Voice Conversion

Shariq A. Mobin
Redwood Center for Theoretical Neuroscience
UC Berkeley
shariq.mobin@gmail.com

Joan Bruna
Department of Statistics
UC Berkeley
joan.bruna@gmail.com

Abstract

The human auditory system is able to distinguish the vocal source of thousands of speakers, yet not much is known about what features the auditory system uses to do this. Fourier Transforms are capable of capturing the pitch, or harmonic structure, of the speaker but this alone proves insufficient at identifying speakers uniquely. The remaining structure, often referred to as timbre, is critical to identifying speakers yet little is understood about it. In this paper we use recent advances in neural networks in order to manipulate the voice of one speaker into another by transforming not only the pitch of the speaker, but the timbre. We review generative models built with neural networks as well as architectures for creating neural networks that learn analogies. Our preliminary results converting voices from one speaker to another are encouraging.

1 Introduction

When audiologists describe what makes the sound of one person to the next sound different they first refer to the pitches of the speakers and second to the timbre of speakers. While pitch is well described by the harmonic structure little is known about what timbre is. Often the pitches can be identical but the way it sounds is completely different. This can heard when a trumpet and piano play the same note, while they have same pitch, the timbre is what gives rise to the strong disparity in perception. One can think about vocal signals as being an entanglement of two factors - what the speaker is saying and who is saying it. The vocal signal is a non-stationary process which causes the disentanglement of these two factors to be very difficult. In this paper we will explore if it possible to hold one of these two factors constant and alternate the other. That is, we will see if it is possible to convert the speaker of a vocal signal while holding the word spoken constant.

2 Background

2.1 Constant Q-Transform

In theory, we could train our neural network with the raw audio waveformm as input. However, in the signal processing community some type of frequency analysis of the waveform is often analyzed as this transformation makes explicit the harmonic structure of the signal. Here, we apply a constant-Q filter bank wavelet transformation (CQT)to the audio signal. This transformation has a number of desirable properties, the most important are:

1. The transformation uses logarithmic scaling in frequency. This is very useful when the sound wafeform spans many octaves as in the case of waveforms from the human vocal system.
2. The CQT transformation has very high temporal resolution and low spectral resolution for high frequencies whereas the converse is true for low frequncies. This is very similar to the

transformation the basilar membrane of the cochlea in the human auditory system performs on the sound waveform.

2.2 Deep Visual Analogy Making

Deep Visual Analogy Networks [2] are a recent neural network architecture that has been able to achieve incredible results rotating sprites in the image domain. The goal of the network is to make analogies: "A is to B as C is to D". That is, given A, B, and C as input we would like to predict D. An example would be: "groom is to bride as king is to queen". The approach taken by this model is to learn an embedding of the input such that solving these analogies is easy, e.g. linear:

$$\Phi(D) - \Phi(C) \approx \Phi(B) - \Phi(A)$$

This embedding can be visualized in Figure 1. In practice the relationship does not have to be linear, the relationship can further be approximated by more neural network layers, as in the case of our model. A visualization of the network can be seen in Figure 2.

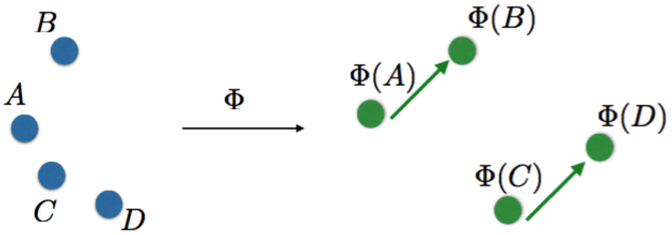

Figure 1: By learning an embedding operator, Φ, we are able to linearize the analogy "A is to B as C is to D"

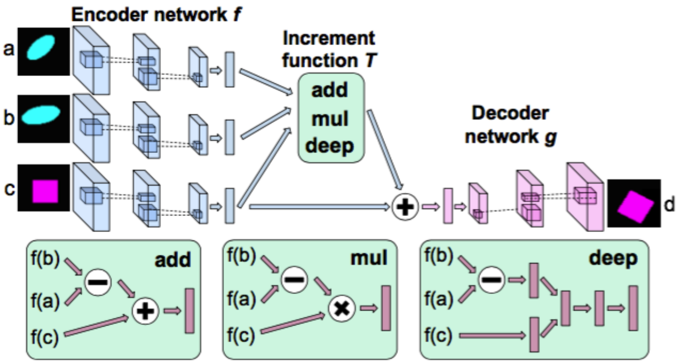

Figure 2: A visualization of the Convolutional Neural Network used in the Visual Analogy Network

Here our objective function is:

$$E = \sum_{(a,b,c,d)} \frac{1}{2} ||d - g(\Phi(b) - \Phi(a) + \Phi(c)||^2$$

2.3 Generative Adversarial Networks

Generative adversarial networks (GANs) [1] are a recent neural network architecture that allow for very good generative models. These networks have been used in the image domain to create very convincing images of a variety of objects [**?**]. The basic idea is to use one neural network that is a generator and use another neural network as a discriminator. The networks are adversarial in the sense that the generative model is trying to imitate the distribution of some true distribution, e.g. images, while the discriminative network is trying to classify images as coming from the true distribuitoin or the generative, fake, distribution. This is articulated in Figure 3.

