ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/343568932

A Perceptually-Motivated Approach for Low-Complexity, Real-Time Enhancement of Fullband Speech

Preprint · August 2020

DOI: 10.48550/arXiv.2008.04259

CITATIONS
0

READS
3,037

6 authors, including:

Jean-Marc Valin
Google Inc.
135 PUBLICATIONS **4,569** CITATIONS
SEE PROFILE

Neerad Phansalkar
Amazon
6 PUBLICATIONS **602** CITATIONS
SEE PROFILE

Ritwik Giri
Amazon Web Services
59 PUBLICATIONS **1,444** CITATIONS
SEE PROFILE

Karim Helwani
Technische Universität Berlin
53 PUBLICATIONS **588** CITATIONS
SEE PROFILE

All content following this page was uploaded by Jean-Marc Valin on 14 August 2020.

The user has requested enhancement of the downloaded file.

A Perceptually-Motivated Approach for Low-Complexity, Real-Time Enhancement of Fullband Speech

Jean-Marc Valin, Umut Isik, Neerad Phansalkar, Ritwik Giri, Karim Helwani, Arvindh Krishnaswamy

Amazon Web Services

{jmvalin, umutisik, neeradp, ritwikg, helwk, arvindhk}@amazon.com

Abstract

Over the past few years, speech enhancement methods based on deep learning have greatly surpassed traditional methods based on spectral subtraction and spectral estimation. Many of these new techniques operate directly in the the short-time Fourier transform (STFT) domain, resulting in a high computational complexity. In this work, we propose PercepNet, an efficient approach that relies on human perception of speech by focusing on the spectral envelope and on the periodicity of the speech. We demonstrate high-quality, real-time enhancement of fullband (48 kHz) speech with less than 5% of a CPU core.

Index Terms: speech enhancement, pitch filtering, postfilter

1. Introduction

Over the past few years, speech enhancement methods based on deep learning have greatly surpassed traditional methods based on spectral subtraction [1] and spectral estimation [2]. Many of these techniques operate directly on the short-time Fourier transform (STFT), estimating either magnitudes [3, 4, 5] or ideal ratio masks (IRM) [6, 7]. This typically requires a large number of neurons and weights, resulting in a high complexity. It also partly explains why many of those methods are restricted to 8 or 16 kHz. The use of the STFT also brings up a trade-off with the window length – long windows can cause musical noise and reverb-like effects, whereas short windows do not provide sufficient frequency resolution for removing noise between pitch harmonics. These problems can be mitigated by the use of complex ratio masks [8] or time-domain processing [9, 10, 11], at the cost of further increasing complexity.

We propose PercepNet, an efficient approach that relies heavily on human perception of speech signals and improves on RNNoise [12]. More precisely, we rely on the perception of audio in critical bands (Section 2) and on the perception of tones and noise (Section 3) with a new acausal comb filter. The deep neural network (DNN) model we use is trained using perceptual criteria (Section 4). We propose a novel envelope postfilter (Section 5) that further improves the enhanced signal.

The PercepNet algorithm operates on 10-ms frames with 40 ms of look-ahead and can enhance 48 kHz speech in real time using just 4.1% of an x86 CPU core. We show that its quality significantly exceeds that of RNNoise (Section 6).

2. Signal Model

Let $x\left(n\right)$ be a clean speech signal, the signal captured by a hands-free microphone in a noisy room is given by

$$y\left(n\right) = x\left(n\right) \star h\left(n\right) + \eta\left(n\right) \ , \tag{1}$$

