COMPARING SIGNAL ESTIMATION TECHNIQUES FROM MAGNITUDE-ONLY SPECTROGRAMS

Michele Perrone, Paolo Sani

Department of Electronics, Information Technology and Bio-engineering (DEIB), Politecnico di Milano
Piazza Leonardo Da Vinci 32, 20122 Milan, Italy
[michele.perrone,paolo1.sani]@mail.polimi.it

ABSTRACT

Is it possible to reconstruct an audio waveform from its magnitude-only spectrogram? It is a known fact that phase information plays a crucial role from the perceptual standpoint. In this paper we present a comparison between different techniques designated for the phase reconstruction problem. The techniques include both hand-crafted and data-driven methods, spanning from the Griffin-Lim algorithm to Generative Adversarial Networks. To the best of our knowledge, this is the first comparison carried out across multiple datasets including speech, music and urban sounds. The reconstructions are evaluated with relative, absolute, and perceptual metrics. For speech signals, Generative Adversarial Networks provide authentic results, but they have a strong dependency from the audio genre they are trained upon. Other techniques perform better on music and urban sounds, while having slight artifacts on speech signals.

Index Terms**—** Phase reconstruction, vocoder, spectrogram, deep neural networks, iSTFT, digital signal processing

1. INTRODUCTION

Problem formulation In the field of Digital Signal Processing (DSP), it is common practice to manipulate signals in the frequency domain. For this task, the Short-Time Fourier Transform (STFT) is often used. The STFT is obtained by computing the Fourier Transform over short segments of time of a signal. This makes it possible to deal with spectral properties that evolve over time. Given how the Fourier Transform is defined, the result of the STFT is obtained in the form of complex numbers in geometric representation, i.e. $x + i * j$, where z is the complex number, x is the real part, j is the imaginary part, and i is the complex unit. However, signal processing tasks often take advantage of the polar form, i.e. $z = re^{i\phi}$, where where r is the absolute value of z, and ϕ is the argument of z. In the DSP field, the absolute value r is frequently called magnitude, and the argument ϕ is called phase. Usually, the magnitude spectrogram contains the majority of the information of the original signal. For this reason, many signal processing algorithms are designed to take advantage of this fact by discarding the phase and reducing their overall computational cost. This is particularly valid for applications where the original signal does not need to be reconstructed, e.g. acoustic scene classification systems. However, applications where the reconstruction is needed, e.g. noise reduction systems, have to deal with the absence of phase information in order to avoid unpleasant artifacts [1]. In order to address this problem, many techniques have been developed during the years.

State of the art In [2], the authors provide a comprehensive summary of phase reconstruction methods. The authors identify the use of prior knowledge of target signals as a key difference between methods. Deep Neural Networks (DNNs) are capable of automatically discovering the structure of signals in the training dataset and utilize the obtained knowledge for phase reconstruction. For this reason, DNNs have been intensively studied in the most recent years. An interesting subdivision between DNN-based methods is presented in [3]. Here, the phase reconstruction algorithms are classified in four main classes: traditional signal processing methods such as the Griffin-Lim algorithm (GLA), Autoregressive DNNs, Non-Autoregressive DNNs, and Generative Adversarial Networks (GANs).

The GLA [4] is probably the most widespread of the pure signal processing methods. Though very efficient and lightweight, the main issue with these type of algorithms is the introduction of different degrees of artifacts. Autoregressive models such as WaveNet [5] are capable of generating high quality speech and music samples at the cost of slow computation. This is due to their nature, as audio samples are generated sequentially. This fact makes Autoregressive models unsuitable for real-time applications. This last problem is partially addressed by Non-Autoregressive models, which can be take advantage of parallel computation exploit modern Deep Learning hardware. Among these, WaveGlow [6] is one of the most prominent examples. WaveGlow is a flow-based generative model based on Glow [7]. It produces high-quality audio compared to WaveNet. However, it requires many parameters for its deep architecture with over 90 layers. UniGlow[1] and SqueezeWave [8] are lighter Non-Autoregressive models based on WaveGlow. A particular kind of Non-Autoregressive models are GANs, which are one of the most predominant deep generative models. MelGAN [3], Hifi-GAN [9] and UnivNet [10] are some examples of these architectures. These networks have a more lightweight and efficient inference compared to most Autogressive and Non-Autoregressive models, but their adversarial structure makes their training significantly more costly. Finally, Deep Griffin-Lim [2] can be reported as an example of hybrid iterative-DNN approach, combining GLA-inspired layers and a trainable DNN.

