Low-Light Speech Intelligibility Enhancement Using Adaptive Spectral Filtering and Signal Energy Analysis

Dr.C. Jeyalakshmi[1], drcjeyalakshmiece@krce.ac.in
Faculty, Department of ECE, K. Ramakrishnan College of Engineering, Tamil Nadu, India
Shangamitra.S[2], Shasty Priyan.N. A[3], Sheik Abdulla.M[4], Shivadharshini.B[5], Shivani.E[6]
shangamitra734@gmail.com[2], shastypriyan28@gmail.com[3], sheikabdulla.moo7@gmail.com[4],
shivadharshinib@gmail.com[5], shivanielavarasu245@gmail.com[6]
Students, Department of ECE, K. Ramakrishnan College of Engineering, Tamil Nadu, India

Abstract: Speech is often the only dependable communication channel when visibility is poor, because the visual cues such as lip movement and gesture that normally support listening are unavailable in low-light environments. This paper presents a lightweight method for improving the intelligibility of noisy speech recorded under such conditions. The method combines signal energy analysis with adaptive spectral filtering. Short-time frame energy is tracked to estimate a noise floor and to flag speech-present frames. This flag steers a recursive noise power estimate and a Wiener-type gain whose over-subtraction factor, smoothing memory and gain floor change from frame to frame according to the estimated frame SNR. The approach is implemented in Python with NumPy and SciPy and evaluated on ten synthetic speech-like utterances mixed with white, pink and machinery noise at 0, 5 and 10 dB input SNR. It is compared with spectral subtraction and a static-noise Wiener filter. Intelligibility is estimated with a simplified STOI-style envelope-correlation measure, and SNR and segmental SNR are also reported. The proposed method raised the intelligibility measure in all nine conditions, by 0.045, 0.036 and 0.021 on average at 0, 5 and 10 dB, whereas both baselines stayed close to the unprocessed input. The baselines, however, achieved larger SNR gains, showing that noise suppression and intelligibility are not the same objective. The method is simple, runs frame by frame, and suits embedded and real-time use. Results on synthetic signals are preliminary and need confirmation on recorded speech and listening tests.

Key Words: Speech Enhancement, Speech Intelligibility, Adaptive Spectral Filtering, Signal Energy Analysis, Noise Floor Estimation, Wiener Filter.

I. INTRODUCTION

Spoken communication depends on more than sound. In ordinary conditions listeners use the speaker's lips, jaw and face to resolve sounds that are hard to hear, and this visual information is known to raise recognition scores considerably when background noise is present [12]. In low-light settings such as night-time field work, underground mines, search and rescue operations, dim control rooms and night surveillance, that visual support largely disappears. The listener, or the automatic speech system standing in for the listener, must recover the message from the audio alone. The same recordings are usually degraded by engine noise, ventilation, mains hum and wind, so the audio channel is under pressure at the moment it is most needed.

Speech enhancement has been studied for several decades, and classical spectral-domain methods remain attractive because they are cheap, need no training data and can run on small processors [1], [2], [10]. Many of them are designed to maximise noise reduction or SNR. A recurring finding, however, is that lower noise power does not guarantee that more words are understood, because aggressive gain rules also remove weak consonant energy and add distortion [13], [14]. For the low-light use case the goal is therefore not a cleaner-sounding signal as such, but a signal in which speech structure is preserved while masking noise is reduced.

This paper proposes a small pipeline built around two ideas. First, *signal energy analysis* is used to decide, frame by frame, how much of the observed energy can be trusted as speech and how much is background. Second, *adaptive spectral filtering* uses that decision to vary how hard the filter works: strongly in noise-only frames, gently in speech dominated frames. The study covers the full path from signal acquisition and short-time analysis to filtering, reconstruction and objective evaluation, and it reports both the gains and the limitations found.

II. LITERATURE REVIEW

Early single-channel enhancement was based on subtracting an estimate of the noise magnitude spectrum from the noisy spectrum. Boll's spectral subtraction [1] is the standard reference, and Berouti et al. [2] added over-subtraction and a spectral floor to reduce the "musical noise" that simple subtraction produces. Wiener-type filtering [3], [10] derives a frequency-dependent gain from the ratio of speech to noise power. The decision-directed estimate of the a priori SNR introduced with the MMSE short-time spectral amplitude estimator [4] and its log-spectral variant [5] reduced musical noise further and became the

basis of many later systems. These methods work well when the noise is stationary and its spectrum is known, and they degrade when it is not.

