Applied Acoustics 110 (2016) 128–136

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Stereophonic channel decorrelation using a binaural masking model

Hefei Yang [a,b], Jie Wang [c], Chengshi Zheng [a,b,*], Xiaodong Li [a,b]

[a] Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, 100190 Beijing, China
[b] Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China
[c] Institute of Acoustics and Lighting Technology, Guangzhou University, 510006 Guangzhou, China

ARTICLE INFO

Article history:
Received 29 March 2015
Received in revised form 23 January 2016
Accepted 20 March 2016
Available online 31 March 2016

Keywords:
Channel decorrelation
Binaural masking
Sinusoidal injection

ABSTRACT

Traditional methods often only use monaural masking models to decorrelate input signals for stereo acoustic echo cancellation. Whereas, it seems more reasonable to use binaural masking models for the following two reasons. First, stereo signals are heard by two ears rather than just one. Second, psychoacoustic researchers have already shown that there are obvious masking level differences between binaural masking models and monaural masking models. By studying binaural masking level difference models, we first introduce a simplified binaural masking model for stereo acoustic echo cancellation. Considering that the interaural time difference is dominant at low frequencies ($\leqslant$1.5 kHz) and the interaural level difference is a major cue at higher frequencies, we propose to use different signal decorrelation schemes at these two frequency bands. In the low-frequency band, a pitch-driven sinusoidal injection scheme is proposed to maintain the interaural time difference, where the amount of injection is determined by the proposed binaural masking model. In the high-frequency band, a modified sinusoidal phase modulation scheme is applied to make a trade-off between preserving the interaural level difference and decorrelating the stereophonic input signals. Assessment results show that the proposed method can effectively improve the non-unique problem and retain good speech quality.

1. Introduction

In stereo acoustic echo cancellation (SAEC) (see Fig. 1), the stereo signals, $x_1(n)$ and $x_2(n)$, are usually highly correlated such that the adaptive filters have non-unique solutions [1]. As a result, the filter misalignment may be large even when the acoustic echo is well cancelled. In this case, the echo cancellation degrades rapidly when the transfer functions between the loudspeakers and the microphones change.

To settle the non-unique problem in SAEC, lots of decorrelation algorithms have already been proposed. They can be roughly classified into three categories. The first one was to add uncorrelated signals to each channel, such as the independent noise injection/modulation methods, and the nonlinear technique (see [1–5] and references therein). The second one was to remove some signal components from one of the two stereophonic channels, where these methods are often based on some psychoacoustic effects, such as the missing fundamental phenomenon [6] and the spectral dominant effect [7]. The last one was to modulate the stereo signal itself, such as the all-pass filtering solution [8], the time reversal skill [9] and the sinusoidal phase modulation (SPM) technique [10]. Note that, compared with the first two categories of algorithms, the modulation methods do not alter the power spectral densities of the stereo signals. In order to achieve better performance, some researchers combined different algorithms together [11,12].

It is well-known that all of these existing decorrelation algorithms may cause some spectral and/or phase distortion, which can degrade speech quality dramatically. In practical applications, it is important to find some guidelines to make a good trade-off between speech quality and interchannel decorrelation. The most popular way is to control the amount of distortion according to some psychoacoustic models [13,14], where the main idea is to ensure that the additional distortion is inaudible. To the best of our knowledge, only the monaural masking models (Mo-MM) were applied in previous studies [13,14]. Unfortunately, these Mo-MM motivated approaches conflict with the SAEC system. Because the stereo signals are heard by two ears rather than just one, it seems more reasonable to use binaural masking models (Bi-MM) for the SAEC system. Note that the binaural cues has already been successfully applied to some fields, such as binaural noise reduction [15], speech recognition [16], and speech dereverberation [17].

* Corresponding author at: Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, 100190 Beijing, China.
E-mail address: cszheng@mail.ioa.ac.cn (C. Zheng).

http://dx.doi.org/10.1016/j.apacoust.2016.03.016
0003-682X/

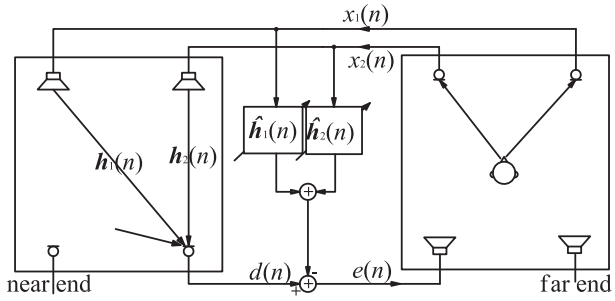

Fig. 1. Schematic diagram of an SAEC system, considering one of the two microphone channels in the stereo system.

Due to the contribution of spatial cues, it is more difficult to mask a sound in binaural hearing systems than in monaural hearing systems. In other words, the masking threshold of the Bi-MM is generally much lower than that of the Mo-MM, where the threshold difference is called binaural masking level difference (BMLD). In this paper, we first discuss the difference between the Mo-MM and the Bi-MM to show the importance of exploiting the Bi-MM in SAEC. Then a Bi-MM applicable to SAEC is built on the basis of the study of BMLD. After that, we propose a pitch-driven sinusoidal injection (PDSI) technique, where the amount of injection is determined by the Bi-MM. As we all know, the stereophonic perception depends more on interaural time difference (ITD) at frequencies below 1.5 kHz and more on interaural level difference (ILD) at higher frequencies [13]. Thus it is better to use various decorrelation techniques in different frequency bands. The proposed Bi-MM-based PDSI is applied to the low-frequency band to better preserve the ITD. On the other hand, an SPM [10] is applied to the high-frequency band, so as to make a promising trade-off between maintaining the ILD and decorrelating the stereo signals. Simulation results verify the better performance of the proposed algorithm.

