INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Speech enhancement based on harmonic estimation combined with MMSE to improve speech intelligibility for cochlear implant recipients

Dongmei Wang, John H. L. Hansen

CRSS Lab, Dept. Electrical Engineering
The University of Texas at Dallas
Richardson, Texas 75080
dongmei.wang@utdallas.edu, john.hansen@utdallas.edu

Abstract

In this paper, a speech enhancement algorithm is proposed to improve the speech intelligibility for cochlear implant recipients. Our method is based on combination of harmonic estimation and traditional statistical method. Traditional statistical based speech enhancement method is effective only for stationary noise suppression, but not non-stationary noise. To address more complex noise scenarios, we explore the harmonic structure of target speech to obtain a more accurate noise estimation. The estimated noise is then employed in the MMSE framework to obtain the gain function for recovering the target speech. Listening test experiments show a substantial speech intelligibility improvement for cochlear implant recipients in noisy environments.

Index Terms: speech enhancement, cochlear implant, harmonic structure, noise estimation, MMSE

1. Introduction

Cochlear implant (CI) devices are able to assist the individuals with severely hearing loss to recover some level of hearing ability. Currently, CI recipients are able to achieve a relatively high level of speech intelligibility in quiet environments. However, when they present in noisy backgrounds, their speech intelligibility drops dramatically. Previous research has shown that the speech reception threshold (SRT) of CI listeners is typically 15 to 25dB higher than normal hearing listeners in noisy backgrounds [1, 2, 3]. Therefore, developing effective speech enhancement algorithms is essential to improve speech perception in noise for CI recipients.

Speech is a type of over-redundant signal. Normal hearing listeners usually have no difficulties in understanding speech in noise, even at a negative signal-to-noise ratio (SNR) level, depending on noise types. On the contrary, CI recipients are only able to decode limited amounts of spectral and temporal information from the speech which are delivered by a few CI encoding channels (16 or 22) [4, 5]. This leaves CI recipients unable to discriminate target speech from noise interference.

To improve speech perception in noise for CI recipients, both single and multiple microphone based speech enhancement algorithms have been developed in previous research [6, 7]. Multiple microphone speech enhancement algorithms are able to significantly increase speech intelligibility for CI recipients in both stationary and non-stationary noise. However, strict assumptions such as speech and noise sources must be spatially separated, have to be satisfied. In some real scenarios, such as a diffusive sound field, target speech and noise coming from the same direction, or if reverberation exists, multiple channel algorithms have very limited benefit. Thus, single microphone speech enhancement to improve speech perception for CI recipients still remains as an open problem. In addition, single microphone algorithm can incorporate with multiple microphone algorithms to improve speech intelligibility in noise [8].

For single microphone based methods, existing research has been focused on i) developing noise reduction algorithm before the CI encoding [1, 9, 10, 11, 12] and ii) optimizing the CI channel selection to deliver higher SNR speech components [13, 14, 15, 16, 17]. In the former cases, noise reduction is performed as a pre-processing to improve speech representation. Such approaches include spectral subtraction method [1, 9], wiener filter [18], subspace methods [10]. In the latter cases, strategies are designed to manipulate the CI encoding channels to efficiently deliver the speech signal to the electrode array for the improved auditory neural stimulation. Both binary mask [14] and soft mask [13, 15, 16] based studies have demonstrated that SNR-dependent channel selection is more desirable than the current default n-of-m channel selection strategy [19]. Non-negative matrix factorization (NMF) method has also been used in channel selection for CI device based on speech sparsity characteristics [17].

Traditional statistical based speech enhancement algorithms are desirable for CI devices because of the real time processing [20]. However, these methods are only effective for stationary noise. Their benefits are moderated or disappear in the fluctuating noises. In this study, we propose a single microphone speech enhancement algorithm based on harmonic estimation combined with MMSE to improve speech intelligibility for CI subjects. On one hand, MMSE method is efficient to reduce the stationary noise along time dimension. On the other hand, we explore the harmonic structures of target speech to accurately estimate the fluctuating noise along frequency

