Grenze International Journal of Engineering and Technology, June Issue

Efficient Speech Denoising Techniques for Adverse Acoustic Conditions

Maneesh Kumar Singh[1], Siow Yong Low[2], Neeraj Sharma[3], C. Chandra Sekhar Reddy[4], B. Anirwan Sai Reddy[5] and R. Deekshith[6]
[1]MITS (Deemed to Be) University, Madanapalle, A.P India maneeshideal@gmail.com
[2]School of Electronics and Computer Science,University of Southampton Malaysia, Johor Malaysia
sy.low@soton.ac.uk
[3]Defence Institute of High-Altitude Research, Defense Research & Development Organization, Govt. of India.
neeraj.sharma.dgre@gov.in
[4-6]Madanapalle Institute of Technology & Science, Madanapalle, A.P India.
Email: chittachandrasekhar64@gmail.com, anirwananirwan882@gmail.com, deekshith0421@gmail.com

***Abstract*— Nowadays, the real time noise suppression systems struggle by introducing temporal lag and perceptual artifacts in the speech assistive devices. This paper proposes an real time auditory-aware noise spectral tracking methodology which merges both the critical bandwidth analysis and the SNR refinement architecture. This proposed method uses decision-directed recursive technique with a Joint Maximum a Posteriori (JMAP) correction which minimizes the frame delay by incorporating frequency-dependent masking threshold derived from Bark-scale critical bands. It applies a noise-to-masking ratio adaptive weight in the noise estimator to modulate the noise spectral suppression depending on the level of the a posteriori SNR values. Results using TIMIT speech database degraded by stationary white noise to highly non stationary babble noise including factory, street, and Himalayan Snow fall noise interference at various noise levels (0 - 20 dB SNR) demonstrated that the developed noise estimator outperformed in terms of PESQ (up to 3.65), STOI (up to 0.985), and segmental SNR against the state of the art noise estimation techniques available till now.**

***Keywords*— Speech enhancement · noise estimation · perceptual weighting · Bark scale · sigmoid adaptation · JMAP refinement · auditory masking.**

I. INTRODUCTION

In Speech assistive devices, the intelligibility improvement remains challenging task especially in a real time noisy conditions. Low intelligibility is critical in communication systems, hearing aids, and assistive devices and impose a constrain to communicate from the noisy conditions. In real time situations where noise is also in the channel with speech, it ubiquitously degrades the signal quality during the transmission [7]. To address this issue, the Speech Enhancement techniques leverage the frequency-domain STFT processing.

Traditional speech enhancement methods uses the voice activity detection (VAD) and continuous tracking approaches. In VAD-based approaches, falter at sub-5 dB SNR in non-stationary environments [1]. However, in Non-VAD recursion methods such as Minimum Statistics exhibits a huge tracking lag which can be seen in improved minima controlled recursive averaging (IMCRA), speech presence probability (SPP) weights although

Grenze ID: 01.GIJET.12.2.92_26

offer small improvement yet these methods assume the standard Gaussian statistics that non-stationary noise spectrum clearly violates [2, 10]. Later the MMSE-Bayesian unbiased estimators reduce tracking complexity but employ fixed thresholds to achieve faster tracking [4]. Using Sigmoid function [23], successfully tackled the non stationary noise spectrum by providing the smooth transitions, but it treated all frequencies uniformly and ignoring the very important human auditory system.

On the other hand, deep learning methods achieve better tracking statistics but demand high computa-tional resources which is not possible in constraining in the small size hearing aid devices. The main critical omission across state of the art methods is the psycho-acoustic integration of human auditory system, critical band frequency selectivity, absolute thresholds, and simultaneous frequency masking phenomena [15]. As a result, the noise below masking thresholds is inaudible whilst aggressive suppression of masked components introduces huge suppression of speech component which results distortion and musical noise.

The following Main contributions has been achieved from this paper as follows

- A dual-stage SNR refinement: DD smoothing (α = 0.96) followed by JMAP correction (balance 0.999), reducing frame lag.
- Perceptual modeling via 24 critical bands (Bark scale) with frequency-dependent absolute thresholds and masking computations.
- Noise-to-masking ratio (NMR) piecewise weight adaptively controlling suppression across frequency.
- SNR- and frequency-dependent smoothing coefficient enabling aggressive tracking at low SNR, conser-vative smoothing at high SNR.
- Comprehensive validation on TIMIT with white, factory, street, babble noise confirming PESQ/STOI/SegSNR superiority.