2. Comparative studies of the binaural masking model and the monaural masking model

2.1. Relationship between the Bi-MM and the Mo-MM

As illustrated in Fig. 2, the Bi-MM and the Mo-MM differ from each other in the following way: we can use the Mo-MM only when the masker source and the masked source are located in the same direction with reference to the listener; otherwise the Bi-MM should be used. For the Mo-MM, because the masker and the masked sources are of the same interaural relationship, the listener can distinguish the two sources by means of only monaural cues [18], e.g., magnitude. For the Bi-MM, however, the two sources are interaurally different, thus not only monaural cues but also spatial cues, such as ITD and ILD [19], can contribute to the unmasking of the masked signal [20].

With the help of the spatial cues, it is more difficult to mask a source in binaural hearing than in monaural hearing [21]. That is to say, the masking threshold of the Bi-MM is generally lower than that of the Mo-MM. Let $\mathrm{SMR}_{\mathrm{bi}}$ and $\mathrm{SMR}_{\mathrm{mo}}$ denote the masking threshold of the Bi-MM and that of the Mo-MM, respectively. We have

$$\mathrm{SMR}_{\mathrm{bi}}(k) = \mathrm{SMR}_{\mathrm{mo}}(k) - \mathrm{BMLD}(k), \tag{1}$$

where k is the frequency index. BMLD could be quite large. For example, it reaches as large as 9 dB when the masker source and the masked source are separately located at 0° and 60° [22].

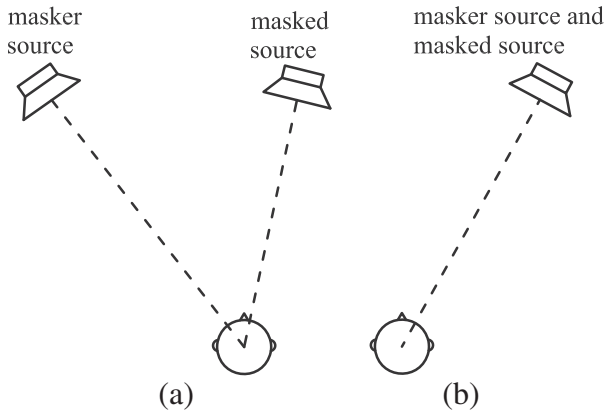

Fig. 2. Illustration of hearing situations, where the Bi-MM and the Mo-MM are separately used in (a) and (b). The masker and the masked sources are located in different directions in (a), while they are in the same direction in (b).

Moreover, BMLD could become larger when the masker source moves further away from the masked source [22].

In [13,14], two independent noises are injected for the stereophonic channel decorrelation. In such a case, we should use the Bi-MM instead of the Mo-MM for the following reason. The high correlation between the stereophonic channels leads to the fact that the masker source (the original stereo signal) and the masked source (the injected noise) are positioned in different directions. However, the traditional methods often use the Mo-MM, which may cause some audible distortion for the SAEC system due to that the BMLD could be quite high. To solve this problem, we propose to consider the Bi-MM for decorrelation.

2.2. A Simplified Bi-MM for SAEC

To be applicable to SAEC situations, a Bi-MM is needed. Unfortunately, until now, there are not any exact and closed-form expressions of the Bi-MM for SAEC. In this paper, efforts are made to establish a closed-form Bi-MM considering the SAEC system in Fig. 3. The listener sits at the sweet spot which together with the two loudspeakers make up an equiangular triangle [23], i.e. the directions of the loudspeakers are $\theta_1 = -\theta_2 = -30^\circ$; $H_c(k), H_s(k)$ and $H_o(k)$ represent the transfer functions from a certain loudspeaker to the head center, the ear in the same side and the ear in the opposite side, respectively; θ_x and θ_y stand separately for the azimuths of the masker and the masked sources; m is the frame index; let $\mathbf{x}_1(m) = [x_1(mL_0), x_1(mL_0+1), \ldots, x_1(mL_0+L-1)]^T$ and $\mathbf{x}_2(m) = [x_2(mL_0), x_2(mL_0+1), \ldots, x_2(mL_0+L-1)]^T$, respectively, be the mth frame of $x_1(n)$ and $x_2(n)$, where L denotes the frame length, L_0 is the frame shift and T symbolizes the matrix transpose, then $X_1(m,k)$ and $X_2(m,k)$ are the short-time Fourier transforms (STFT) of $\mathbf{x}_1(m)$ and $\mathbf{x}_2(m)$, respectively, where k is the frequency index; similarly, $Y(m,k)$ is the STFT of the mth frame of the injected signal $\mathbf{y}(m)$. $\mathbf{y}(m)$ is a signal consisting of multiple sinusoidal components, where the frequencies of these sinusoids are determined by the fundamental frequency of the original stereo signals. Since it is added to only one channel ($X_2(m,k)$ in this paper), $\theta_y = \theta_2 = 30^\circ$. In the case of the masker source, $-30^\circ \leqslant \theta_x \leqslant 30^\circ$ due to the widely accepted rule that the acoustic content can only be produced in directions between the two loudspeakers in a stereo system [23].

Let $X_c(m,k)$ and $Y_c(m,k)$ be the head-center signals presented separately from the masker source and the masked source. Hence

$$\begin{cases} X_c(m,k) = X_1(m,k)\cdot H_c(k) + X_2(m,k)\cdot H_c(k), & \text{(a)} \\ Y_c(m,k) = Y(m,k)\cdot H_c(k). & \text{(b)} \end{cases} \tag{2}$$

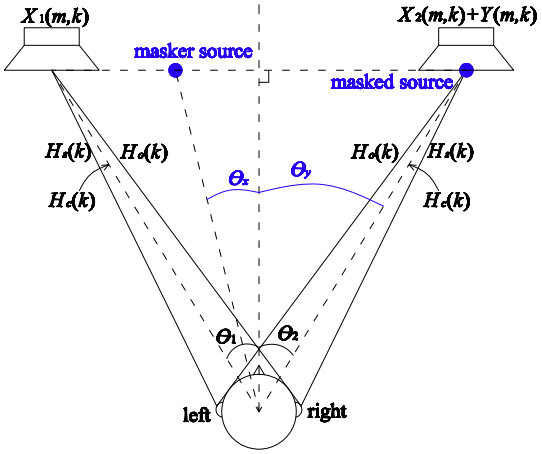

Fig. 3. An SAEC system where the listener is at the sweet spot of the stereo system, i.e. $\theta_1 = -\theta_2 = -30^\circ$.

