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

Improved Codebook-based Speech Enhancement based on MBE Model

Qizheng Huang, Changchun Bao, Xianyun Wang

Speech and Audio Signal Processing Laboratory, Faculty of Information Technology, Beijing University of Technology, Beijing, China, 100124

huangqizheng@emails.bjut.edu.cn, baochch@bjut.edu.cn, b201402001@emails.bjut.edu.cn

Abstract

This paper provides an improved codebook-based speech enhancement method using multi-band excitation (MBE) model. It aims to remove the noise between the harmonics, which may exist in codebook-based enhanced speech. In general, the proposed system is based on analysis-with-synthesis (AwS) framework. During the analysis stage, acoustic features are extracted including pitch, harmonic magnitude and voicing from noisy speech. These parameters are obtained on the basis of the spectral magnitudes obtained by codebook-based method. During the synthesis stage, different synthesis strategies for voiced and unvoiced speech are employed. Besides, this paper introduces speech presence probability to modify the codebook-based Wiener filter so that more accurate acoustic parameters can be obtained. The proposed system can eliminate noise not only between the harmonics, but also in the silent segments, especially in low SNR noise environment. Experiments show that, the performance of the proposed method is better than traditional codebook-based method for different types of noise.

Index Terms: speech enhancement, MBE model, codebooks, analysis-with-synthesis, acoustic features

1. Introduction

Speech enhancement plays an important role in speech communication due to its wide applications. But it still faces enormous challenges because of the variability of noise. Conventional speech enhancement methods such as spectral subtraction [1], Wiener filter [2] and statistical model-based methods [3] perform well for the stationary noises, but the performance of them degrades rapidly in the non-stationary noise environment.

Codebook-based method stores the prior information of speech and noise spectral envelope in the codebooks. The concerned short-term predictor (STP) parameters including AR coefficient and AR gain are estimated online. In [4, 5], STP parameters are first estimated by maximum-likelihood and Bayesian MMSE method, respectively. Then Wiener filter is constructed using the estimated parameters. The codebooks contain a priori information about speech and noise, therefore the enhanced speech of the codebook-based method has a smaller spectral distortion, especially in non-stationary noise conditions. But the codebook-based method is only used to model the spectral envelope, so there is a lot of residual noise between adjacent harmonics. Furthermore, there is a large amount of residual noise in the silent segments due to the inaccuracy of STP parameters.

In recent years, the analysis-with-synthesis (AwS) method is a hotspot of speech enhancement. It has two stages, namely analysis stage and synthesis stage. During the analysis stage, acoustic features are extracted from speech signals, such as pitch and harmonic magnitude. During the synthesis stage, enhanced speech is synthesized by the features. Compared with conventional speech enhancement algorithms, AwS methods can preserve the harmonic structure better and eliminate more noises. In [6], a speech enhancement method which is based on pure speech features reconstruction is proposed. The system uses the maximum a posteriori method to estimate clean spectral envelope from noisy spectral envelope, and the enhanced speech is synthesized using sinusoidal model. In addition, harmonic plus noise model [7] also uses the AwS framework. It extracts four auditory parameters including pitch, spectral envelope, spectral gain, and voicing mixed function. The spectral envelope is obtained by tracking line spectrum frequency trajectory through Kalman filtering. Among the AwS methods, MBE is a significant model. It takes into account the different way of speech production for voiced portion and unvoiced portion, so high quality speech can be synthesized.

This paper is based on the MBE model [8] and codebook-based methods [4, 5]. The proposed system uses AwS framework and aims at removing noise between the harmonics and generating clean harmonics. At first, the linear prediction (LP) parameter codebooks of speech and noise are trained offline. Then the initial enhancement spectrum is obtained by codebook-based method. Subsequently, the acoustic features, such as pitch, harmonic magnitude and voiced/unvoiced (V/UV) decision are extracted from the enhanced spectrum. Finally, the voiced and unvoiced speech signals are synthesized by different strategies.

The remainder of the paper is organized as follows. Section 2 explains the signal model and Bayesian codebook-based speech enhancement method. Section 3 explains the proposed framework in detail. Experiments and results are shown in Section 4 and finally, Section 5 concludes the paper.

