker identity. Once the model is trained, we bias the generation process so that it sounds like a specific speaker. This speaker doesn't have to be in the training data.

2 Related Work

Our work builds upon recent developments in neural network based speech synthesis [Sotelo et al., 2017, Ping et al., 2017, Shen et al., 2017, Van Den Oord et al., 2016]. Specifically, our model architecture closely resembles attention-based speech synthesis models, which map sequences of phonemes or characters to intermediate audio representations e.g. vocoder [Morise et al., 2016][Sotelo et al., 2017] or spectogram [Shen et al., 2017][Ping et al., 2017]. The representation is post-processed via either signal processing based methods e.g. Griffin-Lim spectrogram inversion [Griffin and Lim, 1984], World/Straight vocoder [Morise et al., 2016] or neural vocoders [Sotelo et al., 2017][Shen et al., 2017][Ping et al., 2017] to get raw audio. Our model is very closely related to neural multi-language multi-speaker parametric speech synthesis model described by Li and Zen [2016]. Our approach enables speaker style transfer across languages present in the training dataset, which is different from generalising to new unseen languages with small amount of data. Also Li and Zen [2016] uses union of language-dependent linguistic feature set to represent text input to their model. As opposed to them, we use International Phonetic Alphabet (IPA) [Meier, 2016] to map text input across all languages to a universal representation.

Our model is able to accomplish zero-shot accent transfer, which is very similar to zero-shot machine translation, done by grounding the input from different languages to a common neural representation space, followed by decoding in the audio space [Johnson et al., 2016].

3 International Phonetic Alphabet (IPA)

In this work we use the International Phonetic Alphabet (IPA) [Meier, 2016] to represent the text information. The IPA is designed to represent only those qualities of speech that are part of oral language: phones, phonemes, intonation and the separation of words and syllables. Note that the IPA is language independent. It thus provides a common representation for the text input across several languages to be conditioned upon for speech generation. In the following, we will show that our model is able to generalize the phonetic information learned across multiple languages. This reduces the complexity of text conditioning and improves data efficiency when we train the model with multilingual data.

4 Model

At the core of our Text-To-Speech system lies an attention-based encoder-decoder architecture similar to [Sotelo et al., 2017] [Shen et al., 2017] [Bahdanau et al., 2014]. More specifically, we use a bidirectional recurrent conditioning encoder, a recurrent decoder and a location based attention mechanism very similar to the one used in [Graves, 2013].Discrete conditioning information i.e. IPA sequences and speaker information, are modeled with randomly initialized embedding layers that are trained together with the rest of the model.

The training data consists of audio-transcript pairs. The transcript is translated into its IPA equivalent before being fed to the model and the audio is transformed into an intermediate representation (e.g. WORLD vocoder parameters or spectrogram). Each speaker within the training dataset only speaks a single language. However, at synthesis time, we are able to take any combination of speaker and language, and produce natural sounding speech in the voice of the speaker and in the accent matching that of the language.

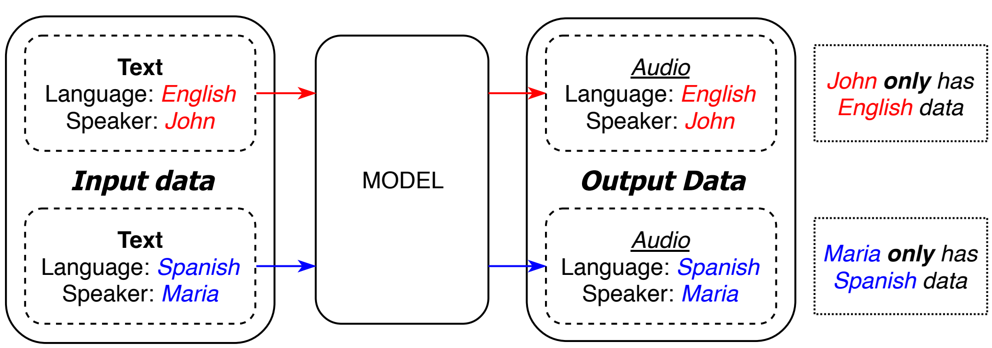
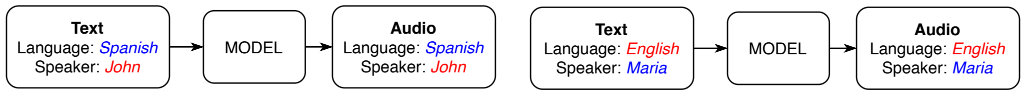


Figure 1: Training process. The model is trained to map text to audio based on *(transcript-audio)* pairs. Maria is a speaker for whom the model only sees Spanish data, John is a speaker for whom the model only sees English data

During inference, the model is able to generate Spanish audio in the style of an English speaker and vice-versa.



(a) Generating Spanish audio in the style of the English speaker

(b) Generating English audio in the style of the Spanish speaker

Figure 2: Inference process. The model is able to interpolate and generate any *(speaker, language)* combination.

