Harmonic Structure Mask for Speech Enhancement Using Sparsity Regularization

Haonan Wang* Kenta Iwai† and Takanobu Nishiura†
* Graduate School of Information Science and Engineering, Ritsumeikan University, Shiga, Japan.
E-mail: is0389sf@ed.ritsumei.ac.jp
† College of Information Science and Engineering, Ritsumeikan University, Shiga, Japan.
E-mail: iwai18sp@fc.ritsumei.ac.jp
† E-mail: nishiura@is.ritsumei.ac.jp

***Abstract*—Harmonic structure, an important characteristic of speech signals, has been utilized in various speech processing applications, such as dereverberation, fundamental frequency (f_0) estimation, voice activity detection (VAD), phase reconstruction and source separation. This paper presents a harmonic structure mask for those speech enhancement applications based on sparsity regularization via convex optimization. Specifically speaking, we first derive a harmonic structure mask of the noisy speech using f_0 and VAD estimations, then use this mask to protect harmonic components of speech during the sparsity regularization process. The proposed mask benefits from the additional harmonic information, leading to better protection of harmonic components. Numerical experiments show that the proposed mask can improve speech quality and intelligibility compared to the previous work.**

I. Introduction

Enhancing noisy single-channel speech corrupted by additive noise only from the noisy observation has been an active topic in speech processing [1]. These works can be classified as spectral subtraction [2], minimum mean-square error (MMSE) estimation [3], [4], machine learning [5] and sparsity regularization [6]–[8] methods. Though the aforementioned methods have yielded good performance, few of them took harmonic structure into account. As a unique characteristic of speech as well as many other audio signals such as musical instruments, harmonic structure has been extensively used in many speech processing applications, such as dereverberation [9] , fundamental frequency (f_0) estimation [10], voice activity detection (VAD) [11], phase reconstruction [12] and source separation [13]. In this paper, we describe theses studies taking harmonic structure into consideration as "harmonicity-aware". The majority of harmonicity-aware speech processing applications are based on the assumption that speech is a weighted superposition of several sinusoids at integer multiples of f_0 including itself. This assumption can derive a binary harmonic structure mask on the time-frequency domain, which illustrates the locations on the time-frequency domain where harmonic components may exist.

In this paper, we use this mask to bring harmonicity into speech enhancement methods based on sparsity regularization via complex optimization (e.g., [6], [7]). In particular, we propose to design an element-wise mask (matrix) based on a smoothed harmonic mask and insert it into the regularization model from the previous work. This mask is calculated from the binary harmonic mask and the window function of the short-time Fourier transform (STFT). This smoothed harmonic mask can control the sparsity of the estimation, leading to better protection of harmonic sturcture. To evaluate the proposed mask, a recently proposed optimization model presented in [6]–[8] is chosen as the baseline method. Numerical experiments compare the performances between the method with and without the proposed mask. Results show that the proposed mask helps to outperform the original method with better speech quality and intelligibility at various SNRs conditions under several noise types.

II. Relation to Prior Work

Some previous works have succeed in bringing harmonicity into their methods [14]–[18]. Here we briefly introduce these works. Ref. [14] proposed a harmonic-regeneration using a non-linear operation to artificially create a fully harmonic noisy observation and used it as additional information to preserve harmonic components. Ref. [15] combined the Wiener filtering with harmonic information and presented an harmonicity-aware MMSE-optimal estimator. Ref. [16] took advantage of harmonic phase reconstruction and presented a harmonicity-phase-aware amplitude estimator. Ref. [17] extended the conventional hidden Markov model (HMM)-based MMSE estimator to enhance the harmonic components for voiced speech. These works, as well as many other related ones, introduced harmonicity to their models as an independent prior. In contrast to these works, we focus on bringing harmonicity into speech enhancement methods based on sparsity regularization in a much simpler way, which is the harmonic structure mask.

Generally, two approaches are taken to introducing harmonicity into convex optimization for speech enhancement. The first one is inserting a harmonicity-aware prior, which is difficult because of the unpredictable weightings for each harmonic components are unknown, meaning that it is unrealistic to calculate the cost (or "distance") between the estimation and the desired amplitude of each harmonic. This is the reason why we take the other approach, adding a mask to support the convergence, in this study. The proposed mask requires no additional prior, and just simply lead the convergence into a

harmonicity-protective direction. This simplicity indicates the possibility that various modifications of the proposed mask can be easily developed based on specific purposes.