Motivation To the best of our knowledge, there are no publications detailing in-depth comparisons between different techniques. For this reason, we propose a comparison between several DNN methods as well as the popular GLA. Furthermore, the performance of the methods is evaluated on three different datasets representing speech, music and urban sounds in order to test their generalization capabilities.

Paper structure The paper is structured as follows. Section 2 provides a formulation of the problem and a brief introduction of

[1] See https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/tts_uniglow

the pipeline used for the comparison. Section 3 goes into the details of the pipeline and describes the experimental setup. Section 4 provides the results of the comparison, and section 5 presents the conclusions and future developments.

2. METHOD

The testing pipeline is structured as follows. For each dataset, we select a number of random audio samples, for which the spectrogram is computed and stored. Every technique is then tasked with the reconstruction the audio signal, starting from a version of the spectrogram where the phase information is absent. The reconstructions obtained by the various techniques are stored and are passed on to the metrics. The metrics then evaluate all the signals, and their output is stored in a data structure. The data structure is then used to carry out comparisons across datasets, techniques and metrics. The following section goes into the details of this pipeline by listing and commenting on all techniques, datasets and metrics used for the comparison.

3. EXPERIMENTAL SETUP

The tested algorithms are mostly DNNs, with the exception of GLA. We create a common pipeline, to initially extract the magnitude spectrograms required as input by the algorithms as well as a phase deprived signal, for comparison purposes. All the networks require a Mel-Spectrogram, except for GLA and Deep Griffin-Lim (DeepGL), which require a linear spectrogram. Furthermore, DeepGL computes the spectrogram after adding a uniform noise to the input audio signal. Once the spectrograms are obtained, each model performs its computation and returns the reconstructed audio waveform. MelGAN, HiFi-GAN, UniGlow, SqueezeWave and UnivNet are part of the nVidia Neural Modules toolkit (NeMo)[2] and are trained exclusively with speech data. Since there are no readily available trained models for DeepGL, we decided to train two versions of DeepGL, the first one with speech signals only, and a second one with a combination of speech and music signals. From now on, we will refer to the former as DeepGL-biased, and to the latter as DeepGL-unbiased.

3.1. Dataset

We use three different datasets, the main one being LJSpeech [11], which consists in 13,100 short audio clips of a single speaker reading passages from 7 non-fiction books. This is the same training dataset that the NeMo models are trained upon. GTZAN [12] is a very representative music dataset, as it consists on a collection of 10 music genres, each with 100 audio files. Lastly, we included UrbanSound8k [13], which is a collection of 8,732 urban sounds from 10 different classes. For testing, we consider a subset of 20 different audio files for each dataset. In the case of GTZAN and UrbanSound8k, the files are equally sampled between classes.

3.2. Techniques

Griffin-Lim As previously mentioned, GLA is a traditional signal processing method. The first step of the algorithm is the creation of a complex matrix, where the magnitudes are the values of the magnitude-only spectrogram, and the phase is initialized with uniform random noise. The iSTFT of this matrix is then computed, and a time series is obtained. The next step consists in performing the STFT of the time series and comparing it with the initial input STFT. The comparison is performed with a metric. The phase is then modified according to the computed metric. This process is repeated iteratively until reaching convergence.

Deep Griffin-Lim Similarly to GLA, this method is based on iterative repetition of a block. This block is composed by a trainable DNN part and a non-trainable GLA-inspired part. The main advantage of this approach is that during the training phase, only one block has to be trained. This same block is then repeated for an arbitrary number of times in the inference phase. For our tests, we set the number repetitions to 70. Another important characteristic of DeepGL is that the input spectrogram is computed upon the original signal with the addition of uniform noise. The authors of [2] claim that training and testing in a noise reduction condition suits well the phase reconstruction problem.