Suppose $P_{xc}(m,k) = |X_c(m,k)|^2$ and $P_{yc}(m,k) = |Y_c(m,k)|^2$. In order to make $Y(m,k)$ be sufficiently masked, we have

$$10\lg\frac{P_{yc}(m,k)}{P_{xc}(m,k)} \leqslant \mathrm{SMR}_{\mathrm{bi}}(k). \tag{3}$$

Thus the injection level of $Y(m,k)$ is limited by (2b) and (3). It also shows that $\mathrm{SMR}_{\mathrm{bi}}(k)$ is the key parameter.

As seen from (1), $\mathrm{SMR}_{\mathrm{bi}}(k)$ can be calculated as the difference between $\mathrm{SMR}_{\mathrm{mo}}(k)$ and $\mathrm{BMLD}(k)$. On one hand, we set $\mathrm{SMR}_{\mathrm{mo}}(k)$ to be $\mathrm{SMR}_{\mathrm{m,max}}$ (see Fig. 4(a)), provided that the injected frequencies of $Y(m,k)$ could be close enough to the speech harmonics [24,25]. According to the auditory masking, $\mathrm{SMR}_{m,\max} \in [-11,-15]$ dB [26], thus $\mathrm{SMR}_{\mathrm{mo}}(k) = -15$ dB is recommended.

On the other hand, it is not so easy to decide a proper $\mathrm{BMLD}(k)$ for the SAEC system. The existing knowledge on $\mathrm{BMLD}(k)$ is not adequate to cover all the possible situations of the SAEC system, since the known information is limited. In the special case that $\theta_x = 0^\circ$, some $\mathrm{BMLD}(k)$ values are drawn in Fig. 4(b) according to the experimental results in [22]. As is shown, $\mathrm{BMLD}(k)$ varies with not only the source azimuths but also frequencies. Since $\mathrm{BMLD}(k)$ takes a maximum at around 200–300 Hz [22]. Therefore, the highest curve (315 Hz curve) in Fig. 4(b) is applied to the entire spectrum since it could feature the relationship between the maximum $\mathrm{BMLD}_{\max}$ and θ_y for $\theta_x = 0^\circ$.

To cover all the SAEC situations, further efforts need to be made to work out the dependence of $\mathrm{BMLD}_{\max}$ on both θ_x and θ_y. Since $\theta_x \neq \theta_y$, the azimuth difference results in that the signal-to-masker ratio at either the left or the right ear is greater than that at the head center [27]. We define this enhancement at a certain ear as the signal-to-masker-ratio improvement $\triangle_{\mathrm{SMR}}(k)$. $\triangle_{\mathrm{SMR}}(k)$ is determined by both θ_x and θ_y and is believed to account for most of the $\mathrm{BMLD}(k)$ [21,27]. A logarithmic model is assumed between $\mathrm{BMLD}_{\max}$ and the average value $\overline{\triangle_{\mathrm{SMR}}}$ of $\triangle_{\mathrm{SMR}}(k)$ over the low-frequency band ([0,1.5] kHz), which can characterize the relationship between $\mathrm{BMLD}_{\max}$ and the source azimuths θ_x and θ_y indirectly. After a nonlinear regression analysis based on the least square principle utilizing the 315 Hz data in Fig. 4(b), a regression Bi-MM is obtained with a mean square error of 0.4, which is expressed as:

$$\mathrm{BMLD}_{\max} = 10.9 \cdot \lg(1.1\overline{\triangle_{\mathrm{SMR}}} + 1.0)(\mathrm{dB}), \tag{4}$$

where $\overline{\triangle_{\mathrm{SMR}}}$ is derived from the source azimuths using the MIT head-related transfer functions (HRTF) [28]. As depicted in Fig. 5(a), the regression Bi-MM (4) fits the original data very well.

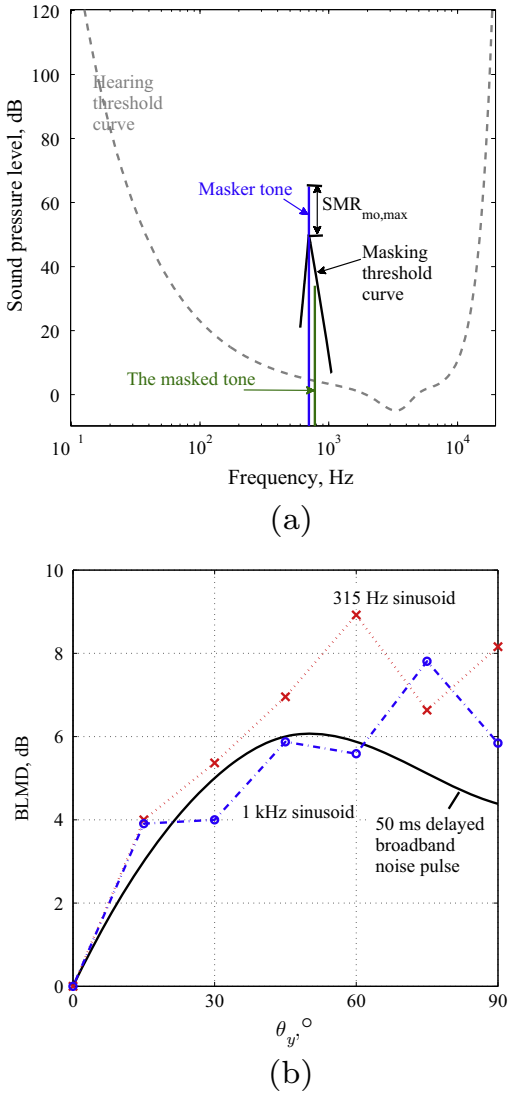

Fig. 4. Auditory masking curves. (a) Mo-MM masking curve of tone by tone. The masking threshold curve applies to all masker frequencies with a horizontal shift. (b) The BMLD as a function of θ_y for $\theta_x = 0^\circ$. Note that the legends explain the signal types used as the masked signal.