where $\eta\left(n\right)$ is the additive noise from the room, $h\left(n\right)$ is the impulse response from the talker to the microphone, and $\star$ denotes the convolution. Furthermore, the clean speech can be expressed as $x\left(n\right) = p\left(n\right) + u\left(n\right)$, where $p\left(n\right)$ is a locally periodic component and $u\left(n\right)$ is a stochastic component (here we consider transients such as stops as part of the stochastic component). In this work, we attempt to compute an enhanced signal $\hat{x}\left(n\right) = \hat{p}\left(n\right) + \hat{u}\left(n\right)$ which is as perceptually close to the clean speech $x\left(n\right)$ as possible. Separating the stochastic component $u\left(n\right)$ from the environmental noise $\eta\left(n\right)$ is a very hard problem. Fortunately, we only need $\hat{u}\left(n\right)$ to *sound* like $u\left(n\right)$, which can be achieved by filtering the mixture $u\left(n\right) \star h\left(n\right) + \eta\left(n\right)$ to have the same spectral envelope as $u\left(n\right)$. Since $p\left(n\right)$ is periodic and the noise is assumed not to have strong periodicity, $\hat{p}\left(n\right)$ should be easier to estimate. Again, we mostly need $\hat{p}\left(n\right)$ to have the same spectral envelope and the same period as $p\left(n\right)$.

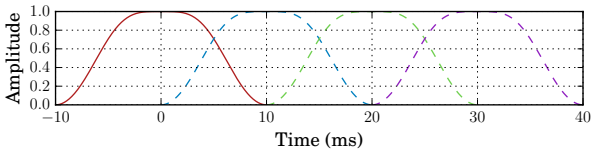

Figure 1: *The current window being synthesized is shown in solid red. We use three windows of look-ahead (shown in dashed lines) such that samples up to time $t = 40\,ms$ are used to compute the audio output up to $t = 0$.*

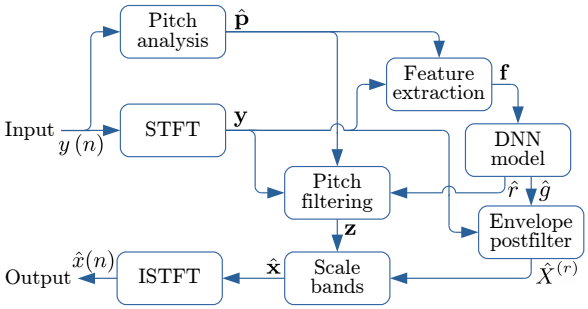

Figure 2: *Overview of the PercepNet algorithm.*

We seek to construct an enhanced signal with the same 1) spectral envelope, and 2) frequency-dependent periodic-to-stochastic ratio, as the clean signal. For both these properties, we use a resolution that matches human perception.

We use the short-time Fourier transform (STFT) with 20-ms windows and 50% overlap. We use the Vorbis window function [13] – which satisfies the Princen-Bradley perfect reconstruction criterion [14] – for analysis and synthesis, as shown in Fig. 1. An overview of the algorithm is shown in Fig. 2.

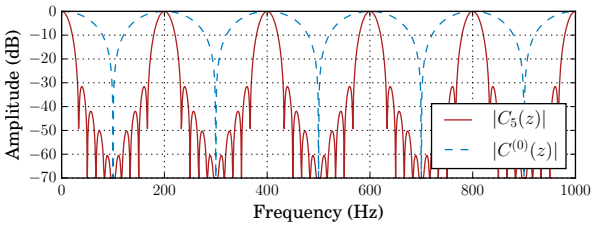

Figure 3: *Frequency response of the proposed comb filter (red) vs the filter used in [12] (blue) for a pitch of 200 Hz.*

2.1. Bands

The vast majority of noise signals have a wide bandwidth with a smooth spectrum. Similarly, both the periodic and the stochastic components of speech have a smooth spectral envelope. This allows us to represent their envelope from 0 to 20 kHz using 34 bands, spaced according to the human hearing equivalent rectangular bandwidth (ERB) [15]. To avoid bands with just one DFT bin, we impose a minimum band width of 100 Hz.

For each band of the enhanced signal to be perceptually close to the clean speech, both their total energy and their periodic content should be the same. In this paper, we denote the complex-valued spectrum of the signal $x\,(n)$ for band b in frame ℓ as $\mathbf{x}_b\,(\ell)$. We also denote the L_2-norm of that band as $X_b\,(\ell)$.

