COMPLEX-VALUED MUSICAL SOURCE LOCALIZATION

Andy M. Sarroff[1] & Michael Casey[1,2]
Departments of [1]Computer Science and [2]Music
Dartmouth College
Hanover, NH USA
sarroff@cs.dartmouth.edu

ABSTRACT

We suggest using complex-valued neural networks (CVNNs) for music information tasks such as source localization. We provide results on a blind localization task with binaural mixtures of two musical sources. All computations are performed directly on the time-domain audio waveform using the complex field, which we argue is more suitable than a real-valued representation for modeling fine-scale temporal structure such as phase.

1. INTRODUCTION

We consider the problem of localizing one or more sources in a mix directly from the stereo waveform using complex-valued neural networks (CVNNs). Source localization from mixed music signal is relevant to music information and discovery tasks in which we wish to blindly extract mixing parameters from produced musical recordings. This work has a few novel contributions. First, our computational model acts directly on the audio waveform. Most music signal processing and modeling techniques require that the signal is transformed to the frequency domain via the Short-Time Fourier Transform (STFT). We suggest that we can perform dimensionality reduction by first learning an alternative yet informative non-Fourier complex-valued transformation of the real-valued audio signal.

Second, all computations are performed on the complex field. Whilst CVNNs have been studied since the late 1980s, the literature on neural networks is heavily dominated by the signal processing community, especially with respect to nonlinear adaptive filters, as can be observed in several recent monographs [1, 4, 6, 7]. We suggest that CVNNs may be useful in music discovery tasks where the fine-scale temporal structure is impossible to model without considering signal phase. CVNNs have not been significantly investigated in the literature with respect to audio.

Finally we show how CVNNs may be used to infer the virtual azimuthal positions of one or more binaurally mixed musical sources. We experimentally show that CVNNs may achieve reasonably good performance at inferring source positions at virtual locations at 360 degrees around the head without using hand-chosen features or physical modeling. By contrast existing methods rely on psychoacoustic models or inverse head-related transforms (HRTFs).

2. BACKGROUND

Computational models for source localization usually attempt to estimate interaural time differences (ITD) and interaural level differences (ILD) from the STFT of the signal. For instance Raspaud, Viste, and Evangelista propose an algorithm that jointly estimates ILD and ITD from the STFT using parameters related to a physical model of a specific HRTF or average HRTF [8]. Mandel, Weiss, and Ellis provide an expectation maximization algorithm to estimate ITD and ILD [5]. Woodruff and Wang use Gaussian mixture models for modeling binaural cues and augment the model with monaural cues such as pitch tracking [11].

Neural networks are typically defined to perform their computations using the field of real numbers. Yet wave related signals such as musical audio have properties that naturally lend themselves to complex representations. The ITD and ILD cues that are explicitly modeled from the STFT are latent in the time domain signal. We suggest that complex-valued networks can (1) capture latent ITD and ILD cues directly from the waveform; and (2) do so with a lower-dimensional representation than is required with the STFT.

3. MODEL ARCHITECTURE AND TRAINING

We utilize the Wirtinger calculus [10] for back-propagating the gradient of a mean-squared error loss function. Let the target and inferred signals be denoted respectively as $\mathbf{y}, \hat{\mathbf{y}}$ and denote complex conjugation using the symbol $\bar{\cdot}$. The loss function is defined as:

$$\mathcal{L}(\mathbf{e}) = |\mathbf{e}|^2 = \mathbf{e}\bar{\mathbf{e}}\,, \tag{1}$$

where the error $\mathbf{e}$ is variably defined as:

$$\mathbf{e}^{\text{linear}} = \mathbf{y} - \hat{\mathbf{y}} \qquad \text{or} \tag{2}$$

$$\mathbf{e}^{\log} = \log\left[\frac{\mathbf{y}}{\hat{\mathbf{y}}}\right]\,. \tag{3}$$

© Andy M. Sarroff[1] & Michael Casey[1,2]. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).
Attribution: Andy M. Sarroff[1] & Michael Casey[1,2]. "Complex-Valued Musical Source Localization", Extended abstracts for the Late-Breaking Demo Session of the 17th International Society for Music Information Retrieval Conference, 2016.

We first construct a linear complex-valued autoencoder to learn a low-dimensional embedding of the training dataset. The autoencoder learns to approximately reproduce its input at its output by optimizing Equation 2, which has been a standard method for optimizing complex adaptive filters since reported in [3].

The activations of the autoencoder's hidden layer are provided as the input to a nonlinear recurrent inference network. The network has a hidden layer utilizing a logarithmic activation; a hidden linear recurrent layer; and a normalization operator at the last layer placing outputs on the unit circle. Savitha, Suresh, Sundararajan, and Saratchandran claim that the error function shown in Equation 2 may over-penalize magnitude error while under-penalizing phase error [9] and we therefore use their error function (Equation 3).

4. EXPERIMENT

We trained and tested our architecture using one or two musical sources convolved with HRTFs [1] representing azimuthal angles placed 360 degree around the head. In all cases we used only zero-elevation transfer functions.

