SIP (2018), vol. 7, e5, page 1 of 12 © The Authors, 2018.

doi:10.1017/ATSIP.2018.7

ORIGINAL PAPER

Noise masking method based on an effective ratio mask estimation in Gammatone channels

FENG BAO AND WALEED H. ABDULLA

In computational auditory scene analysis, the accurate estimation of binary mask or ratio mask plays a key role in noise masking. An inaccurate estimation often leads to some artifacts and temporal discontinuity in the synthesized speech. To overcome this problem, we propose a new ratio mask estimation method in terms of Wiener filtering in each Gammatone channel. In the reconstruction of Wiener filter, we utilize the relationship of the speech and noise power spectra in each Gammatone channel to build the objective function for the convex optimization of speech power. To improve the accuracy of estimation, the estimated ratio mask is further modified based on its adjacent time–frequency units, and then smoothed by interpolating with the estimated binary masks. The objective tests including the signal-to-noise ratio improvement, spectral distortion and intelligibility, and subjective listening test demonstrate the superiority of the proposed method compared with the reference methods.

Keywords: CASA, Noise masking, Ratio mask estimation, Convex optimization

Received 1 November 2017; Revised 6 April 2018

I. INTRODUCTION

Speech enhancement is a focused topic in the speech signal-processing area. The noise reduction or noise masking is often concerned in the speech enhancement. They aim to remove or mask a certain amount of background noise from noisy speech and make the enhanced speech have a better quality and a higher intelligibility. A lower speech intelligibility in the background noise remains a major complaint of the comfort and hearing fatigue by listeners. Although the state-of-the-art monaural speech enhancement algorithms have achieved an appreciable suppression of the noise and improved speech quality certainly, it is still a challenge for them, thus far, to improve the intelligibility of noise-degraded speech.

Monaural speech enhancement, i.e., speech enhancement from single-microphone recordings, is particularly challenging due to an infinite number of solutions. From the application point of view, monaural speech enhancement is perhaps most desirable compared with multi-microphone solutions, since a monaural system is less sensitive to room reverberation and spatial source configuration. In the past several decades, many monaural speech enhancement methods have been proposed, such as spectral-subtraction [1, 2], Wiener filtering [3], and statistical-model-based methods [4] enhanced in the frequency domain. These approaches are more suitable to handle the stationary noise (e.g., white or car noise) rather than non-stationary noise (e.g., babble and street noises). The musical noise and vacuum feeling are usually caused by these typical methods. In the face of these problems, a very effective de-nosing method [5] was proposed with Generalized Gamma Prior (GammaPrior), which was extended from the minimum mean-square error (MMSE) estimation of discrete Fourier transform (DFT) magnitude. In this method, two classes of generalized Gamma distributions were adopted for the complex-valued DFT coefficients. The better performance on subjective and objective tests was achieved compared with typical Wiener filter and statistical-model-based methods. Zoghlami proposed an enhancement approach [6] for noise reduction based on non-uniform multi-band analysis. The noisy signal spectral band is divided into subbands using a gammatone filterbank, and the sub-bands signals were individually weighted according to the power spectral subtraction technique and the Ephraim and Malah's spectral attenuation algorithm. This method is a kind of sub-band spectral subtraction or Wiener filter method.

However, the difficulties of non-stationary noise are still not solved very well based on the above methods. Thus, the baseline Hidden Markov Models (HMMs) [7] and Codebook-driven [8] methods with *a priori* information (i.e., spectral envelops or spectral shapes) about speech and noise were proposed to overcome the situation of non-stationary noise. Some revised methods based on codebook and HMM were proposed in recent years. For example, the Sparse Autoregressive Hidden Markov Models (SARHMM) method [9] modeled linear prediction gains of speech and

Department of Electrical and Computer Engineering, University of Auckland, Auckland 1010, New Zealand

Corresponding author:
F. Bao,
Email: fbao026@aucklanduni.ac.nz

noise in non-stationary noise environments. The likelihood criterion was adopted to find the model parameters, combined with *a priori* information of speech and noise spectral shapes by a small number of HMM states. The sparsity of speech and noise modeling helped to improve the tracking performance of both spectral shape and power level of non-stationary noise. Another very new codebook-based approach with speech-presence probability (SPP) [10] also achieved a very good result on speech enhancement of non-stationary noise environment. This kind of codebook-based method with SPP (CBSPP) utilized the Markov process to model the correlation between the adjacent code-vectors in the codebook for optimizing Bayesian MMSE estimator. The correlation between adjacent linear prediction (LP) gains was also fully considered during the procedure of parameter estimation. Through the introduction of SPP in the codebook constrained Wiener filter, the proposed Wiener filter achieved the goal of noise reduction.

The aforementioned methods are focused on noise reduction. These noise reduction methods often utilize a gain function into each time–frequency (T–F) bin to suppress the noise based on a T–F representation of the noisy speech. The usage of the gain function over all T–F bins can be considered as an attenuation of noise magnitude in each T–F bin. These methods generally derive their gains as a function of the short-time signal-to-noise ratio (SNR) in the respective T–F bin, that is, speech and noise powers at each T–F bin need to be estimated. With respect to the noise masking, the Computational Auditory Scene Analysis (CASA) [11, 12] is considered as an effective approach. By incorporating auditory perception model (i.e., Gammatone filterbank), it could mask the noise based on an estimation of the binary mask in each T–F unit that includes the different number of T–F bins concerned in noise reduction methods. Because the result of binary mask only corresponds to value 0 dominated by noise or value 1 dominated by the speech in each T–F unit, these CASA methods based on the binary mask often wrongly remove the background noise in a weak T–F unit dominated by the speech and seriously affect the hearing quality. A good solution for the shortage of binary mask is ideal ratio mask (IRM) [13, 14] that a priori knowledge of speech and noise is known in advance in the derivation of the mask. The IRM could be considered as a soft decision that the mask values continuously vary from 0 to 1 instead of a hard decision that the mask value is 0 or 1 derived from the ideal binary mask (IBM) [15, 16]. The IRM is more reasonable to handle the situations that speech energy is larger or less than noise in each T–F unit. These ratio mask estimators [13, 14] operated in DFT domain. It is a kind of Weiner filtering solution that the transfer function of Weiner filter can be obtained by estimating speech and noise powers.

Because of the huge and complicated training process of speech and noise priori information, the noise reduction methods based on HMM and codebook or deep learning-based noise ratio masking method may have some limitations on solving the practical issues. Particularly worth mentioning is that the noise types and priori information are not easy and unrealistic to predict in advance.

Therefore, we propose a noise-masking method without any priori information and assumption of speech and noise signals. This proposed method can better mimic the hearing perception properties of the human being to improve the intelligibility of the enhanced speech. The soft decision factor, ratio mask, is used to resynthesize speech signal so that it can avoid the wrong elimination of weak speech T–F units caused by the binary mask. Furthermore, the ratio mask in our proposed method is a kind of ratio between the estimated speech and noise powers. Considering the superiority of solving the minimization problem, the speech power is estimated by convex optimization [17, 18] in the proposed method. For further compensating the powers of weak speech components, the estimated ratio mask is modified and interpolated to recover parts of speech components.

The remainder of this paper is organized as follows. In Section II, we present the overall principle of the proposed method. The performance evaluation is described in Sections III and IV provides the conclusions.