Because the noise spectrum is rarely known in advance, much work has addressed noise tracking. Martin's minimum-statistics approach [6] follows the minimum of a smoothed power spectrum over a sliding window, so no explicit speech detector is needed. Cohen's improved minima-controlled recursive averaging [7] adds a speech-presence probability that controls how quickly the noise estimate adapts. Both ideas motivate the energy-based tracking used here, which is deliberately simpler: the smoothed frame energy is tracked over a window and compared with a margin.

A second thread concerns what objective measures actually predict. Loizou and Kim [13] and Hu and Loizou [14] showed that many conventional enhancement algorithms improve quality ratings or SNR yet give no intelligibility benefit, mainly because of gain-induced distortion of low-energy speech segments. The short-time objective intelligibility measure (STOI) of Taal et al. [8] correlates short-time temporal envelopes of clean and processed speech in one-third-octave bands and is widely used as an intelligibility predictor. Textbook treatments of these topics are given by Loizou [9].

More recently, supervised deep-learning methods such as mask estimation networks [11] and generative approaches such as SEGAN [15] have achieved strong results on benchmark data. They require large training sets and considerable computation and may generalise poorly to noise types absent from training, which limits their use on small embedded devices or in unusual acoustic environments. Classical adaptive filters therefore remain relevant when computation, latency or training data are limited.

The gap addressed here is a compact, training-free method that adapts its own aggressiveness to the estimated condition of each frame and is judged by an intelligibility-oriented measure rather than SNR alone.

III. MATERIALS AND METHODS

The pipeline has six stages, shown in Fig. 1: signal acquisition, framing with the short-time Fourier transform (STFT), signal energy analysis, adaptive spectral filtering, signal reconstruction and objective evaluation.

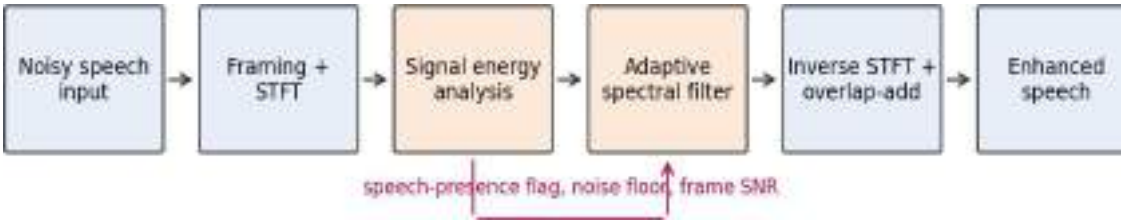


Fig.1 Block diagram of the proposed enhancement pipeline

A. Signal acquisition and test material

Speech is captured by a microphone as an acoustic pressure signal, converted to a voltage, and digitised. All signals here are mono at a sampling rate of 16 kHz. To keep the experiment fully reproducible and the clean reference exactly known, ten synthetic speech-like utterances of 4 s each were generated in Python. Each utterance contains a 0.4 s leading silence followed by a sequence of voiced syllables and occasional fricative bursts. A voiced syllable is produced by a pulse train with a time-varying fundamental frequency of 100 to 210 Hz, a spectral tilt, and three cascaded resonators at vowel-like formant frequencies. A fricative is band-pass noise between 3.5 and 7 kHz. These signals reproduce harmonic structure, formants, syllabic amplitude modulation and high-frequency consonant energy, but they are not real speech, and this limitation is discussed in Section IV.

Three noise types represent the low-light conditions of interest. *White noise* models sensor and electronic noise. *Pink noise* (1/f spectrum) models diffuse environmental noise. *Machinery noise* combines a 50 Hz mains hum with harmonics and a low-pass noise component whose level is modulated slowly at 0.7 Hz, which makes it non-stationary. Noise is scaled to give an input SNR of 0, 5 or 10 dB over the whole utterance, producing 9 conditions and 90 test signals. Random seeds are fixed.

B. Framing and short-time Fourier transform

The noisy signal $y(n)$ is divided into frames of 512 samples (32 ms) with a hop of 128 samples (75% overlap) and weighted by a Hann window. The STFT $Y(k,t)$ is computed for frequency bin k and frame t. A 32 ms frame is short enough for speech to be treated as quasi-stationary and long enough to resolve pitch harmonics. The 75% overlap with a Hann window satisfies the overlap-add condition for perfect reconstruction when no modification is applied; this was verified numerically (reconstruction SNR above 150 dB).