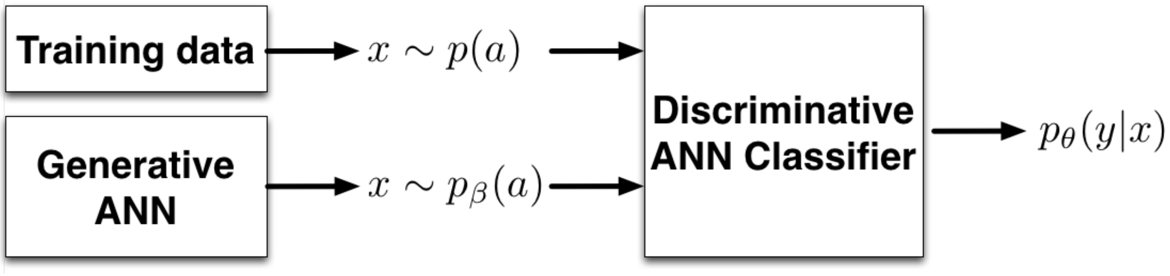

Figure 3: A visualization of the the Generative Adversarial Network idea

The goal then is to solve a minimax problem:

$$\min_{\beta} \max_{\theta} \Big[\mathbb{E}_{x \sim p(a)} \log\, p_\theta(y = \text{‘real’}|x) + \mathbb{E}_{x \sim p_\beta(a)} \log\, p_\theta(y = \text{‘fake’}|x) \Big]$$

2.4 Our Model - Conditionals Generative Adversarial Networks

$$\min_{\beta} \max_{\theta} \Bigg[\sum_{(w,s)} \qquad \mathbb{E}_{x \sim p(a|W=w,S=s)} \log\, p_\theta(W = w, S = s|x) \; + \\ \mathbb{E}_{x \sim p_\beta(a|W=w,S=s)} \log\, p_\theta(w = \text{‘fake’}, s = \text{‘fake’}|x) \Bigg]$$

2.5 Results

2.6 Conclusion

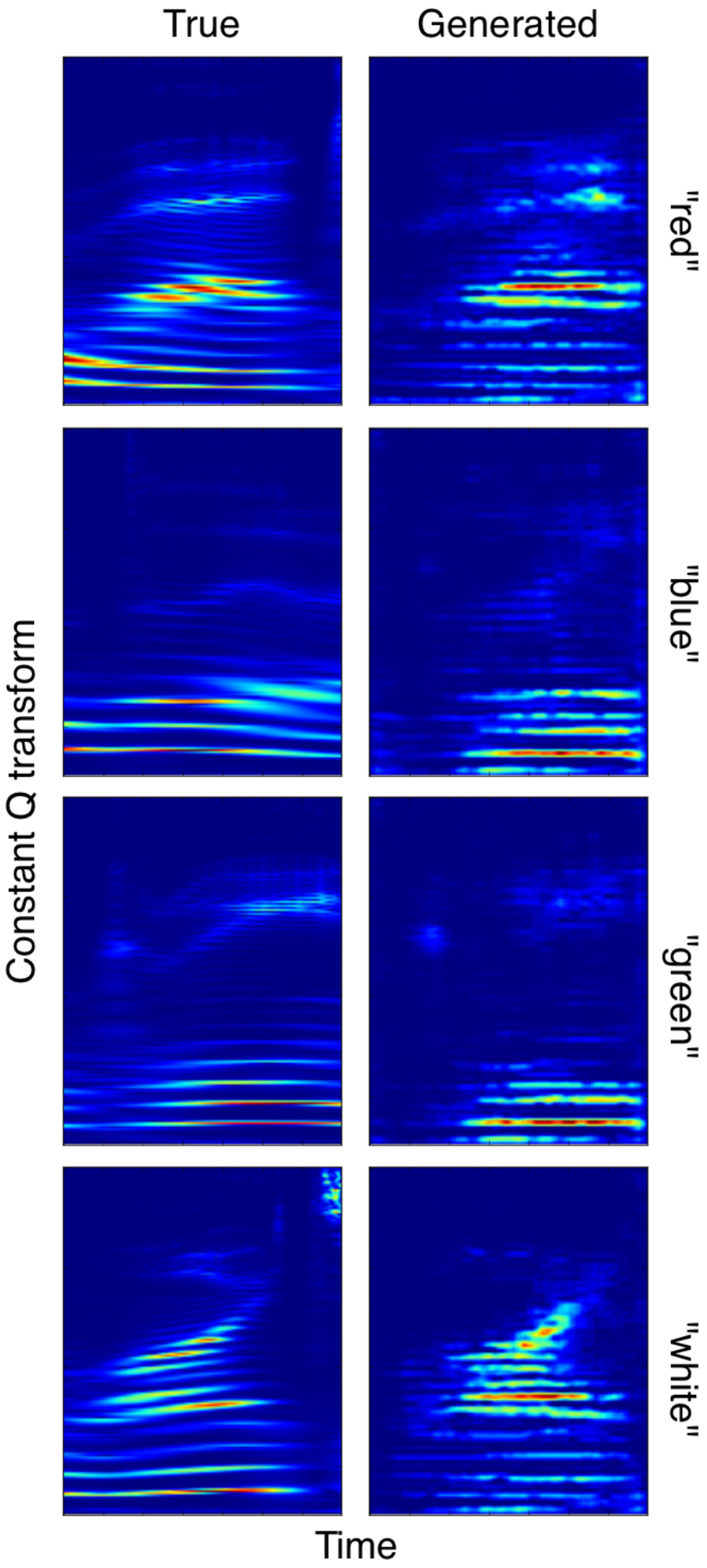

Figure 4: The left column corresponds to samples of Speaker 2 saying the color indicated on the row from the training data. The right column corresponds to generates samples from our model of the same speaker and color.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[2] Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. In *Advances in Neural Information Processing Systems*, pages 1252–1260, 2015.