II. Noise Tracking System Modeling

The speech in observed noisy situation is modeled as additive uncorrelated mixture:

$$y(t) = x(t) + n(t). \tag{1}$$

STFT segmentation (frame length $L = 256$ samples, hop $S = 128$, FFT size 1024, $f_s = 8$ kHz, Ham-ming window) produces frequency-bin representations. The noisy spectrum satisfies PSD additivity under independence:

$$\lambda Y(k, n) = \lambda S(k, n) + \lambda D(k, n). \tag{2}$$

A posteriori and *a priori* SNR are defined:

$$\gamma(k, n) = \frac{|Y(k, n)|^2}{\lambda_D(k, n)}, \quad \xi(k, n) = \frac{\lambda_S(k, n)}{\lambda_D(k, n)}. \tag{3}$$

A. Dual-Stage A Priori SNR Refinement

Stage 1 (Decision-Directed): Classical DD recursion:

$$\xi_{dd}(k, n) = \alpha \frac{|\hat{S}(k, n-1)|^2}{\hat{\lambda}_D(k, n-1)} + (1-\alpha)\max[\gamma(k, n) - 1, 0], \tag{4}$$

Stage 2 (JMAP Refinement): Instantaneous speech PSD:

$$\lambda_{S,\text{inst}}(k, n) = \max[|Y(k, n)|^2 - \hat{\lambda}_D(k, n), 0]. \tag{5}$$

Hybrid estimate via balance factor $\alpha_{\text{jmap}} = 0.999$:

$$\xi(k, n) = \alpha_{\text{jmap}}.\xi_{dd}(k, n) + (1 - \alpha_{\text{jmap}}).\frac{\lambda_{S,\text{inst}}(k, n)}{\hat{\lambda}_D(k, n) + \epsilon}. \tag{6}$$

This reduces DD lag while maintaining stability.

(k, n) + ϵ

B. Perceptual Modeling: Critical Bands and Masking Bark Scale Mapping: Perceptual frequency (B in Bark):

$$B(f) = 13 \arctan(0.00076f) + 3.5 \arctan\left(\frac{f}{7500}\right)^2 . \quad (7)$$

The 8 kHz bandwidth spans approximately 17.4 Bark. Dividing into $N_b = 24$ uniform Bark intervals yields frequency-mapped critical band edges.

Absolute Threshold: Frequency-dependent hearing floor (ISO 226 simplification) is piecewise constant: 40 dB SPL for $f < 100$ Hz down to 5 dB SPL at 1–4 kHz range, then rising to 15 dB SPL above 4 kHz.

Masking Threshold: For bin k in critical band i:

$$T_{\text{mask}}(k, n) = \max\left[0.1 \cdot P_{\text{band}}(i, n), T_{\text{abs}}(i)\right], \quad (8)$$

where P_{band} is the average noisy power within the band.

C. Noise-to-Masking Ratio and Weighting

NMR quantifies auditory prominence of noise:

$$\text{NMR}(k, n) = \frac{\hat{\lambda}_D(k, n)}{0.9999 \cdot T_{\text{mask}}(k, n)}. \quad (9)$$

Piecewise perceptual weight:

$$\omega(k, n) = \begin{cases} 0.4 + 0.4 \cdot \text{NMR}, & \text{NMR} < 0.3 \\ 0.65 + 0.35 \cdot \text{NMR}, & 0.3 \le \text{NMR} < 1.0 \\ 1.0, & \text{NMR} \ge 1.0 \end{cases} \quad (10)$$

Masked regions (NMR < 0.3) apply moderate suppression (0.4–0.52), partially masked bands (0.3–1.0) increase to 0.65–1.0, and unmasked regions enforce maximum attenuation (1.0).

D. Adaptive Smoothing and Sigmoid Update

Frequency-dependent smoothing factors:

$\eta 1 = \exp(-0.704)$, (11)

$\eta 2 = \exp(-0.44)$, (12)

$\eta 3 = \exp(-0.0001467)$

selected based on frequency: η_1 for $f \le 50$ Hz, η_2 for $50 < f \le 3000$ Hz, η_3 for $f > 3000$ Hz.

SNR-dependent factor $\alpha_h(k, n)$ ranges 0.4–3.0 across 8 thresholds (–15 to 20 dB in 5 dB steps), enabling aggressive tracking at low SNR.