C. Signal energy analysis

For each frame the mean spectral power is expressed in decibels:

$$E(t) = 10 \log [(1/K) \Sigma |Y(k,t)|^2] \quad (1)$$

where K is the number of bins. $E(t)$ is smoothed with a five-frame moving average to give $E(t)$. The noise floor $F(t)$ is the s minimum of E over the preceding 150 frames (about 0.43 s), raised by 1.5 dB to compensate for the downward bias of as minimum tracker. A frame is flagged as speech-present when

$$E(t) > F(t) + 3 \text{ dB} \ldots\ldots\ldots (2)$$

and the frame-level SNR estimate is

$$\Delta(t) = \text{clip}(E(t) - F(t), 0, 30) \text{ dB}.$$

These three quantities (flag, floor and Δ) are passed to the filter.
No training data or explicit voice activity detector is used.

D. Adaptive spectral filtering

Let $\lambda(k,t)$ be the noise power estimate, initialised from the first eight frames. The estimate is updated recursively under control of the energy-analysis flag. In frames without speech, λ follows the observed power quickly but in frames with speech it is allowed to fall quickly when the observed power is below it, and to rise only very slowly:

$$\lambda(t) = 0.90\, \lambda(t-1) + 0.10\, |Y(t)|^2 \text{ (no speech)} \quad (3)$$

$$\lambda(t) = 0.999\, \lambda(t-1) + 0.001\, |Y(t)|^2 \text{ (speech, power above } \lambda) \quad (4)$$

(when the observed power is below λ in a speech frame, a 0.95/0.05 mixture is used instead).

The a posteriori SNR is $\gamma = |Y|^2/\lambda$, and the a priori SNR is obtained with the decision-directed rule

$$\xi(k,t) = \beta\, |\hat{S}(k,t-1)|^2/\lambda(k,t) + (1 - \beta) \max[\gamma(k,t) - 1, 0] \quad (5)$$

with memory $\beta = 0.92$ in speech frames, so that onsets are tracked quickly, and $\beta = 0.98$ in noise-only frames, for strong smoothing. The gain is a Wiener-type rule with an over-subtraction factor that depends on the frame SNR:

$$G(k,t) = \xi / (\xi + \alpha(t)), \quad \alpha(t) = \text{clip}(2.0 - 0.07\, \Delta(t), 1.0, 2.0) \quad (6)$$

so that filtering is strongest ($\alpha = 2$) when the frame SNR is low and relaxes to the ordinary Wiener rule ($\alpha = 1$) when it is high. The gain is smoothed across three neighbouring frequency bins to limit isolated spectral peaks and is then bounded below by a floor that is −12 dB in speech frames, to keep weak speech components and avoid distortion, and −18 dB in noise-only frames, to remove steady noise. The floor, memory and over-subtraction factor are the adaptive elements. The parameter values were selected from a small grid on a separate development set (five utterances with different seeds, three noise types, 5 dB input SNR) and were then fixed for all test conditions.

E. Signal reconstruction

The enhanced spectrum is $\hat{S}(k,t) = G(k,t)Y(k,t)$, while preserving the phase of the noisy signal. The enhanced signal is then reconstructed in the time domain using the inverse STFT and the overlap-add method. For each frame, the process involves one FFT, one gain calculation, and one inverse FFT. This makes the method computationally efficient and suitable for real-time implementation.

F. Baselines and evaluation metrics

Two reference methods were implemented with the same STFT settings. *Spectral subtraction* uses a fixed noise estimate from the first ten frames, an over-subtraction factor of 2 and a spectral floor of 0.02 [1], [2]. *Static Wiener* uses the same fixed noise estimate with decision-directed a priori SNR (memory 0.98) and a gain floor of 0.1 [3], [4]. Note that both baselines benefit from the leading silence in the test signals, which gives them an accurate noise estimate; in practice such a silence is not always available.