http://dx.doi.org/10.21437/Interspeech.2017-78

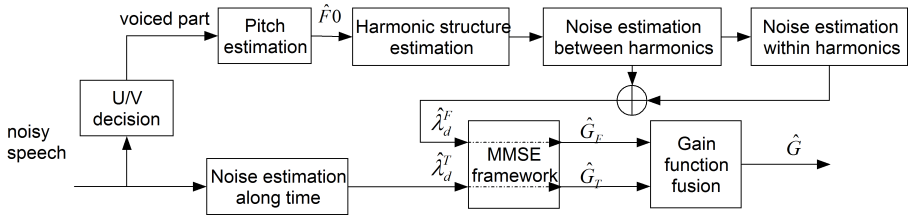


Figure 1: *Algorithm Overview*

dimension. The noise estimation from both the time and frequency dimensions are employed in the MMSE framework for speech enhancement [21]. Moreover, the harmonic structure estimation is based on the noise robust pitch estimation from our previous study [22, 23].

2. The Proposed Speech Enhancement Algorithm

2.1. Algorithm Overview

The overall algorithm overview is shown in Fig. 1. Two types of noise estimation are included: i) noise estimation along the frequency dimension in the voiced speech segments, ii) noise estimation along the time dimensions in both voiced and unvoiced segments. Specifically, along time dimension, the assumption of stationary noise is made during noise tracking. Along the frequency dimension, the noise estimation is based on exploring harmonic structure. The estimated noise variance from both frequency and time dimensions ($\hat{\lambda_d^T}$ and $\hat{\lambda_d^F}$) will be inputting to the MMSE framework to obtain the gain functions ($\hat{G_T}$ and $\hat{G_F}$). Finally, the above two gain functions will be fused into one form ($\hat{G}$) for the target speech estimation based on the MMSE [21].

2.2. Harmonic structure estimation

Speech spectrum of voiced speech is comprised of a series of harmonic partials with frequencies which are integer multiple times of fundamental frequency (F0). Given the estimated F0s of the target speech, the harmonic structures can be approximated by selecting the noisy spectrum peaks which are near the ideal harmonic frequencies ($kF0$). In practice, the observed harmonic partials usually deviates from the ideal frequency due to the instability of the glottal pulse sequence/shape during speech production. Therefore, we set the deviation threshold $\triangle f_H$ differently depending on the frequency band, shown as follows,

$$\Delta f_H = \begin{cases} 20, & f < 500Hz \\ 30, & 500Hz \leq f < 2000Hz \\ 45, & f \geqslant 2000Hz \end{cases} \quad (1)$$

The F0 estimation for noisy speech is based on harmonic feature classification based methods [22, 23]

2.3. Noise estimation based on harmonic structure

Based on harmonic model, we infer that: i) the noisy spectrum between the harmonic partials are usually dominated by noise; ii) the noisy spectrum within the harmonic partials are dominated by speech. Accordingly, we perform the noise estimation between harmonics (BH) and within harmonics (WH) separately.

First, harmonic spectrum for the target speech is generated by convolving the harmonic peak vector with the spectrum of a short-term analysis hamming window, shown as follow,

$$\boldsymbol{S_H}(f) = \boldsymbol{S_{win}}(f) * \sum_{k=1}^{K} a_H^k \cdot \boldsymbol{\delta}(f - f_H^k), \quad (2)$$

where a_H^k and f_H^k are the amplitude and frequency of the kth order harmonic peak, $\boldsymbol{S_{win}}$ is the spectrum vector of a hamming window, and $\boldsymbol{\delta}(f)$ is a delta function. Next, the generated harmonic spectrum amplitude $|\boldsymbol{S_H}|$ are reduced from the noisy speech spectrum to obtain the initial estimated noise spectrum $\hat{\boldsymbol{A}}_{\boldsymbol{n}}^{\boldsymbol{0}}$, shown as below,

$$\hat{\boldsymbol{A}}_{\boldsymbol{n}}^{\boldsymbol{0}} = \max\left(|\boldsymbol{S_n}| - |\boldsymbol{S_H}|, 0\right). \quad (3)$$

