Do WaveNets Dream of Acoustic Waves?

Kanru Hua

University of Illinois, U.S.A.

khua5@illinois.edu

Abstract

Various sources have reported the WaveNet deep learning architecture being able to generate high-quality speech, but to our knowledge there haven't been studies on the interpretation or visualization of trained WaveNets. This study investigates the possibility that WaveNet understands speech by unsupervisedly learning an acoustically meaningful latent representation of the speech signals in its receptive field; we also attempt to interpret the mechanism by which the feature extraction is performed. Suggested by singular value decomposition and linear regression analysis on the activations and known acoustic features (e.g. F0), the key findings are (1) activations in the higher layers are highly correlated with spectral features; (2) WaveNet explicitly performs pitch extraction despite being trained to directly predict the next audio sample and (3) for the said feature analysis to take place, the latent signal representation is converted back and forth between baseband and wideband components.

Index Terms: WaveNet, interpretation, visualization, feature extraction

1. Introduction

Until fairly recently, researches on the digital synthesis of speech signals have mainly been focusing on the manipulation of manually-engineered audio and speech features. With the advent of deep neural network (DNN) architectures directly modeling the waveform [1, 2], it is now possible to construct systems automating a significant portion of the signal processing tasks in speech analysis-synthesis.

In particular, the WaveNet architecture [1] and several recently proposed WaveNet-derived convolutional networks [3, 4] have demonstrated near-natural quality ratings in single-speaker speech synthesis. However, the high computational demand for training and evaluating the said networks remains a limiting factor on the applications. In addition, attempts on certain tasks such as multi-speaker speech synthesis [3] and denoising [5] still show opportunities for enhancements. It is desirable to gain an understanding of the learned models to set forth possible topological changes and other improvements on the networks as well as training techniques.

It is well-known that the weights in first layer of a trained convolution neural network often reveal interpretable features [6]. For the visualization of features in the subsequent layers, various approaches have been proposed to find a representative input that maximizes the activation of a particular hidden unit by means of gradient descent or deconvolutional networks [7, 8]. A similar class of methods that relates the learned network to features in the input space is to assign a local measure of how much the variation of the input affects the final output, along the input dimensions for a given output [9]. Despite the reported successes of pattern discovery in the weight or input space on image recognition/classification models, it is not clear on how to meaningfully apply such methods to time-series pre-diction, especially when a visualization in the frequency domain is desired for its relevance to human perception.

Another related area is the interpretation of recurrent neural networks (RNN), since the dilated module outputs in a WaveNet are also time series, similar to the hidden states in a RNN. Empirical studies in the said area have mostly focused on the task of language modeling: Strobelt, et al. [10] designed a tool for inspecting reoccuring hidden state patterns associated with a subsequence from the inputs; Karpathy, et al. [11] tested the hypothesis that the hidden cells contain interpretable information by color-coding the input text according to a sequence of activation values. However, our initial attempt to apply such pattern discovery methods on WaveNet was hindered by the large amount of data generated by the network due to the sampling rate and number of layers much greater than a typical RNN setup for natural language processing tasks.

This study attempts to answer the following questions: whether or not a WaveNet model trained on speech data *understands* the signal by performing feature extraction, and how is feature extraction done inside the trained network. To circumvent the difficulties listed above, we hypothesize that WaveNet is able to *unsupervisedly* learn features (e.g. fundamental frequency and band energy) amenable to human understanding. This hypothesis is verified by training linear mappings from the activations of a layer to features extracted from the input signals. This paper starts with a brief review of the original WaveNet architecture in Section 2. The experiment setup is described in Section 3 and the results are discussed in Section 4. We conclude the study by summarizing the findings and implications for future works in Section 5.

