INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Speech Denoising with Auditory Models

Mark R. Saddler[1,2,3,†], *Andrew Francl*[1,2,3,†], *Jenelle Feather*[1,2,3], *Kaizhi Qian*[4], *Yang Zhang*[4], *Josh H. McDermott*[1,2,3]

[1] Department of Brain and Cognitive Sciences, MIT, USA
[2] McGovern Institute for Brain Research, MIT, USA
[3] Center for Brains, Minds and Machines, MIT, USA
[4] MIT-IBM Watson AI Lab, IBM Research, USA

{msaddler, francl, jhm}@mit.edu

Abstract

Contemporary speech enhancement predominantly relies on audio transforms that are trained to reconstruct a clean speech waveform. The development of high-performing neural network sound recognition systems has raised the possibility of using deep feature representations as 'perceptual' losses with which to train denoising systems. We explored their utility by first training deep neural networks to classify either spoken words or environmental sounds from audio. We then trained an audio transform to map noisy speech to an audio waveform that minimized the difference in the deep feature representations between the output audio and the corresponding clean audio. The resulting transforms removed noise substantially better than baseline methods trained to reconstruct clean waveforms, and also outperformed previous methods using deep feature losses. However, a similar benefit was obtained simply by using losses derived from the filter bank inputs to the deep networks. The results show that deep features can guide speech enhancement, but suggest that they do not yet outperform simple alternatives that do not involve learned features.

Index Terms: speech enhancement, denoising, deep neural networks, cochlear model, perceptual metrics

1. Introduction

Recent advances in speech enhancement have been driven by neural network models trained to reconstruct speech sample-by-sample [1, 2, 3, 4, 5, 6, 7, 8]. These methods provide substantial benefits over previous approaches, but nonetheless leave room for improvement. The resulting processed speech usually contains audible artifacts, and noise removal is usually incomplete at lower SNRs.

A parallel line of work has explored the use of deep artificial neural networks as models of sensory systems [9, 10]. Although substantial discrepancies remain [11, 12], such trained neural networks currently provide the best predictive models of brain responses and behavior in both the visual and auditory systems [9, 13]. The apparent similarities between deep supervised feature representations and representations in the brain raises the possibility that such representations could be used as perceptual metrics. Such metrics have been successfully employed in image processing [14], but are not widely used in audio applications.

Deep feature losses for denoising were previously proposed in [15, 16, 17, 18, 19], but were explored only for relatively high signal-to-noise ratios (SNRs), a single task and network, or were not compared to baseline methods using the same transform architecture. Additionally, direct comparisons have not been made to simpler losses derived from conventional filter banks. It was thus unclear the extent to which deep feature losses could improve on simpler approaches, and what choices in the feature training would produce the best results. The goal of this paper was to directly compare deep perceptual losses to alternative losses, and to explore the conditions in which benefits might be achieved. We found that deep feature losses produced more natural denoising compared to waveform losses, but that a similar benefit could be achieved using a loss derived from standard filter bank representations.

2. Methods

There were two components to our denoising approach (Figure 1). The first component was a recognition network trained to recognize either speech or environmental sounds. Once trained, this network was used to define deep feature losses. Speech recognition is a natural choice in this context, but it also seemed plausible that more general-purpose audio features learned for environmental sound recognition might help to achieve natural-sounding audio even in speech applications. The input to the network was the output of a filter bank modeled on the human cochlea.

The second component was a waveform-to-waveform audio transform whose parameters were adjusted via gradient descent to minimize a loss function (evaluated on features of the recognition network, or the outputs of a filter bank, or on the waveform). We used a Wave-U-Net [20], which has been found to perform comparably to WaveNet [21] based on objective metrics of noise reduction, but which can be specified with many fewer parameters and run with a much lower memory footprint. Code, models, and audio examples are available at: http://mcdermottlab.mit.edu/denoising/demo.html.

2.1. Recognition Networks

The recognition networks took as input simulated cochlear representations of 2s sound clips (audio sampled at 20 kHz). The cochlear model consisted of a bank of 40 bandpass filters whose frequency tuning mimics that of the human ear (evenly spaced on an Equivalent Rectangular Bandwidth scale [22]), followed by half-wave rectification, downsampling to 10 kHz, and 0.3 power compression [23].