II. THE PRINCIPLE OF THE PROPOSED METHOD

Figure 1 describes the main block diagram of the proposed noise-masking method. First, the input noisy speech is decomposed into 128 channels by using Gammatone filterbank [19, 20]. Then, the signal of each channel is windowed in time domain and a fast Fourier transform (FFT) is done for this windowed signal to obtain the power spectrum of noisy speech. The feature extraction module is utilized to

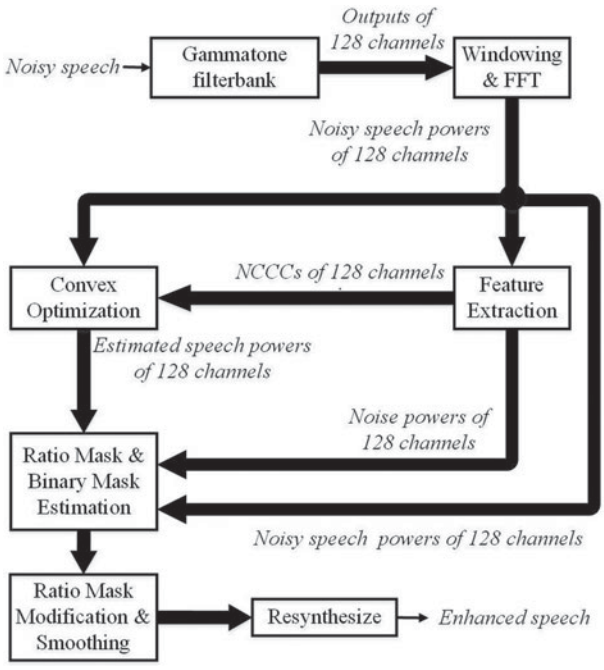

Fig. 1. The Block diagram of the proposed method.

calculate the noise power spectrum by Minima Controlled Recursive Averaging (MCRA) [21] and normalized cross-correlation coefficient (NCCC) [22] between the spectra of noisy speech and noise. The NCCC is used to represent the proportion of noise power in noisy speech, which will contribute to the convex optimization of speech power. The objective function used for convex optimization is built based on noisy speech power and NCCC in each channel, and minimized by the gradient descent method [23]. The speech power is estimated by minimizing the objective function. After that, the powers of estimated speech and noise are used to estimate the ratio mask. This ratio mask is then modified based on the adjacent T–F units and further smoothed by interpolating with the estimated binary mask for increasing the accuracy of ratio mask estimation. Finally, the enhanced speech is resynthesized from the smoothed ratio masks [24].

Based on above block diagram, we mainly describe four key parts. In Section II-A, the speech synthesis mechanism will be described, which is a basic orientation of the proposed work. The estimation method of the ratio mask and binary mask concerned in speech synthesis are given in Sections II-B and II-C, respectively. The speech power estimation closely related to the ratio mask and binary mask estimation is deduced in Section II-D.

A) Speech synthesis mechanism

In the proposed method, the enhanced signal is resynthesized in time domain based on CASA model [24], which is different from the frequency domain synthesis method, such as the Wiener filter method. In the last stage of the speech enhancement system, the target speech is synthesized by means of the estimated ratio mask and filter responses of noisy speech signal in each Gammatone channel.

The Gammatone filter response of arbitrary channel is, $G_c[y(t)]$,

$$G_c[y(t)] = y(t) * g_c(t), \tag{1}$$

such as c is the Gammatone channel index and t is the time index. The symbol $*$ is the convolution operation by Gammatone filter [19]. $g_c(t)$ is a Gammatone filter impulse response [20] described as:

$$g_c(t) = \begin{cases} t^{a-1}\exp(-2\pi bt)\cos(2\pi f_c t), & t \geq 0 \\ 0, & \text{else} \end{cases} \tag{2}$$

where $a = 4$ is the order of the filter, b is the equivalent rectangular bandwidth, which increases as the center frequency f_c increases.

Then, the first filtering response of the noisy speech signal, $g_c(t)$, is reversed in time domain again and further filtered by the Gammatone filter, that is, $F_c[y(t)]$,

$$F_c[y(t)] = \overline{G_c[y(t)]} * g_c(t), \tag{3}$$

where $\overline{G_c[y(t)]}$ represents the time-reverse operation of $G_c[y(t)]$.

These twice time-reverse operations eliminate the phase difference between the filter outputs of the Gammatone channels. The phase-corrected output from each filter channel is then divided into time frames by a raised cosine window for the overlap-and-add. Figure 2 shows the block diagram of the speech synthesis mechanism.

The signal magnitude in each channel is then weighted by the corresponding ratio mask value at that time instant. The weighted filter responses are then summed across all Gammatone channels to yield a reconstructed speech waveform as follows:

$$\hat{x}(t) = \sum_{C} M(c,t) \cdot \overline{F_c[y(t)]} \cdot W(c,t), \tag{4}$$

where c and C are the Gammatone channel index and number, respectively. $M(c,t)$ is the estimated ratio mask. $\overline{F_c[y(t)]}$ is the time-reverse signal of $F_c[y(t)]$. $W(c,t)$ is a raised cosine window. $\hat{x}(c,t)$ is the resynthesized speech signal.

B) Ratio mask estimation

In our noise-masking method, the noisy speech signal with 4 kHz bandwidth is decomposed into the T–F units by a 128-channel Gammatone filterbank [19, 20] whose center frequencies are quasi-logarithmically spaced from 80 to 4000 Hz. In the proposed method, we assume that the clean speech and noise are additive and statistically independent in each gammatone channel. Thus, the powers of noise and speech in each channel can be expressed as follows:

$$P_y(c,m) = P_x(c,m) + P_d(c,m), \tag{5}$$

where c is the channel index, m is the frame index, $P_y(c,m)$, $P_x(c,m)$, and $P_d(c,m)$ indicate the powers of the noisy speech, clean speech, and noise in the cth channel of the mth frame, respectively. The ratio mask can be estimated as follows:

$$V_R(c,m) = \frac{\hat{P}_x(c,m)}{\hat{P}_x(c,m) + \hat{P}_d(c,m)}, \tag{6}$$

where $V_R(c,m)$ is the initial ratio mask estimation in the cth channel of the mth frame. $\hat{P}_x(\cdot)$ and $\hat{P}_d(\cdot)$ are the estimated powers of speech and noise, respectively. The speech power estimation will be given in Section II-D and the noise power is obtained by the MCRA method [21].

As equation (6), the ratio mask changes from 0 to 1 rather than 0 or 1 happened in the binary mask. The ratio mask value will close to 0, if the current T–F unit is dominated by the noise, that is, the noise power is larger than speech power. On the contrary, the ratio mask approximates to 1, if the speech dominates the current T–F unit. Thus, the soft discrimination factor, ratio mask, can keep parts of speech component in the weak voiced fragments by using a smaller ratio value instead of 0 caused by the binary mask estimation.

The speech components at low frequency are more important than that of at high frequency because the

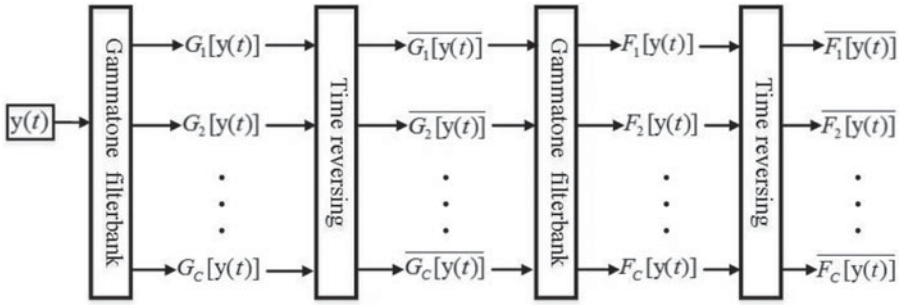

Fig. 2. The Block diagram of the speech synthesis mechanism.

information of low frequency contributes to more speech intelligibility. Therefore, the ratio mask at low frequency is modified by its adjacent T–F units to further preserve the speech energy in the proposed method. The modified ratio mask, $\tilde{V}_R$, is obtained as

$$\tilde{V}_R(c,m) = \begin{cases} V_a(c,m), & \text{if } c \in [1,50] \\ V_R(c,m), & \text{otherwise} \end{cases}, \tag{7}$$

where

$$V_a(c,m) = \frac{V_R(c+2,m) + V_R(c+1,m) + V_R(c,m)}{3}, \tag{8}$$

here $V_a(c,m)$ represents the average of ratio masks in three adjacent T–F units of the same frame. The purpose that we average the adjacent T–F units is to eliminate the outliers units and keep the speech energy. The initial ratio mask defined in equation (6) is only modified below the 50th channel which corresponds to the center frequencies of 636 Hz, because the speech components are more important below this frequency based on the hearing perception. Thus, the basic idea of equation (7) is to recover the partial speech energy that has been masked in the initial ratio mask estimation.