2. WaveNet

WaveNet [1] is an autoregressive discrete-time signal modeling network that predicts the next sample by conditioning on a certain number of previously generated samples. To synthesize speech from text inputs, the prediction is conditioned on an extra set of variables containing the linguistic context. Figure 1 shows the model structure implementing the said computation. The main element for computation is a set of dilated convolution layers, each of which defines the following operation,

$$\mathbf{y}[t] = \sum_{n=0}^{N} W_n \mathbf{x}[t - nk] \qquad (1)$$

where $\mathbf{x}$ and $\mathbf{y}$ are the input and the output sequences of vectors; W is the tensor for weights; k is the dilation factor and N is the convolution size. In the original WaveNet setup, N is set to 1 for every layer and k grows as an integer power of 2 so that the topology becomes effectively a binary tree. The same block consisting of dilated convolution layers of exponentially growing sizes is repeatedly cascaded for a few times, reaching a combined receptive field covering a few hundred milliseconds.
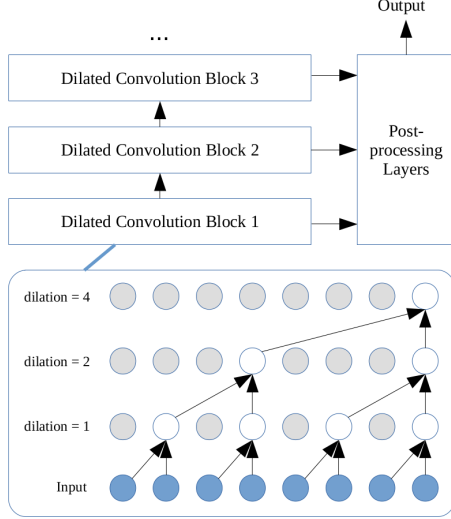
Figure 1: *A simplified overview of the WaveNet architecture. The bottom panel outlines the data flow in the first dilated convolution block in each time step.*

Each layer also implements an activation gating operation combining the results from two dilated convolutions and a residual connection. The input to the next layer can be computed as,

$$\mathbf{y} = \mathbf{x} + \tanh\left(W_f * \mathbf{x} + V_f \mathbf{c}\right) \sigma \left(W_g * \mathbf{x} + V_g \mathbf{c}\right) \quad (2)$$

where W_f and W_g are the filtering and gating tensors. Optionally, the tensors V_f and V_g can incorporate an additional input sequence $\mathbf{c}$ that carries the linguistic information.

To estimate a conditional distribution on the waveform amplitude, the speech data is quantized to 8 bits and the output layer is formulated as a softmax function over 256 categories. The outputs from the convolution layers are combined by a post-processing block consisting of a few fully-connected layers. Quantization artifacts can be reduced by a non-linear quantization scheme [12].

3. Experiment Setup

We hypothesize that at each time step, interpretable speech features can be derived from linear combinations of the outputs of a dilated convolution layer, in the same way that matrix multiplication is performed in a convolutional neural network. The relation between the latent representation formed in a WaveNet and speech features can be tested by regression analysis with a properly chosen objective function.

3.1. WaveNet Models

To eliminate correlations caused by the features' dependency on linguistic inputs, this study specifically investigates unconditioned WaveNet models, i.e., the case when $\mathbf{c} = 0$ in (2). Speaker-dependent models with identical configurations (Table 1) are trained on four speakers (2 male and 2 female) from the CMU Arctic speech database [13].

For each speaker, the trained WaveNet is evaluated on the test set and the activations from all hidden layers are stored for regression analysis; regression models are trained on 60 sentences from the test set and evaluated on the rest 20 sentences. Unless otherwise noted, the regression linearly projects a 64-dimensional activation vector (according to the layer size) with an extra bias term to the target feature space.

Table 1: *Model and Training Configuration*

Sample rate	16 kHz
Layer size	64 units
Dilation in each block	$1, 2, 4, 8, 16, 32, 64, 128, 256, 512$
Number of blocks	5 (total receptive field of 320 ms)
Optimization	Adam, learning rate $= 0.001$
Training set	1052 sentences, around 1 hour
Test set	80 sentences

3.2. Features

This study considers the following types of speech features: speech waveform (prior to 8-bit quantization), fundamental frequency (F0), band energy (dB), and STFT magnitude spectrogram (both wideband and narrowband).

The list includes both rapidly varying features (waveform and wideband spectra) and quasi-stationary features (F0, band energy and narrowband spectra). Details on computing the target features are listed as the follows.