[†] equal contribution

http://dx.doi.org/10.21437/Interspeech.2021-1973

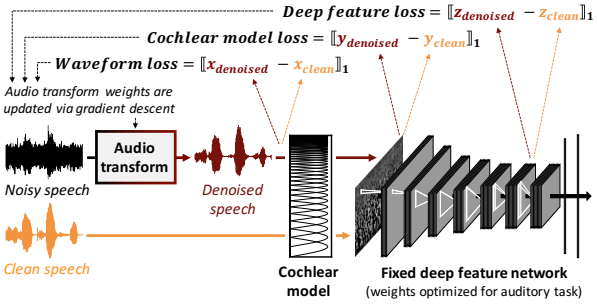

Figure 1: *Schematic of audio transform training.*

2.1.1. Recognition Network Architectures

We used three feed-forward CNN architectures for the recognition networks. Each consisted of stages of convolution, rectification, batch normalization, and weighted average pooling with a hanning kernel to minimize aliasing [24, 12]. The three architectures were selected based on word recognition task performance from 3097 randomly-generated architectures varying in number of convolutional layers (from 4 to 8), size and shape of convolutional kernels, and extent of pooling. The selected architectures had 6 (arch1) or 7 (arch2,3) convolutional layers.

2.1.2. Recognition Network Training

The recognition networks were trained to perform either word recognition or environmental sound recognition. For the speech task, each training example was a speech excerpt (from the Wall Street Journal [25] or Spoken Wikipedia Corpora [26]). The task was to recognize the word overlapping with the center of the clip [13, 12] (out of 793 word classes sourced from 432 unique speakers, with 230,357 unique clips in the training set and 40,651 segments in the validation set). For the environmental sound recognition task, each training example was a non-speech YouTube soundtrack excerpt (from a subset of 718,625 AudioSet examples [27]), and the task was to predict the AudioSet labels (spanning 516 categories in our dataset).

The three network architectures were trained on each task until performance on the validation set task plateaued. Word task classification accuracies for the three architectures were: arch1 $=$ 90.4%, arch2 $=$ 88.5%, and arch3 $=$ 80.6%. AudioSet task AUC values were: arch1 $=$ 0.845, arch2 $=$ 0.861, and arch3 $=$ 0.869.

2.2. Audio Transforms

2.2.1. Wave-U-Net Architecture

The Wave-U-Net architecture was the same as in [21]: 12 layers in the contracting path, a 1-layer bottleneck, and 12 layers in the expanding path. All layers utilized 1D convolutions with learned filters and LeakyReLU activation functions. There were 24 filters in the first layer, and the number of filters increased by a factor of 2 with each successive layer prior to the bottleneck.

2.2.2. Deep Feature Losses

The recognition networks were used to define a deep feature loss function as the $L1$ distance between network representations of noisy speech and clean speech. The total loss for a single recognition network and single training example was the sum of the $L1$ distances between the noisy speech and clean speech activations for each convolutional layer, weighted to approximately balance the contribution of each layer.

2.2.3. Cochlear Model Losses

We also trained transforms using losses derived from the cochlear model that provided input to the recognition network, as well as variants of the model that varied in i) the number of filters (5, 10, 20, 40, 80 and 160 filters, evenly spaced on an ERB-scale [22], with bandwidths scaled to tile the spectrum in all cases), ii) the dependence of filter bandwidth on frequency (linearly-spaced and 'reversed', with broad low-frequency filters and narrow high-frequency filters, opposite to what is found in the ear), and iii) in their phase invariance (subband envelopes computed by lowpass-filtering the rectified subbands; cutoff of 100 Hz).

2.2.4. Wave-U-Net Training

Out of concern that the audio transform might overfit to idiosyncrasies of any individual recognition network, we trained some transforms on losses computed simultaneously from an ensemble of three different networks (arch1,2,3), and some on just a single network (arch1).

In all cases the Wave-U-Net was trained on speech superimposed on non-speech AudioSet excerpts (the same corpora used to train the recognition networks) with SNR drawn uniformly from $[-20, +10]$ dB. AudioSet excerpts were used as the training 'noise' as they were highly varied and diverse. All Wave-U-Net models were trained with the ADAM optimizer for 600,000 steps (batch size=8, learning rate=10^{-4}).