The binary mask can be deemed as a hard decision and is easy to keep more speech components due to its binary discriminant. Also, parts of the noise components are not masked enough and kept at the same time. Comparatively speaking, the ratio mask has a good ability of masking noise, but it simultaneously damages some speech components. Thus, combining the advantages of both ratio mask and binary mask, the linear interpolation between them given in equation (9) is utilized to smooth the ratio mask, when the binary mask value is 1. The smoothed ratio mask is obtained by

$$\hat{V}_R(c,m) = \begin{cases} \eta \cdot \tilde{V}_R(c,m) + (1-\eta) \cdot V_B(c,m), \\ \qquad\qquad \text{if } V_B(c,m) = 1 \\ \tilde{V}_R(c,m), \\ \qquad\qquad \text{if } V_B(c,m) = 0 \end{cases}, \tag{9}$$

where $\eta = 0.2$ is a smoothing factor obtained by massive listening test. Meanwhile, we also use the average HIT-False Alarm rate (HIT-FA) objective test [25] to determine the credibility of $\eta = 0.2$. Figure 3 shows the HIT-FA score curve versus the smoothing factor η which varies from 0 to 1.0. The speech signal is subjected to five types of noises (white, babble, office, street, and factory1 noise) under three kinds of SNRs (0, 5, and 10 dB). The highest point appears in the condition of $\eta = 0.2$ for three SNRs. Thus, in the paper, we set the factor η to 0.2. $V_B(c,m)$ is the estimated binary mask value [25] that will be introduced in the following Section II-C and $\tilde{V}_R(c,m)$ is the modified ratio mask given in equation (7).

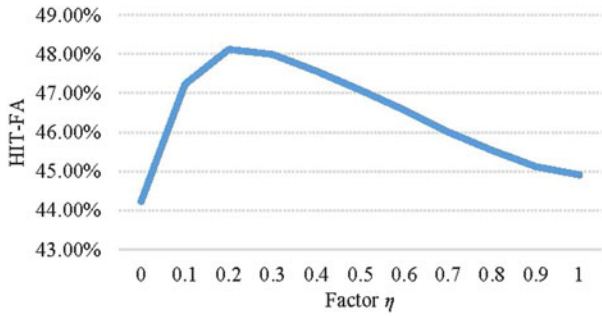

Fig. 3. Average HIT-FA score histogram with respect to factor η.

C) Binary mask estimation

Due to the power estimation error, the estimated powers of speech and noise do not meet equation (5) any more. So, we introduce a factor to solve this issue, that is, the noisy speech power can be expressed as

$$P_y(c,m) = \hat{P}_x(c,m) + \sigma(c,m) \cdot \hat{P}_d(c,m), \tag{10}$$

where $P_y(c,m)$, $\hat{P}_x(c,m)$, and $\hat{P}_d(c,m)$ are the noisy speech power, estimated speech power and estimated noise power in the cth channel of the mth frame, respectively. $\sigma(c,m)$ is a factor to balance equation (10). Thus, the power estimation errors of the speech and noise are compensated by the factor $\sigma(c,m)$. When the factor $\sigma(c,m)$ increases, the noise components will be reduced in the current T–F unit. It means that the current T–F unit is dominated by speech components, when $\sigma(c,m)$ has a larger value. Oppositely, when the factor $\sigma(c,m)$ has a smaller value, the noise

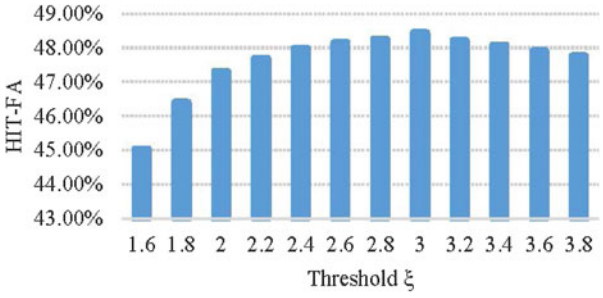

Fig. 4. Average HIT-FA score histogram with respect to threshold ξ.

components dominate the current unit. Based on equation (10), the factor $\sigma(c,m)$ can be given as

$$\sigma(c,m) = \frac{P_y(c,m) - \hat{P}_x(c,m)}{\hat{P}_d(c,m)}. \tag{11}$$

The difference of $P_y(c,m)$ and $\hat{P}_x(c,m)$ in the numerator of equation (11) corresponds to the noise power derived from the speech power estimation. The denominator is the noise power obtained by MCRA method [21]. The factor $\sigma(c,m)$ makes the estimated powers, $\hat{P}_x(\cdot)$ and $\hat{P}_d(\cdot)$, meet equation (5). The $\sigma(c,m)$ can be considered as a boundary factor [25] to distinguish the speech or noise T–F unit. In our experiments, if factor $\sigma(c,m)$ is larger than the threshold ξ, the current T–F unit is dominated by speech components and labeled as value 1. Otherwise, the noise components dominate the current T–F unit labeled as value 0. Thus, the binary mask $V_B(c,m)$ used in equation (9) can be determined as

$$V_B(c,m) = \begin{cases} 1, & if\ \sigma(c,m) > \xi \\ 0, & \text{otherwise} \end{cases}. \tag{12}$$

By a HIT-FA objective test based on equation (12), we found that HIT-FA has a good result when the threshold ξ is chosen as 3. Figure 4 shows the average HIT-FA score histogram about threshold ξ with the value from 1.6 to 3.8 based on five types of noises (white, babble, office, street, and factory1 noise) under three kinds of SNRs situations (0, 5 and 10 dB).

D) Speech power estimation

The enhanced speech is resynthesized by the estimated ratio mask defined in equation (9), which relies on the powers of the estimated speech and noise. Thus, the speech power is the key point of our proposed noise-masking method. The speech power estimation can be deemed as a minimizing problem. We apply the convex optimization [17] that its local optimal solution easily matches the global optimum to solve the minimization problem.

For real and positive speech power vector $\mathbf{p}_x \in \Re^n$ composed of 128 channels, if objective function $J: \Re^n \to \Re$ is convex, $J(\mathbf{p}_x)$ has the minimum value with respect to $\mathbf{p}_x$. The vector $\mathbf{p}_x$ can be estimated by the minimization problem without constrains. The optimal solution can be reached when we optimize each element of $\mathbf{p}_x$ individually. Thus, the optimal value $\hat{P}_x(c,m)$ of $P_x(c,m)$ can be obtained as follows:

$$\hat{P}_x(c,m) = \arg\min_{p_x} J\Big[P_x(c,m)\Big]. \tag{13}$$

The above minimization problem is a kind of convex optimization with respect to variable $P_x(c,m)$ based on objective function $J[P_x(c,m)]$. This convex optimization can be easily implemented by the gradient descent method. Here, the objective function $J(\cdot)$ is built as follows:

$$J\Big[P_x(c,m)\Big] = \sum_{c=1}^{128}\Big[P_y(c,m) - P_d(c,m) - P_x(c,m)\Big]^2 + \lambda \cdot \varphi\Big(P_x(c,m)\Big). \tag{14}$$

The first term in equation (14) is defined in the sense of minimum mean-square error, which is completely convex, i.e., the square error between $P_y(\cdot)$ and $P_d(\cdot) + P_x(\cdot)$ should equal to 0, when $P_d(\cdot)$ and $P_x(\cdot)$ are correctly estimated. Since the estimation errors with respect to $P_d(\cdot)$ and $P_x(\cdot)$ exist in practical situation, the second term in equation (14) is introduced for the error constrain. The function $\varphi(\cdot)$ is called the regularization or penalty function. Here we use the l_1 norm as a penalty function, i.e., $\varphi(P_x) = \sum_{c=1}^{128}|P_x|$. The $\lambda > 0$ is the regularization parameter. By varying the parameter λ, we can trace out the optimal trade-off solution of equation (14). Due to the non-negativity of λ and $\varphi(\cdot)$, equation (14) is also a completely convex function. It means that the approximative value $\hat{P}_x(\cdot)$ can be estimated by minimizing equation (14).