2. Codebook-based Bayesian method

We consider an additive noise model, where speech and noise are independent,

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

where $y(n)$, $x(n)$ and $w(n)$ represent the sampled noisy speech, clean speech and noise, respectively. As a priori information, the spectral shape codebooks of clean speech and noise are trained offline in advance, respectively.

The aim of codebook-based method is to estimate STP parameters, namely LP coefficient and excitation variance. Assuming $\theta_x = (\alpha_{x_0},...,\alpha_{x_p})$ and $\theta_w = (\alpha_{w_0},...,\alpha_{w_q})$ are the LP coefficients of clean speech and noise with p and q

http://dx.doi.org/10.21437/Interspeech.2017-109

being the respective LP-model orders. Let $\theta=[\theta_x,\theta_w,\sigma_x^2,\sigma_w^2]$, where σ_x^2 and σ_w^2 are the excitation variances of clean speech and noise.

The STP parameter vector θ can be estimated by [5]

$$\hat{\theta}=\frac{1}{N_x N_w}\sum_{i,j=1}^{N_x,N_w}\theta_{ij}^{'}\frac{p(\mathbf{y}\mid\theta_x^i,\theta_w^j,\sigma_{x,ij}^{2,ML},\sigma_{w,ij}^{2,ML})p(\sigma_{x,ij}^{2,ML})p(\sigma_{w,ij}^{2,ML})}{p(\mathbf{y})} \tag{2}$$

where $\mathbf{y}=[y(0)y(1)...y(N-1)]^T$ denotes the observed vector of noisy samples for the current frame and N is the frame length. $p(\mathbf{y})$ serves as a normalization term and can be obtained as

$$p(\mathbf{y})=\frac{1}{N_x N_w}\sum_{i,j=1}^{N_x,N_w}p(\mathbf{y}\mid\theta_x^i,\theta_w^j,\sigma_{x,ij}^{2,ML},\sigma_{w,ij}^{2,ML})p(\sigma_{x,ij}^{2,ML})p(\sigma_{w,ij}^{2,ML}) \tag{3}$$

where N_x, N_w are the sizes of speech and noise codebook, respectively. The modeled noisy spectrum is defined as

$$\hat{P}_y=\hat{\sigma}_x^2/\mid\hat{A}_x(\omega_k)\mid^2+\hat{\sigma}_w^2/\mid\hat{A}_w(\omega_k)\mid^2 \tag{4}$$

where $\hat{A}_x(\omega_k)$ and $\hat{A}_w(\omega_k)$ are the spectrum corresponding to $\hat{\theta}_x$, $\hat{\theta}_w$, respectively. They are given by

$$\hat{A}_x(\omega_k)=\sum_{k=0}^{p}\alpha_{x_k}e^{-j\omega k},\ \hat{A}_w(\omega_k)=\sum_{k=0}^{q}\alpha_{w_k}e^{-j\omega k} \tag{5}$$

Using the equivalence of the log-likelihood and the Itakura-Saito distortion measure, we can obtain

$$p(\mathbf{y}\mid\theta_x^i,\theta_w^j,\sigma_{x,ij}^{2,ML},\sigma_{w,ij}^{2,ML})=C\exp(-d_{IS}(P_y,\hat{P}_y^{ij,ML})) \tag{6}$$

where C is a constant which has no influence on computing STP parameters. And the IS measure between the observed noisy spectral envelope and estimated noisy spectral envelope is defined as

$$d_{IS}(P_y,\hat{P}_y)=\frac{1}{2\pi}\int_0^{2\pi}(\frac{P_y(\omega)}{\hat{P}_y(\omega)}-\ln(\frac{P_y(\omega)}{\hat{P}_y(\omega)})-1)d\omega \tag{7}$$

Besides, $\theta_{ij}^{'}=[\theta_x^i,\theta_w^j,\sigma_{x,ij}^{2,ML},\sigma_{w,ij}^{2,ML}]$, where θ_x^i and θ_w^j are the i^{th} speech codebook and the j^{th} noise codebook entries, $\sigma_{x,ij}^{2,ML}$ and $\sigma_{w,ij}^{2,ML}$ are the maximum-likelihood estimates of speech and noise excitation variances which can be obtained by [4]

$$\mathbf{C}[\sigma_x^2\ \sigma_w^2]^T=\mathbf{D} \tag{8}$$

where the matrices $\mathbf{C}$ and $\mathbf{D}$ are given in [4].