Sigmoid parameters: $\xi_{\text{opt}} = 50.12$ (17 dB), yielding slope $\sigma_a \approx 0.9804$, center $\sigma_c \approx 4.010$

Furthermore, the smoothing factor is depending on the current a posteriori SNR γ(k, n) to enable aggressive tracking of the noise spectrum in a very low-SNR conditions and adds a conservative smoothing when the SNR condition is high. The SNR-dependent factor is:

$$\alpha_h(k, n) = \begin{cases} 0.4, & f_m \le 50 \text{ Hz and } \gamma(k, n) \le 0.0316, \\ 0.7, & 0.0316 < \gamma(k, n) \le 0.10, \\ 1.0, & 0.10 < \gamma(k, n) \le 0.316, \\ 1.5, & 0.316 < \gamma(k, n) \le 1.0, \\ 1.75, & 1.0 < \gamma(k, n) \le 3.162, \\ 2.0, & 3.162 < \gamma(k, n) \le 10.0, \\ 2.5, & 10.0 < \gamma(k, n) \le 31.62, \\ 3.0, & \gamma(k, n) > 31.62. \end{cases} \quad (14)$$

These thresholds correspond to a posteriori SNR values of -15 dB, -10 dB, -5 dB, 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB, respectively.

Sigmoid Membership Function The sigmoid membership function computes a soft decision indicating the probability that the current observation is dominated by noise:

$$G_\gamma(k, n) = \left[\frac{1}{1 + \exp\left(-\sigma_a \left(\gamma(k, n)^{1/\alpha_h(k,n)} - \sigma_c\right)\right)}\right]^{1/\omega(k,n)}. \quad (15)$$

The exponent $1/\alpha_h(k, n)$ modulates the sensitivity of the sigmoid function to variations in $\gamma(k, n)$. Higher values of $\alpha_h(k, n)$ (corresponding to higher SNR) reduce the effective slope, making the decision boundary smoother and reducing abrupt transitions.

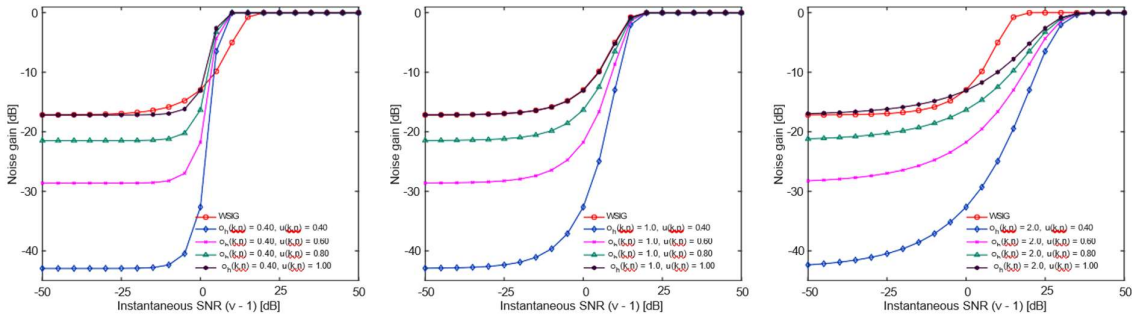

Fig. 1: Noise Gain performance for varying αh(k, n) & ω(k, n) for different instantaneous SNR conditions.

Perceptually Weighted Soft Decision The perceptual weight ω(k, n) is incorporated into the soft decision by applying an exponent:

The exponent 1/ω(k, n) amplifies the sigmoid output when ω(k, n) < 1, making the noise update more aggressive in frequency bins where the noise is perceptually masked. Conversely, when ω(k, n) = 1 (unmasked regions), the soft decision is applied directly.