In equation (14), $P_y(\cdot)$ is the noisy speech power obtained in time domain. We assume that the noise has been pre-estimated by the MCRA [21] method. Thus, the objective function (14) only has one variable, $P_x(\cdot)$. Then, we further utilize the relationship of noise and noisy speech to obtain the proportion of noise power within the noisy power. Multiplying noisy power by this proportion, we can get a modified objective function as follows

$$J\Big[P_x(c,m)\Big] = \sum_{c=1}^{128}\Big[P_y(c,m) - \rho(c,m)\cdot P_y(c,m) - P_x(c,m)\Big]^2 + \lambda \cdot \sum_{c=1}^{128} P_x(c,m), \tag{15}$$

where $\rho(c,m)$ is the normalized cross-correlation coefficient [22] between noise and noisy speech spectra calculated in the frequency domain. The $\rho(c,m)\cdot P_y(c,m)$ is considered as an approximative value of noise power. Actually, the factor $\rho(c,m)$ indicates the percentage of noise components within the noisy speech signal. Therefore, we apply this coefficient to represent the proportion of noise power in noisy speech signal, instead of using noise power spectrum directly. The common situation of noise overestimation can make the estimated speech power negative, which is

impossible for the real application. The usage of normalized cross-correlation coefficient cleverly avoids noise overestimation, because $\rho(c,m)$ varies from 0 to 1. Therefore, the $\rho(c,m) \cdot P_y(c,m)$ is always smaller than noisy speech power, $P_y(c,m)$.

Figure 5 shows an examples that two NCCCs vary with the time with respect to the channel index and frame index in five channels. The channel indexes are 20, 50, 70, 100, and 120 which correspond to the center frequencies of 237, 636, 1077, 2195, and 3432 Hz, respectively. Figure 5(a) describes the NCCC, $\chi(c,m)$, between the true noise and noisy speech, and Fig. 5(b) shows the NCCC, $\rho(c,m)$, between of the estimated noise and noisy speech. From Fig. 5, we can find that each Gammatone channel has different NCCC value because the signal energy of each channel is different, where more energy is concentrated at low frequencies. The estimated $\rho(c,m)$ approximately matches to the ideal ratio trajectory $\chi(c,m)$. Thus, it is confident to apply the estimated noise with MCRA to obtain the proportion of noise

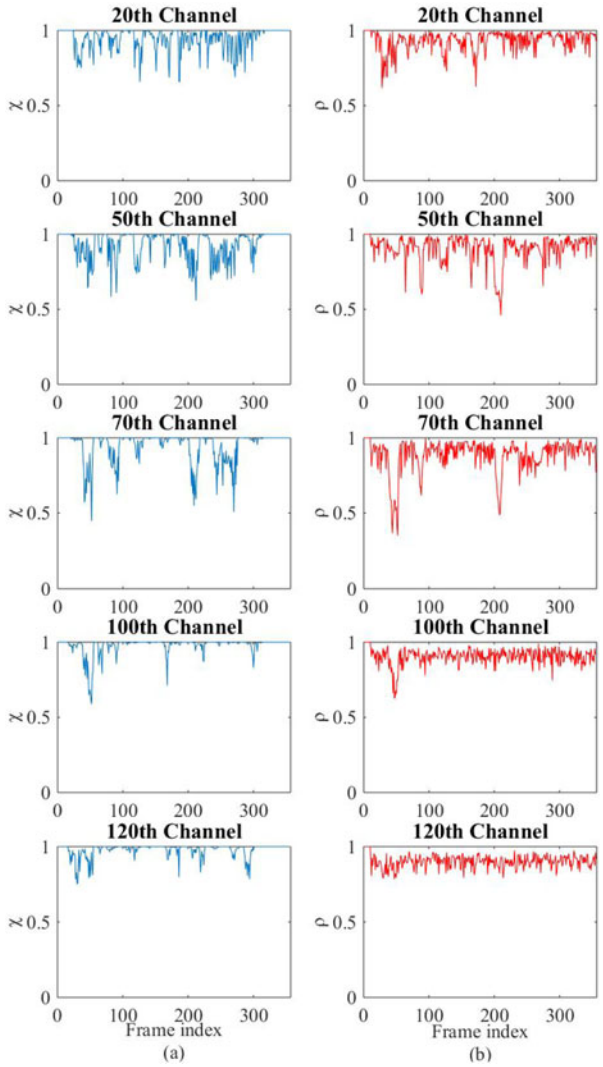

Fig. 5. An example of normalized cross-correlation coefficient in different channels (Input SNR = 5 dB, white noise). (a) True noise condition. (b) Estimated noise condition.

Table 1. Iterative algorithm of $\hat{P}_x$.

Input: $P_y(c,m)$ and $\rho(c,m)$
Output: $\hat{P}_x(c,m)$
For each frame and channel of noisy speech
$k = 0$
iterative step size $\delta = 0.1$
If *iterative error* $> \theta$
$\nabla = \frac{dJ(\hat{P}_x)}{d\hat{P}_x}$
$\hat{P}_x^{(k+1)} \longleftarrow \hat{P}_x^{(k)} - \delta \cdot \nabla$
iterative error$= J(\cdot)^{(k+1)} - J(\cdot)^{(k)}$
$k = k + 1$
End if *error* $\leq \theta$
Return $\hat{P}_x^{(k)}(c,m)$

power in noisy speech. The $\rho(c,m)$ in the unvoiced fragments is larger than that in voiced fragments, because the signal components in the unvoiced segments more like the noisy components. Moreover, $[1-\rho(c,m)] \cdot P_y(c,m)$ can ensure the difference between noisy and noise signals is not negative since $\rho(c,m)$ is less than or equal to 1 based on the following normalized cross-correlation:

$$\rho(c,m) = \frac{\sum_{l=1}^{L} Y(c,m,l) \cdot \hat{D}(c,m,l)}{\sqrt{\sum_{l=1}^{L} Y^2(c,m,l) \cdot \sum_{l=1}^{L} \hat{D}^2(c,m,l)}}, \quad (16)$$

where L is the number of FFT points, l is the frequency index, and $Y(c,m,l)$ and $\hat{D}(c,m,l)$ are the spectral magnitude of noisy speech and the estimated noise in the cth channel of the mth frame, which are calculated by FFT with the size of 256 and MCRA method [21], respectively.

Equation (15) can be solved by the gradient descent method [23] to calculate the approximative value of the estimated speech power, $\hat{P}_x(c,m)$, in each channel of each frame. The complete algorithm framework is presented in Table 1. The input of iteration algorithm is noisy power and $\rho(c,m)$. The output of iteration algorithm is the optimal solution of speech power. By taking derivation of $J(\cdot)$ in equation (15) with respect to $\hat{P}_x(c,m)$, we can get the gradient ∇ of objective function. Then, moving $\hat{P}_x(c,m)$ to the direction of the negative gradient to obtain the $(k+1)$th iteration solution. After that, the iterative error of adjacent two iterations is computed to determine if the iteration is over. The iteration will stop, if the iterative error is smaller than threshold θ where it is set to 1 based on objective and subjective tests. Otherwise, the iteration will keep going until convergence.

III. EXPERIMENTS AND RESULTS

In this section of performance evaluation, we discuss the enhanced results of our proposed method named as ConvexRM. The test clean speeches are selected from TIMIT [26] database including 50 utterances. Five types of noises from NOISEX-92 [27] noise database are used, which include white, babble, office, street, and factory1 noises. The input SNR is defined as −5, 0, 5 and 10 dB, respectively.

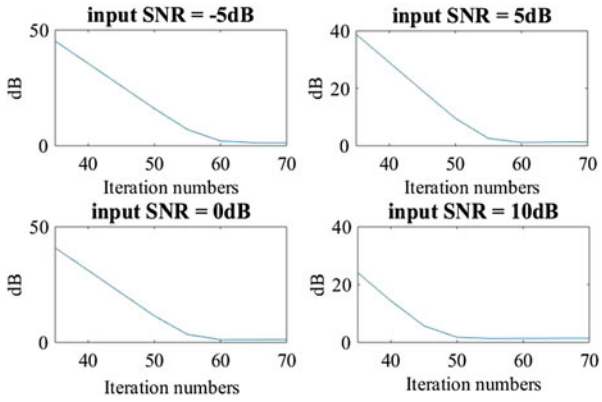

Fig. 6. Power error comparison of speech against the number of Iteration.