2.2.5. Baselines

We used two baseline models, both trained to explicitly reconstruct clean speech waveforms from noisy speech waveforms drawn from the same training set described above. The first was a previously described WaveNet [5] and the second was the Wave-U-Net [21] used with the deep feature and filter losses.

We also compared our results to those of a previously published denoising transform trained with a deep feature loss [15], using both the pre-trained model made available by Germain et al. and a Wave-U-Net that we trained on our dataset using the feature loss from [15] (deep network features trained on the DCASE 2016 [28] environmental sound challenge).

2.3. Evaluation

We evaluated the trained models on 40 speech excerpts (from a separate validation set) superimposed separately on each of four types of noise signals: speech-shaped Gaussian noise, auditory scenes from the DCASE 2013 dataset [29], instrumental music from the Million Song Dataset [30], and 8-speaker babble made from public-domain audiobooks (librivox.org). These noise sources were chosen to be distinct from those in the training set, and to span a variety of noise types to assess the generality of the trained transforms.

2.3.1. Human Perceptual Evaluation and Objective Metrics

We evaluated the audio transforms by conducting perceptual experiments on Amazon Mechanical Turk. Participants first completed a screening task to help ensure that they were wearing headphones or earphones [31]. The participants who passed this screening task then rated the naturalness of a set of processed

Table 1: *Experiment 1 results. Reported metrics are averaged across the five tested SNR levels. Higher is better for all metrics.*

Model name	Loss function	Natural.	PESQ	STOI	SDR
Cochlear model (N=40; human)	40 ERB-spaced subbands	**4.43**	1.55	0.75	7.16
A123	AudioSet features (arch123)	**4.43**	1.66	0.77	4.06
A1+W1	AudioSet + Word features (arch1)	4.36	**1.68**	0.79	6.18
A123+W123	AudioSet + Word features (arch123)	4.33	1.67	0.77	4.18
A1	AudioSet features (arch1)	4.33	1.65	0.78	3.63
W123	Word features (archs123)	4.24	1.67	**0.79**	6.64
W1	Word features (arch1)	4.22	1.63	0.77	3.30
Random1	Random features (arch1)	3.91	1.57	0.78	5.64
Random123	Random features (arch123)	3.84	1.57	0.77	5.08
Germain DeepFeatures	DCASE features from [15]	3.83	1.47	0.77	6.72
Germain (pre-trained)	DCASE features from [15]	2.36	1.14	0.64	0.93
Waveform (Wave-U-Net)	Waveform	4.17	1.51	0.76	**7.35**
Waveform (WaveNet)	Waveform	3.72	1.40	0.75	6.00
Unprocessed input		2.67	1.15	0.70	0.21

speech signals, presented seven at a time in a MUSHRA-like paradigm. Listeners could listen to each clip as many times as they wished and then gave each a numerical rating on a scale of 1-7. Listeners were provided with anchors corresponding to the ends of the rating scale (1 and 7). The anchor at the high end was always the original clean speech. The low-end anchor was 4-bit-quantized speech (an example of very high distortion). To help ensure that participants were using the scale as instructed, each experiment included 3 catch trials where two of the stimuli were the two anchors. In order to be included in the analysis, participants had to rate all instances of the high and low anchors as 7 and 1, respectively.

We ran two identically structured experiments to evaluate all of our audio transforms. Experiment 1 compared various deep feature losses to baselines and contained all of the conditions listed in Table 1. Experiment 2 compared losses derived from different cochlear filter banks and contained all of the conditions listed in Table 2. 49 and 105 participants met the inclusion criteria for Experiments 1 and 2, respectively.

We also used three standard objective measures for evaluation: perceptual evaluation of speech quality (PESQ) [32], short-time objective intelligibility measure (STOI) [33], and the signal-to-distortion ratio (SDR) [34].