Noise PSD Update The noise PSD is updated using the perceptually weighted soft decision sN (k, n): λˆD(k, n) = ω(k, n)λˆD(k, n − 1) + (1 − ω(k, n))|Y (k, n)|2. (16)

This intermediate estimate blends the previous noise PSD estimate with the current noisy power, weighted by the perceptual weight. The final noise PSD update is:

λˆD(k, n) = sN (k, n)λˆD(k, n − 1) + (1 − sN (k, n))λˆD(k, n). (17) Expanding this expression:

Simplifying:

λˆD(k, n) = sN (k, n)λˆD(k, n − 1)
+ (1 − sN (k, n)) hω(k, n)λˆD(k, n − 1)
+(1 − ω(k, n))|Y (k, n)|2
λˆD(k, n) = [sN (k, n) + (1 − sN (k, n))ω(k, n)] λˆD(k, n − 1)
+ (1 − sN (k, n))(1 − ω(k, n))|Y (k, n)|2. (18)

TABLE I: PESQ PERFORMANCE COMPARISON UNDER DIFFERENT NOISE TYPES [FRAME SIZE 32MS]

Algorithm	White Noise					Factory Noise					Babble Noise				
	0dB	5dB	10dB	15dB	20dB	0dB	5dB	10dB	15dB	20dB	0dB	5dB	10dB	15dB	20dB
Noisy	1.5026	1.8029	2.1377	2.4912	2.8380	2.1960	2.5382	2.8694	3.1908	3.4907	1.6787	2.0196	2.3688	2.7273	3.0606
IMCRA	2.0132	2.4062	2.7516	3.0299	3.3034	2.6884	3.0063	3.2937	3.5721	3.8250	1.8386	2.2381	2.6229	2.9811	3.3142
MMSE-BC	2.0211	2.4190	2.7699	3.0875	3.3882	2.7230	3.0373	3.3197	3.5848	3.8397	1.8404	2.2167	2.5809	2.9632	3.3005
WSIG	2.0254	2.4286	2.7786	3.0933	3.3901	2.7392	3.0527	3.3328	3.5939	3.8442	1.8683	2.2463	2.6217	2.9851	3.3174
Proposed	2.0167	2.4214	2.7928	3.1185	3.4231	2.7754	3.0945	3.3805	3.6460	3.8877	1.9053	2.3053	2.6689	3.0429	3.3703
Algorithm	Street Noise					Traffic Noise					Snowfall Noise				
Noisy	1.8623	2.1895	2.5137	2.8259	3.1306	2.2972	2.6266	2.9598	3.2817	3.6025	1.9580	2.3048	2.6658	3.0088	3.3334
IMCRA	2.2576	2.5982	2.9434	3.2537	3.5191	2.6873	2.9973	3.2991	3.6040	3.8772	2.1098	2.4517	2.7867	3.1154	3.4066
MMSE-BC	2.2959	2.6370	2.9863	3.2918	3.5622	2.6998	3.0255	3.3464	3.6628	3.9288	2.1058	2.4498	2.7851	3.1021	3.3969
WSIG	2.3079	2.6541	3.0041	3.3061	3.5707	2.7208	3.0441	3.3606	3.6712	3.9352	2.2701	2.6312	2.9769	3.3036	3.6110
Proposed	2.3548	2.7025	3.0443	3.3538	3.6142	2.7615	3.0995	3.4243	3.7331	3.9944	2.2780	2.6526	3.0063	3.3409	3.6446

III. EXPERIMENTAL EVALUATION

To evaluate the performance of the proposed perceptually weighted sigmoid noise estimator, comprehensive experiments were conducted on the TIMIT speech database [25]. A total of 12 phonetically balanced speech

sentences were selected, comprising six male and six female speakers. Each sentence has a duration of approximately 4.5 seconds. The speech signals were down-sampled from 16 kHz to 8 kHz to match the sampling rate used in the proposed method. To simulate realistic recording conditions, 0.5 seconds of initial silence was prepended to each speech sample.
Results demonstrate consistent PESQ advantages in babble (up to +0.37 at 15 dB) and street noise. STOI preservation in 50–2000 Hz critical band validates auditory modeling. Segmental SNR improvements confirm noise tracking fidelity without speech distortion.