The sampling rate of the noisy speech signal is 8 kHz. We apply segmental SNR (SSNR) improvement measure [28], log spectral distortion (LSD) measure [29], and short-time objective intelligibility (STOI) [30] to evaluate the objective quality of the enhanced signal. Meanwhile, the Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) listening test [31] is utilized to measure the subjective performance. To put our results in perspective, the GammaPrior [5], SARHMM [9], and CBSPP [10] are selected as reference enhancement methods and their related ready-made batch program and source codes are used for tests. The IRM is the ideal situation that the clean speech and noise signals are known in advance.

A) Experiment setup

In the Section II-D, the objective function (15) was solved by the gradient descent method. The iteration times should be seriously considered. A large number of iterations can not only obtain a better estimated result but also make the enhanced system very complicated. However, the enhanced performance may be degraded, if the number of iterations is unreasonably constrained. Therefore, we analyze the average iteration error in all 128 channels between 30 and 70 iterations to ensure the reasonable number of iterations. An example of average errors in terms of "dB" is expressed in Fig. 6. In this error comparison, we use four kinds of SNRs under five types of noises to determine the number of iterations. From Fig. 6, we can find that the power error reaches a lower value when the number of iterations is larger than 60. Therefore, during the convex optimization of speech power, the number of iterations in the gradient descent method is set to 60, which adequately satisfies the convergent condition.

Although the proposed method needs to iteratively estimate the speech power, we set a fixed iterative number of 60 to reduce the computation complexity. The SARHMM and CBSPP not only need the prior information and big database but also cost a lot of time to search the mapping pairs during the online enhancement. However, the iterative estimation of our proposed method is very simple with a small iteration number. Our proposed method is a slightly more complex than the GammaPrior approach, but the proposed method does not need the additive Gaussian noise assumption and the complexity of it is tolerable.

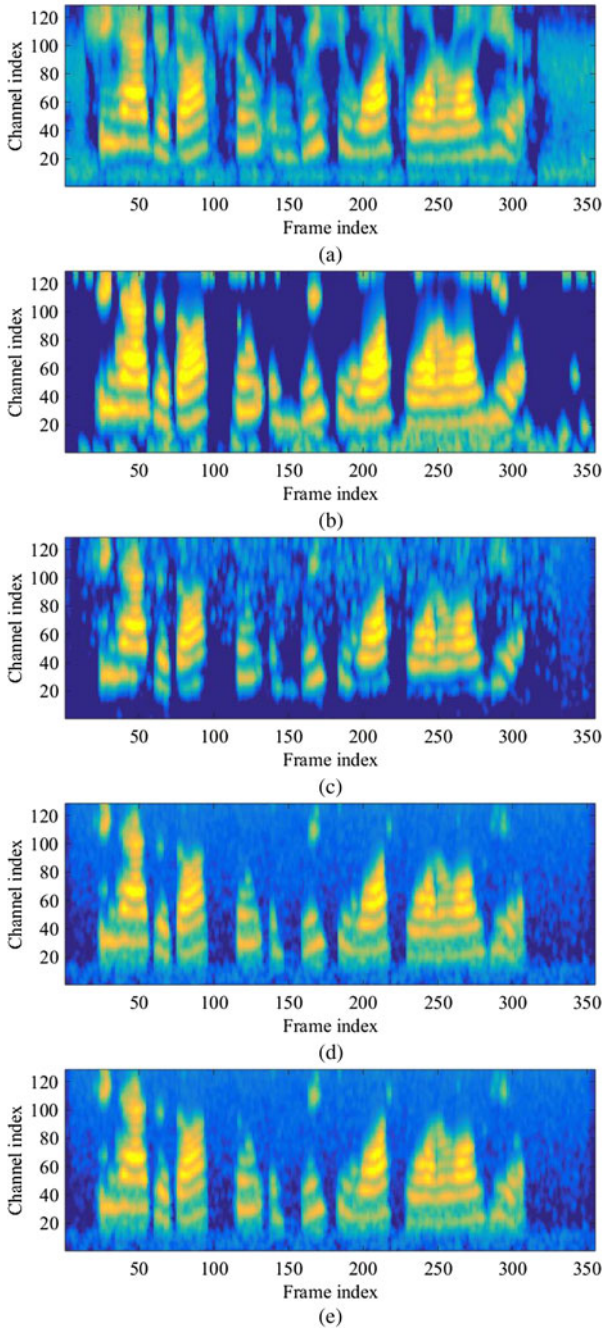

Fig. 7. The cochleogram comparison (Input SNR = 5 dB, factory1 noise). (a) The cochleogram resynthesized by idea ratio mask; (b) The cochleogram resynthesized by the estimated binary mask V_B; (c) The cochleogram resynthesized by the initial ratio mask V_R; (d) The cochleogram resynthesized by the modified ratio mask with adjacent T–F units $\widetilde{V}_R$; (e) The cochleogram resynthesized by the smoothed ratio mask with binary mask $\hat{V}_R$.

An example of cochleogram comparison based on ratio mask estimations is given by Fig. 7 to observe the estimation performance. Figure 7(a) is the cochleogram based on the ideal ratio mask that the clean speech and noise are known in advance. Figure 7(b) is the cochleogram obtained through the estimated binary mask by equation

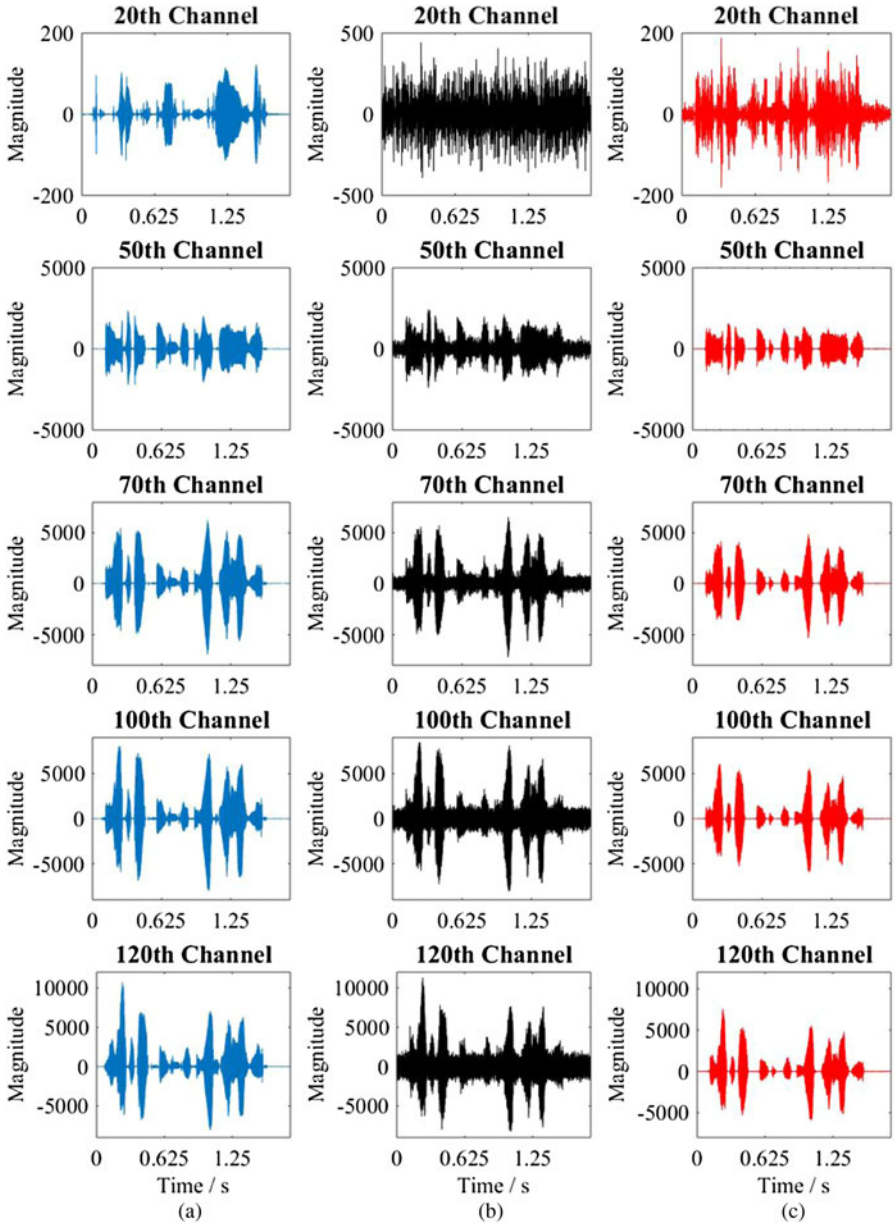

Fig. 8. Speech waveform comparison of five channels (Input SNR = 5 dB, white noise). (a) Clean speech; (b) Noisy speech; (c) Enhanced speech.