Eventually, the estimated AR coefficients can be used to construct a Wiener filter to obtain the enhanced speech

$$H(\omega_k)=\frac{\hat{\sigma}_x^2/\mid\hat{A}_x(\omega_k)\mid^2}{\hat{\sigma}_x^2/\mid\hat{A}_x(\omega_k)\mid^2+\hat{\sigma}_w^2/\mid\hat{A}_w(\omega_k)\mid^2} \tag{9}$$

3. MBE Framework

Motivated by the high quality speech synthesized by MBE speech coding, we introduce the MBE AwS framework in this paper. In the MBE model, speech spectrum is divided into multiple sub-bands and each sub-band is sent for voicing judgment. Voiced sub-band is generated by periodic excitation, and unvoiced sub-band is generated by random noise. In this paper, the enhanced spectrum by codebook-based method is considered as an input parameter for MBE model.

In general, the proposed MBE speech enhancement framework can be divided into two stages, namely, speech analysis and speech synthesis. Fig. 1 shows the proposed MBE AwS framework. The following will describe the framework in detail.

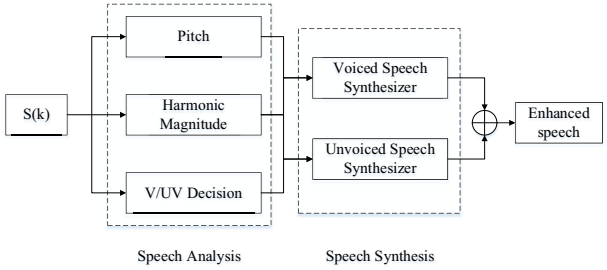

Figure 1: *The proposed MBE framework.*

3.1. Speech analysis

During the speech analysis stage, several acoustic parameters are estimated including pitch period, harmonic magnitude and V/UV decision. First of all, the harmonic magnitude corresponding to each candidate pitch period is calculated. The harmonic magnitude is obtained by [8]

$$A_m(\tau)=\frac{\sum_{k=a_m(\tau)}^{b_m(\tau)}\mid S(k)\mid\mid W(\tau,k)\mid}{\sum_{k=a_m(\tau)}^{b_m(\tau)}\mid W(\tau,k)\mid^2} \tag{10}$$

where τ is a set of candidate pitch period. $S(k)$ is the enhanced spectrum by the codebook-based method and $W(\tau,k)$ is a periodic excitation spectrum with period τ, which can be replaced by the frequency response of the window function. Besides, $a_m(\tau)$ and $b_m(\tau)$ are the upper and lower bounds of the frequency bins in m^{th} harmonic band. Pitch τ_{opt} is obtained by minimizing the following error function $\alpha(\tau)$ [8]

$$\alpha(\tau)=\frac{\sum_{m=1}^{M(\tau)}\sum_{k=a_m(\tau)}^{b_m(\tau)}[\mid S(k)\mid-A_m(\tau)\mid W(\tau,k)\mid]^2}{\sum_{m=1}^{M(\tau)}\sum_{k=a_m(\tau)}^{b_m(\tau)}\mid S(k)\mid^2} \tag{11}$$

where $M(\tau)$ is the number of harmonics.

After the optimal pitch period τ_{opt} is fixed, a set of harmonic magnitude corresponding to τ_{opt} is determined automatically.