III. HARMONIC STURCTURE MASK

A. Harmonic speech model and notation

We assume that clean speech $x[n]$ is corrupted by additive noise $d[n]$ and the noisy observation is $y[n] = x[n] + d[n]$. The STFT coefficients of the noisy observation are denoted as

$$Y[k,\tau] = X[k,\tau] + D[k,\tau], \tag{1}$$

where k, τ are indexes of frequencies and time frames respectively. $Y[k,\tau], X[k,\tau]$ and $D[k,\tau]$ are the complex STFT coefficients for $y[n], x[n]$ and $d[n]$ respectively. Based on the following hypotheses [12],

- Voice speech is a weighted superposition of several sinusoids at integer multiples of f_0 including itself, the harmonic frequencies are $f[h,\tau] = (h+1)f_0[\tau]$, where h denotes the index of harmonic components.
- Each harmonic component dominates the frequency bands in its direct neighborhood and the influence of other harmonic components on this neighbor can be neglected.

we define the frequency indexes of harmonic components as

$$k^*[h,\tau] = \arg\min_{\kappa} |\kappa - \frac{Kf[h,\tau]}{f_s}|, \tag{2}$$

where $Kf[h,\tau]/f_s$ is a non-integer value for mapping harmonic frequencies into integer indexes. K, f_s are the maximum of frequency indexes and the sampling-rate. Hence, the binary mask of harmonic components is represented as

$$H[k,\tau] = \begin{cases} 1, (k = k^*[h,\tau] \wedge v[\tau] = 1) \\ 0, (\text{otherwise}) \end{cases}, \tag{3}$$

where $v[\tau]$ shows whether τ^{th} frame is a voiced frame (1 for true, 0 for false). The binary mask $H[k,\tau]$ illustrates the appearance of harmonic components on the time-frequency domain. However, due to the analysis window function $w[n]$ used in STFT, harmonic components leak their power into the neighbored frequency bands. Therefore, we can calculate a more practical harmonic mask, which is smoothed (or blurred) by the cyclic convolution between the binary mask $H[k,\tau]$ and the frequency response of $w[n]$. The smoothed harmonic mask is denoted by

$$H^{'}[:,\tau] = W[k] \circledast H[:,\tau], \tag{4}$$

where $W[k]$ is the amplitude response of $w[n]$ and $\circledast$ denotes the convolution operator along with the frequency direction. The smoothed harmonic mask $H'[k,\tau]$ is an "element-wise map" representing "how much power do harmonic components have for every single time-frequency cell". In this paper, the proposed harmonicity-aware parametrization takes advantage of $H'[k,\tau]$.

B. Sparsity control by proposed harmonic structure mask

It is known that the speech amplitude spectrogram is sparse especially in those time-frequency cells without harmonics. This characteristic has been used in many convex-optimization-based speech enhancement applications [6]–[8]. Generally, most of the convex-optimization-based speech enhancement applications can be simplified as

$$\hat{S} = \arg\min_{S} \frac{1}{2} \|Y - S\|_{\mathrm{F}}^2 + \|S\|_{1,\Lambda} + \mathcal{G}(S), \tag{5}$$

where $Y, S \in \mathbb{R}_+^{K\times T}$ are the noisy and estimation amplitude spectrograms. $\mathcal{G}(S)$ is a cost function varies from different purposes. $\|S\|_F$ is the Frobenius norm of S. $\Lambda \in \mathbb{R}_+^{K\times T}$ denotes an element-wise regularization parameter matrix for $\|S\|_{1,\Lambda}$. This element-wisely parameterized ℓ-1 norm is defined as

$$\|S\|_{1,\Lambda} = \sum_{k=1}^{K}\sum_{\tau=1}^{T} \Lambda[k,\tau]S[k,\tau], \tag{6}$$

where $K \times T$ is the size of the STFT coefficient matrix.