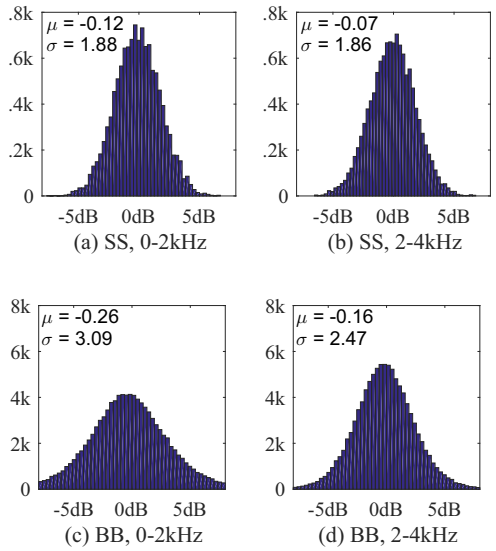
Here, we call the noise spectrum within the harmonics main lobe as "WH noise", and the noise spectrum between the harmonics main lobes as "BH noise". The bandwidth of the main lobe for each harmonics is set as $2/3$ of that of the corresponding short-term hamming window spectrum.

In the BH frequency range, noise is the dominant component which is seldom overlapped with speech. Thus, the initial estimated noise $\hat{\boldsymbol{A}}_{\boldsymbol{n}}^{\boldsymbol{0}}$ in this frequency range is considered as the estimated noise spectrum, shown as,

$$\hat{\boldsymbol{A}}_{\boldsymbol{BH}}(f) = \hat{\boldsymbol{A}}_{\boldsymbol{n}}^{\boldsymbol{0}}(f), \quad (4)$$

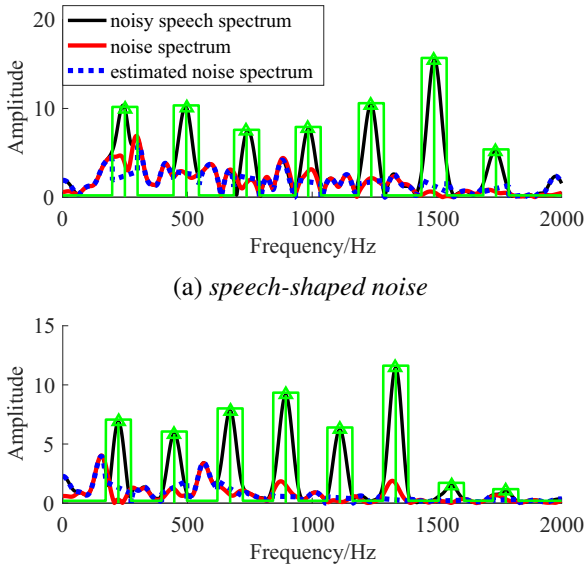
where $f \in \left[kF0 + \frac{1}{2}f_{mb}, (k+1)F0 - \frac{1}{2}f_{mb}\right]$, and f_{mb} is the bandwidth of the harmonic main lobe. However, in the WH frequency range, speech has the dominant energy. The initiated estimated noise spectrum in this frequency range is not as accurate which requires re-estimation.

To estimate the WH noise, we made an assumption that noise spectrum is continuously distributed along the frequency dimension. Given the noise variance in the



(a) SS, 0-2kHz (b) SS, 2-4kHz (c) BB, 0-2kHz (d) BB, 2-4kHz

Figure 2: *Histogram of energy ratio between neighboring frequency bands. SS: speech-shaped noise, BB: babble noise*



(a) *speech-shaped noise*

(b) *babble noise*

Figure 3: *Noise estimation based on harmonic structure*

near frequency bands, the noise variance in the current frequency band can be approximated based on interpolation technique. Fig. 2 presents the statistical histogram of the logarithmic energy ratio between neighboring frequency bands. Both the bandwidth and the frequency shift are set to 100Hz. We include two different frequency ranges: 0-2kHz and 2k-4kHz. Two types of noise are considered: speech-shaped noise and babble noise. The mean and standard variance values are shown along with the histogram. From Fig. 2 we see that the mean values of the logarithmic energy ratio between neighboring bands are around 0dB in all four cases. The maximum spread value for both types of noises is less than 5dB. This analysis indicates the continuity of the noise energy between adjacent frequency bands.

Therefore, we propose to estimate the WH noise variance using the estimated BH noise variance with linear interpolation method, shown as below,

$$\hat{\boldsymbol{A}}_{\boldsymbol{WH}}(f) = \hat{a}^L_{BH} + \left(\hat{a}^R_{BH} - \hat{a}^L_{BH}\right) \cdot \frac{f - f^L_{BH}}{f^R_{BH} - f^L_{BH}} \quad (5)$$

where $f \in \left[kF0 - \frac{1}{2}f_{mb}, kF0 + \frac{1}{2}f_{mb}\right]$, f^L_{BH} and f^R_{BH} are the edge frequencies of left and right neighboring BH band near the current harmonic partial, $\hat{a}^L_{BH}$ and $\hat{a}^R_{BH}$ are the average amplitude of estimated BH noise spectrum in left and right adjacent frequency band.