The method presented above only allows us to generate speech for speakers that are present in the training dataset. In order to generate multilingual speech for a new speaker, we fine-tune the trained model on a small amount of new speaker's data, which can be as little as merely 300 seconds of utterances. It is important to preventing catastrophic forgetting during fine-tuning. We carried out a careful and exhaustive search hyperparameter search to obtain robust performance across many speakers. Crucially, we apply a smaller learning rate to the encoder and decoder parts of the models, and a higher one for the speaker embedding. This improved speaker fidelity considerably.

After fine-tuning, the model is able to generate any text in any language[1] with the new speaker's vocal identity.

5 Experiments

We conduct experiments on our models trained in two distinct settings. First, we train our model with data in two languages (Bilingual Model). Second, we train our model with data in six languages (Multilingual Model).

For these experiments, we used several datasets. We used an internal English dataset composed of approximately 20000 speakers, with about 10 utterances per speaker. We also used the TIMIT dataset [Garofolo et al., 1993] and DIMEx100 [Pineda, 2009]. DIMEx100 is a Spanish dataset composed of 100 Spanish native speakers, with about 60 2-seconds utterances per speaker.

For all the experiments we provide audio samples[2] rather than an exhaustive quantitative analysis.

5.1 Bilingual Model

We use a bilingual dataset (English and Spanish) composed of TIMIT and DIMEx100 to train this model. We concatenate to this data from the speaker we want to clone. We are able to generate bilingual speech for both English and Spanish speakers within the dataset. We also fine-tune the trained model on new speaker's data to generate bilingual speech for the new speaker. We only need data from the new speaker in a **single** language.

5.2 Multilingual model

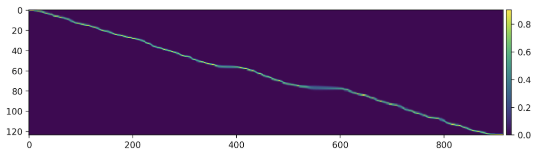
For this experiment we train the model with our internal multi-speaker dataset to which we add data from 5 single-speaker audiobooks from the CSS10 dataset [Kyubyong Park, 2018], each in a distinct language. We show that the model is able to generate in any language for any speaker in the dataset. The model also shows robust performance on new, out-of-sample speakers after the fine-tuning step (see figure 3).

References

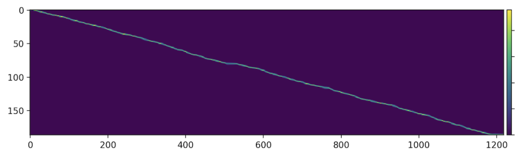
Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014. URL `http://arxiv.org/abs/1409.0473`.

[1] Included in the training dataset.

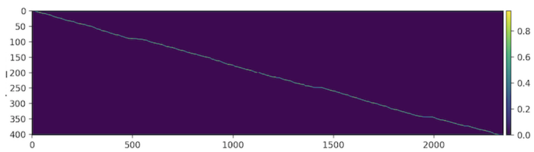
[2] https://everyone-speaks-every-language.github.io/



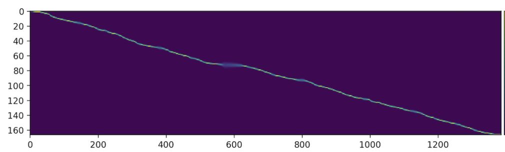
(a) Attention plot for a sentence in English



(b) Attention plot for a sentence in Spanish



(c) Attention plot for a sentence in French



(d) Attention plot for a sentence in German

Figure 3: Attention plots for the same sentence in Englsih, Spanish, French and German (same speaker)

J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. Darpa timit acoustic phonetic continuous speech corpus cdrom, 1993.

Alex Graves. Generating sequences with recurrent neural networks. *CoRR*, abs/1308.0850, 2013. URL `http://arxiv.org/abs/1308.0850`.

Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. Google's multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*, 2016.

Tommy Mulc Kyubyong Park. Css10: A collection of single speaker speech datasets for 10 languages. `https://github.com/Kyubyong/css10`, 2018.

Bo Li and Heiga Zen. Multi-language multi-speaker acoustic modeling for lstm-rnn based statistical parametric speech synthesis. In *INTERSPEECH*, pages 2468–2472, 2016.

Paul Meier. *International Phonetic Alphabet (IPA) Charts*. Paul Meier Dialect Services, Lawrence, KS, 2016. URL `http://www.paulmeier.com/ipacharts/`. bibtex: meier_international_2016.

Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE TRANSACTIONS on Information and Systems*, 99(7):1877–1884, 2016.

Hayde Cuetara Javier Galescu Lucian Juárez Janet Llisterri Joaquim Pérez Patricia Villaseñor-Pineda Luis Pineda, Luis Castellanos. The corpus dimex100: Transcription and evaluation. *Language Resources and Evaluation. 44. 347-370. 10.1007/s10579-009-9109-9*, 2009.

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: 2000-speaker neural text-to-speech. *arXiv preprint arXiv:1710.07654*, 2017.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. *CoRR*, abs/1712.05884, 2017. URL `http://arxiv.org/abs/1712.05884`.

Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. Char2wav: End-to-end speech synthesis. 2017.

Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *SSW*, page 125, 2016.