The concept of the proposed parametrization is to control the sparsity of the estimation by modifying the original parameters using $H^{'}[k,\tau]$, in terms of better harmonic protection. Basically, the bigger $\Lambda[k,\tau]$ is, the more sparse $\hat{S}[k,\tau]$ will be. Therefore, it is an extremely simple and natural way to use $H^{'}[k,\tau]$ to modify $\Lambda[k,\tau]$, because $H^{'}[k,\tau]$ represents "which time-frequency cell should be sparse and how much sparsity should be retained." The proposed harmonicity-aware parametrization $\Lambda'[k,\tau]$ is proposed as

$$\Lambda'[k,\tau] = (1 - \frac{H'[k,\tau]}{\max\{H'\}})\Lambda[k,\tau], \tag{7}$$

where $\max\{*\}$ is the operator for extracting the maximum. For those time-frequency cells with high $H^{'}[k,\tau]$ values, $\Lambda'[k,\tau]$ becomes extremely small, even 0, leading to the protection effect of harmonics. On the contrary, for others with low $H^{'}[k,\tau]$ values, $\Lambda'[k,\tau]$ approximately remains unchanged, leading to the noise reduction effect. We define this modification mask as

$$\mathcal{H}[k,\tau] = 1 - \frac{H'[k,\tau]}{\max\{H'\}}. \tag{8}$$

Figure 1 visualizes the harmonic masks and the proposed modification mask.

IV. NUMERICAL EXPERIMENT

A. Baseline optimization model

To evaluate the proposed mask, we choose an optimization model a recently proposed speech enhancement method based on sparsity regularization [6] as the baseline. The optimization model is formulated as

$$\hat{S} = \arg\min_{S} \frac{1}{2} \|Y - S\|_{\mathrm{F}}^2 + \|S\|_{1,\Omega} + \|\nabla_\tau S\|_{1,\Theta}, \tag{9}$$

where ∇_τ is the temporal derivative operator. $\Omega, \Theta \in \mathbb{R}_+^{K\times T}$ are the original element-wise parameter matrixes proposed in [6]. This optimization model maximizes the sparsity of speech

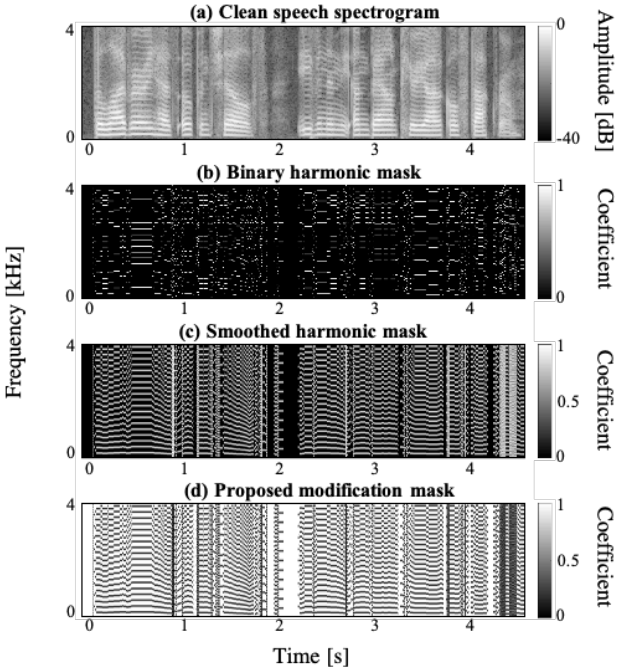

Fig. 1. Illustration of the proposed parametrization. (a): Clean speech spectrogram $|X[k,\tau]|$. (b): Binary harmonic mask $H[k,\tau]$ representing where harmonic components exist. (c): Smoothed harmonic mask $H^{'}[k,\tau]$ calculated using (4). (d): The proposed parametrization modification mask $\mathcal{H}[k,\tau]$.

amplitude spectrogram as well as its temporal derivative simultaneously, both of which should be sparse in those time-frequency cells without harmonic components.

The original parametrization Ω is based on the noise power estimation from L silent frames at the beginning of the noisy observation, defined as

$$\Omega[k,:] = \frac{1}{L}\sum_{\tau=1}^{L}|Y[k,\tau]|, \tag{10}$$

where L is the number of silent frames. As shown in this equation, Ω is temporally constant, and the parameters of each frequency band are determined by the average amplitude of noise. For those frequency bands with higher noise amplitude estimations, Ω penalizes them more compared with those bands with lower noise power. This noise-estimation-based parametrization has been widely used in many other studies [2], [3], [8], [19]. However, one of the weakness of the noise-estimation-based parametrization strategy is that it cannot handle time-varying noise such as speech babble noise properly. To solve this issue, in this paper, the proposed method improves its performance by incorporating harmonic information to make sure the algorithm will not suppress harmonics even when the accuracy of noise estimation is not satisfying.