Fig. 3 demonstrates an example of the noise estimation based on harmonic structure. Fig. 3a is for the speech-shaped noise case, and Fig. 3b is for the babble noise case. It can be seen that both the BH and WH noise estimation is almost consistent as the original noise.

2.4. Noise tracking along time dimension

The noise tracking along time dimension is performed based on a minimum statistics algorithm with optimal smoothing [24], assuming noise is stationary. In practice, noise variance is estimated in the beginning quiet section and updated during later unvoiced segments.

2.5. Gain function estimation

The noise variance estimated along frequency dimension is employed in the MMSE framework to generate the harmonic-based gain function $\hat{G_F}$ [25, 20]. Alternatively, the time dimension estimated noise is used to generate a gain function $\hat{G_T}$.

Then $\hat{G_F}$ and $\hat{G_T}$ are combined with weights to obtain the optimal gain function for the target speech [21], shown as,

$$\hat{G} = \hat{G_F} \cdot \frac{\hat{\lambda_F}}{\hat{\lambda_F} + \hat{\lambda_T}} + \hat{G_T} \cdot \frac{\hat{\lambda_T}}{\hat{\lambda_F} + \hat{\lambda_T}}. \quad (6)$$

where $\hat{\lambda_F}$ and $\hat{\lambda_T}$ are the frequency- and time-dimension based noise variances which are computed from Sec. 3.3 and Sec. 3.4.

3. Experiments and Results

3.1. Experiment setting

Listening test is carried out with CI subjects to evaluate the performance of the proposed algorithm. The target speech materials are comprised of sentences from the IEEE database [26]. Babble noise was used to corrupt the sentences at the SNR of 0dB, 5dB and 10dB. Six post-lingually deafened cochlear implants users, with a mean

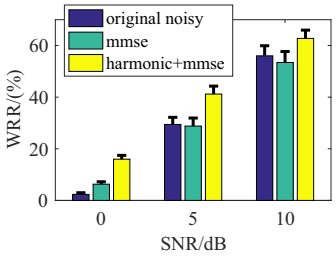


Figure 4: *Average word recognition rate results*

of 7.5 years implant use, participated in the listening test. Subjects were paid an hourly wage for their participation.

The listening task involved sentence recognition by CI subjects who were seated in a soundproof room (Acoustic System, Inc). The sentences were played to the CI subjects through a loudspeaker placed at a distance of 80cm in front of the subject. The sound pressure level of the speech sentences from the loudspeaker was set as fixed 65dB through out the test. The subjects were fitted with their daily strategy during the test.

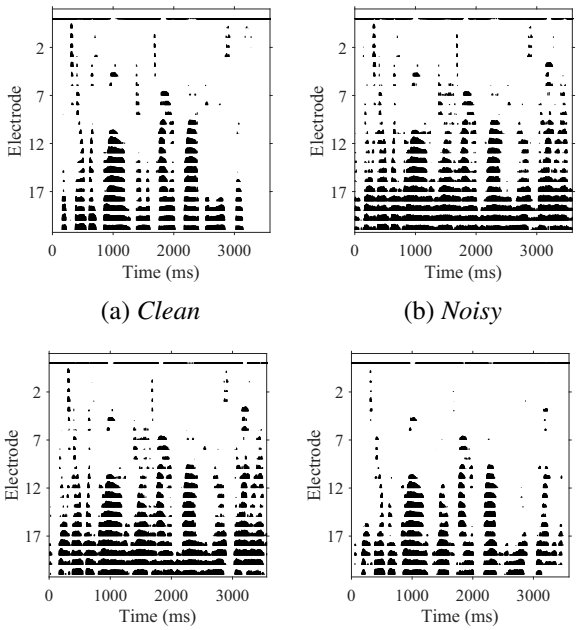
Each subject participated in a total of 9 test conditions (3 SNR levels $\times$ 3 processing conditions). Two IEEE sentence lists were used per test condition, and each IEEE sentence list includes 10 sentences. The order of test conditions was randomized across subjects. Subjects were given a 5-min break every 30 min during the test sessions to avoid listening fatigue.

