Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)
Doctoral Consortium Track

End-to-End Signal Factorization for Speech: Identity, Content, and Style

Jennifer Williams
Centre for Speech Technology Research, The University of Edinburgh, United Kingdom
j.williams@ed.ac.uk

Abstract

Preliminary experiments in this dissertation show that it is possible to factorize specific types of information from the speech signal in an abstract embedding space using machine learning. This information includes characteristics of the recording environment, speaking style, and speech quality. Based on these findings, a new technique is proposed to factorize multiple types of information from the speech signal simultaneously using a combination of state-of-the-art machine learning methods for speech processing. Successful speech signal factorization will lead to advances across many speech technologies, including improved speaker identification, detection of speech audio deep fakes, and controllable expression in speech synthesis.

1 Introduction

Abstract representations of speaker identity, such as *i*-vectors and *x*-vectors, are used in many different speech technologies including speaker identification (SID), language identification (LID), audio deep fake detection, voice conversion (VC), and text-to-speech (TTS) synthesis. These representations are designed to maximize the variance between speakers, while minimizing within-speaker differences (recording device, mood, age, etc). It has recently been shown that these representations contain additional information including gender, age, spoken content, and session variability [Raj *et al.*, 2019]. Factorizing each information type would not only improve the purity of the speaker representations, but it will also lead to additional embeddings such as speaking style and emotion that can be used in many different speech technologies [Williams and King, 2019].

2 Methodology

2.1 Overview

The goal of this dissertation is to provide an end-to-end machine learning method that deconstructs the speech signal into representations that can be used in various speech technology tasks. This task of deconstructing, also known as factorization or disentanglement, allows for representing parts of the signal in terms of components. These components can then be re-assembled in different ways depending on the application. For example, in a VC task it is necessary to retain speech content and style but change only the speaker identity. In another example, in TTS synthesis it is necessary to retain the speaker identity but change the content and style. There has been recent work to separately factorize specific pieces of information from the speech signal. However, those learned representations are often noisy. There are currently no methods to learn this multiple-component factorization in a single system, which is what this work seeks to do. Three types of factorization that are modeled in this end-to-end system are:

- Who: speaker identity
- What: utterance content
- How: speaking style/emotion

2.2 Stacked Encoders

The generative autoencoder network called VQ-VAE learns a discrete abstract latent space for audio while overcoming a common problem of posterior collapse [van den Oord *et al.*, 2017]. While it is possible to generate audio from a discrete set of VQ codes, a major improvement of this model is to add more structure to the latent space [Ding and Gutierrez-Osuna, 2019]. This additional structure is so rich that it can represent spoken utterance content in the form of phonemes. Versions of the model can also learn to generate speech audio with variable speaking style [Wang *et al.*, 2019]. This dissertation proposes to stack multiple VQ-VAEs wherein each encoder learns to represent different information in the latent space.

2.3 Decoder

In order to generate speech audio using an autoencoder paradigm, this requires a decoder that can be conditioned on a set of discrete embeddings as input (such as from VQ-VAE) and generate high-fidelity audio samples as output. A major advancement comes from using WaveRNN as a decoder. It employs a narrow RNN layer as well as sparsified weight matrices and operates on batches of samples per step in training which improves training speed [Kalchbrenner *et al.*, 2018]. Importantly, WaveRNN can be conditioned locally or globally using any combination of discrete VQ embeddings, and generate high-quality speech output. This means that the VQ-VAE encoders can be stacked so that different types of information is represented simultaneously, which is the ultimate

objective of this work. From these stacked encoders and WaveRNN decoder, it is then possible to use techniques such as Deep InfoMax (DIM) [Hjelm *et al.*, 2018] to guide and evaluate how well the encoders are separating the information that is being factorized (speaker identity, content, and style). DIM provides a framework for evaluating representations based on maximizing mutual information between the input and output of an encoder.