Three objective measures were computed against the clean reference: (i) overall *SNR* in dB; (ii) *segmental SNR*, averaged over 32 ms frames in which the clean signal is active and limited to −10 to 35 dB; and (iii) an intelligibility estimate, called here the *STOI-proxy*. The STOI-proxy follows the structure of STOI [8]: 15 one-third-octave band envelopes from 150 Hz are computed for the clean and processed signals, and the correlation between them is calculated over 30-frame segments after normalisation and clipping, then averaged over bands and segments. It is a simplified re-implementation, not the reference STOI code, so its values should be compared with each other and not with published STOI scores. Higher values indicate better predicted intelligibility (maximum 1).

IV RESULT AND DISCUSSION

Fig. 2 shows the signal energy analysis for one example utterance in machinery noise at 5 dB. The smoothed energy follows the syllabic structure, while the tracked floor stays near the energy level of the leading silence and the shortest pauses and moves only slightly during speech. Frames above the floor by more than 3 dB are flagged as speech-present. The flag is conservative: short gaps between syllables often remain flagged because the noise itself keeps the energy above the margin. As a result, the noise estimate is refreshed mainly in the leading and trailing silences and in longer pauses, and is otherwise held almost constant, which is the intended behaviour for slowly varying noise but would respond slowly to a sudden change in the noise level.

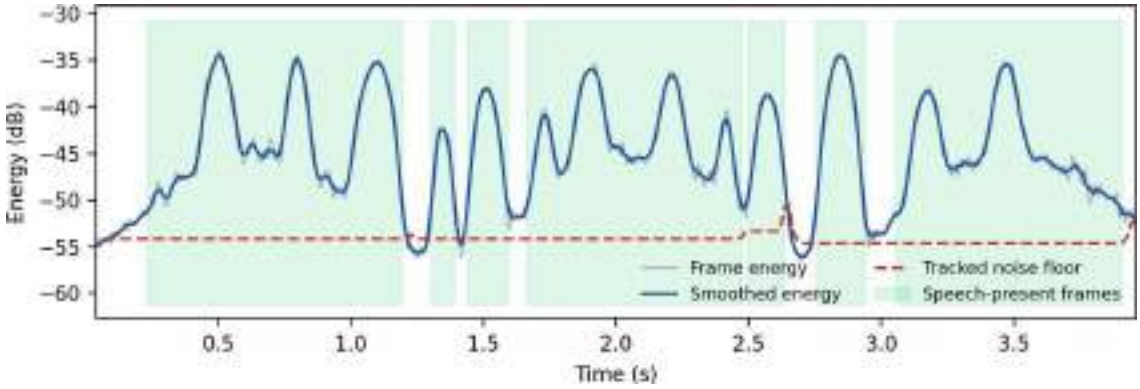


Fig.2 Frame energy, tracked noise floor and speech-present decision (machinery noise, 5 dB)

Fig. 3 compares spectrograms of the clean, noisy and enhanced versions of the same utterance. In the noisy signal the low-frequency harmonics of the vowels in the 200 to 1500 Hz range and the first formants are partly covered by the machinery noise. After enhancement the harmonic stripes and formant bands are visibly clearer and the broadband background is lowered, while the fricative bursts at high frequency are preserved. Residual noise is still visible in the low-frequency region, because the gain floor is deliberately kept above zero in speech frames.