UniGlow UniGlow is a simplified vocoder based on WaveGlow, which is an Non-Autoregressive model. With respect to WaveGlow, the parameter reduction amounts to $12\times$.

SqueezeWave SqueezeWave is also a modified version of WaveGlow. It is the lightest DNN model tested here and can perform real-time computations on edge devices.

MelGAN, Hifi-GAN, UnivNet These three networks are all GANs. Starting from MelGAN, which is the first GAN vocoder ever introduced, Hifi-GAN improves the results with the use of two Discriminators: a multi-scale and a multi-period Discriminator. This way, the Generator network is forced to learn periodic patterns that are present in signals. UnivNet improves previous work with the use of a Discriminator that employs spectrograms of multiple resolutions as the input.

3.3. Metrics

Performance evaluations are based on different metrics and conditions: in particular, we perform the evaluation with comparisons of raw audio audio waveforms as well as phase, perceptual quality comparisons, and an absolute perceptual quality evaluation. The root mean square error (RMSE) and normalized root mean square error (NRMSE) are used for waveform and phase comparison. The perceptual evaluation of speech quality (PESQ) and perceptual evaluation of audio quality (PEAQ) are used for a relative perceptual comparison. Mean Opinion Score Net (MOSNet) is used as an absolute perceptual quality assessment.

RMSE The RMSE is calculated with the following equation:

$$\text{RMSE} = \sqrt{\frac{1}{K} \cdot \sum_{i=1}^{K} (y - \hat{y})^2} \tag{1}$$

where K is the length of the audio excerpt expressed in samples, y is the original audio, and $\hat{y}$ is the reconstructed audio. Since all audio clips are normalized in the range $[-1, 1]$, the RMSE can take values from 0 (no error) to 1 (maximum error).

NRMSE The NRMSE is computed as follows:

$$\text{NRMSE} = \frac{\sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} (y - \hat{y})^2}}{N} \tag{2}$$

where $N = \max(y) - \min(y)$ is the normalization factor.

[2] See https://catalog.ngc.nvidia.com/orgs/nvidia/containers/nemo

PESQ The perceptual evaluation of speech quality is a methodology standardized by the International Telecommunication Union (ITU) under recommendation ITU-T P.862[3] with the goal of testing the quality of speech signals. It has been developed by taking into account the physiology of human hearing, and it outputs values that follow the mean opinion score (MOS) scale, ranging from 1 (bad) to 5 (excellent).

PEAQ The perceptual evaluation of audio quality is also standardized by the ITU and published as recommendation ITU-R BS.1387[4]. Similarly to PESQ, it uses a model of human hearing in order to provide an estimate of the degradation of an audio signal in comparison to the reference signal. However, PEAQ is not focused on speech, since it has been developed to evaluate the degradation of generic audio transmissions. In our tests, we use the implementation presented in [14]. PEAQ outputs its results in terms of the Objective Difference Grade (ODG), which ranges from -4 (impairment very annoying) to 0 (impairment imperceptible).

MOSNet MOSNet [15] is an audio quality assessment model based on deep learning. Specifically, its architecture is that of a Convolutional and Recurrent Neural Network. It is trained to predict human ratings of converted speech and returns values that range from 1 (worst perceived quality) to 5 (best perceived quality). While MOSNet cannot replace a quality assessment given by humans, the results presented in [15] show a fair correlation between the mean opinion scores of human listeners and those predicted by the proposed model.

4. RESULTS

This section presents the results that we consider most relevant for the comparison. In addition to these, the reader can listen to the audio excerpts and corresponding reconstructions from a dedicated web page[5]. The page also includes additional plots that are not featured in this section.