Backward to the SAEC system in Fig. 3, $\theta_y = 30^\circ$ and $-30^\circ \leqslant \theta_x \leqslant 30^\circ$. $\overline{\triangle_{\mathrm{SMR}}}|_{\theta_y=30^\circ}$ is therefore estimated as a function of θ_x. As seen in Fig. 5(b), the extreme case occurs when $\theta_x = -\theta_y = -30^\circ$, where $\overline{\triangle_{\mathrm{SMR}}}$ is about 4.7 dB. Substitute this value into (4) and the $\mathrm{BMLD}_{\max}$ is attained as 8.6 dB. Thus we set $\mathrm{SMR}_{\mathrm{bi}}(k)$ to its minimum $\mathrm{SMR}_{\mathrm{bi}}(k) = \mathrm{SMR}_{mo}(k) - \mathrm{BMLD}_{\max} = -23.6$ dB over the whole spectrum.

Although the proposed Bi-MM is derived from the SAEC system in Fig. 3, note that it can also be applied to other SAEC situations with a simple adjustment on the $\mathrm{BMLD}_{\max}$.

2.3. Determination of the initial-phase difference

Let α and β denote the signal-to-masker differences of the initial phase and the interaural phase difference (IPD), respectively. Previous studies have already shown that α and β have an obvious impact on the masking threshold $\mathrm{SMR}_{\mathrm{bi}}$ [29–32]. Some phenomena can be found in previous experiments for $\beta = 0$ and $\beta = \pi$ [30–36]: when α rises from 0° to 90°, $\mathrm{SMR}_{\mathrm{bi}}|_{\beta=0}$ increases monotonically with a high rate; while $\mathrm{SMR}_{\mathrm{bi}}|_{\beta=\pi}$ presents sort of decreasing trend

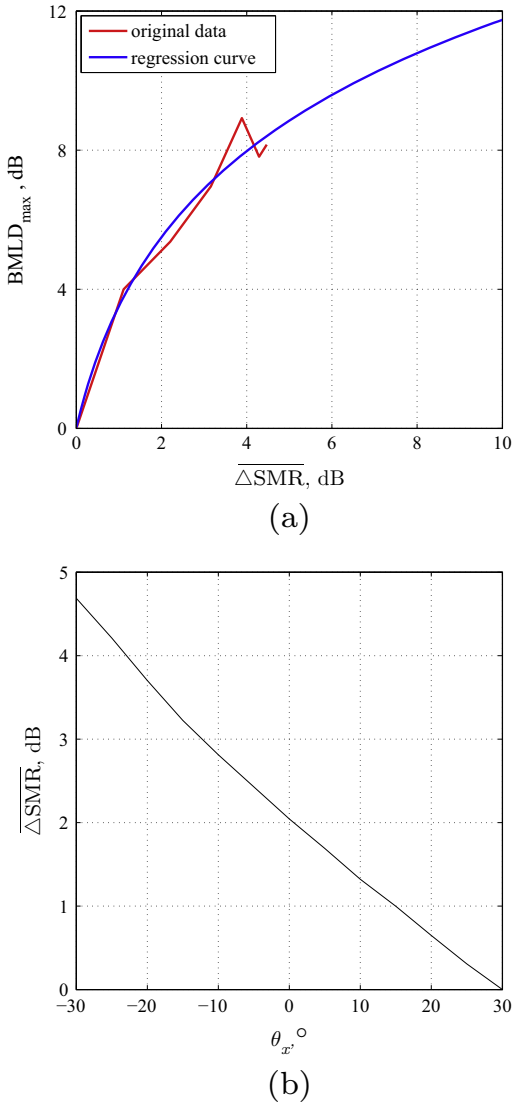

Fig. 5. The determination of a proper masking level difference. (a) The regression model of $BMLD_{max}$ as a logarithmic function of $\overline{\triangle_{SMR}}$. The original data is also shown to make contrast. (b) The $\overline{\triangle_{SMR}}$ as a function of θ_x while $\theta_y = 30°$.

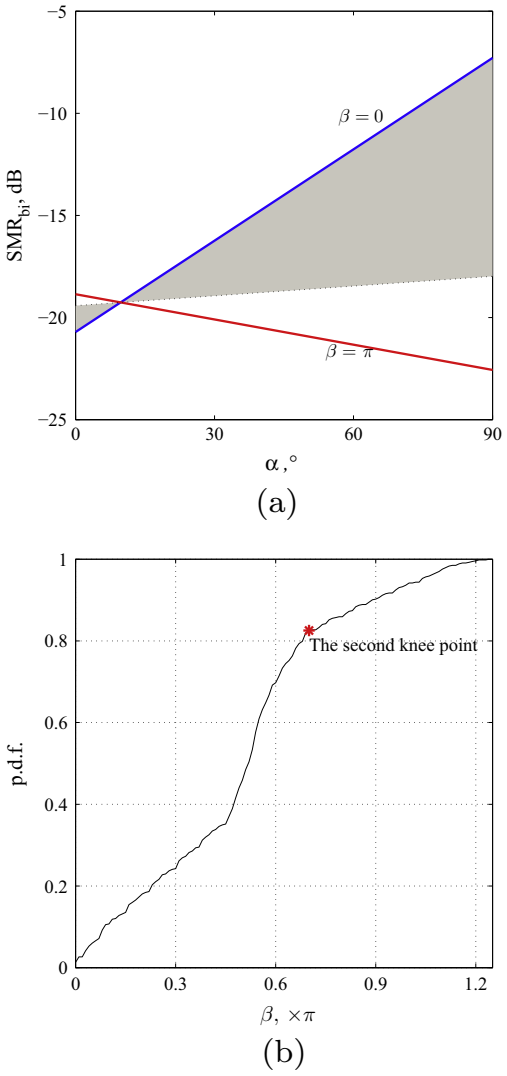

Fig. 6. The determination of the initial-phase difference α. (a) The dependence of SMR_{bi} on α, which varies with β. (b) The p.d.f. of β in the Bi-MM for SAEC.