3.2.1. Details on Feature Extraction

For the reconstruction of waveform, it is important to consider both input and target signals, which differ by an 1-sample offset and would respectively indicate the representation's capability at linearizing the discrete data and predicting the next sample.

Log fundamental frequency is extracted from the input data using adYANGsaf [14]. Our preliminary experiments showed that warping the frequency in linear or log scale does not significantly affect the results.

Without assuming that WaveNet learns an energy representation on an auditory frequency scale, we extract band energy from the input signal using a 2nd-order Butterworth filterbank with 20 channels uniformly spanned from 0 to 8 kHz. Preliminary experiments showed that a higher Spearman's correlation coefficient is achieved using a log-scale energy representation.

Finally, we attempt to reconstruct STFT magnitude spectra from activation vectors at the output end of each layer. Wideband spectra are taken using a 4 ms Blackman window and the window length for narrowband spectra is 32 ms.

3.2.2. Objectives for Regression

Ordinary least squares (OLS) regression is applied to reconstruct waveform, log F0, and log band energy features. Following a recent study on directly modeling magnitude spectrum using DNNs [15], we formulate the problem of magnitude spectra estimation as minimizing the KL divergence. However in the case of this study, Itakura-Saito distance based regression was found to better discriminate voiced and unvoiced speech. The said regression can be easily done in 2 Newton iterations, initialized by an OLS regression on log magnitude spectra.

4. Findings and Discussions

A theoretical limit on the accuracy of feature reconstruction can be derived by comparing the receptive field at a given layer to the minimal duration required to characterize a certain feature. The result is a useful reference in the following analysis when characterizing the change in accuracy across layers. For example, a reliable estimation of the fundamental frequency requires at least two periods of time-domain signal, which correspond to around 160 samples for female voices at a 16 kHz sample rate. So it is expected that the characterization of F0 will appear no
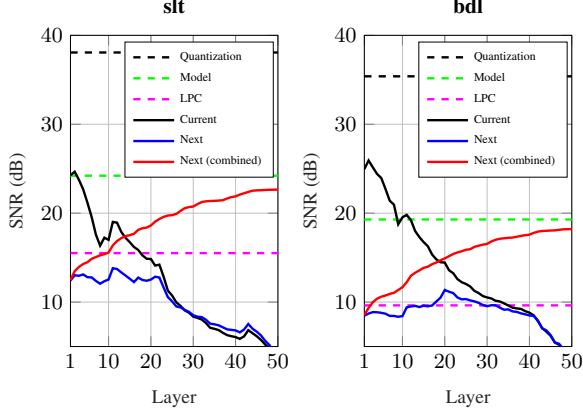
Figure 2: *Regression errors (in signal-to-noise ratios) of waveforms linear-transformed from WaveNet activations. Dashed lines represent reference levels; solid lines represent the regression results. Although the layer-wise SNR ("Current" and "Next") decreases, the multi-layer prediction ("Next combined") becomes more accurate as more layers are involved.*

Figure 3: *Correlation between activation-derived F0 and reference F0. Plateaus are observed with increasing correlation.*

Figure 4: *Correlation between activation-derived band energy and reference energy for speaker slt (female) and bdl (male).*

earlier than layer 7 ($\log_2 160 \approx 7.32$). Similar deductions can be made on other features.

Due to limited space, some figures shown in this section only include speaker *slt* (female) and *bdl* (male). The author has verified that similar trends can be observed on the rest of the speakers being tested.

4.1. Waveform Regression from Hidden Layers

Figure 2 shows the signal-to-noise ratio, measured in decibels, of both current-sample and next-sample reconstructions from the activations. For predicting the next sample, we also tested larger regression models with all the activations below a certain layer as the input ("combined"). The SNR due to quantization effects, the final output from the model, and a 512-order linear predictive filter representing the best linear autoregressive model are plotted for reference.