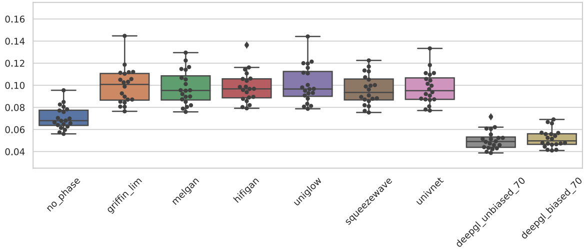

Figure 1: Speech waveform reconstruction evaluation with **RMSE**

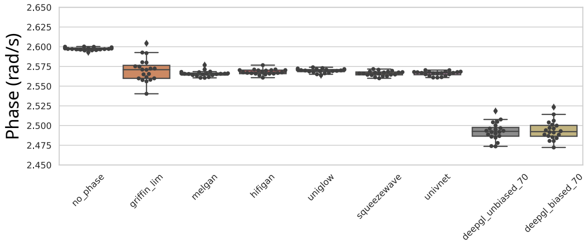

Figure 2: Speech phase reconstruction evaluation with **RMSE**

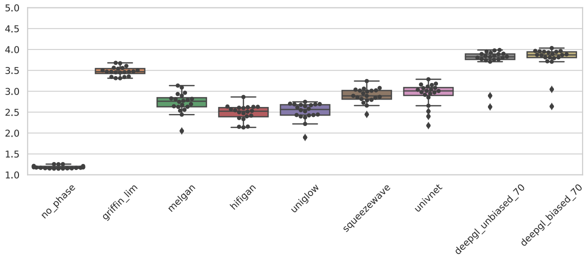

Figure 3: Speech waveform reconstruction evaluation with **PESQ**

Speech We first consider the waveform RMSE in Fig. 1 (lower values are better). It can be seen that the lowest error value is obtained by the unbiased DeepGL, followed closely by its biased version. Interestingly enough, the signal without phase information scores better than all remaining techniques, which leads us to the conclusion that this metric cannot be fully trusted in the context of phase reconstruction. A better idea is given by Fig. 2, which shows the RMSE calculated between the original signal phase and reconstructed signal phase. As expected, in this case the signal without phase has the maximum error. DeepGL remains the best performing technique, while the results of the other methods remain too close to each other to assess which performs better. The results of PESQ in Fig. 3 are in line with those of 2 for the DeepGL and no-phase signals, but this time a clear differentiation can be observed within the remaining techniques. It is interesting to lead that according to PESQ, Griffin-Lim takes the lead on all deep learning techniques except for DeepGL. The evaluation given by MOSNET in Fig. 4 does not seem to be conclusive, except for the lowest score that is attributed to the no-phase signals. However, the difference with the other techniques is not so well marked. This may be due to the fact that the MOSNET model has not been trained with audio degradation that include phase artifacts. We choose not to include the NRMSE plots, both for the waveform and the phase, because the normalization does not seem to bring advantages. This is probably due to the fact that the ranges of the audio files does not vary significantly across the excerpts. It is interesting to notice that the metrics results do not resemble very well the listening test on the reconstructed audios conducted by the authors. GANs seem to obtain the cleanest result, while DeepGL presents some clearly audible artifacts.

Music We observe from Fig. 5 that the waveform RMSE is lowest for DeepGL, regardless of the biased or unbiased model. Similarly to the case of speech, there is not a very noticeable distinction between the other techniques with this metric. This changes when we look at the phase RMSE in Fig. 6. As expected, the no-phase version has the highest error, while DeepGL has a clear advantage over all techniques. This is confirmed by PESQ in Fig. 7 and PEAQ in Fig. 8. The lowest scores are given to MelGAN, HiFiGAN, UniGlow, SqueezeWave, and UnivNet. The reason for such a poor performance is that these models have been trained exclusively on speech. It is interesting to note that DeepGL-biased does not seem to suffer from this problem. This is probably due to the fact that DeepGL is not purely a DNN model, but it rather combines blocks without prior knowledge of the target signal with a trainable DNN block.