The last acoustic feature to be estimated in speech analysis stage is V/UV decision. When calculating the harmonic magnitude and pitch, it is assumed that each harmonic band is voiced. It will make the synthetic spectrum in voiced portion close to the original spectrum, while in unvoiced portion is very different from the original spectrum. By using the difference and a given threshold, we can get the V/UV decision. The difference function of m^{th} harmonic band D_m is defined as [8]

$$D_m = \frac{\sum_{k=a_m(\tau_{opt})}^{b_m(\tau_{opt})} [|S(k)| - |\hat{S}(\tau_{opt},k)|]^2}{\sum_{k=a_m(\tau_{opt})}^{b_m(\tau_{opt})} |S(k)|^2} \quad (12)$$

where $|\hat{S}(\tau_{opt},k)|$ is the magnitude of reconstruct spectrum, can be obtained by

$$|\hat{S}(\tau_{opt},k)| = A_m(\tau_{opt})|W(\tau_{opt},k)|,\ a_m(\tau_{opt}) \le k \le b_m(\tau_{opt}) \quad (13)$$

If D_m is lower than the given threshold T_m, the harmonic band is considered as voiced, otherwise the harmonic band is considered as unvoiced. In practice, T_m is set to 0.2 which can obtain a well performance.

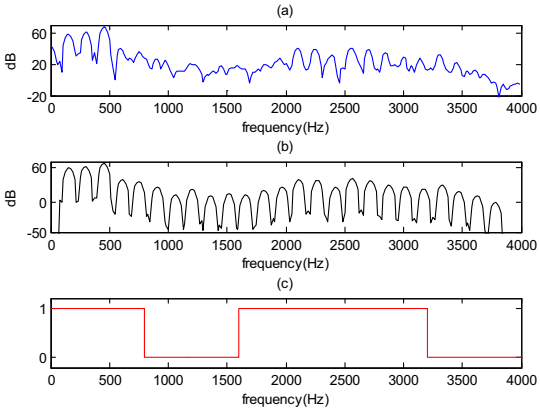

Figure 2: *An example of V/UV decision. (a) Codebook-based method enhanced spectrum, (b) Excitation spectrum and (c) V/UV decision.*

Fig. 2 (a) and (b) show the enhanced spectrum of codebook-based method and periodic excitation spectrum, respectively. In Fig. 2 (c), '1' denotes voiced and '0' denotes unvoiced. As shown in the figure, if the matching error is small, the harmonic band is declared as voiced, otherwise is unvoiced.

3.2. Speech synthesis

During the synthesis stage, voiced and unvoiced speech signals are synthesized using different strategies. Each harmonic of which is declared as voiced corresponds to a sinusoidal oscillator defined by the magnitude, frequency and phase. The m^{th} harmonic of voiced speech can be expressed as

$$s_v^m(n) = a_m(n)\cos[\theta_m(n)],\ 0 \le n < N \quad (14)$$

where $a_m(n)$ and $\theta_m(n)$ are magnitude function and phase function of the m^{th} harmonic, respectively. Here the phase function is obtained by sampling phase spectrum of noisy speech at each integer multiples of pitch frequencies.

By using the method of time domain synthesis, it is possible to make the synthesized speech smooth at the boundaries of the frame. The magnitude function is obtained by linear interpolation of each harmonic magnitude

$$a_m(n) = \hat{A}_m(-1) + \frac{n}{N}[\hat{A}_m(0) - \hat{A}_m(-1)] \quad (15)$$

where $\hat{A}_m(-1)$ and $\hat{A}_m(0)$ are m^{th} harmonic magnitude of previous and current frame, respectively. Similarly, the pitch also needs to do a similar linear interpolation between frames.

The final voiced speech is obtained by adding the sinusoidal function of each harmonic

$$\hat{s}_v(n) = \sum_{m=1}^{M(\tau_0)} 2s_v^m(n) \quad (16)$$

For unvoiced speech, it is synthesized by harmonics which are declared as unvoiced. First, a random Gaussian white noise is generated and Fourier transform is applied. Then, the spectrum corresponding to the voiced sub-band components are set to zero. Finally, an inverse FFT is applied to get the final unvoiced speech. Fig. 3 shows the block diagram of unvoiced speech synthesis.