3. Results

3.1. Deep Feature Losses Yield Improved Denoising

The best-performing systems trained with deep perceptual feature losses outperformed both waveform-based baselines. The

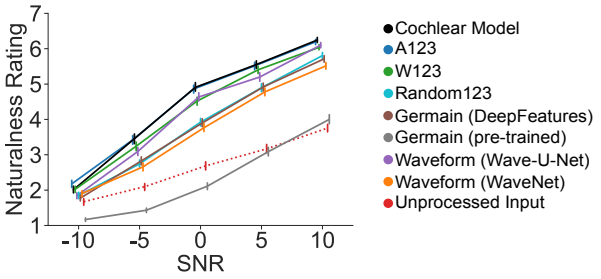

Figure 2: *Rated naturalness vs. SNR for speech processed by Wave-U-Nets trained on deep feature losses, in addition to baseline models trained to reconstruct clean speech waveforms, and two versions of a related prior method [15]. Error bars plot SEM (across 49 participants).*

average objective and subjective evaluation results are shown in Table 1. Human listeners found the speech processed by the deep feature models to be more natural than the speech processed by the baseline models. We plot the naturalness results in more detail (Figure 2) for two of the best-performing models trained on each of AudioSet features (A123) and word recognition features (W123), as well as a model trained on random features (Random123), the two baselines, and the two versions of the denoising network from [15].

3.2. Learned vs. Random Deep Features

The benefit of deep feature losses was specific to models trained with learned features. Audio transforms trained to reconstruct random features did not produce better naturalism than the baseline WaveNet, and performed worse overall than the baseline Wave-U-Net (Figure 2; Table 1).

3.3. Comparison to Previous Deep Feature Systems

Our best-performing deep feature-based systems also outperformed previously published systems with deep feature losses. The pre-trained system from Germain et al. [15] generalized poorly to our test set. Furthermore, the Wave-U-Net we trained using the deep feature loss from [15] also performed worse than the baseline Wave-U-Net. These findings suggest that the features used for the perceptual loss are important, and that the DCASE task used in [15] may not have produced sufficiently general features.

3.4. Effect of Task Used to Train Deep Features

The best results occurred for features trained on the environmental sound recognition task – naturalism was consistently higher than for features trained on word recognition (Figure 2; Table 1). However, all of the models trained with feature losses from our recognition networks produced more natural-sounding speech than the baselines, and than the systems trained with DCASE features based on [15]. There was no obvious benefit from training on features from three different networks.

3.5. Cochlear Model Losses Match Deep Feature Losses

Although deep features produced better performance than baselines trained using waveform losses, we found that we could reproduce their benefit using losses derived from the cochlear model inputs to the recognition networks. Based on rated naturalness, the transform trained with this 'cochlear' loss per-

Table 2: *Experiment 2 results. Reported metrics are averaged across the five tested SNR levels. Higher is better for all metrics.*

Model name	Loss function	Natural.	PESQ	STOI	SDR
Cochlear model (N=20)	20 ERB-spaced subbands	**4.33**	1.54	0.77	**7.61**
Cochlear model (N=40; human)	40 ERB-spaced subbands	4.30	1.55	0.75	7.16
Cochlear model (N=160)	160 ERB-spaced subbands	4.26	1.60	0.77	7.51
Cochlear model (N=10)	10 ERB-spaced subbands	4.22	1.49	0.76	7.08
Cochlear model (N=80)	80 ERB-spaced subbands	4.21	1.53	0.74	6.69
Cochlear model (N=5)	5 ERB-spaced subbands	3.93	1.42	0.75	6.02
Cochlear model (N=40; linear)	40 linearly-spaced subbands	4.32	1.51	0.76	6.82
Cochlear model (N=40; env.)	Envelopes of 40 ERB subbands	4.16	1.59	0.75	6.94
Cochlear model (N=40; reverse)	40 reverse-ERB-spaced subbands	4.08	1.47	0.73	4.73
A123	AudioSet features (arch123)	4.27	**1.66**	**0.77**	4.06
Waveform (Wave-U-Net)	Waveform	4.17	1.51	0.76	7.35
Unprocessed input		2.47	1.15	0.70	0.21

formed just as well as our best model trained with deep feature losses (Table 1).

3.6. Effect of Filter Bank Characteristics

The benefit of the cochlear loss depended to some extent on the filter characteristics (Table 2; Figure 3, left). Worse performance was obtained with a 'reversed' filter bank, with wide filters at low frequencies and narrower filters at high frequencies, opposite to that of the ear. Using the envelope of the filter outputs also produced worse performance (counter to the hypothesis that phase invariance might be critical). However, filters that were linearly spaced along the frequency axis worked about as well as those modeled on the ear.