[3] See `https://www.itu.int/rec/T-REC-P.862`

[4] See `https://www.itu.int/rec/R-REC-BS.1387`

[5] See `https://micheleperrone.it/spectrogramplayer`

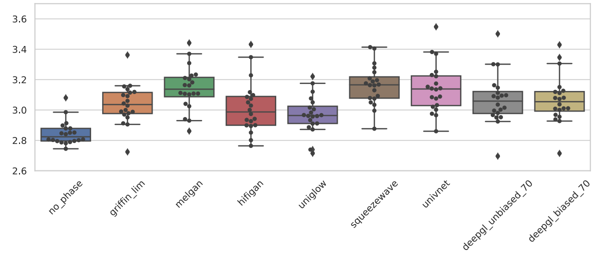

Figure 4: Speech waveform reconstruction evaluation with **MOSNET**

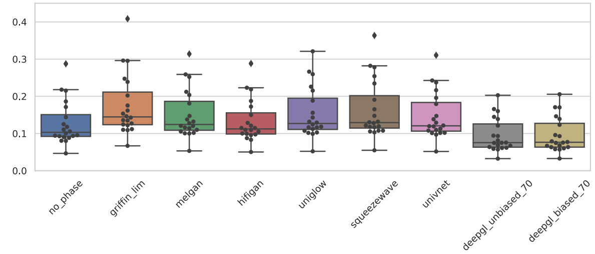

Figure 5: Music waveform reconstruction evaluation with **RMSE**

Urban sounds For the urban dataset, we decide to show only the phase RMSE plot in Fig. 9, which in our opinion gives a hint as to the perceptual quality obtained by the various techniques. Similarly to the case of music, MelGAN, HiFiGAN, UniGlow, SqueezeWave and UnivNet produce results that are perceptually very distant from the original audio clips. On the other hand, Griffin-Lim outputs convincing results. DeepGL is the model with best performance.

5. CONCLUSION

In this paper, we present a comparison between a selection of techniques for the reconstruction of phase information from magnitude-only spectrograms. The comparison is carried out with datasets comprising speech, music, and urban sounds. According to the metrics, the techniques with best performance are DeepGL and Griffin-Lim, regardless of the audio genre. However, in the case of speech signals, these results are not confirmed by the subjective listening tests of the authors. As a matter of fact, GANs, UnivNet and HiFi-GAN in particular, seem to provide the cleanest results. Apart from the case of speech, the additional computational cost of MelGAN, HiFi-GAN, UnivNet and WaveGlow does not seem justified with respect to the other approaches. SqueezeWave is the only lightweight technique based solely on DNNs, but its performance is always inferior to Griffin-Lim and DeepGL. Of all the techniques tested, Griffin-Lim can be considered as a good all-round solution, given its low computational cost and its good results across a wide range of audio signals. DeepGL is the best-sounding technique for music and urban sounds, but it comes at the additional cost of training the model on a large dataset; on the other hand, results show that training DeepGL on a specific audio genre does not impair significantly its performance when used on other types of audio signals.

Future work will include the comparison of DeepGL with GANs trained on a more heterogeneous dataset, comprising a wide variety of sounds from speech, music and acoustic scenes.