Θ controls the sparsity of the time derivative of S, defined as

$$\Theta[:,:] = \frac{1}{TK}\sum_{k=1}^{K}\sum_{\tau=1}^{T}|\nabla_\tau Y|. \tag{11}$$

Θ calculates the average of the total variation (TV) of the noisy amplitude spectrogram Y and all elements of Θ are identical. This unified value represents the TV level of the noisy amplitude spectrogram, which is highly relevant to the noise type (if the input SNRs are the same). For example, those non-stationary noise types (e.g., speech babble noise) with intense time fluctuation may result in a higher Θ than stationary noise (e.g., white noise). Θ penalizes the TV of S to achieve better performance.

By introducing the proposed harmonic structure mask into (9), we get

$$\hat{S} = \arg\min_{S}\frac{1}{2}\|Y-S\|_{\mathrm{F}}^2 + \|S\|_{1,\mathcal{H}\odot\Omega} + \|\nabla_\tau S\|_{1,\mathcal{H}\odot\Theta}, \tag{12}$$

where $\odot$ is the Hadamard product. As shown in this equation, the proposed parametrization just simply modifies the original parameters Ω, Θ by element-wise multiplication by $\mathcal{H}$, and requires no additional prior to the optimization model. Furthermore, note that for those unvoiced frames, the values of proposed modification mask $\mathcal{H}$ equal to 1, meaning the proposed method maximize the penalty of noise compared to those time-frequency cells containing harmonics. This mechanism is considerably safe since only noise exists in unvoiced frames.

Here we briefly describes that (9) and (12) are solved by the alternating direction method of multipliers (ADMM) [20], an iterative convex optimization solver.

B. Experiment and result

To evaluate the performance of the proposed mask, we carried out objective evaluations evaluating speech quality and intelligibility improvements by the perceptual evaluation of speech quality (PESQ) [21] and the short-time objective intelligibility (STOI) [22]. Twenty utterances (10 of female and 10 of male) were randomly selected from the TIMIT database [23] sampled at 8 kHz / 16 bit. The speech was degraded by white Gaussian, speech babble and factory noise from the NOISE-92X database [24] at various input SNRs (-3 dB, 0 dB, 3 dB, 6 dB, 9 dB). STFT used 32 ms Hamming window with $1/4$ shift and 2048 point discrete Fourier transform. f_0 & VAD were estimated by [25], [26] respectively.

The evaluation targets are as follows: **Base.** : Solving (9). **Prop.** : Solving (12) with estimated $\mathcal{H}$. **Prop.*** : Solving (12) with oracle $\mathcal{H}$. The estimated and oracle $\mathcal{H}$ mean that $\mathcal{H}$ is calculated by f_0 & VAD derived from the noisy observations and clean speech, respectively.

The experimental results are summarized in Table I. Results show that the proposed mask can help (9) to perform better and result in higher PESQ and STOI. For all types of noise, the proposed mask with the estimated f_0 & VAD has almost the same performance compared with the baseline in terms of

TABLE I
THE AVERAGE PESQ AND STOI. **Underlined bold font** MEANS THE BEST PERFORMED METHODS AND **bold font** FOR THE SECOND ONES.

		White noise					babble noise					Factory noise				
		-3 dB	0 dB	3 dB	6 dB	9 dB	-3 dB	0 dB	3 dB	6 dB	9 dB	-3 dB	0 dB	3 dB	6 dB	9 dB
PESQ	Noisy	1.37	1.52	1.70	1.89	2.09	1.61	1.87	2.05	2.24	2.43	1.55	1.74	1.94	2.14	2.35
	Base.	1.83	2.07	2.22	2.45	2.67	1.65	2.04	2.13	2.35	2.56	**1.80**	**2.06**	2.28	2.48	2.68
	Prop.	**1.85**	**2.08**	**2.31**	**2.51**	**2.70**	**1.71**	**2.09**	**2.18**	**2.39**	**2.60**	1.78	2.05	**2.29**	**2.50**	**2.70**
	Prop.*	**2.12**	**2.31**	**2.48**	**2.65**	**2.80**	**1.91**	**2.11**	**2.30**	**2.50**	**2.67**	**2.06**	**2.29**	**2.44**	**2.62**	**2.79**
STOI	Noisy	0.57	0.64	0.72	0.78	0.84	**0.59**	0.65	0.72	0.78	0.83	0.57	0.64	0.71	0.78	0.84
	Base.	0.58	0.67	0.75	0.82	0.87	0.55	0.63	0.71	0.78	0.84	0.56	0.64	0.72	0.79	0.85
	Prop.	**0.61**	**0.71**	**0.78**	**0.84**	**0.88**	0.58	**0.67**	**0.74**	**0.80**	**0.85**	**0.59**	**0.68**	**0.76**	**0.82**	**0.87**
	Prop.*	**0.65**	**0.72**	**0.79**	**0.84**	**0.88**	**0.60**	**0.68**	**0.75**	**0.81**	**0.88**	**0.62**	**0.69**	**0.76**	**0.82**	**0.87**