It is seen that, with the exception of the first few layers, the single-layer waveform prediction/reconstruction achieves a SNR only comparable to high-order linear prediction and the performance decreases in upper layers. However, within a mere 2 dB margin, the combined prediction based on activations from all layers approaches the model's final prediction. Such a trend suggests the possibility that the dilated convolution layers transform the raw signal into a more abstract representation and complementary information is derived at different layers. Further evidences can be gathered in the following analysis on slowly varying features.

4.2. F0 Regression from Hidden Layers

In agreement with the theoretical limit derived in the beginning of this section, a significant increase in Spearman's correlation coefficient on log F0 can be observed for all speakers near layer 7 shown in Figure 3. Notably, the first plateau is reached one layer later in the case of speaker *jmk*, who has the lowest average F0. The consistent trend that a high correlation can be reached at a layer index close to the theoretical limit implies the importance of pitch estimation to feature analysis within the WaveNet model.

The correlation coefficient peaks around 0.9 after reaching a second plateau starting between layer 15 and 17, which belong to the se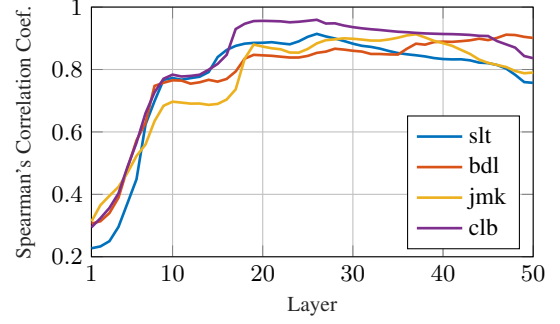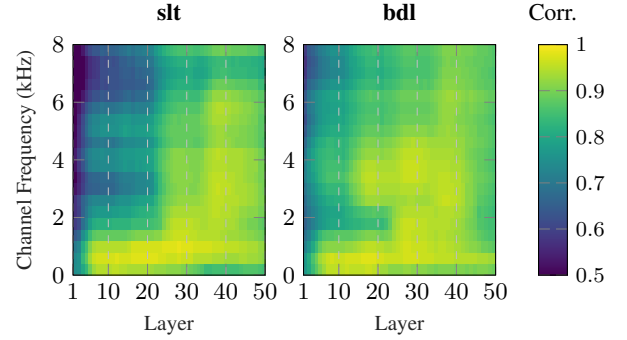cond convolution block (see Figure 1 for a graphical illustration). Considering that WaveNet maintains a reasonably good reconstruction of the input waveform in the second convolution block (Figure 2), it can be inferred that a refined F0 analysis is performed in the second half of the block. This claim is also supported by the fact that the dilation factors in layers between 15 and 17 correspond to the 125 - 500 Hz frequency range, which relates to the first one or two speech harmonics.

4.3. Band Energy Regression from Hidden Layers

Similar to the experiment on log F0, the quality of band energy (in dB) reconstructed from the activations is evaluated in a similar way but repeated for different frequency bands. With the exception of the first few layers below the receptive field limit, high correlation (greater than 0.9) is consistently observed at all layers for frequencies below 1 kHz. The pattern above 1 kHz is more complicated with a low initial correlation and multiple plateaus spanning across convolution block boundaries and rises are often observed in the middle of a block (e.g. layer 15, 25, and 35.) At this point it is unclear why the contents in mid-high frequency, covering most of the speech formants and spectral cues important to consonant perception [16] are not identified until a much later stage in the network.