2.2. Gains

From the magnitude of the noisy speech signal in band b, we compute the ideal ratio mask, i.e. the gain that needs to be applied to $\mathbf{y}_b$ such that it has the same energy as $\mathbf{x}_b\,(\ell)$:

$$g_b\,(\ell) = \frac{X_b\,(\ell)}{Y_b\,(\ell)}\,. \tag{2}$$

In the case where the speech only has a stochastic component, applying the gain $g_b\,(\ell)$ to the magnitude spectrum in band b should result in an enhanced signal that is almost indistinguishable from the clean speech signal. On the other hand, when the speech is perfectly periodic, applying the gain $g_b\,(\ell)$ results in an enhanced signal that sounds *rougher* than the clean speech; even though the energy is the same, the enhanced signal is less harmonic than the clean speech. In that case, the noise is particularly perceptible due to the fact that tones have relatively little masking effect on noise [16]. In that situation, we use the comb filter described in the next section to remove the noise between the pitch harmonics and make the signal more periodic.

3. Pitch Filtering

To reconstruct the harmonic properties of the clean speech, we use comb filtering based on the pitch frequency. The comb filter can achieve much finer frequency resolution than would otherwise be possible with the STFT (50 Hz using 20-ms frames). We estimate the pitch period using a correlation-based method combined with a dynamic programming search [17].

3.1. Filter

For a voiced speech signal with period T, a simple comb filter

$$C^{(0)}(z) = \frac{1+z^{-T}}{2} \tag{3}$$

introduces zeros at regular interval between harmonics and attenuates the noisy part of the signal by around 3 dB. This provided a small, but noticeable quality improvement in [12]. In this work, we extend the comb filtering to more than one period, including non-causal taps using the following filter:

$$C_M(z) = \sum_{k=-M}^{M} w_k z^{-kT}\,, \tag{4}$$

where M is the number of periods on each side of the central tap and w_k is a window function satisfying $\sum_k w_k = 1$. Using $C_M\,(z)$, the noisy part of the signal is attenuated by $\sigma_w^2 = \sum_k w_k^2$. Although a rectangular window would minimize σ_w^2, we use a Hann window, which shapes the remaining noise to be lower between harmonics. Due to the behavior of tone masking [15], this results in a lower perceptual noise. For $M = 5$, we have $\sigma_w = -9\,\mathrm{dB}$ and the full response is shown in Fig. 3. In practice, since the maximum look-ahead is bounded, we truncate the window w_k to values of kT that are permitted.

The filtering occurs in the time domain, with the output denoted $\hat{p}\,(n)$ since it approximates the "perfect" periodic component $p\,(n)$ from the clean speech. Its STFT is denoted $\hat{\mathbf{p}}_b\,(\ell)$.

3.2. Filtering Strength

The amount of comb filtering is important: not enough filtering results in roughness, whereas too much results in a robotic voice. The strength of the comb filtering in [12] is controlled by a heuristic. In this work, we instead have the neural network learn the strength that best preserves the ratio of periodic to stochastic energy in each band. The equations below describe what that ideal strength should be. Since they rely on properties of the clean speech, they are only used at training time.

We define the pitch coherence $q_{x,b}\,(\ell)$ of the clean signal as the cosine distance between the complex spectra of the signal with its periodic component (both ℓ and b are omitted for clarity)

$$q_x \triangleq \frac{\Re\left[\mathbf{p}^{\mathrm{H}}\mathbf{x}\right]}{\|\mathbf{p}\|\cdot\|\mathbf{x}\|}\,, \tag{5}$$

where $\cdot^{\mathrm{H}}$ denotes the Hermitian transpose and $\Re\,[\cdot]$ denotes the real component. Similarly, we define q_y as the pitch coherence of the noisy signal. Since the ground truth $\mathbf{p}$ is not available, the coherence values need to be estimated. Considering that the noise in $\hat{\mathbf{p}}$ is attenuated by a factor σ_w^2, the pitch coherence of the estimated periodic signal $\hat{\mathbf{p}}$ itself can be approximated as

$$q_{\hat{p}} = \frac{q_y}{\sqrt{(1-\sigma_w^2)\,q_y^2 + \sigma_w^2}}\,. \tag{6}$$