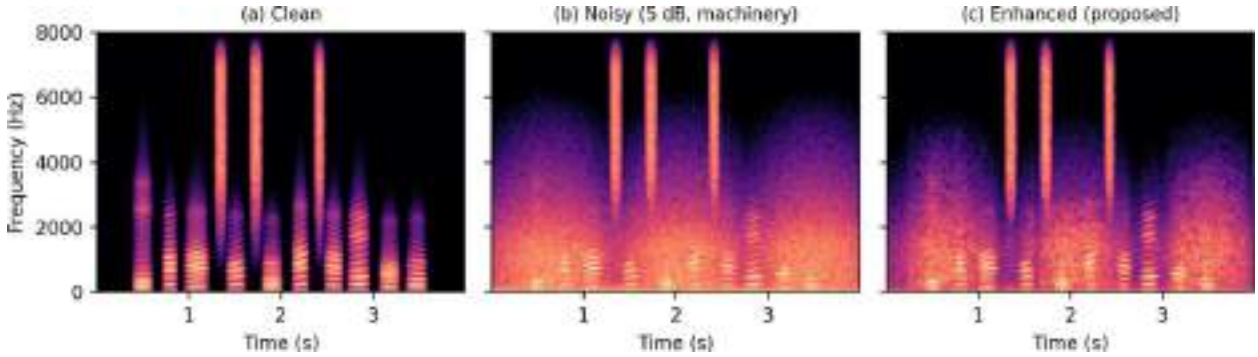


Fig.3 Spectrograms: (a) clean, (b) noisy, (c) enhanced by the proposed method

The intelligibility results are in Table 1 and Fig. 4. The proposed method has the highest STOI-proxy in all nine conditions. Averaged over noise types, its gain over the unprocessed input is **0.045**, **0.036** and **0.021** at 0, 5 and 10 dB, so the benefit is largest in the most difficult condition and shrinks as the input becomes cleaner. The standard deviation of the proposed method's score over the ten utterances did not exceed 0.023 in any condition, so the ordering is consistent across utterances, although the differences are small in absolute terms and should be read as indicative.

Noise	Input SNR	Noisy	Spectral subtraction	Static Wiener	Proposed
White	0 dB	0.745	0.733	0.740	**0.802**
White	5 dB	0.794	0.785	0.796	**0.847**
White	10 dB	0.835	0.828	0.834	**0.868**
Pink	0 dB	0.730	0.724	0.727	**0.787**
Pink	5 dB	0.794	0.794	0.800	**0.834**
Pink	10 dB	0.847	0.848	0.849	**0.869**
Machinery	0 dB	0.701	0.683	0.687	**0.720**

Machinery	5 dB	0.768	0.758	0.761	**0.783**
Machinery	10 dB	0.829	0.822	0.823	**0.836**

Table 1. STOI-proxy (mean of 10 utterances; best value in each row in bold)

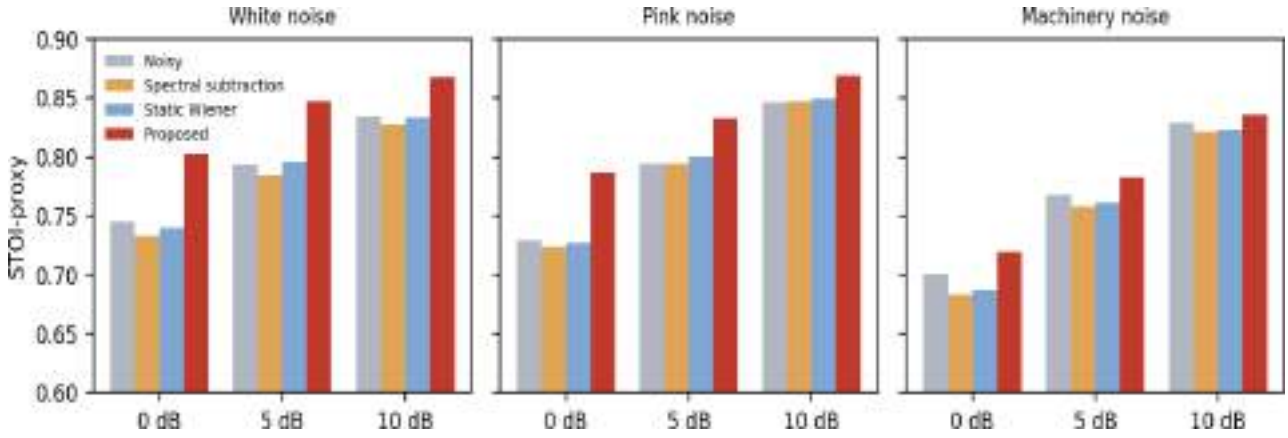


Fig.4 STOI-proxy by noise type, input SNR and method

A different picture appears in the SNR measures (Table 2). Both baselines improve SNR more than the proposed method: the average SNR gains over the three noise types were 7.4, 5.9 and 5.2 dB for static Wiener, 5.4, 4.8 and 4.2 dB for spectral subtraction, and 4.6, 3.7 and 2.2 dB for the proposed method at 0, 5 and 10 dB. Segmental SNR, which is evaluated only where speech is active, follows the same order. At the same time the baselines give a STOI-proxy within about ±0.012 of the unprocessed signal, and for machinery noise it is lower than that of the unprocessed signal in every condition. In other words, removing more noise power did not make the speech envelope more faithful, which agrees with the observations in [13] and [14] that SNR-oriented gain rules distort low-energy speech segments. The proposed method gives up some noise suppression through its higher gain floor in speech frames and its smoothing, and in return preserves the envelope structure that the intelligibility measure rewards.