4.4. Magnitude Spectra Regression from Hidden Layers

Frequency-layer plots on the task of STFT magnitude spectrum reconstruction closely resemble the results shown in Figure 4 and are thus not included in this paper. Examples of STFT spectrograms and their approximated versions are given in Figure 5. Although the spectrum tends to be over-smoothed at higher frequencies, the reconstruction derived from a linear transform of activations preserves the first 5 or 6 harmonics in the narrowband case and some excitation structures can be seen in the
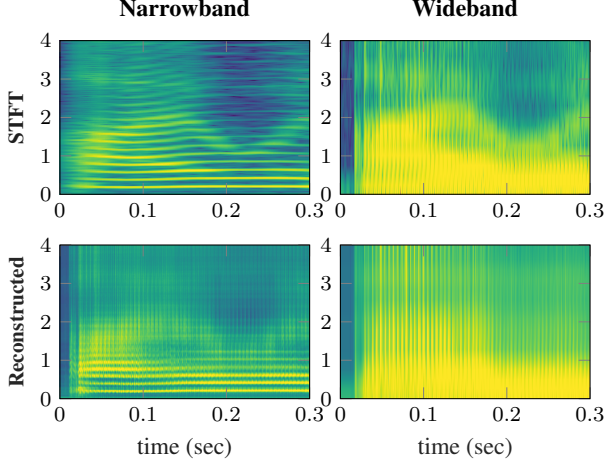
Figure 5: *Examples of spectrograms and their activation-derived versions on speaker slt (female), using outputs from layer 25. Spectral-temporal details are selectively preserved.*

wideband case. Note that both spectrograms are derived from the same set of 64 units in layer 25.

Manual inspection shows that the regression weights mapping activation vectors to log magnitude bins, after proper reordering, visually resemble the harmonic basis for discrete cosine transform. It is thus intriguing to think of the activations as a cepstral representation of speech. However, by taking the Fourier transform of regression weights, we find that the basis is not harmonic. There is yet no evidence supporting the existence of an assignment of interpretable meaning to each of the units.

4.5. Discussions and Further Analysis

Experiments in the previous sections have shown that the activations in each dilated convolution layer encode information about both baseband (i.e. slowly varying) and wideband (i.e. rapidly varying) features. The ability to separate the two components via a linear combination learned by regression implies a certain degree of degeneracy in the matrix of activations.

Previously singular value decomposition (SVD) has been employed to explore the redundancy in learned weights [17]; in this study, we apply SVD on low-pass filtered and high-pass filtered activations, and visualize the sorted singular values to characterize the resource allocation in the activation space. The activations are processed by a 2nd-order Linkwitz-Riley crossover filter [18] with an 80 Hz cutoff frequency and examples for baseband/wideband singular values taken from one of the speakers are shown in Figure 6.

It is immediately observable that a complementary and periodic (along layers) pattern exists across baseband and wideband singular values. The tails of baseband singular values are typically longer near the boundary of a convolution block (layer 10, 20, 30, 40 and 50) while the tails of wideband singular values are longer, though less apparently, near the middle of a convolution block (layer 5, 15, 25, and 35).

The said phenomenon encourages a theory, furthering the claim made in section 4.2, that WaveNet performs analysis on wideband activation components (e.g. reconstructed waveform) in the middle portion of each convolution block and stores the results into the baseband at the end of the block. To verify this theory, we look for evidences that WaveNet *explicitly* computes F0 from the activations in the middle layers. This is done by evaluating the terms inside the non-linear functions in (2) and
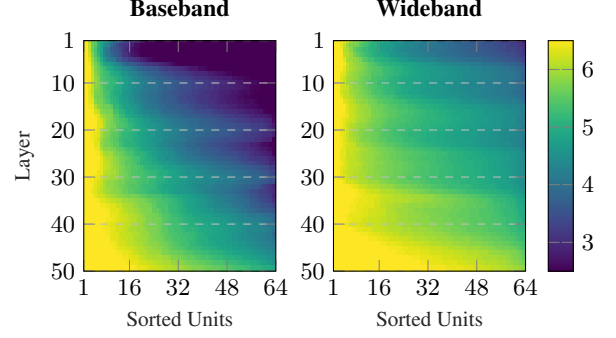

Figure 6: *Sorted log singular values of activation matrices on speaker slt (female) reveal a periodic structure across layers. The pattern is replicable on other speakers as well.*
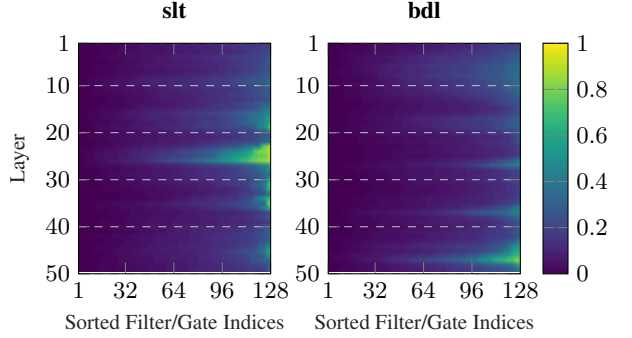