We define the pitch filtering strength $r \in [0,\,1]$, where $r = 0$ causes no filtering to occur and $r = 1$ replaces the signal with $\hat{\mathbf{p}}$. Let $\mathbf{z} = (1-r)\,\mathbf{y} + r\hat{\mathbf{p}}$ be a pitch-enhanced signal, we want the pitch coherence of $\mathbf{z}$ to match the clean signal:

$$q_z = \frac{\mathbf{p}\cdot((1-r)\,\mathbf{y} + r\hat{\mathbf{p}})}{\|\mathbf{p}\|\cdot\|(1-r)\,\mathbf{y} + r\hat{\mathbf{p}}\|} = q_x\,. \tag{7}$$

Solving (7) for r results in

$$r = \frac{\alpha}{1+\alpha}\,, \tag{8}$$

$$\alpha = \frac{\sqrt{b^2 + a\left(q_x^2 - q_y^2\right)} - b}{a}\,, \tag{9}$$

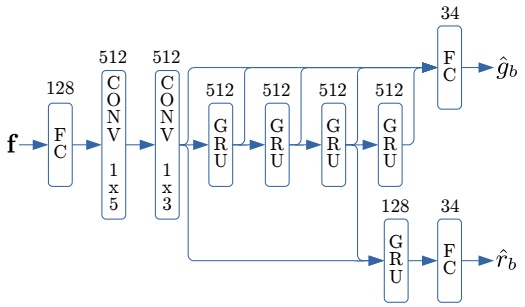

Figure 4: *Overview of the DNN architecture computing the 34 gains $\hat{g}_b$ and 34 strengths $\hat{r}_b$ from the 70-dimensional input feature vector* $\mathbf{f}$*. The number of units on each layer is indicated above the layer type.*

where $a = q_{\hat{p}}^2 - q_x^2$ and $b = q_{\hat{p}} q_y \left(1 - q_x^2\right)$.

In very noisy conditions, it is possible for the periodic estimate $\hat{\mathbf{p}}$ to have a lower coherence than the clean speech in a band ($q_{\hat{p}} < q_x$). In that case, we set $r = 1$ and compute a gain attenuation term that will ensure that the stochastic component of the enhanced speech matches the level of the clean speech (at the expense of making the periodic component too quiet)

$$g^{(\text{att})} = \sqrt{\frac{1 + n_0 - q_x^2}{1 + n_0 - q_{\hat{p}}^2}}, \tag{10}$$

where $n_0 = 0.03$ (or 15 dB) limits the maximum attenuation to the noise-masking-tone threshold [18]. For the normal case ($q_{\hat{p}} \geq q_x$), then $g^{(\text{att})} = 1$.

4. DNN Model

The model uses both convolutional layers (a 1x5 layer followed by a 1x3 layer), and GRU [19] layers, as shown in Fig. 4. The convolutional layers are aligned in time such as to use up to M frames into the future. To achieve 40 ms look-ahead including the 10-ms overlap, we use $M = 3$.

The input features used by the model are tied to the 34 ERB bands. For each band, we use two features: the magnitude of the band with look-ahead $Y_b\left(\ell + M\right)$ and the pitch coherence without look-ahead $q_{y,b}\left(\ell\right)$ (the coherence estimation itself uses the full look-ahead). In addition to those 68 band-related features, we use the pitch period $T\left(\ell\right)$, as well as an estimate of the pitch correlation [20] with look-ahead, for a total of 70 input features. For each band b, we also have 2 outputs: the gain $\hat{g}_b\left(\ell\right)$ approximates $g_b^{(\text{att})}\left(\ell\right) g_b\left(\ell\right)$ and the strength $\hat{r}_b\left(\ell\right)$ approximates $r_b\left(\ell\right)$.

The weights of the model are forced to a $\pm\frac{1}{2}$ range and quantized to 8-bit integers. This reduces the memory requirement (and bandwidth), while also reducing the computational complexity of the inference by taking advantage of vectorization.

4.1. Training Data

We train the model on synthetic mixtures of clean speech and noise with SNRs ranging from -5 dB to 45 dB, with some noise-free examples included. The clean speech data includes 120 hours of 48 kHz speech from different public and internal databases, including more than 200 speakers and more than 20 different languages. The noise data includes 80 hours of various noise types, also sampled at 48 kHz.

To ensure robustness in reverberated conditions, the noisy signal is convolved with simulated and measured room impulse responses. Inspired by [21], the target includes the early reflections so that only late reverberation is attenuated.