(although not always decreasing monotonically), and the changing rate is low. Based on the experimental results of 7 researches [30–36], linear interpolation is carried out on their average values and the blue[1] and the red lines in Fig. 6(a) can be obtained.

In order to find a proper relationship between SMR_{bi} and α, the probability distribution function (p.d.f.) of β is statistically analyzed when taking into account all the possible situations of the SAEC system. The values of β are estimated as the mean IPD difference over [0, 1.5] kHz based on the MIT HRTF database [28]. Fig. 6 (b) gives a picture of the p.d.f. of β, which shows two knee points where the second-order derivative of the curve changes sign. The second knee point lies at about $\beta = 0.7\pi$. Considering that the p.d.f. rises slowly afterwards, and that the p.d.f at this point is greater than 80% which can cover most situations, we assume 0.7π as the upper limit of β. The $SMR_{bi}|_{\beta=0.7\pi}$ curve is then obtained through interpolation, which is drawn in the black broken line in Fig. 6(a). Since we consider $\beta \leqslant 0.7\pi$, the influence of α on $SMR_{bi}|_{\beta}$ would be characterized by lines within the gray zone which go through the crossover point. Thus it is supposed that SMR_{bi} increases with rising α from 0° to 90° in the proposed Bi-MM. Aiming for better masking, $\alpha = 90°$ is suggested.

3. Proposed algorithm

3.1. Bi-MM-based PDSI

The Bi-MM-based PDSI technique (see Fig. 7) is proposed for decorrelation according to the following three considerations. The first one is the pitch-driven motivation. Since speech often exhibits harmonic characteristics [37], the harmonic components would make the main contribution to the interchannel coherence. Our efforts are thus focused on decorrelation at these speech harmonics. The second one is to achieve better decorrelation performance. We propose to inject sinusoidal signals near the speech harmonics instead of adding the wideband noise [13,14], since the injection energy is more concentrated in the sinusoidal injection, which leads to higher decorrelation. The third one is to guarantee the speech quality. To make the injected sinusoids be sufficiently masked in SAEC, both the sinusoidal levels and phases in PDSI are determined by the Bi-MM. Because the Bi-MM takes into account the BMLD, less distortion is expected.

[1] For interpretation of color in Fig. 6, the reader is referred to the web version of this article.

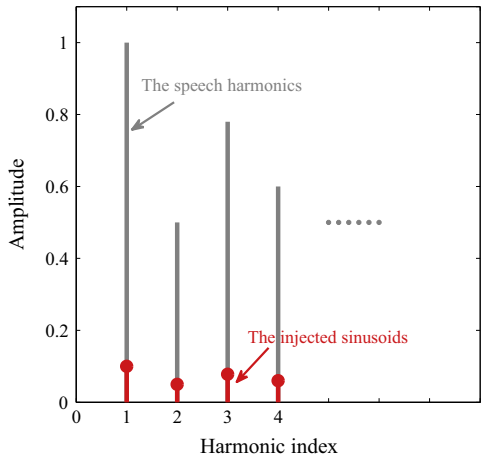

Fig. 7. Overview of the Bi-MM-based PDSI technique. Sinusoidal signals are injected close to the speech harmonics, where the injection levels are determined by the proposed Bi-MM.

The Bi-MM-based PDSI is summarized as

$$\mathbf{x}_2'(m) = \mathbf{x}_2(m) + \mathbf{y}(m) = \mathbf{x}_2(m) + \sum_l A_l(m) \cdot \cos\left(2\pi \cdot l \cdot \frac{k_0(m)}{\mathrm{NFT}} \cdot \mathbf{n}(m) + \varphi_l(m)\right), \tag{5}$$

where $\mathbf{x}_2'(m)$ is the processed version of $\mathbf{x}_2(m)$ and the PDSI is applied to only one channel ($\mathbf{x}_2(m)$ in this paper) so that the interchannel disparity helps reduce coherence; NFT denotes the STFT length; $k_0(m)$ represents the fundamental frequency of $\mathbf{x}_2(m)$; $l = 1, 2, 3, \ldots$ corresponds to the lth harmonic frequency; $\mathbf{n}(m) = [mL_0 mL_0 + 1 \ldots mL_0 + L - 1]$; $A_l(m)$ and $\varphi_l(m)$ signify the magnitudes and initial phases of the injected sinusoids, respectively. In order to reduce speech distortion, $A_l(m)$ and $\varphi_l(m)$ are determined by the Bi-MM as mentioned before. According to Section 2.3, the initial phases of $Y_c(m,k)$ are $\angle\{X_c(l \cdot k_0(m))\} + \alpha$, where $\angle\{\bullet\}$ denotes the phase of a complex number. Using (2b), we further get

$$\varphi_l(m) = \angle\{X_c(l \cdot k_0(m))\} + \alpha - \angle\{H_c(l \cdot k_0(m))\}. \tag{6}$$

As for the sinusoidal magnitudes $A_l(m)$, they can be obtained by combining (3) and (2b), which are expressed as

$$A_l(m) = \frac{\sqrt{P_{xc}(m, l \cdot k_0(m)) \cdot 10^{\mathrm{SMR}_{\mathrm{bi}}(l \cdot k_0(m))/10}}}{|H_c(l \cdot k_0(m))|}. \tag{7}$$