The MedleyDB dataset [2] consists of raw tracks from multitrack recordings across a variety of produced music mixes. We selected two raw audio tracks of a Vivaldi excerpt, sampled at 44100 Hz. The single-source experiment used the double bass track and the two-source experiment additionally used the cello track. The audio was convolved with a subset of the HRTFs. The single-source experiment processed the double bass at azimuthal positions located at every 15 degrees, resulting in 24 stereo convolutions. For the two-source experiment we convolved all combinations of the sources located at 30-degree intervals, resulting in 144 spatial combinations of two sources. The raw audio was normalized before convolution but no additional preprocessing was done.

The convolved stereo waveforms were broken into short sequences of rectangular overlapping windows. This resulted in 11256 sequences of data for the one-source experiment and 67536 sequences for the two-source experiment. Taking the sample rate into account, the models received approximately 58 ms of audio context per inference. The corpus of stereo sequences were randomly partitioned so that 50% was used for training, 25% was used for validation, and 25% was used for testing.

All training was performed on a GPU using a custom fork of the Chainer library. [2] We have modified the library to perform complex-valued gradient descent using the Wirtinger calculus. Both stages of learning were performed with stochastic mini-batch gradient descent. The best performing hyper-parameters were selected according to the validation set. The model was trained until no performance increase was observed over 20 epochs.

The global mean angular error was 37.243 and 60.142 degrees for one and two sources respectively. The dataset consists of a significant number of frames with low average energy. The model's performance is improved greatly by filtering out low-energy frames, yielding 26.946 and 46.563 mean angular error.

The accuracy of inference varies with azimuthal angle. The single-source model has the best accuracy for sources that are located in the front center (11.876 average error) and performs the poorest for sources located at the rear. The two-source model has more difficulty locating sources that are placed near the center.

5. CONCLUSION

We present a model for binaural source localization using complex-valued neural networks. Our approach is data-driven, without any explicit modeling of head-related transfer functions or physical characteristics of the human auditory system. We experimentally showed that such models can reasonably perform azimuthal inference directly from the stereo waveform on a mixture of one or two sources. There are many future directions to to pursue. We intend to perform a closer examination of binaural source localization with respect to more musical sources and different timbres of sources. We are investigating different schemes for improving the rate of convergence for CVNNs, including composing layers with mixtures of heterogenous activation functions. We also intend to examine whether a similar architecture is useful for source separation. The literature on CVNNs and musical audio is sparsely populated and we hope that others find useful contributions in this work toward complex-domain modeling of musical signal.

6. REFERENCES

[1] Igor N. Aizenberg. *Complex-Valued Neural Networks with Multi-Valued Neurons*, volume 353 of *Studies in Computational Intelligence*. Springer, 2011.

[2] Rachel M. Bittner, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam, and Juan Pablo Bello. Medleydb: A multitrack dataset for annotation-intensive MIR research. In Hsin-Min Wang, Yi-Hsuan Yang, and Jin Ha Lee, editors, *Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR 2014, Taipei, Taiwan, October 27-31, 2014*, pages 155–160, 2014.

[3] D.H. Brandwood. A complex gradient operator and its application in adaptive array theory. *Communications, Radar and Signal Processing, IEE Proceedings F*, 130(1):11–16, February 1983.

[4] A. Hirose. *Complex-Valued Neural Networks: Advances and Applications*. John Wiley & Sons, Inc., 2013.

[5] Michael I. Mandel, Ron J. Weiss, and Daniel P. W. Ellis. Model-based expectation maximization source separation and localization. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):382–394, February 2010.

[1] Retrieved online from `http://recherche.ircam.fr/equipes/salles/listen/download.html`.

[2] `http://chainer.org/`.

[6] Danilo P Mandic and Vanessa Su Lee Goh. *Complex Valued Nonlinear Adaptive Filters: Noncircularity, Widely Linear and Neural Models*. Wiley, Chichester, U.K., 2009.

[7] Tohru Nitta and Tohru Nitta. *Complex-valued Neural Networks: Utilizing High-dimensional Parameters*. Information Science Reference - Imprint of: IGI Publishing, Hershey, PA, 2009.

[8] Martin Raspaud, Harald Viste, and Gianpaolo Evangelista. Binaural source localization by joint estimation of ILD and ITD. *IEEE Trans. Audio, Speech & Language Processing*, 18(1):68–77, 2010.

[9] Ramaswamy Savitha, Sundaram Suresh, N. Sundararajan, and P. Saratchandran. A new learning algorithm with logarithmic performance index for complex-valued neural networks. *Neurocomputing*, 72(16-18):3771–3781, 2009.

[10] W. Wirtinger. Zur formalen Theorie der Funktionen von mehr komplexen Veränderlichen. *Mathematische Annalen*, 97(1):357–375, 1927.

[11] John Woodruff and DeLiang Wang. Binaural localization of multiple sources in reverberant and noisy environments. *IEEE Trans. Audio, Speech & Language Processing*, 20(5):1503–1512, 2012.