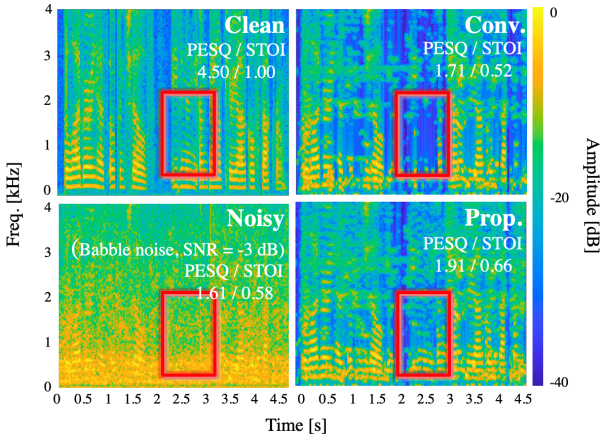

Fig. 2. Spectrograms.

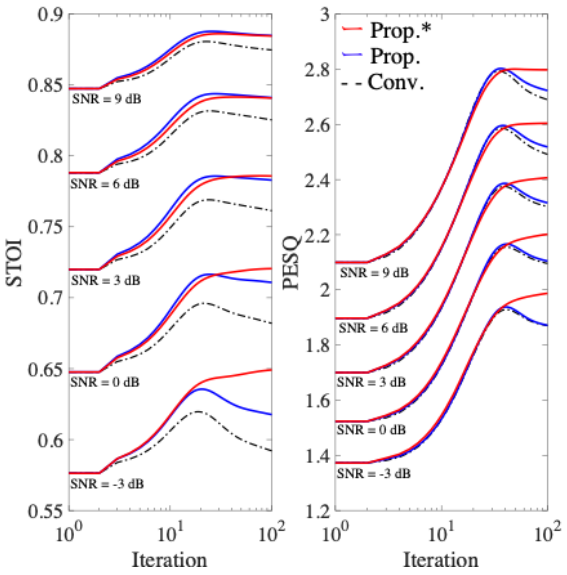

Fig. 3. The converging behaviors under white noise at various input SNRs.

PESQ. However, given the oracle f_0 & VAD, the proposed method completely outperforms all of the other comparison targets, which experimentally proves that the concept of improving speech quality by controlling sparsity of harmonic and non-harmonic time-frequency cells. As for the STOI results, regardless of the accuracies of f_0 & VAD estimations, the proposed mask has better speech intelligibility than the baseline. This is because that, STOI is designed with a time-frequency weighting process which predicts higher speech intelligibility scores if harmonics of speech are protected [22].

As the proof of the protection effect of the proposed mask, Fig. 2 shows the difference between spectrograms of the evaluation targets. From Fig. 2 We can see that the corrupted harmonic components in the noisy spectrogram are completely compressed in the baseline method. Whereas, with the help of the proposed mask, the vanished harmonic components are preserved.

Besides the numerical results, an example of converging behaviors under white noise is illustrated in Fig. 3. The converging behaviors of the baseline model achieved relatively good performance for both PESQ and STOI but degraded soon. This phenomenon implies that the baseline model cannot stop the optimization model from penalizing some of the vital harmonic components too much, leading to the degraded speech quality and intelligibility. However, the proposed mask can protect harmonic components and lead the optimization process towards a speech quality and intelligibility improving direction, despite the fact that there is no harmonicity-aware prior in this optimization model in the first place, and simply achieve that with a single mask insertion. Clearly, the performance of the proposed method is heavily affected by the accuracy of f_0 estimation, which indicates that the resolution of time-frequency analysis matters. Currently, the proposed method needs relatively high frequency resolution to ensure the accuracy of f_0 estimation, and this could be considered as one of the limitation of the proposed method.