Now we briefly prove the effectiveness of the PDSI in decorrelation. Before decorrelating, we have

$$\gamma_0^2(k) = \frac{\left|E\{X_1^*(m,k) \cdot X_2(m,k)\}\right|^2}{E\{|X_1(m,k)|^2\} \cdot E\{|X_2(m,k)|^2\}}, \tag{8}$$

where γ_0 is the interchannel coherence coefficient (ICCC) [38] of the original stereo signals; $E\{\bullet\}$ represents the mathematical expectation; and * denotes complex conjugate. Taking into account the pitch-detection error, there is a difference between $k_0(m)$ and the true speech pitch. Since this difference is small [24] and varying, $k_0(m)$ is supposed to change within the main lobe of the true pitch. In such a case, the injected sinusoids and the original stereo signals are independent. Therefore, after the PDSI processing, the ICCC becomes

$$\gamma^2(k) = \gamma_0^2(k) \Big/ \left\{1 + \frac{E\left\{A_l^2(m) \cdot \delta(k - l \cdot k_0(m))\right\}}{E\{|X_2(m,k)|^2\}}\right\}, \tag{9}$$

where $\delta(\bullet)$ is the impulse function. As easily seen from (9), $\gamma^2(k) < \gamma_0^2(k)$ holds true for at least all harmonic frequencies. For other frequencies, note that coherence could also be somewhat reduced due to the spectral leakage.

Eq. (9) reveals that $\gamma^2(k)$ is lower when $E\{|X_2(m,k)|^2\} < E\{|X_1(m,k)|^2\}$ than its contrary. In other words, better decorrelation can be achieved when $Y(m,k)$ is injected into the channel with lower power. Another conclusion is that the ICCC decreases with rising $A_l(m)$, i.e. the injection level. However, we cannot arbitrarily increase the $A_l(m)$, due to that it may result in audible distortion. (7) gives a clear guideline on choosing the injection level.

3.2. Algorithm description

The proposed algorithm is depicted in Fig. 8. As can be seen from this figure, the Bi-MM-based PDSI is only applied to the low-frequency band ([0, 1.5] kHz) to maintain ITD after considering that the ITD plays a dominant role in the stereophonic perception of this frequency band. On the other hand, an modified SPM scheme is applied to the high-frequency band. Since the ILD is more important in this frequency band, the SPM can make a promising trade-off between preserving ILD and reducing interchannel coherence. Note that the SPM is not quite suitable for low-frequency stereophonic decorrelation, due to that only a slight phase modulation can be applied at these frequencies to preserve speech quality. Thus, by combining the Bi-MM-based PDSI with the SPM, we can expect better trade-off between decorrelation and speech quality throughout the whole spectrum for the SAEC system. Detailed description of the proposed system is given below.

3.2.1. Low-frequency band

Low-frequency band decorrelation is achieved through the Bi-MM-based PDSI technique. Based on the stereo signals and the proposed Bi-MM, we calculate the masking threshold for the SAEC system in the first stage. In the second stage, the PDSI technique is carried out in one channel (the right channel in this paper) of the stereo signals, where the amount of injection is determined by the masking threshold.

Focusing on the low-frequency band, the proposed method can be summarized into the following five steps:

(1) Extract the pitch $k_0(m)$ of $\mathbf{x}_2(m)$, where the weighted autocorrelation method [24] is utilized in this paper.
(2) Calculate $X_c(m,k)$ using (2a). And $P_{xc}(m,k) = |X_c(m,k)|^2$ is computed.
(3) Compute the sinusoidal magnitudes $A_l(m)$ by using (7), where $l \cdot k_0(m) \leqslant 1.5$ kHz is applied to determine the maximum value of l. According to Section 2.2, $\mathrm{SMR}_b(k) = -23.6$ dB is adopted for the whole spectrum.
(4) Obtain the initial phases $\varphi_l(m)$ using (6), where $\alpha = 90^\circ$ is employed.
(5) Construct sinusoidal signals based on $k_0(m), A_l(m)$ and $\varphi_l(m)$, and then inject these signals into $\mathbf{x}_2(m)$ according to (5).

3.2.2. High-frequency band

In the high-frequency band, an SPM scheme is carried out. Since we have computed $X_i(m,k)$, with $i = 1, 2$ representing the left and the right channels, respectively, we propose to perform the SPM in the frequency domain, which is slightly different from the traditional subband-domain SPM [10]. The frequency-domain SPM is expressed as

$$X_i''(m,k) = X_i(m,k) \cdot \exp\{j \cdot (-1)^i \cdot \Theta(m,k)\}, \tag{10}$$

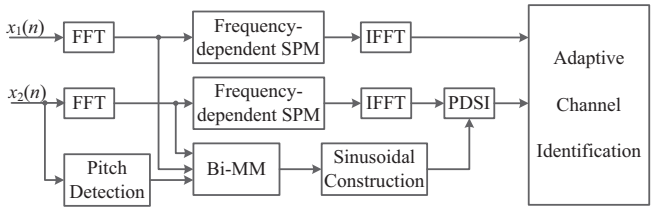

Fig. 8. Block diagram of the proposed decorrelation algorithm for SAEC.

where $j=\sqrt{-1}$ is the imaginary unit; $(-1)^i$ represents that the modulation phases of the two channels are opposite to each other, which helps decorrelate the stereo signals; $\Theta(m,k)$ stands for the modulation phase.

To further enhance the decorrelation, $\Theta(m,k)$ varies over time in the manner $\Theta(m,k)=\vartheta(k)\cdot\sin(2\pi mLk/\mathrm{NFT})$. As we know, higher values of $\vartheta(k)$ lead to better decorrelation at the cost of larger distortion [10]. To make a good trade-off between high decorrelation and good speech quality, $\vartheta(k)$ values in [10] are chosen. Note that $\vartheta(k)$ is set to be zero in the low-frequency band in this paper.

3.2.3. Implementation of the proposed algorithm

As shown in Fig. 8, for the full-band decorrelation, we need to perform the Bi-MM-based PDSI in the low-frequency band and the modified SPM scheme in the high-frequency band. The implementation of the proposed algorithm is summarized in Algorithm 1.

Algorithm 1. The implementation procedures of the proposed algorithm.