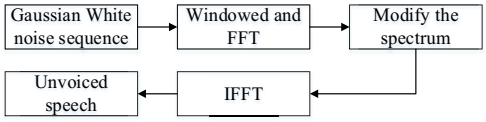

Figure 3: *Unvoiced speech synthesis.*

The enhanced speech is obtained by summing the voiced and unvoiced speech. Because the phase spectrum of the synthesized speech is discrete, it is necessary to apply a FFT transform and replace the previous phase with the original noisy phase spectrum.

3.3. Modified codebook-based enhanced spectrum

In order to make the estimation of acoustic features more accurate, this paper combines the codebook-based Wiener filter with speech presence probability before MBE framework. Let $H_1(k)$ denote that speech is present in frequency bin k and $p(H_1(k))$ denote the priori speech presence probability in frequency bin k. Since the codebook-based method is only used to model the spectral envelope, it is necessary to estimate the speech presence probability based on the frequency bin. Here, Minima Controlled Recursive Averaging (MCRA) algorithm [9] is employed to estimate the power spectrum of noise, $P_w(k)$. Then the posteriori speech presence probability in frequency bin k can be estimated as [10]

$$p(H_1(k)|y) = \frac{p(H_1(k))}{p(H_1(k)) + (1 - p(H_1(k)))(1 + \xi'(k))\exp(-\nu'(k))} \quad (17)$$

where

$$\xi'(k) = \xi(k) / p(H_1(k)),\ v'(k) = \gamma(k)\xi'(k) / (\xi'(k)+1) \quad (18)$$

where $\xi(k)$ denotes the priori SNR which is estimated by Decision-Directed (DD) method [11], $\gamma(k)$ denotes the posteriori SNR and $p(H_1(k))$ is set to 0.5.

Above all, the codebook-based Wiener filter can be improved as

$$H'(\omega_k) = p(H_1(k) \mid y) \frac{\hat{\sigma}_x^2 / |\hat{A}_x(\omega_k)|^2}{\hat{\sigma}_x^2 / |\hat{A}_x(\omega_k)|^2 + \hat{\sigma}_w^2 / |\hat{A}_w(\omega_k)|^2} \quad (19)$$

4. Results

In this section, we compare the performance of proposed method with two traditional codebook methods which come from [4] and [5]. For comparison, we name them as Ref. A and Ref. B, respectively. The test set is selected from NTT database and down-sampled to 8 kHz. It consists of nine utterances, four males and five females. The speech is corrupted by babble, white, f16 and factory2 noise from NOISEX-92 database at 0dB, 5dB and 10dB SNR. The size of samples per frame is 256 with 50% overlapped. For codebook-based approach, a 5-bit speech codebook and 3-bit noise codebooks for each type of noise are trained. Besides, the order of AR models for speech and noise is 10.

In order to demonstrate the effectiveness of the proposed method, the comparison of spectrograms between the proposed method and the reference methods is presented in Fig. 4.

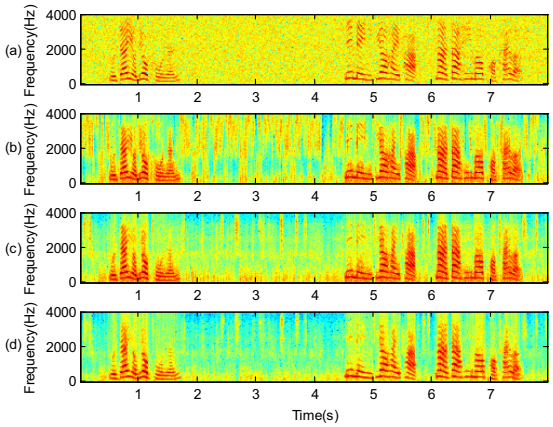

Figure 4: *Spectrograms of (a) noisy speech (White noise 10dB), (b) Ref. A, (c) Ref. B, and (d) proposed.*

From the spectrograms, it is obvious that the proposed method can generate clean harmonics and remove the noise between harmonics. In addition, because the speech presence probability is used, the noise in silent segments can be further removed.