IV. Result Analysis

The experimental results demonstrate that the proposed framework consistently achieves superior enhance-ment performance across all evaluated noise conditions and SNR levels. As tabulated in 1, the proposed method yields noticeable improvements in PESQ scores compared to conventional approaches, particularly under low SNR conditions (0 dB and 5 dB). This improvement can be directly attributed to the adap-tive SNR estimation strategy formulated in (15), which effectively balances instantaneous and smoothed spectral information. Unlike traditional estimators that exhibit delayed response, the proposed formulation enables faster adaptation to abrupt noise variations, thereby improving robustness in highly non-stationary environments such as babble and street noise.
The derived gain function in (15) is closely examined and from the gain plot with varying parameters (α, ω) clearly reveals that incorporating the perceptual masking controls the noise suppression. This is done by assigning higher noise spectral suppression while preserving masked regions. This resulted in the reduction of unnecessary over estimation of speech spectral components. In contrast, existing state of the art methods perform either an over-suppression or leave residual noise component, i.e., resulting in degraded perceptual quality.
The adaptive changes in smoothing constant is governed by Eq. (14) and contributes to the Sigmoid-based noise controlling. This allows frequency-dependent adaptation to ensure that low-frequency components are preserved with minimal distortion, which are essential for speech intelligibility, whilst high-frequency noise component is more aggressively attenuated. This adaptive controlling is beneficial in highly non-stationary real time noisy environments such as babble of Himalayan Snow fall noise, where spectral characteristics vary significantly over time. As evident from table 1, 2 and 3, the proposed method maintains stable improvement across all frequency bands without introducing any abrupt distortion over the existing state of the art methods.

TABLE II: STOI Performance Comparison under Different Noise Types [frame size 32ms]

Algorithm	White Noise					Factory Noise					Babble Noise				
	0dB	5dB	10dB	15dB	20dB	0dB	5dB	10dB	15dB	20dB	0dB	5dB	10dB	15dB	20dB
Noisy	0.4958	0.6404	0.7581	0.8396	0.8891	0.6727	0.7684	0.8408	0.8887	0.9174	0.4655	0.6210	0.7511	0.8384	0.8872
IMCRA	0.5288	0.6880	0.7933	0.9080	0.9566	0.7117	0.8335	0.9172	0.9563	0.9762	0.4816	0.6375	0.7644	0.8577	0.9397
MMSE-BC	0.5512	0.6883	0.8428	0.9245	0.9613	0.7060	0.8337	0.9178	0.9567	0.9771	0.4883	0.6362	0.7624	0.8518	0.9345
WSIG	0.5502	0.6885	0.8473	0.9225	0.9587	0.7040	0.8349	0.9140	0.9537	0.9747	0.4865	0.6338	0.7600	0.8509	0.9389
Proposed	0.5653	0.6965	0.8457	0.9438	0.9745	0.7196	0.8367	0.9321	0.9682	0.9854	0.5026	0.6557	0.7789	0.8666	0.9555
Algorithm	Street Noise					Traffic Noise					Snowfall Noise				
Noisy	0.5449	0.6653	0.7700	0.8474	0.8959	0.6837	0.7696	0.8384	0.8868	0.9173	0.6007	0.7299	0.8231	0.8807	0.9132
IMCRA	0.5912	0.7276	0.8516	0.9216	0.9592	0.7131	0.8197	0.9107	0.9533	0.9758	0.6182	0.7461	0.8372	0.8916	0.9226
MMSE-BC	0.5932	0.7391	0.8649	0.9271	0.9618	0.7141	0.8208	0.9093	0.9550	0.9783	0.6219	0.7393	0.8338	0.8917	0.9232
WSIG	0.5914	0.7381	0.8621	0.9228	0.9584	0.7141	0.8212	0.9081	0.9535	0.9768	0.6321	0.7663	0.8482	0.8978	0.9309
Proposed	0.6093	0.7402	0.8764	0.9454	0.9745	0.7247	0.8246	0.9140	0.9660	0.9848	0.6479	0.7784	0.8608	0.9075	0.9410

Overall, the strong approval and alignment between theoretical noise estimator design and experimen-tal validations establishes that the proposed method effectively outperforms across diverse real time noisy conditions and making the noise is well-suited for real time practical speech assistive applications.

V. Conclusion

The proposed noise framework clearly indicated the performances in real time noisy conditions by combining statistical estimation with auditory modeling principles effectively. The dual-stage SNR refinement integra-tion

with perceptual masking-based weighting, and adaptive smoothing mechanisms controls the system to achieve consistent improvements in both the speech quality and intelligibility across a wide range of highly non stationary noise conditions. The experimental findings clearly demonstrated that the proposed method is not only outperforming over the existing state of the art methods in terms of PESQ, STOI, and segmen-tal SNR, but also maintaining the consistent behavior under highly non-stationary and extreme low SNR environments.

REFERENCES

[1] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhance-ment," IEEE Sig. Proc. Lett., vol. 9, no. 1, pp. 12–15, Jan. 2002.