3.2. Results

The speech intelligibility performance of CI subjects is measured in terms of average word recognition rate (WRR). The results of average WRR in different conditions are presented in Fig. 4. The standard error of the mean (SEM) of WRR results is also shown along with the average value. From Fig. 4 we see that the Harmonic + MMSE approach shows benefits for CI subjects at all SNR levels in the babble noise condition. In contrast, the MMSE processing only shows minor benefit at the SNR of 0dB.

We also carry out a two-way ANOVA on the WRR results to investigate the significance of processing conditions at different SNR levels. The ANOVA results for 0dB and 5dB are $[F(2,17) = 21.26, p < 0.0003]$ and $[F(2,17) = 16.21, p < 0.0007]$ respectively indicating significant interaction between different processing conditions. However, the ANOVA result for 10dB is $[F(2,17) = 2.67, p < 0.1181]$, indicating no significant difference between different processing conditions.

The *Post hoc* comparisons between different processed conditions are carried out at 0dB and 5dB which have been shown significant statistics by ANOVA analysis. Specifically, when the SNR is 0dB, the *Post hoc* results show significant difference between MMSE + Harmonic processed condition and the original noisy condition. Significant difference is also shown between MMSE + Harmonic processed condition and the only MMSE processed condition ($p < 0.0002$ and $p < 0.003$). However, there is no significant difference between MMSE processed condition and the original unprocessed condition ($p < 0.2078$). When the SNR is 5dB, the similar statistic results are found as 0dB.



(a) *Clean* (b) *Noisy*

(c) *MMSE* (d) *Harmonic + MMSE*

Figure 5: *Electrodogram: babble noise, SRN = 0dB*

Fig. 5 shows an example of electrodogram for different processing conditions. Electrodogram represents the output of the cochlear implant devices. Comparing the electrodograms among different processing conditions, we see that much more residual noise exists in MMSE processed speech than Harmonic+MMSE processed speech. Harmonic structure estimation is able to provide more accurate noise tracking, thus removing the fluctuating noise, whereas MMSE method is not as effective.

4. Conclusion and discussion

A speech enhancement algorithm based on harmonic estimation combined with MMSE framework is proposed in this paper to improve the speech intelligibility for CI recipients. The proposed Harmonic + MMSE processing is able to distinguish speech harmonics from noise interference. In this way, the non-stationary noise can be estimated and removed accurately. The listening test results demonstrate the potential benefit to CI subjects by the proposed method in terms of the WRR performance. It further indicates that F0 is a substantial cue for speech perception in fluctuating noise for CI recipients

For the future research, the investigation of tradeoff between noise reduction and speech distortion is warranted for different noise condition regarding speech intelligibility for CI recipients

5. References

[1] Hochberg I., A. Boothroyd, and M Weiss, "Effects of noise and noise suppression on speech perception by cochlear implant users," *Ear and Hearing*, vol. 13, no. 4, pp. 263–271, Aug. 1992.

[2] J. Wouters and J. Vanden Berghe, "Speech recognition in noise for cochlear implantees with a two-microphone monaural adaptive noise reduction system," *Ear and Hearing*, vol. 22, no. 5, pp. 420–430, Oct. 2001.

[3] A. Spriet, L. Van Deun, K. Eftaxiadis, J. Laneau, M. Moonen, B. van Dijk, A. van Wieringen, and Wouters J., "Speech understanding in background noise with the two-microphone adaptive beamformer beam in the nucleus freedom cochlear implant system," *Ear and Hearing*, vol. 28, no. 1, pp. 62–72, Feb. 2007.

[4] J. F. Patrick, P. A. Busby, and P. J. Gibson, "The development of the nucleus freedom cochlear implant system," *Trends Amplify*, vol. 10, no. 4, pp. 175–200, Dec. 2006.

[5] P. C. Loizou, "Speech processing in vocoder-centric cochlear implants," in *Cochlear and Brainstem Implants, Otorhinolaryngol*, A. R. Moller, Ed., vol. 64, pp. 109–143. Karger, Basel, 2006.

[6] K. Kokkinakis, B. Azimi, Y. Hu, and D. R. Friedland, "Single and multiple microphone noise reduction strategies in cochlear implants," *Trends in Amplification*, vol. 16, no. 2, pp. 102–116, Jun. 2012.