Worse performance was also obtained using only five filters (scaled to cover the frequency spectrum), but good results were obtained provided at least 10 filters were used (Figure 3, right). This result suggests that splitting the audio up into multiple frequency channels is sufficient to replicate the benefit of deep features provided there are enough channels with reasonably sensible frequency tuning.

3.7. Objective Metrics

The models trained on deep recognition features also performed better than the baselines according to PESQ and STOI. Notably, this advantage was not evident when measured with SDR. The filter bank-trained models showed the opposite trend – better performance as measured by SDR, and worse via PESQ and STOI (Table 2). These differences suggest that the filter bank and deep feature losses are not fully interchangeable despite having similar effects on overall naturalness. The results also underscore the limitations of objective metrics for capturing human perception of altered speech.

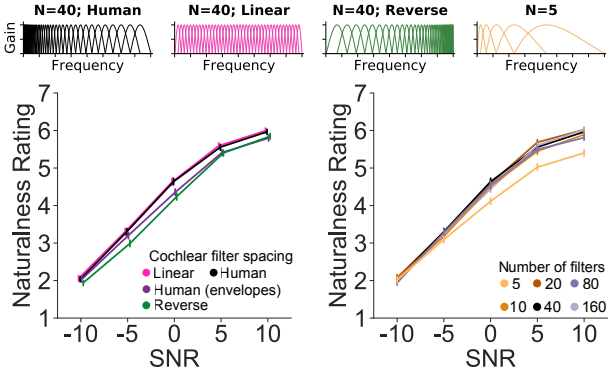

Figure 3: *Rated naturalness vs. SNR for speech processed by Wave-U-Nets trained on cochlear model losses with different filter banks (select examples depicted above). Error bars plot SEM (across 105 participants).*

4. Discussion

Prior work has proposed denoising based on deep feature losses [15, 16, 17, 18, 19], but has not evaluated it relative to methods using simpler waveform- or subband-based losses. We found that deep recognition features could be used to train denoising systems that outperform waveform-based methods, but that their benefit could be matched using a standard one-layer auditory filter bank. The results thus provide no evidence that deep features provide a unique benefit for denoising.

Although deep neural networks yield the best current models of biological sensory systems [9, 10], our results indicate that these similarities are not yet sufficient to produce audio enhancement algorithms above and beyond what can be obtained from simple filter bank models. However, it is possible that building better models of human perceptual systems will also yield feature losses [35, 36] that would better transfer their perceptual benefits to humans, and produce benefits relative to simpler approaches. It also remains possible that the audio quality is limited more by the audio transform than the feature loss. More expressive transforms, or transforms with stronger generative constraints, might yield a clearer benefit of deep features.

The benefits of deep feature and cochlear model losses relative to waveform-based losses were clear from the ratings of human listeners, but were less evident in the objective metrics we tested (PESQ, STOI, SDR). This result indicates that optimizing for auditory model-based losses may provide perceptual benefits that conventional objective metrics are poorly suited to measuring, and suggests to the potential value of auditory model features as new objective metrics.

In sum, we found that audio transforms trained to modify noisy speech so as to reconstruct deep feature representations of clean speech produce better denoising performance than transforms trained to reconstruct clean speech waveforms, as measured by the ratings of human listeners. However, a similar benefit was obtained using one-layer auditory filter banks, suggesting the importance of multi-channel, overcomplete representations rather than learned features per se.

5. Acknowledgements

The authors thank Ray Gonzalez for developing the training dataset, John Cohn and the Oak Ridge National Laboratory for use of Summit, and the MIT-IBM Watson AI Lab for funding.

6. References

[1] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," *arXiv preprint arXiv:1609.07132*, 2016.

[2] S. Pascual, A. Bonafonte, and J. Serra, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[3] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *arXiv preprint arXiv:1708.07524*, 2017.

[4] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florêncio, and M. Hasegawa-Johnson, "Speech enhancement using bayesian wavenet." in *Interspeech*, 2017, pp. 2013–2017.

[5] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5069–5073.