Input SNR	Metric	Spectral subtraction	Static Wiener	Proposed
0 dB	SNR gain (dB)	5.40	7.44	4.64
	Segmental SNR (dB)	3.39	5.39	2.41
5 dB	SNR gain (dB)	4.85	5.94	3.70
	Segmental SNR (dB)	7.42	8.92	6.03
10 dB	SNR gain (dB)	4.18	5.16	2.19
	Segmental SNR (dB)	11.68	12.87	9.40

Table 2. SNR gain and segmental SNR averaged over the three noise types

Several limitations should be noted. First, the test signals are synthetic. They have realistic harmonic and formant structure but lack the variability of real speakers, coarticulation, and the weak consonants that matter most for word recognition, so the absolute values and possibly the ranking could change on recorded speech. Second, the STOI-proxy is a simplified measure and no listening tests were carried out, so no claim is made about human word recognition. Third, the baselines use a fixed noise estimate from the leading silence, which favours them for stationary noise; a comparison with a tracking baseline such as minimum statistics [6] or improved minima-controlled recursive averaging [7] would be a fairer test. Fourth, residual noise remains audible and visible after processing, and the energy-based speech flag reacts slowly to abrupt changes in the noise level. Finally, no real low-light recordings or audio-visual measurements were used; “low-light” here defines the intended application (audio-only communication) rather than a property of the signals. Within these limits, the results suggest that adapting the filter’s aggressiveness to the estimated frame condition is a useful way to favour intelligibility over raw noise reduction. The method is training-free, has few parameters, and its per-frame cost is low, which supports its use in embedded, low-power or real-time systems.

V. CONCLUSION

This study presented a lightweight speech enhancement method for low-light, audio-only communication that couples signal energy analysis with adaptive spectral filtering. Frame energy is used to track a noise floor and to flag speech-present frames, and this information adapts the noise estimate, the over-subtraction factor, the decision-directed memory and the gain floor of a Wiener-type filter. On synthetic speech-like signals in white, pink and machinery noise at 0, 5 and 10 dB, the method gave the highest STOI-proxy in all nine conditions, with an average gain of 0.021 to 0.045 over the unprocessed input, while spectral subtraction and a static Wiener filter stayed near the input level despite larger SNR gains. The results confirm that SNR improvement and intelligibility improvement are different goals and that gain limits matter for the second one.

The method is sensitive to its parameters, leaves residual low-frequency noise, and has been tested only on synthetic material with a simplified intelligibility measure. Future work should evaluate the method on recorded speech corpora with the reference STOI and PESQ implementations, add tracking baselines, run listening tests, extend the noise tracker to respond faster to abrupt changes, and examine combination with learned noise estimators or audio-visual cues where some light is available.

REFERENCES

[1] Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2), 113–120.

[2] Berouti, M., Schwartz, R., & Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 208–211.

[3] Lim, J. S., & Oppenheim, A. V. (1979). Enhancement and bandwidth compression of noisy speech. *Proceedings of the IEEE*, 67(12), 1586–1604.

[4] Ephraim, Y., & Malah, D. (1984). Speech enhancement using a minimum mean-square error short-time spectral amplitudeestimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6), 1109–1121.

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

[6] Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing*, 9(5), 504–512.

[7] Cohen, I. (2003). Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Transactions on Speech and Audio Processing*, 11(5), 466–475.

[8] Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2011). An algorithm for intelligibility prediction of time–frequencyweighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7), 2125–2136.

[9] Loizou, P. C. (2013). *Speech Enhancement: Theory and Practice* (2nd ed.). CRC Press.

[10] Scalart, P., & Filho, J. V. (1996). Speech enhancement based on a priori signal to noise estimation. *Proc. IEEE ICASSP*, 629–632.

[11] Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10), 1702–1726.

[12] Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 26(2), 212–215.

[13] Loizou, P. C., & Kim, G. (2011). Reasons why current speech-enhancement algorithms do not improve speech intelligibilityand suggested solutions. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1), 47–56.

[14] Hu, Y., & Loizou, P. C. (2007). Subjective comparison and evaluation of speech enhancement algorithms. *Speech Communication*, 49(7–8), 588–601.

[15] Pascual, S., Bonafonte, A., & Serrà, J. (2017). SEGAN: Speech enhancement generative adversarial network. *Proc. Interspeech*, 3642–3646.