(12). Figure 7(c) is the cochleogram acquired via the initial estimation of ratio mask by equation (6). Figure 7(d) is the cochleogram based on the modified ratio mask by adjacent T–F units and the cochleogram of Fig. 7(e) is the final estimated ratio mask interpolated with the binary mask in equation (9). The speech components at low frequency usually contain some useful information and, so it is unrealistic to totally remove the background noise at low frequency. Thus, in Figs 7(d) and 7(e), we deliberately keep the very little energy of the T–F units as the floor noise below the 12th Gammetone channels that correspond to the center frequencies of 120 Hz. The binary mask estimation loses some speech T–F units at low frequency and keeps parts of speech units at the middle frequency. The initial ratio mask has relatively good results on remaining the speech components at the middle frequency but misses too many speech T–F units at low frequency. By using the modified algorithm given by equation (7), the ratio masks at low frequency are recovered. Moreover, Combining the feature of better preservation of speech T–F units at the middle frequency of binary mask, the smoothed ratio mask given by equation (9) recovers and enlarges the edge of speech T–F units. Based on the above analysis, the final estimated ratio mask ($\hat{V}_R$) obtains a better performance and is considered as the key point in our proposed method (ConvexRM).

To further observe the noise-masking performance in each Gammatone channel, Fig. 8 shows an example that the waveforms vary with the time in five channels. The clean speech, noisy speech and enhanced speech are shown in Figs 8(a)–8(c), respectively. From this figure, we can find that the enhanced speech waveforms match the clean speech waveforms well in the 50th, 70th, 100th, and 120th channels. For the 20th channel, the enhanced speech has more background noise than clean speech, because Gammatone filter

prefers to reinforce the energy at low frequency. Actually, this energy level of noise does not affect the hearing quality too much.

To demonstrate the effectiveness of the proposed method with 60 iterations using the gradient descent method, the spectrograms of the enhanced speech by different methods are shown in Fig. 9 for verifying the noise masking.

From the Fig. 9, we can find that the proposed ConvexRM method masks more background noise and keeps more speech components than other three reference approaches, respectively. The GammaPrior algorithm removes the least noise in these methods. Although the SARHMM wipes off noise components to a certain degree, it also loses parts of speech components in high frequency. As the CBSPP approach, the parts of weak speech components are eliminated after enhancement. Our ConvexRM method has better results on noise masking and speech energy reservation. To further observe the performance of noise elimination, The SSNR, LSD, STOI, and MUSHRA tests are discussed in the next subsections.

B) SSNR improvement test results

The average SSNR measurement [28] is often utilized to evaluate the denoising performance of speech enhancement method. The input SNR and the output SNR are defined as follows, respectively. The average SSNR improvement is obtained by subtracting S_{in} from S_{out}.

$$S_{in} = 10 \cdot \log_{10} \frac{\sum_{n=0}^{N-1} x^2(n)}{\sum_{n=0}^{N-1} [x(n) - y(n)]^2}, \tag{17}$$

$$S_{out} = 10 \cdot \log_{10} \frac{\sum_{n=0}^{N-1} x^2(n)}{\sum_{n=0}^{N-1} [x(n) - \hat{x}(n)]^2}, \tag{18}$$

where N is the number of samples. $x(n)$ is the original clean speech, $y(n)$ represents the input noisy speech and $\hat{x}(n)$ denotes the enhanced signal.

The average SSNR improvement of various enhancement methods for stationary and non-stationary noise conditions are presented in Table 2 for different input SNRs (i.e., −5, 0, 5, and 10 dB). As Table 2, the ConvexRM method generally shows a higher value than other three reference methods. To give more details, the proposed ConvexRM is a little higher than the CBSPP method in babble and office noises. Both CBSPP and SARHMM approaches are better than the GammaPrior on the performance of reducing the background noise.

C) LSD test results

During the signal enhancement, although parts of background noises are removed, the signal may also distort at the same time. Therefore, in order to further check the spectrum distortion of the enhanced signal, the LSD measure [29] is employed to evaluate the objective quality of the enhanced speech. It measures the similarity between the clean speech spectrum and the enhanced speech spectrum,

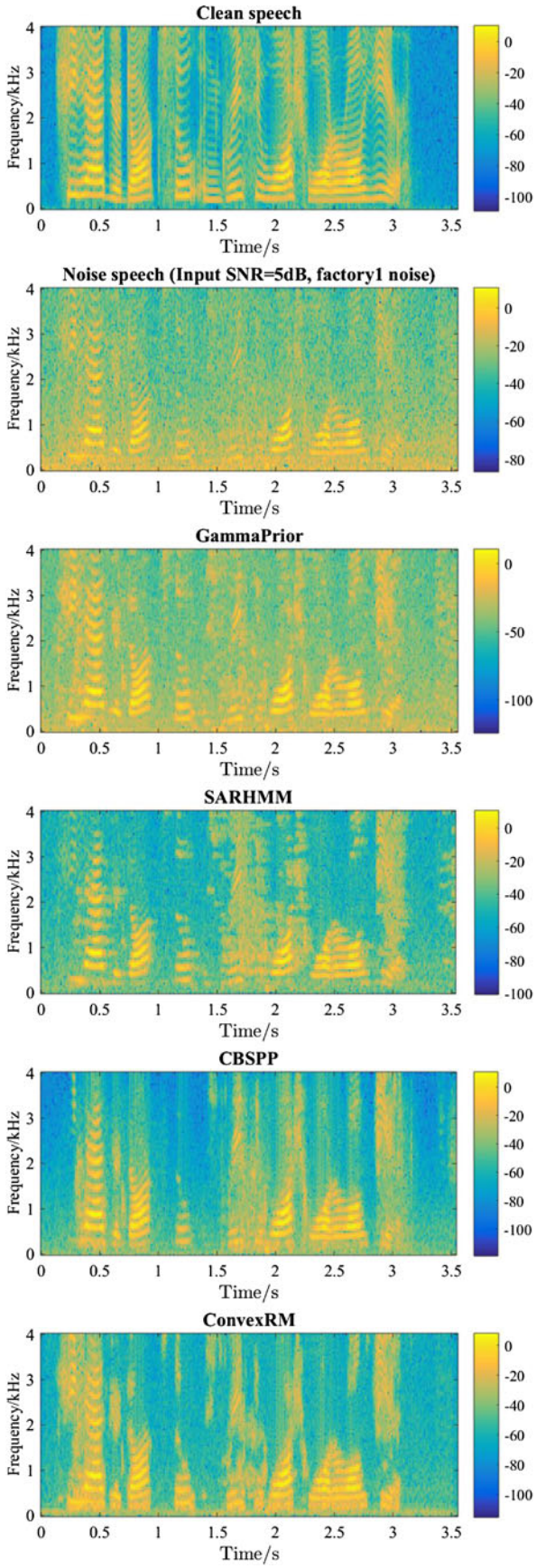

Fig. 9. Spectrogram comparison (Input SNR = 5 dB, factory1 noise), (*"She had your dark suit in greasy wash water all year"*).

Table 2. SSNR improvement results.