For all time frame m:
(1) Apply FFT to $\mathbf{x}_i(m)$, with $i=1,2$, to obtain $X_i(m,k)$;
(2) Perform frequency-dependent SPM using (10) to obtain $X_i''(m,k)$, with $i=1,2$, where $\vartheta(k)$ equals to 0 at frequencies below 1500 Hz and takes the values in [10] at higher frequencies;
(3) Apply IFFT to $X_i''(m,k)$, with $i=1,2$, to obtain $\mathbf{x}_i''(n)$;
(4) Detect the pitch $k_0(m)$ of $\mathbf{x}_2(m)$ via the weighted autocorrelation method presented in [24];
(5) Compute the head-center signal $X_c(m,k)$ of the masker source using (2a), and $P_{xc}(m,k)=|X_c(m,k)|^2$;
(6) Calculate sinusoidal magnitudes $A_l(m)$ using (7), where $\mathrm{SMR}_b(k)=-23.6$ dB for all frequencies and $l\in[1\quad\lfloor 1500/k_0(m)\rfloor]$;
(7) Estimate the initial phase $\varphi_l(m)$ using (6) with $\alpha=90^\circ$ and $l\in[1\quad\lfloor 1500/k_0(m)\rfloor]$;
(8) Construct sinusoids employing $k_0(m)$, $A_l(m)$ and $\varphi_l(m)$;
(9) Carry out the PDSI using (5), where $\mathbf{x}_2(m)$ is replaced by $\mathbf{x}_2''(n)$ calculated in the step (3).

4. Simulation

The SAEC system in Fig. 1 is simulated with a sample rate of 16 kHz. The rooms in the near end and the far end are of the same size $4\times3\times3$ m^3 with a reverberation time of 128 ms. The two loudspeakers are located at (1,2,1.2) m and (3,2,1.2) m. The two microphones are placed at (1.8,1,1.2) m and (2.2,1,1.2) m. Corresponding to Fig. 3, the listener is seated at $(2,2-\sqrt{3},1.2)$ m, which is the sweet spot of the two-loudspeaker stereo system. The room impulse responses are simulated using the image method [39] with a length of 1024 samples. The NLMS [40] adaptive algorithm is exploited with filter order $N=1024$ and step size $\mu=0.4$. In the simulation, a white Gaussian noise is added to get a signal-to-noise ratio of 30 dB. A piece of test signal is adopted lasting about 2 min. It is taken from the SQAM disk [41] and is composed of 6 speech tracks (track 49–54), containing 3 female tracks and 3 male tracks.

Since the proposed algorithm uses the Bi-MM to control the injection levels of the sinusoidal signals for stereophonic decorrelation, we will refer to it as Bi-MM-PDSI. In order to validate the Bi-MM-PDSI method, performance comparisons are made with other four decorrelation methods listed below:

(1) Mo-MM-PDSI method: Mo-MM-based PDSI is conducted for decorrelation in the low-frequency band. While the high-frequency band processing is the same as that in the Bi-MM-PDSI method.
(2) Mo-MM-WNI method: Wideband noise injection is applied to the low-frequency band according to a Mo-MM presented in [14], while the SPM is applied to the high-frequency band.
(3) SPM method: The SPM technique is performed in the whole spectrum, where the modulation magnitude is the same as that in [10].
(4) NLT method: The NLT method using the half-wave rectification [1].

Except for the NLT method which is carried out point by point, all the other methods are implemented frame by frame. The frame length is $L=512$ samples with the frame shift $L_0=256$. Each segment is multiplied by the Hanning window.

Performances are evaluated in terms of interchannel decorrelation, the improvement of the non-unique problem, speech distortion and the stereophonic perception. Correspondingly, five objective measurements are obtained including the ICCC, the filter misalignment [3], the perceptual evaluation of speech quality (PESQ) [42], the low-frequency ITD and the high-frequency ILD. The results are presented in the following four parts.

4.1. ICCC

The ICCC is computed to assess the decorrelation performance. In calculating the ICCC, the mathematical expectation is approximated as the average over all frames.

For the low-frequency band decorrelation, the Bi-MM-PDSI method is compared with the Mo-MM-PDSI and the Mo-MM-WNI methods, since they all take advantage of psychoacoustic effects in this frequency band. As shown in Fig. 9(a), the Bi-MM-PDSI method has a moderate decorrelation over the low-frequency band. The Mo-MM-PDSI method is of the lowest ICCCs among the three competitive methods, which is reasonable since the highest level is injected according to the Mo-MM-based PDSI. The Mo-MM-WNI method can decorrelate the stereo signals only at very low frequencies. Because both the Mo-MM-PDSI and the Mo-MM-WNI methods use the Mo-MM to control the injection level, their large difference in ICCC reveals that the PDSI plays an important role in reducing the interchannel coherence.

The decorrelation over the whole spectrum is pictured in Fig. 9 (b). The Mo-MM-PDSI and the Mo-MM-WNI methods are not evaluated any more after considering that their differences from the Bi-MM-PDSI method only lie in the low-frequency band. The SPM and the NLT methods are taken into account. The ICCCs of the SPM and the Bi-MM-PDSI methods have some features in common: the ICCCs are high in the low-frequency band but get much lower in the high-frequency band. Compared with the Bi-MM-PDSI method, the ICCC of the SPM method is close to 1 at frequencies below about 1 kHz, due to that only a slight phase modulation could be applied. This is the main reason why we propose to use the Bi-MM-based PDSI in the low-frequency band. In contrast, the decorrelation of the NLT method is quite limited in the whole spectrum.