Figure 7: *One-to-one correlation between filter/gating results and F0 on speaker slt (female) and bdl (male).*

computing the Spearman's correlation coefficient with log F0 extracted from the input speech. Interestingly, the results in Figure 7 are in well-agreement with the theory and a maximal Spearman's correlation coefficient above 0.8 is observed for both speakers, despite that the WaveNet model is trained in an unsupervised setup.

5. Conclusions

We have evaluated the capability of unsupervisedly trained WaveNets to encode speech features in the activations. The findings point to a compact representation relating to spectral-temporal details of the speech signal, formed by means of iteratively analyzing the rapidly changing signals and storing the results in a temporally smooth component. This iterative behavior is encouraged by the stacking of convolution blocks with increasing dilation factors.

Although introducing architectural changes with the goal of quality improvement is beyond the scope of this study, the findings suggest several possibilities worth exploring in future studies. It is tempting to augment the dilated convolution with recurrent low-pass filters to better separate baseband features from the noisy activations, reducing the (frequency-wise) degeneracy of activation matrices. In addition, we expect an increase in the frequency resolution of reconstructed spectrograms (Figure 5) if the layer size can be larger.

6. Acknowledgments

7. References

[1] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1702.07825*, 2017.

[2] K. Tokuda and H. Zen, "Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis," in *Proc. ICASSP*, 2015, pp. 4215–4219.

[3] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, J. Raiman, S. Sengupta, and M. Shoeybi, "Deep Voice 2: Multi-speaker neural text-to-speech," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 2962–2970.

[4] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skeery-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," *arXiv preprint arXiv:1712.05884*, 2017.

[5] D. Rethage, J. Pons, and X. Serra, "A Wavenet for speech denoising," *arXiv preprint arXiv:1706.07162*, 2017.

[6] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to hand-written zip code recognition," *Neural Computation 1*, no. 4, pp. 541–551, 1989.

[7] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," *ECCV 2014 Lecture Notes in Computer Science*, 2014, pp. 818–833.

[8] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.R. Müller, "Evaluating the Visualization of What a Deep Neural Network Has Learned," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 11, pp. 2660–2673, 2017.

[9] G. Montavon, W. Samek, K. R. M uller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, Feb. 2018.

[10] H. Strobelt, S. Gehrmann, H. Pfister, and A. M. Rush, "LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 667–676, 2018.

[11] A. Karpathy, J. Johnson, and F. Li, "Visualizing and understanding recurrent networks," *arXiv preprint arXiv:1506.02078*, 2015.

[12] ITU-T, Rec. G.711, "Pulse Code Modulation (PCM) of voice frequencies," International Telecommunication Union-Telecommunication Standardisation Sector, 1988.

[13] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, Pittsburgh, 2004.

[14] K. Hua, "Improving YANGsaf F0 estimator with adaptive Kalman filter," in *Interspeech*, Stockholm, 2017.

[15] S. Takaki, H. Kameoka, and J. Yamagishi, "Direct modeling of frequency spectra and waveform generation based on phase recovery for DNN-based speech synthesis," in *Interspeech*, Stockholm, 2017.

[16] F. Li, A. Menon, and J. B. Allen, "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," *The Journal of the Acoustical Society of America*, vol. 127, no. 4, pp. 2599–2610, 2010.

[17] J. Xue, J. Li, and Y. Gong, "Restructuring of deep neural network acoustic models with singular value decomposition," in *Interspeech*, Lyon, 2013.

[18] S. H. Linkwitz, "Active crossover networks for noncoincident drivers," *Journal of the Audio Engineering Society*, vol. 24, no. 1, pp. 2–8, 1976.

[19] N. Adiga, V. Tsiaras, and Y. Stylianou, "On the use of WaveNet as a Statistical Vocoder," in *ICASSP*, 2018.