Noise type	Methods	−5 dB	0 dB	5 dB	10 dB
White	IRM	19.89	15.03	12.77	9.32
	GammaPrior	10.01	8.95	7.62	5.94
	SARHMM	13.06	10.72	8.52	6.34
	CBSPP	16.01	13.02	10.07	6.86
	ConvexRM	16.19	13.48	10.42	7.09
Babble	IRM	15.12	14.12	11.51	7.99
	GammaPrior	7.24	6.86	5.65	4.47
	SARHMM	8.50	7.43	6.31	5.04
	CBSPP	10.49	9.03	7.41	5.04
	ConvexRM	10.78	9.29	7.53	5.40
Office	IRM	16.12	14.89	12.82	9.02
	GammaPrior	9.33	8.61	7.46	6.05
	SARHMM	10.96	9.55	7.41	6.30
	CBSPP	11.28	9.47	7.74	5.62
	ConvexRM	11.37	9.78	7.85	6.33
Street	IRM	18.29	16.12	13.92	11.65
	GammaPrior	11.89	10.96	9.54	7.54
	SARHMM	12.81	11.54	9.98	7.43
	CBSPP	12.92	10.92	8.67	7.34
	ConvexRM	15.85	13.19	10.51	8.06
Factory1	IRM	17.23	15.73	13.12	10.65
	GammaPrior	9.14	8.09	6.84	5.32
	SARHMM	10.33	8.86	7.38	5.62
	CBSPP	12.93	10.60	8.09	5.63
	ConvexRM	13.00	10.87	8.45	5.78

and is defined as

$$l = \frac{1}{M}\sum_{m=0}^{M-1}\sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\left[10\cdot\log_{10}\frac{|\hat{X}(m,k)|^2}{|X(m,k)|^2}\right]^2}, \quad (19)$$

where k is the index of frequency bins. $K = 512$ is the FFT size. m is the frame index, and M is the total number of frames. $|X(m,k)|$ denotes the clean speech amplitude of DFT coefficients, and $|\hat{X}(m,k)|$ denotes the enhanced speech amplitude of DFT coefficients.

The LSD test results are given in Table 3. From Table 3, we can see that the ConvexRM method has less distortion than other methods in −5, 0, 5, and 10 dB conditions. Additionally, the SARHMM method almost shares the same level of spectral distortion with our proposed system in 10 dB situation that our method is still better than other two reference approaches. It means that our ConvexRM method masks more background noise and causes less spectral distortion.

D) STOI test results

STOI [30] denotes a correlation of short-time temporal envelopes between clean and enhanced speech, and has been shown to be highly correlated to human speech intelligibility score. The STOI measure is derived based on a correlation coefficient between the temporal envelops of the clean and enhanced speech in short-time regions and the score of STOI ranges from 0 to 1. The higher the STOI value is, the more the intelligibility has. Table 4 describes the STOI results that the enhanced signal by ConvexRM holds the highest intelligibility among all methods, especially in the low SNR conditions (−5 and 0 dB). Three reference

Table 3. LSD results.

Noise type	Methods	−5 dB	0 dB	5 dB	10 dB
White	Noisy	14.58	12.85	11.48	8.66
	IRM	5.29	4.67	4.01	3.12
	GammaPrior	9.34	7.98	6.75	5.94
	SARHMM	7.78	7.07	6.46	5.92
	CBSPP	9.18	8.17	8.13	7.69
	ConvexRM	7.15	6.69	6.33	5.82
Babble	Noisy	10.87	9.03	7.44	6.04
	IRM	5.01	4.12	3.23	2.34
	GammaPrior	8.09	6.83	5.77	5.05
	SARHMM	7.84	6.76	5.85	5.01
	CBSPP	8.09	7.40	6.93	5.78
	ConvexRM	7.39	6.58	5.73	4.97
Office	Noisy	10.01	8.27	6.75	5.45
	IRM	4.45	3.67	3.05	2.21
	GammaPrior	7.23	6.06	5.21	4.60
	SARHMM	7.03	6.09	5.34	4.61
	CBSPP	7.35	6.73	6.32	5.42
	ConvexRM	6.56	5.79	5.16	4.58
Street	Noisy	9.17	7.53	6.18	5.05
	IRM	3.72	3.06	2.46	2.09
	GammaPrior	6.10	5.18	4.87	4.27
	SARHMM	6.62	5.74	4.95	4.29
	CBSPP	6.56	6.23	5.68	5.01
	ConvexRM	5.98	5.28	4.82	4.24
Factory1	Noisy	12.43	10.46	8.61	7.01
	IRM	5.59	4.88	3.75	3.05
	GammaPrior	8.35	7.02	5.96	5.35
	SARHMM	8.14	7.02	6.04	5.36
	CBSPP	8.48	7.02	6.04	5.36
	ConvexRM	7.28	6.49	5.87	5.34

methods even are lower than the noisy signal in several situations(e.g., 10 dB white, −5, and 0 dB office noise) because they lose parts of speech components in the weak speech fragments. All the methods obtained the high scores under street noise. The reason is that the street noise given in NOISEX-92 database is a kind of relatively stationary noise and the more noise energy is usually existed at the low frequency. Actually, the street noise is also easier to process than white and much easier than the babble or office noise.

E) Subjective listening test results

In our experiments, the MUSHRA listening test [31] is used to evaluate the subjective quality of the enhanced speech. The MUSHRA listening test consists of several successive experiments. Each experiment aims to compare a high-quality reference speech (i.e., clean speech) to several test speech signals sorted in random order, in which the subjects are provided with the signals under test as well as one clean speech and hidden anchor. Each subject needs to grade the whole test speech signals on a quality scale between 0 and 100.

During the MUSHRA test, we used as hidden anchor a speech signal having an SNR of 0 dB less than the noisy speech to be enhanced. Seven male and seven female listeners participated in the tests, and each listener did the test two times. The listeners were allowed to listen to each test

Table 4. STOI results.

Noise type	Methods	−5 dB	0 dB	5 dB	10 dB
White	Noisy	0.52	0.64	0.76	0.86
	IRM	0.77	0.80	0.84	0.90
	GammaPrior	0.52	0.64	0.76	0.85
	SARHMM	0.51	0.63	0.74	0.83
	CBSPP	0.52	0.63	0.73	0.81
	ConvexRM	0.56	0.68	0.79	0.86
Babble	Noisy	0.51	0.63	0.74	0.83
	IRM	0.76	0.79	0.83	0.89
	GammaPrior	0.51	0.63	0.74	0.82
	SARHMM	0.49	0.61	0.72	0.81
	CBSPP	0.49	0.61	0.72	0.79
	ConvexRM	0.52	0.65	0.77	0.85
Office	Noisy	0.63	0.72	0.80	0.86
	IRM	0.78	0.81	0.85	0.92
	GammaPrior	0.61	0.71	0.79	0.85
	SARHMM	0.55	0.66	0.75	0.82
	CBSPP	0.60	0.66	0.77	0.82
	ConvexRM	0.65	0.75	0.82	0.88
Street	Noisy	0.75	0.81	0.86	0.90
	IRM	0.84	0.88	0.91	0.95
	GammaPrior	0.76	0.82	0.86	0.90
	SARHMM	0.68	0.75	0.81	0.87
	CBSPP	0.74	0.79	0.88	0.86
	ConvexRM	0.80	0.85	0.88	0.92
Factory1	Noisy	0.51	0.63	0.75	0.84
	IRM	0.75	0.80	0.84	0.91
	GammaPrior	0.51	0.64	0.75	0.84
	SARHMM	0.47	0.61	0.73	0.83
	CBSPP	0.46	0.60	0.72	0.81
	ConvexRM	0.53	0.67	0.78	0.87

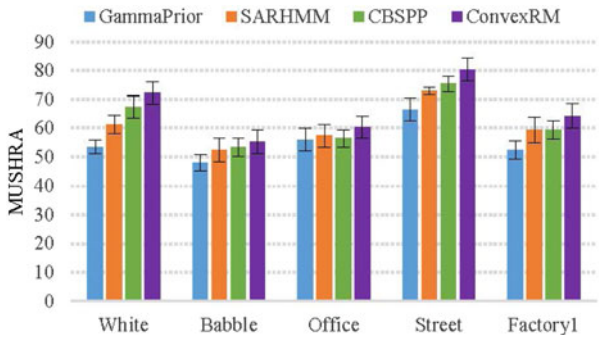

Fig. 10. The MUSHRA results for five types of noises.

speech several times and had access to the clean speech reference. The ten test utterances used were contaminated by aforementioned five types of noises using a 5 dB input SNR. A statistical analysis of the test results was conducted for the different de-noising methods under five noise conditions.