[2] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," IEEE Trans. Speech Audio Process., vol. 11, no. 5, pp. 466–475, Sept. 2003.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral ampli-tude estimator," IEEE Trans. Acoust., Speech, Signal Process., vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[4] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 4, pp. 1383–1393, May 2012.

[5] J. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in Proc. Int. Conf. Spoken Lang. Process., Sydney, Australia, Dec. 1998, pp. 2819–2822.

[6] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in Proc. IEEE ICASSP, Dallas, TX, USA, Mar. 2010, pp. 4266–4269.

[7] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligi-bility and suggested solutions," IEEE Trans. Audio, Speech, Lang. Process., vol. 19, no. 1, pp. 47–56, Jan. 2011.

[8] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model," EURASIP J. Appl. Sig. Process., vol. 2005, no. 7, pp. 1110–1126, 2005.

[9] R. Martin, "Spectral subtraction based on minimum statistics," in Proc. EUSIPCO, Edinburgh, UK, Sept. 1994, pp. 1182–1185.

[10] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," IEEE Trans. Speech Audio Process., vol. 9, no. 5, pp. 504–512, July 2001.

[11] R. Martin, "Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors," in Proc. IEEE ICASSP, Orlando, FL, USA, May 2002, vol. 1, pp. 253–256.

[12] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," IEEE Trans. Acoust., Speech, Signal Process., vol. 28, no. 2, pp. 137–145, Apr. 1980.

[13] K. Paliwal, W. Kamil, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," Speech Commun., vol. 52, no. 5, pp. 450–475, May 2010.

[14] K. Paliwal, B. Schwerin, and W. Kamil, "Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator," Speech Commun., vol. 54, no. 2, pp. 282–305, Feb. 2012.

[15] E. Plourde and B. Champagne, "Auditory-based spectral amplitude estimators for speech enhancement," IEEE Trans. Audio, Speech, Lang. Process., vol. 16, no. 8, pp. 1614–1623, Nov. 2008.

[16] T. F. Quatieri, Discrete-Time Speech Signal Processing: Principles and Practice. Upper Saddle River, NJ: Prentice Hall, 2002.

[17] Y. Rongshan, "A low-complexity noise estimation algorithm based on smoothing of noise power estimation and estimation bias correction," in Proc. IEEE ICASSP, Taipei, Taiwan, Apr. 2009, pp. 4421–4424.

[18] M. K. Singh, S. Y. Low, S. Nordholm, and Z. Zhuquan, "Bayesian noise estimation in the modulation domain," Speech Commun., vol. 96, pp. 81–92, Feb. 2018.

[19] J. Sohn and N. Kim, "Statistical model-based voice activity detection," IEEE Sig. Proc. Lett., vol. 6, no. 1, pp. 1–3, Jan. 1999.

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio, Speech, Lang. Process., vol. 19, no. 7, pp. 2125–2136, Sept. 2011.

[21] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 26, no. 10, pp. 1702–1726, Oct. 2018.

[22] P. C. Yong, S. Nordholm, and H. Dam, "Noise estimation based on soft decisions and conditional smoothing for speech enhancement," in Proc. Int. Workshop Acoust. Signal Enhancement, Aachen, Germany, Sept. 2012, pp. 1–4.

[23] P. C. Yong, S. Nordholm, and H. Dam, "Optimization and evaluation of sigmoid function with a priori SNR estimate for real-time speech enhancement," Speech Commun., vol. 55, no. 2, pp. 358–376, Feb. 2013.

[24] E. Zwicker, "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," J. Acoust. Soc. Amer., vol. 33, no. 2, p. 248, Feb. 1961.

[25] J. S. Garofolo, "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM," National Institute of Standards and Technology, Tech. Rep., 1990.

[26] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum es-timation and an adaptive threshold," IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 2, pp. 412–424, Mar. 2006.

[27] Y. D. Cho, K. Naimi, and A. Kondoz, “Improved statistical voice activity detection based on a smoothed statistical likelihood ratio,” in Proc. IEEE ICASSP, Salt Lake City, UT, USA, May 2001, pp. 737–740.
[28] S. Pascual, A. Bonafonte, and J. Serra, “SEGAN: Speech enhancement generative adversarial network,” arXiv preprint arXiv:1703.09452, 2017.
[29] D. Rethage, J. Pons, and X. Serra, “A wavenet for speech denoising,” arXiv preprint arXiv:1706.07162, 2017.