3 Preliminary Results

3.1 Encoding Speech Environments

Recent work in [Williams and Rownicka, 2019] has shown it is possible to create representations of the recording environment using the speech signal. These elements of session variability include room size, quality of recording device, distance between speaker and microphone, and reverberation. The information is encoded into the speech signal and can be factorized out and it is helpful for detecting various types of deep fake attacks for speech audio.

3.2 Disentangling Speaker Factors

[Williams and King, 2019] showed that speaking style and emotion can be extracted from speaker representations using a technique of stacked encoders. Each encoder learned different information. It is proposed that disentanglement directly from the speech signal, instead of intermediary representations, would yield more robust factorization that can be applied easily across speech technology applications.

3.3 Automatic Speech Quality Assessment

Very recent work in [Williams *et al.*, 2020] has shown a way to train a neural network that can predict speech audio quality that correlates with human judgements. This technique is useful for evaluating synthetic speech, as well as learning to detect deep fakes. The models can also be adapted to evaluate specific aspects of quality such as speaker similarity or speaking style.

4 Discussion and Impacts

This dissertation builds on previous work and preliminary results to achieve a robust factorization of the speech signal for speaker identity, content and speaking style. All of these factors are critical for advancing the field of speech technology. In TTS synthesis, it allows control of expressiveness while maintaining high-quality speech output. TTS technology is important because it overlaps with assistive technology for those with disordered speech, screen readers for the blind, second language learning, speech-to-speech translation, storytelling, audio books, and others. The speaker representations that are achieved by this work may be considered more pure as they would have non-relevant information removed. Purified speaker representations have the capacity to lead to more secure speaker identification technology. And the ability to factorize speech content from voice characteristics forms the core of voice privacy technology. Voice privacy is a growing field that is closely related to speaker identification and speech audio deep fakes.

Acknowledgments

This work was supported in part by the EPSRC Centre for Doctoral Training in Data Science, funded by the UK Engineering and Physical Sciences Research Council (grant EP/L016427/1) and the University of Edinburgh. This work was also supported through an internship at the National Institute for Informatics in Tokyo, Japan. The author thanks Simon King, Junichi Yamagishi, Erica Cooper, Yi Zhao, and Xin Wang for their guidance and mentorship.

References

[Ding and Gutierrez-Osuna, 2019] Shaojin Ding and Ricardo Gutierrez-Osuna. Group Latent Embedding for Vector Quantized Variational Autoencoder in Non-Parallel Voice Conversion. In *Proc. Interspeech 2019*, pages 724–728, 2019.

[Hjelm *et al.*, 2018] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning Deep Representations by Mutual Information Estimation and Maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[Kalchbrenner *et al.*, 2018] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient Neural Audio Synthesis. In *International Conference on Machine Learning*, pages 2410–2419, 2018.

[Raj *et al.*, 2019] Desh Raj, David Snyder, Daniel Povey, and Sanjeev Khudanpur. Probing the information encoded in x-vectors. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2019, Singapore, December 14-18, 2019*, pages 726–733. IEEE, 2019.

[van den Oord *et al.*, 2017] Aäaron van den Oord, Oriol Vinyals, et al. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems*, pages 6306–6315, 2017.

[Wang *et al.*, 2019] Xin Wang, Shinji Takaki, Junichi Yamagishi, Simon King, and Keiichi Tokuda. A Vector Quantized Variational Autoencoder (VQ-VAE) Autoregressive Neural F0 Model for Statistical Parametric Speech Synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:157–170, 2019.

[Williams and King, 2019] Jennifer Williams and Simon King. Disentangling style factors from speaker representations. *Proc. Interspeech 2019*, pages 3945–3949, 2019.

[Williams and Rownicka, 2019] Jennifer Williams and Joanna Rownicka. Speech replay detection with x-vector attack embeddings and spectral features. *Proc. Interspeech 2019*, pages 1053–1057, 2019.

[Williams *et al.*, 2020] Jennifer Williams, Joanna Rownicka, Pilar Oplustil, and Simon King. Comparison of Speech Representations for Automatic Quality Estimation in Multi-Speaker Text-to-Speech Synthesis. *Speaker Odyssey*, 2020.