We improve the generalization of the model by applying a different random second-order pole-zero filter to both the speech and the noise. We also apply the same random spectral tilt to both signals to better generalize across different microphone frequency responses. To achieve bandwidth-independence, we apply a low-pass filter with a random cutoff frequency between 3 kHz and 20 kHz. This makes it possible to use the same model on narrowband to fullband audio.

4.2. Loss function

We use a different loss function for the gain and for the pitch filtering strength. For the gain, we consider that the perceptual loudness of a signal is proportional to its energy raised to a power $\gamma/2$, where we use $\gamma = 0.5$. For that reason, we raise the gains to the power γ before computing the metrics. In addition to the squared error, we also use the fourth power to overemphasize the cost of making large errors (e.g. completely attenuating speech):

$$\mathcal{L}_g = \sum_b \left(g_b^\gamma - \hat{g}_b^\gamma\right)^2 + C_4 \sum_b \left(g_b^\gamma - \hat{g}_b^\gamma\right)^4, \tag{11}$$

where we use $C_4 = 10$ to balance between the L_2 and L_4 terms.

Although simple, the loss function in (11) implicitly incorporates many of the characteristics of the improved loss function proposed in [22], including scale-invariance, SNR-invariance, power-law compression, and non-linear frequency resolution.

For the pitch filtering strength, we use the same principle as for $\mathcal{L}_g$ but evaluating the loudness of the noisy component of the enhanced speech. Since the comb filter with strength r_b attenuates the noise by a factor $(1 - r_b)$, we use the strength loss

$$\mathcal{L}_r = \sum_b \left((1 - r_b)^\gamma - (1 - \hat{r}_b)^\gamma\right)^2. \tag{12}$$

Since the enhancement is not overly sensitive to errors in the value of $\hat{r}_b$, we do not use a fourth power term.

5. Envelope Postfiltering

To further enhance the speech, we slightly deviate from the gains $\hat{g}_b$ produced by the DNN. The deviation is inspired by the formant postfilters [23] often used in CELP codecs. We intentionally de-emphasize noisier bands slightly further than they would be in the clean signal, while overemphasizing clean bands to compensate. This is done by computing a warped gain

$$\hat{g}_b^{(w)} = \hat{g}_b \sin\left(\frac{\pi}{2}\hat{g}_b\right), \tag{13}$$

which leaves $\hat{g}_b$ essentially unaffected for clean bands, while squaring it (like the gain of a Wiener filter) for very noisy bands. To avoid over-attenuating the enhanced signal as a whole, we also apply a global gain compensation heuristic computed as

$$G = \sqrt{\frac{(1 + \beta)\,\frac{E_0}{E_1}}{1 + \beta\left(\frac{E_0}{E_1}\right)^2}}, \tag{14}$$

where E_0 is the total energy of the enhanced signal using the original gain $\hat{g}_b$ and E_1 is the total energy when using the

warped gain $\hat{g}_b^{(w)}$. We use $\beta = 0.02$, which results in a maximum theoretical gain of 5.5 dB for clean bands. Scaling the final signal for the frame by G results in a perceptually cleaner signal that is about as loud as the clean signal. The band energy after that postfilter is given by

$$\hat{X}_b = G\hat{g}_b^{(w)} Y_b \,. \tag{15}$$

When listening to the enhanced speech through loudspeakers in a room, the impulse response of the room is added back to the signal such that it blends with any speech coming from the room. However, when listening through headphones, the lack of any reverberation can make the enhanced signal sound overly *dry* and unnatural. This is addressed by enforcing a minimum decay in the energy, subject to never exceeding the energy of the noisy speech:

$$\hat{X}_b^{(r)}(\ell) = \min\left(\max\left(\hat{X}_b(\ell), \delta\hat{X}_b^{(r)}(\ell-1)\right), \hat{Y}_b(\ell)\right), \tag{16}$$

where δ is chosen to be equivalent to a reverberation time $T_{60} = 100\,\text{ms}$.

After the frequency-domain enhanced speech is converted back to the time domain, a high-pass filter is applied to the output. The filter helps eliminating some remaining low-frequency noise and its cutoff frequency is determined by the estimated pitch of the talker [20] to avoid attenuating the fundamental.