[6] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[7] A. Pandey and D. Wang, "A new framework for cnn-based speech enhancement in the time domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, 2019.

[8] T.-A. Hsieh, H.-M. Wang, X. Lu, and Y. Tsao, "WaveCRN: An efficient convolutional recurrent neural network for end-to-end speech enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 2149–2153, 2020. [Online]. Available: https://doi.org/10.1109/lsp.2020.3040693

[9] D. L. Yamins and J. J. DiCarlo, "Using goal-driven deep learning models to understand sensory cortex," *Nature neuroscience*, vol. 19, no. 3, p. 356, 2016.

[10] A. Kell and J. H. McDermott, "Deep neural network models of sensory systems: windows onto the role of task constraints," *Current Opinion in Neurobiology*, vol. 55, pp. 121–132, 2019.

[11] R. Rajalingham, E. B. Issa, P. Bashivan, K. Kar, K. Schmidt, and J. J. DiCarlo, "Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks," *Journal of Neuroscience*, vol. 38, no. 33, pp. 7255–7269, 2018.

[12] J. Feather, A. Durango, R. Gonzalez, and J. H. McDermott, "Metamers of neural networks reveal divergence from human perceptual systems," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[13] A. J. Kell, D. L. Yamins, E. N. Shook, S. V. Norman-Haignere, and J. H. McDermott, "A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy," *Neuron*, vol. 98, no. 3, pp. 630–644, 2018.

[14] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE CVPR 2018*, 2018, pp. 586–595.

[15] F. G. Germain, Q. Chen, and V. Koltun, "Speech Denoising with Deep Feature Losses," in *Proc. Interspeech 2019*, 2019, pp. 2723–2727. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1924

[16] R. Pilarczyk and W. Skarbek, "Multi-objective noisy-based deep feature loss for speech enhancement," in *Photonics Applications in Astronomy, Communications, Industry, and High-Energy Physics Experiments 2019*, R. S. Romaniuk and M. Linczuk, Eds., vol. 11176, International Society for Optics and Photonics. SPIE, 2019, pp. 858 – 865. [Online]. Available: https://doi.org/10.1117/12.2536967

[17] J. Su, Z. Jin, and A. Finkelstein, "Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," *arXiv preprint arXiv:2006.05694*, 2020.

[18] T.-A. Hsieh, C. Yu, S.-W. Fu, X. Lu, and Y. Tsao, "Improving perceptual quality by phone-fortified perceptual loss for speech enhancement," *arXiv preprint arXiv:2010.15174*, 2020.

[19] S. Kataria, J. Villalba, and N. Dehak, "Perceptual loss based speech denoising with an ensemble of audio pattern recognition and self-supervised models," *arXiv preprint arXiv:2010.11860*, 2020.

[20] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," *arXiv preprint arXiv:1806.03185*, 2018.

[21] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-u-net," *arXiv preprint arXiv:1811.11307*, 2018.

[22] B. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103–138, 1990.

[23] J. H. McDermott and E. P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis," *Neuron*, vol. 71, pp. 926–940, 2011.

[24] O. J. Hénaff and E. P. Simoncelli, "Geodesics of learned representations," *arXiv preprint arXiv:1511.06394*, 2015.

[25] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the Workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[26] A. Köhn, F. Stegen, and T. Baumann, "Mining the spoken wikipedia for speech data and beyond," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, 2016.

[27] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[28] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1128–1132.

[29] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: An ieee aasp challenge," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[30] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, 2011.

[31] K. J. Woods, M. H. Siegel, J. Traer, and J. H. McDermott, "Headphone screening to facilitate web-based auditory experiments," *Attention, Perception, and Psychophysics*, vol. 79, pp. 2064–2072, 2017.

[32] I.-T. Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[33] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[34] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, 2006.

[35] I. Ananthabhotla, S. Ewert, and J. A. Paradiso, "Towards a perceptual loss: Using a neural network codec approximation as a loss for generative audio models," in *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France*. ACM, 2019, pp. 1–8.

[36] P. Manocha, A. Finkelstein, Z. Jin, N. J. Bryan, R. Zhang, and G. J. Mysore, "A differentiable perceptual audio metric learned from just noticeable differences," *arXiv preprint arXiv:2001.04460*, 2020.