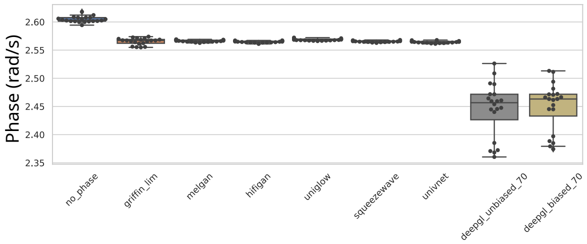

Figure 6: Music phase reconstruction evaluation with **RMSE**

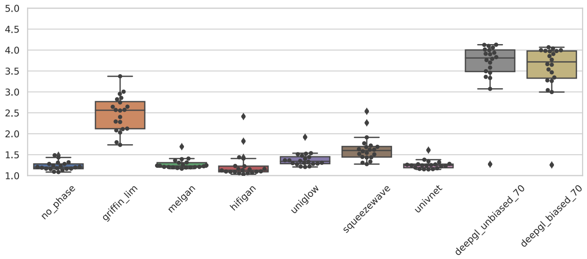

Figure 7: Music waveform reconstruction evaluation with **PESQ**

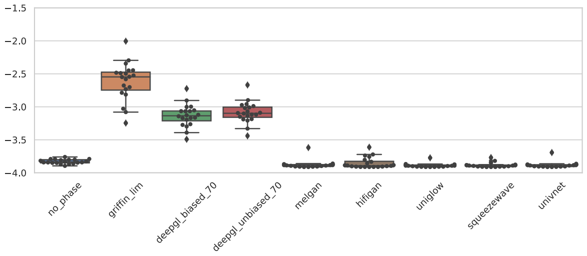

Figure 8: Music waveform reconstruction evaluation with **PEAQ**

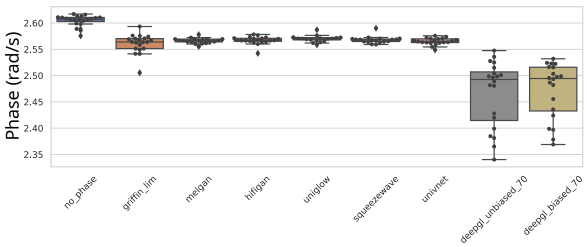

Figure 9: Urban phase reconstruction evaluation with **RMSE**

6. REFERENCES

[1] J. Laroche and M. Dolson, “Phase-vocoder: about this phasiness business,” in *Proceedings of 1997 Workshop on Applications of Signal Processing to Audio and Acoustics*, 1997, pp. 4 pp.–.

[2] Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, “Deep griffin–lim iteration: Trainable iterative phase reconstruction using neural network,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 1, pp. 37–50, 2021.

[3] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, “Melgan: Generative adversarial networks for conditional waveform synthesis,” in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/6804c9bca0a615bdb9374d00a9fcba59-Paper.pdf

[4] D. Griffin and J. Lim, “Signal estimation from modified short-time fourier transform,” in *ICASSP ’83. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 8, 1983, pp. 804–807.

[5] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.

[6] R. Prenger, R. Valle, and B. Catanzaro, “Waveglow: A flow-based generative network for speech synthesis,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[7] D. P. Kingma and P. Dhariwal, “Glow: Generative flow with invertible 1x1 convolutions,” *Advances in neural information processing systems*, vol. 31, 2018.

[8] B. Zhai, T. Gao, F. Xue, D. Rothchild, B. Wu, J. E. Gonzalez, and K. Keutzer, “Squeezewave: Extremely lightweight vocoders for on-device speech synthesis,” *arXiv preprint arXiv:2001.05685*, 2020.

[9] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 17 022–17 033. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/c5d736809766d46260d816d8dbc9eb44-Paper.pdf

[10] W. Jang, D. C. Y. Lim, J. Yoon, B. Kim, and J. Kim, “Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation,” in *Interspeech*, 2021.

[11] K. Ito and L. Johnson, “The lj speech dataset,” https://keithito.com/LJ-Speech-Dataset/, 2017.

[12] G. Tzanetakis, G. Essl, and P. Cook, “Automatic musical genre classification of audio signals,” 2001. [Online]. Available: http://ismir2001.ismir.net/pdf/tzanetakis.pdf

[13] J. Salamon, C. Jacoby, and J. P. Bello, “A dataset and taxonomy for urban sound research,” in *22nd ACM International Conference on Multimedia (ACM-MM’14)*, Orlando, FL, USA, Nov. 2014, pp. 1041–1044.

[14] M. Holters and U. Zölzer, “Gstpeaq – an open source implementation of the peaq algorithm,” in *Proc. of the 18th Int. Conference on Digital Audio Effects (DAFx-15)*, 12 2015. [Online]. Available: https://www.dafx.de/paper-archive/2015/DAFx-15_submission_12.pdf

[15] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, “MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion,” in *Proc. Interspeech 2019*, 2019, pp. 1541–1545.