V. CONCLUSION

In this paper, we introduced a harmonic structure for speech enhancement applications those based on sparsity regularization. We presented a method to incorporate harmonic information into existing optimization models by add a harmonic structure mask. Numerical experiments based on a recently proposed optimization model [6] showed the effectiveness of

the proposed mask. The proposed mask is extremely simple, yet able to bring speech quality and intelligibility improvements to those existing optimization models. Note that the proposed mask is not designed for some specific optimization models, and it can be applied to many other similar works as well.

ACKNOWLEDGMENT

This work is partly supported by JST COI, and JSPS KAKENHI Grant Number JP19H04142.

REFERENCES

[1] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement: A Survey of the State of the Art*, Morgan & Claypool, 2013.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 2, pp. 113–120, April 1979.

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

[4] T. Gerkmann, "MMSE-optimal enhancement of complex speech coefficients with uncertain prior knowledge of the clean speech phase," in *2014 IEEE Int. Conf. Acoust., Speech. Signal Process.*, May 2014, pp. 4478–4482.

[5] Y. Zhao, D. Wang, I. Merks, and T. Zhang, "DNN-based enhancement of noisy and reverberant speech," in *2016 IEEE Int. Conf. Acoust., Speech. Signal Process.*, March 2016, pp. 6525–6529.

[6] S. Kammi and M. R. K. Mollaei, "Noisy speech enhancement with sparsity regularization," *Speech Commun.*, vol. 87, pp. 58–69, 2017.

[7] S. Kammi and M. R. K. Mollaei, "A novel regularization framework for transient noise reduction," *Appl. Acoust.*, vol. 129, pp. 135–143, 2018.

[8] N. Saleem, M. I. Khattak, and M. Shafi, "Unsupervised speech enhancement in low SNR environments via sparseness and temporal gradient regularization," *Appl. Acoust.*, vol. 141, pp. 333–347, 2018.

[9] T. Nakatani, K. Kinoshita, and M. Miyoshi, "Harmonicity-based blind dereverberation for single-channel speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 80–95, Jan. 2007.

[10] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 528–537, March 2010.

[11] T. Fukuda, O. Ichikawa, and M. Nishimura, "Long-term spectro-temporal and static harmonic features for voice activity detection," *IEEE Journal of Selected Topics Signal Process.*, vol. 4, no. 5, pp. 834–844, Oct. 2010.

[12] M. Krawczyk and T. Gerkmann, "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1931–1940, Dec. 2014.

[13] Y. Masuyama, K. Yatabe, and Y. Oikawa, "Phase-aware harmonic/percussive source separation via convex optimization," in *2019 IEEE Int. Conf. Acoust., Speech. Signal Process.*, May 2019, pp. 985–989.

[14] H. Xuchu and Z. Xiaojing, "Speech enhancement using harmonic regeneration," in *2011 IEEE Int. Conf. Acoust., Speech. Signal Process.*, June 2011, vol. 1, pp. 150–152.

[15] M. Krawczyk-Becker and T. Gerkmann, "MMSE-optimal combination of wiener filtering and harmonic model based speech enhancement in a general framework," in *2015 IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2015, pp. 1–5.

[16] Y. Wakabayashi and N. Ono, "Maximum a posteriori estimation of spectral gain with harmonic-structure-based phase reconstruction for phase-aware speech enhancement," in *2018 Asia-Pacific Signal Inform. Process. Assoc. Ann. Summ. Conf.*, Nov. 2018, pp. 1649–1652.

[17] M. E. Deisher and A. S. Spanias, "HMM-based speech enhancement using harmonic modeling," in *1997 IEEE Int. Conf. Acoust., Speech. Signal Process.*, April 1997, vol. 2, pp. 1175–1178.

[18] W. Jin, X. Liu, M. S. Scordilis, and L. Han, "Speech enhancement using harmonic emphasis and adaptive comb filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 356–368, Feb 2010.

[19] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[20] S. Boyd, N. Parikh, C. Neal, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends® Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2011.

[21] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE Int. Conf. Acoust., Speech. Signal Process.*, May 2001, vol. 2, pp. 749–752.

[22] 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, Sep. 2011.

[23] J. S. Garofolo, "Getting started with the darpa timit cd-rom: An acoustic phonetic continuous speech database, national institute of standards and technology (NIST), gaithersburg," *Gaithersburgh, MD, USA*, 1988.

[24] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[25] S. Gonzalez and M. Brookes, "A pitch estimation filter robust to high levels of noise (PEFAC)," in *2011 19th European Signal Processing Conference*, Aug. 2011, pp. 451–455.

[26] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Letters*, vol. 6, no. 1, pp. 1–3, Jan. 1999.