[7] Raphael Koning, *Speech enhancement in cochlear implants*, Ph.D. thesis, Dept. Neurosciences, KU Leuven, 2014.

[8] A. A. Hersbach, K. Arora, Mauger S. J., and Dawson P. W., "Combining directional microphone and single-channel noise reduction algorithms: a clinical evaluation in difficult listening conditions with cochlear implant users," *Ear and Hearing*, vol. 33, no. 4, pp. e13–e19, July - Aug. 2012.

[9] L. P. Yang and Q. J. Fu, "Spectral subtraction-based speech enhancement for cochlear implant patients in background noise," *J. Acoust. Soc. Am.*, vol. 117, no. 3 Pt. 1, pp. 1001–1004, Nov. 2005.

[10] P. C. Loizou, A. Lobo, and Y. Hu, "Subspace algorithms for noise reduction in cochlear implants," *J. Acoust. Soc. Am.*, vol. 118, no. 5, pp. 2791–2793, Nov. 2005.

[11] J. Li, Q. J. Fu, H. Jiang, and M. Akagi, "Psychoacoustically-motivated adaptive beta-order generalized spectral subtraction for cochlear implant patients," in *Proc. ICASSP*, Taipei, Taiwan, Ari. 2009, pp. 4665–4668.

[12] F. Toledo, P. C. Loizou, and A. Lobo, "Subspace and envelope subtraction algorithms for noise reduction in cochlear implants," in *Proc.25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Taipei, Taiwan, Ari. 2003, pp. 2002–2005.

[13] Y. Hu, P. C. Loizou, N. Li, and K. Kasturi, "Use of a sigmoid-shaped function for noise attenuation in cochlear implants," *J. Acoust. Soc. Am.*, vol. 122, no. 4, pp. EL128–EL134, Oct. 2007.

[14] Y. Hu and P. C. Loizou, "Environment-specific noise suppression for improved speech intelligibility by cochlear implant users," *J. Acoust. Soc. Am.*, vol. 127, no. 6, pp. 3689–3695, June 2010.

[15] P. W. Dawson, S. J. Mauger, and A. A. Hersbach, "Clinical evaluation of signal-to-noise ratio based noise reduction in nucleus cochlear implant recipients," *Ear and Hearing*, vol. 32, no. 3, pp. 382–390, May-June 2011.

[16] S. J. Mauger, K. Arora, and P. W. Dawson, "Cochlear implant optimized noise reduction," *J. Neural Eng.*, vol. 9, no. 6, pp. 1–9, Dec. 2012.

[17] H. Hu, M. E. Lutman, S. D. Ewert, G. Li, and S. Bleeck, "Sparse nonnegative matrix factorization strategy for cochlear implants," *Trends in Hearing*, vol. 19, pp. 1–16, Dec. 2015.

[18] F. Chen, Y. Hu, and M. Yuan, "Evaluation of noise reduction methods for sentence recognition by mandarin-speaking cochlear implant listeners," *Ear and Hearing*, vol. 36, no. 1, pp. 61–71, Jan. 2015.

[19] Y. Hu and P. C. Loizou, "A new sound coding strategy for suppressing noise in cochlear implants," *J. Acoust. Soc. Am.*, vol. 124, no. 1, pp. 498–509, July 2008.

[20] P. C. Loizou, *Speech Enhancement: Theory and Practice*, chapter 7, CRC Press, Boca Raton, USA, 1 edition, 2007.

[21] M. Krawczyk-Becker and T. Gerkmann, "MMSE-optimal combination of wiener filtering and harmonic model based speech enhancement in general framework," in *Proc. WASPAA*, New Paltz, NY, Dec. 2015, pp. 1–5.

[22] D. Wang, P. C. Loizou, and J. H. L. Hansen, "F0 estimation in noisy speech based on long-term harmonic feature analysis combined with neural network classification," in *Proc. INTERSPEECH*, Singapore, Sep. 2014, pp. 2258–2262.

[23] D. Wang, C. Yu, and J. H. L. Hansen, "Robust harmonic features for classification-based pitch estimation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 5, pp. 952–964, May 2017.

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

[25] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustic Speech and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[26] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust*, vol. 17, no. 3, pp. 225–246, Sep. 1969.