Figure 10 shows the averaged MUSHRA listening test results with a 95% confidence interval. Most of the listeners preferred to the proposed ConvexRM method over the other methods under five types of noises. SARHMM and CBSPP are the very competing methods to the proposed one. In the situations of white, street and factory1 noises, ConvexRM obtains an obvious preference compared with other three reference approaches. As the conditions of babble and office noise, most listeners still chose the proposed ConvexRM method despite its score has a little decline. From the hearing perception, the enhanced speech by the proposed method is more comfortable and continuous than SARHMM and CBSPP methods. Meanwhile, the proposed method can feel less background noise than GammaPrior method.

IV. CONCLUSIONS

In this paper, we proposed a novel method for noise masking based on an effective estimation of ratio mask in Gammatone domain instead of DFT domain. The convex optimization algorithm was applied to estimate the speech power, combined with an adaptive factor named NCCC. The adjacent T–F units were considered for keeping the speech components at the low frequency. To recover weak speech components, the linear interpolation between the ratio mask and the binary mask was utilized to smooth the estimated ratio mask. By objective measures and subjective listening test, our proposed method has shown a better performance than other three reference methods, especially in the low SNR conditions the improvement of intelligibility is obvious. This also implies that our noise-masking method is effective.

REFERENCES

[1] Boll, S.: Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust., Speech, Signal Process*, **ASSP-27** (2) (1979), 113–120.

[2] Li, C.; Liu, W.J.: A novel multi-band spectral subtraction method based on phase modification and magnitude compensation, in *Proc. IEEE ICASSP*, 2011, 4760–4763.

[3] Loizou, P.C.: Speech Enhancement: Theory and Practice, CRC Press, Boca Raton, FL, USA, 2007.

[4] Ephraim, Y.; Malah, D.: Speech enhancement using a minimum mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.*, **32** (6) (1984), 1109–1121.

[5] Erkelens, J.S.; Hendriks, R.C.; Heusdens, R.; Jensen, J.: Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors. *IEEE Trans. Audio, Speech, Lang. Process.*, **15** (6) (2007), 1741–1752.

[6] Zoghlami, N.; Lachiri, Z.; Ellouze, N.: Speech enhancement using auditory spectral attenuation, in *EUSIPCO 2009*, Scotland, 24–28 August 2009.

[7] Zhao, D.Y.; Kleijn, W.B.: HMM-Based gain modeling for enhancement of speech in noise. *IEEE Trans. Audio, Speech, Lang. Process.*, **15** (3) (2007), 882–892.

[8] Srinivasan, S.; Samuelsson, J.; Kleijn, W.B.: Codebook driven short term predictor parameter estimation for speech enhancement. *IEEE Trans. Audio, Speech, Lang. Process.*, **14** (1) (2006), 163–176.

[9] Deng, F.; Bao, C.C.; Kleijin, W.B.: Sparse hiddenMarkov models for speech enhancement in non-stationary noise environments. *IEEE Trans. Audio, Speech, Lang. Process.*, **23** (11) (2015), 1973–1987.

[10] He, Q.; Bao, F.; Bao, C.C.: Multiplicative update of auto-regressive gains for codebook-based speech enhancement. *IEEE Trans. Audio, Speech, Lang. Process.*, **25** (3) (2017), 457–468.

[11] Hu, G.; Wang, D.L.: Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Trans. Neural Netw.*, **15** (5) (2004), 1135–1150.

[12] Bao, F.; Abdulla, W.H.: A Noise Masking Method with Adaptive Thresholds based on CASA, APSIPA, Jeju, South Korea, 2016.

[13] Wang, Y.; Narayanan, A.; Wang, D.L.: On training targets for supervised speech separation. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, **22** (12) (2014), 1849–1858.

[14] Williamson, D.S.; Wang, Y.X.; Wang, D.L.: Complex ratio masking for monaural speech separation. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, **24** (3) (2016), 483–493.

[15] Madhu, N.; Spriet, A.; Jansen, S.; Koning, R.; Wouters, J.: The potential for speech intelligibility improvement using the ideal binary mask and the ideal Wiener filter in single channel noise reduction systems: application to auditory prostheses. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, **21** (1) (2013), 63–72.

[16] Koning, R.; Madhu, N.; Wouters, J.: Ideal timeÍCfrequency masking algorithms lead to different speech intelligibility and quality in normal-hearing and Cochlear implant listeners. *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, **62** (1) (2014), 331–341.

[17] Boyd, S.; Vandenberghe, L.: Convex Optimization, Cambridge University Press, 2004.

[18] Bao, F.; Abdulla, W.H.: A new IBM estimation method based on convex optimization for CASA. *Speech Commun.*, **97** (2018), 51–65.

[19] Patterson, R.D.; Nimmo-Smith, I.; Holdsworth, J.; Rice, P.: An Efficient Auditory Filterbank based on the Gammatone Function, Appl. Psychol. Unit, Cambridge Univ., Cambridge, UK, 1998.

[20] Abdulla, W.H.: Advance in Communication and Software Technologies, Chapter Auditory Based Feature Vectors for Speech Recognition Systems, WSEAS Press, 2002, pp. 231–236.

[21] Cohen, I.: Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE Signal Process. Lett.*, **9** (1) (2002), 12–15.

[22] Bao, F.; Dou, H.J.; Jia, M.S.; Bao, C.C.: A novel speech enhancement method using power spectra smooth in wiener filtering, in *APSIPA*, 2014.

[23] Gardner, W.A.: Learning characteristics of stochastic gradient-descent algorithms: a general study, analysis, and critique. *Signal Process*, **6** (2) (1984), 113–133.

[24] Weintraub, M.: A Theory and Computational Model of Auditory Monaural Sound Separation. Ph.D. dissertation, Dept. Elect. Eng., Stanford University, Stanford, CA, 1985.

[25] Bao, F.; Abdulla, W.H.: A convex optimization approach for time-frequency mask estimation, in *WASPAA*, 2017, pp. 31–35.

[26] Garofolo, J.S.; Lamel, L.F.; Fisher, W.M.; Fiscus, J.G.; Pallett, D.S.; Dahlgrena, N.L.: DARPA- TIMIT, Acoustic Phone Ticcontinuous Speech Corpus, US Department of Commerce, Washington, DC, 1993 (NISTIR Publication No. 4930).

[27] Varga, A.P.; Steeneken, H.J.M.; Tomlinson, M.; Jones, D.: The NOISEX-92 study on the effect of additive noise on automatic speech recognition. http://spib.rice.edu/spib/select, 1992.

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

[29] Abramson, A.; Cohen, I.: Simultaneous detection and estimation approach for speech enhancement. *IEEE Trans. Audio, Speech, Lang. Process*, **15** (8) (2007), 2348–2359.

[30] Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J.: An algorithm for intelligibility prediction of timeÍCfrequency weighted noisy speech. *IEEE Trans.Audio, Speech, Lang. Process*, **19** (7) (2011), 2125–2136.

[31] Vincent, E.: MUSHRAM: A MATLAB interface for MUSHRA listening tests, [Online]. Available: http://www.elec.qmul.ac.uk/ people/ emmanuelv/mushram, 2005.

Feng Bao received the B.S. and M.S. degrees in Electronic Engineering from Beijing University of Technology in 2012 and 2015, respectively. From October 2015, he started to pursue Doctor degree in the Department of Electrical and Computer Engineering, University of Auckland, Auckland, New Zealand. His research interests are in the areas of speech enhancement and speech signal processing. He is the author or coauthor of over 20 papers in journals and conferences.

Waleed H. Abdulla holds a Ph.D. degree from the University of Otago, New Zealand. He is currently an Associate Professor in the University of Auckland. He was Vice President-Member Relations and Development (APSIPA). He has published more than 170 refereed publications, one patent, and two books. He is on the editorial boards of six journals. He has supervised over 25 postgraduate students. He is the recipient of many awards and funded projects such as JSPS, ETRI, and Tsinghua fellowships. He has received Excellent Teaching Awards for 2005 and 2012. He is also a Senior Member of IEEE. His research interests include human biometrics; signal, speech, and image processing; machine learning; active noise control.