6. Experiments and Results

We evaluate the quality of the enhanced speech with two mean opinion score (MOS) [24] tests conducted using the crowdsourcing methodology P.808 [25]. First, we use the 48 kHz noisy VCTK test set provided in [26] to compare PercepNet to the original RNNoise [12], while also conducting an ablation study. The test includes 824 samples, rated by 8 listeners each, resulting in a 95% confidence interval of 0.04. We also provide PESQ-WB [27] results as a reference for comparison with other methods like SEGAN [9]. The results in Table 1 not only demonstrate a base improvement over RNNoise, but also show that both the pitch filter and the envelope postfilter help improve the quality of the enhanced speech. In addition, subjective testing clearly shows the limitations of PESQ-WB when evaluating the envelope postfilter – even though the subjective evaluation shows a strong improvement from the postfilter, PESQ-WB considers it a degradation. Note that the unusually high absolute numbers in the MOS results are likely due to the fullband samples in that test.

In the second test, the DNS challenge [28] organizers evaluated *blind* test samples processed with PercepNet and provided us with the results in Table 2. The test set includes 150 synthetic samples without reverberation, 150 synthetic samples with reverberation, and 300 real recordings. Each sample was rated by 10 listeners, leading to a 95% confidence interval of 0.02 for all algorithms. Since PercepNet operates at 48 kHz, the 16-kHz challenge test data was internally up-sampled (and later down-sampled) in the STFT domain, avoiding any additional algorithmic delay. The same model parameters were used for both the challenge 16-kHz evaluation and our own 48-kHz VCTK evaluation, demonstrating the capability to operate on speech with different bandwidths. The quality also exceeds that of the baseline [29] algorithm.

The algorithm complexity is mostly dictated by the neural network, and thus the number of weights. For a frame size of 10 ms and 8M weights, the complexity is around 800 MMACS (one multiply-and-accumulate per weight per frame/second). By quantizing the weights with 8 bits, vectorization makes it possible to run the network efficiently. With the default frame size of 10 ms, PercepNet requires 5.2% of one mobile x86 core (1.8 GHz Intel i7-8565U CPU) for real-time operation. Evaluated with a frame size of 40 ms (four internal frames of 10 ms each to improve cache efficiency), the complexity is reduced to 4.1% on the same CPU core with an identical output. Despite a much lower complexity than the maximum allowed by the DNS challenge, PercepNet ranked second in the real-time track.

Table 1: *P.808 MOS results based on internal testing on the VCTK test set at 48 kHz.*

Algorithm	PESQ-WB	**MOS (P.808)**
Noisy	1.97	3.40
SEGAN [9]	2.16	-
RNNoise (original) [12]	2.29	3.70
PercepNet (no pitch, no pf)	2.64	3.81
PercepNet (no pf)	**2.73**	3.91
PercepNet (no pitch)	2.47	3.93
PercepNet	2.54	**4.05**

Table 2: *Challenge official P.808 MOS results. The baseline model is provided by the challenge organizers.*

Algorithm	Synthetic w/o reverb	Synthetic w/ reverb	Real record	**Overall**
Noisy	3.32	2.78	2.97	3.01
Baseline	3.49	2.64	3.00	3.03
PercepNet	**3.92**	**3.16**	**3.51**	**3.52**

Qualitatively, the use of ERB bands – rather than operating directly on frequency bins – makes the algorithm incapable of producing musical noise (*aka* birdie artifacts) in the output. Similarly, the short window used for analysis avoids reverb-like smearing in the time domain. Instead, the main noticeable artifact is a certain amount of *roughness* caused by some noise remaining between pitch harmonics, especially for loud car noise.

7. Conclusion

We have demonstrated an efficient speech enhancement algorithm that focuses on the main perceptual characteristics of speech – spectral envelope and periodicity – to produce high-quality fullband speech in real time with low complexity. The proposed PercepNet model uses a band structure to represent the spectrum, along with pitch filtering and an additional envelope postfiltering step. Evaluation results show significant quality improvements for both wideband and fullband speech and demonstrate the effectiveness of both the pitch filtering and the postfilter. We believe the results demonstrate the benefits of modeling speech using perceptually-relevant parameters.