The objective evaluation for speech enhancement is performed from perceptual evaluation of speech quality (PESQ) [12], segmental SNR (SSNR) [13] and log-spectrum distortion (LSD) [14]. The results are listed in Table 1, Table 2 and Table 3, respectively.

Table 1: *Test Results of Average PESQ.*

Method	0dB	5dB	10dB
Noisy	1.61	1.92	2.30
Ref. A	1.87	2.25	2.55
Ref. B	2.00	2.36	2.65
proposed	2.22	2.50	2.71

Table 2: *Test Results of Average SSNR Improvement.*

Method	0dB	5dB	10dB
Ref. A	9.22	8.30	7.22
Ref. B	12.97	11.72	10.55
proposed	14.61	12.77	10.58

Table 3: *Test Results of Average LSD.*

Method	0dB	5dB	10dB
Noisy	15.06	13.26	11.46
Ref. A	10.68	9.27	8.01
Ref. B	9.66	8.23	6.90
proposed	9.01	7.60	6.35

As can be seen from Table 1, the PESQ score of the proposed method is higher compared with reference methods, especially in low SNR conditions. From the results of SSNR and LSD in Table 2 and Table 3, we can know that the proposed method has less residual noise while has less spectrum distortion. It is due to the AwS framework which can reconstruct clean harmonic. In addition, because of the introduction of speech presence probability, the noise in silent segments can be further removed.

From three objective measures, we can know that the proposed MBE AwS framework has better effect in low SNR conditions. The reason is that the MBE model is based on the extraction of acoustic parameters, and the reconstructed clean speech is obtained by these parameters. So in low SNR conditions, it can remove more noise, especially between the harmonics.

In the MBE speech analysis stage, we need to divide the frequency band according to the harmonics. It should be noted that a harmonic band can have more than one harmonic. It is necessary to make a compromise between the precise judgment of each harmonic and the continuity of the harmonics. In the experiment a harmonic band containing five harmonics can get the best result.

5. Conclusions

In this paper, we propose an improved codebook-based speech enhancement based on MBE model. It uses analysis-with-synthesis framework to reconstruct speech through acoustic features extracted from noisy speech. Compared with traditional codebook-based approach, it can remove noise not only between adjacent harmonics, but also in silent segments. The experiment results show a better performance of the proposed method compared with reference methods under different noise conditions.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61471014, No. 61231015).

7. References

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol. 27, no. 2, pp. 113-120, 1979.

[2] A. Amehraye, D. Pastor, and A. Tamtaoui, "Perceptual improvement of Wiener filtering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, pp. 2081-2084, 2008.

[3] R. Martin, "Speech enhancement based on minimum mean-square error estimation and super-Gaussian priors," *IEEE Transactions on Speech Audio Processing*, vol. 11, no. 5, pp. 845–856, Sep. 2005.

[4] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook Driven Short-Term Predictor Parameter Estimation for Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 163–176, Jan. 2006.

[5] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 441–452, Feb. 2007.

[6] P. Harding and B. Milner, "Speech enhancement by reconstruction from cleaned acoustic features," in *Interspeech Proceedings*, 2011, pp. 1189-1192.

[7] R. Chen, C.-F. Chan, and H. C. So, "Model-based speech enhancement with improved spectral envelope estimation via dynamics tracking," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1324–1336, 2012.

[8] D. Griffin and J. Lim, "Multiband excitation vocoder," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 8, pp. 1223– 1235, 1988.

[9] Cohen, I. "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, 9(1), 12-15. 2002.

[10] P. C. Loizou, Speech enhancement: theory and practice. *Boca Raton, FL, USA: CRC Press*, 2007.

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

[12] "Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs," *ITU-T Recommendation*, P.862, Feb, 2001.

[13] Quackenbush, S. R., Barnwell, T. P., Clements, M. A., "Objective Measures of Speech Quality," *Englewood Cliffs, NJ: Prentice Hall*, 1988.

[14] Abramson, A., Cohen, I., "Simultaneous Detection and Estimation Approach for Speech Enhancement," *IEEE Transactions on Speech Audio Processing*, 15(8), pp. 2348-2359, 2007