8. References

[1] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, 27(2):113–120, 1979.

[2] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33(2):443–445, 1985.

[3] D. Liu, P. Smaragdis, and M. Kim. Experiments on deep learning for speech denoising. In *Proceedings of Fifteenth Annual Con-*

ference of the International Speech Communication Association, 2014.

[4] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee. A regression approach to speech enhancement based on deep neural networks. *IEEE Transactions on Audio, Speech and Language Processing*, 23(1):7–19, 2015.

[5] K. Tan and D. Wang. A convolutional recurrent neural network for real-time speech enhancement. In *Proceedings of INTERSPEECH*, volume 2018, pages 3229–3233, 2018.

[6] A. Narayanan and D. Wang. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7092–7096, 2013.

[7] Y. Zhao, D. Wang, I. Merks, and T. Zhang. Dnn-based enhancement of noisy and reverberant speech. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6525–6529, 2016.

[8] D.S. Williamson, Y. Wang, and D. Wang. Complex ratio masking for monaural speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 24(3):483–492, 2016.

[9] S. Pascual, A. Bonafonte, and J. Serra. SEGAN: Speech enhancement generative adversarial network. *arXiv:1703.09452*, 2017.

[10] D. Rethage, J. Pons, and X. Serra. A wavenet for speech denoising. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5069–5073, 2018.

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

[12] J.-M. Valin. A hybrid DSP/deep learning approach to real-time full-band speech enhancement. In *Proceedings of IEEE Multimedia Signal Processing (MMSP) Workshop*, 2018.

[13] C. Montgomery. Vorbis I specification, 2004.

[14] J. Princen and A. Bradley. Analysis/synthesis filter bank design based on time domain aliasing cancellation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(5):1153–1161, 1986.

[15] B.C.J. Moore. *An introduction to the psychology of hearing*. Brill, 2012.

[16] H. Gockel, B.C.J. Moore, and R.D. Patterson. Asymmetry of masking between complex tones and noise: Partial loudness. *The Journal of the Acoustical Society of America*, 114(1):349–360, 2003.

[17] D. Talkin. A robust algorithm for pitch tracking (RAPT). In *Speech Coding and Synthesis*, chapter 14, pages 495–518. Elsevier Science, 1995.

[18] T. Painter and A. Spanias. Perceptual coding of digital audio. *Proceedings of the IEEE*, 88(4):451–515, 2000.

[19] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*, 2014.

[20] K. Vos, K. V. Sorensen, S. S. Jensen, and J.-M. Valin. Voice coding with Opus. In *Proceedings of the 135^{th} AES Convention*, 2013.

[21] Y. Zhao, D. Wang, B. Xu, and T. Zhang. Late reverberation suppression using recurrent neural networks with long short-term memory. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5434–5438. IEEE, 2018.

[22] H. Erdogan and T. Yoshioka. Investigations on data augmentation and loss functions for deep learning based speech-background separation. In *Proceedings of INTERSPEECH*, pages 3499–3503, 2018.

[23] J.-H. Chen and A. Gersho. Adaptive postfiltering for quality enhancement of coded speech. *IEEE Transactions on Speech and Audio Processing*, 3(1):59–71, 1995.

[24] ITU-T. *Recommendation P.800: Methods for subjective determination of transmission quality*, 1996.

[25] ITU-T. *Recommendation P.808: Subjective evaluation of speech quality with a crowdsourcing approach*, 2018.

[26] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi. Investigating rnn-based speech enhancement methods for noise-robust text-to-speech. In *Proceedings of ISCA Speech Synthesis Workshop (SSW)*, pages 146–152, 2016.

[27] ITU-T. P.862.2: Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs (PESQ-WB). 2005.

[28] C.K.A. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke. The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. *arXiv preprint arXiv:2005.13981*, 2020.

[29] Y. Xia, S. Braun, C.K.A. Reddy, H. Dubey, R. Cutler, and I. Tashev. Weighted speech distortion losses for neural-network-based real-time speech enhancement. *arXiv:2001.10601*, 2020.

View publication stats