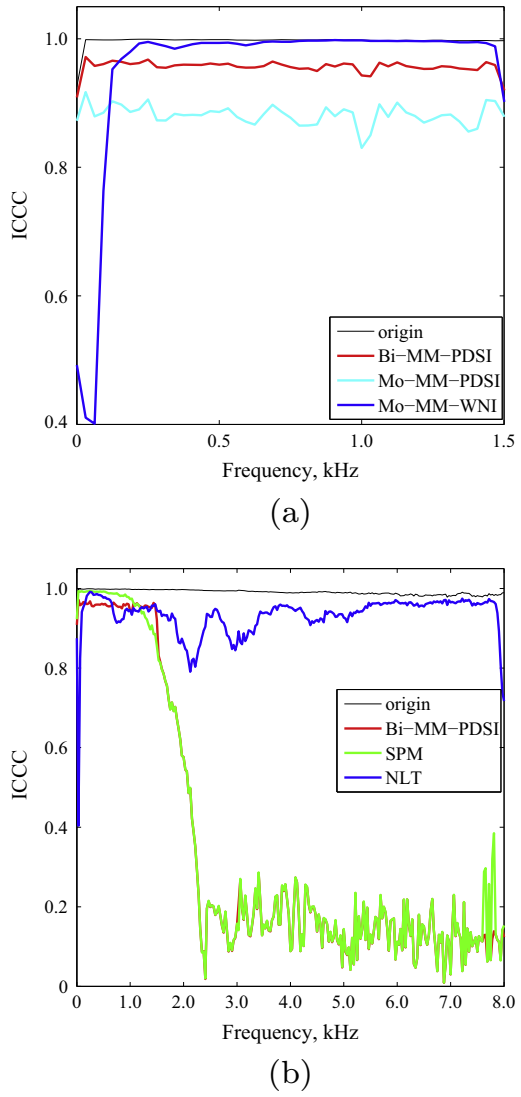

Fig. 9. ICCC in (a) the low-frequency band [0,1.5] kHz and (b) the whole spectrum.

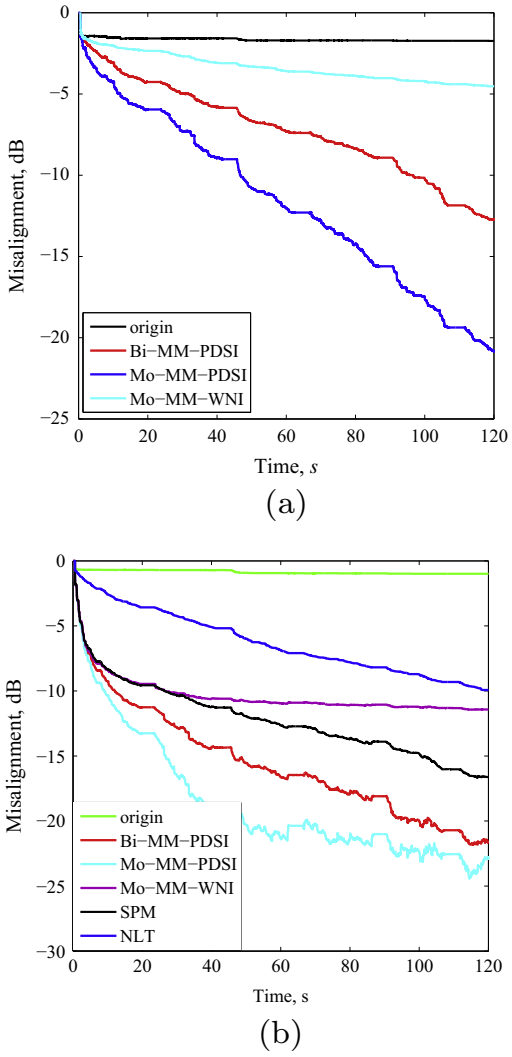

Fig. 10. Filter misalignment in (a) the low-frequency band [0,1.5] kHz and (b) the whole spectrum.

4.2. Misalignment

To show the improvement of the non-unique problem, the filter misalignment is calculated as

$$\zeta(n) = 20 \cdot \lg \frac{\sum_{i=1}^{2} \left\| \mathbf{h}_i - \widehat{\mathbf{h}}_i(n) \right\|_2}{\sum_{i=1}^{2} \|\mathbf{h}_i\|_2} \tag{11}$$

where $\mathbf{h}_i, i = 1, 2$ signify the physical impulse responses, $\widehat{\mathbf{h}}_i(n)$ stands for the adaptive response at time index n, and $\|\cdot\|_2$ denotes 2-norm.

Fig. 10(a) depicts the misalignment values in the low-frequency band, considering the Bi-MM-PDSI together with the Mo-MM-PDSI and the Mo-MM-WNI methods. The low-frequency band misalignment of the Mo-MM-PDSI method is several decibels better than that of the Bi-MM-PDSI method, which results from the lower ICCC values in the low-frequency band (see Fig. 9(a)). For the same reason, the Mo-MM-WNI method performs poor misalignment in the low-frequency band. The misalignment of the Mo-MM-WNI method is almost the worst among these decorrelation algorithms, which can further confirm that using the PDSI can better reduce the filter misalignment than using the WNI.

The misalignment over the entire spectrum is shown in Fig. 10 (b). Through contrast between with and without decorrelation, the effectiveness of decorrelation in improving the non-unique problem in SAEC is confirmed. Among these methods, the Mo-MM-PDSI method provides the best filter misalignment in the whole spectrum, while the Bi-MM-PDSI method is of the second best misalignment. The difference between the misalignments of these two methods becomes smaller in contrast with that in Fig. 10(a). This is due to that, in the high-frequency band, both of them use the frequency-domain SPM method.

4.3. PESQ

In order to measure speech quality, we make use of the PESQ. PESQ ranges from −0.5 to 4.5 and a higher value implies less distortion. Table 1 lists the average PESQs of the left and the right channels of different decorrelation methods.

From the PESQ scores in Table 1, the SPM method has the best performance, not only in the low-frequency band but also in the entire spectrum. The Bi-MM-PDSI method is comparable with the SPM method, however, its misalignment (see Fig. 10(b)) performs much better than that of the SPM method in the entire spectrum. Compared with the Mo-MM-PDSI and the Mo-MM-WNI methods which both exploit the Mo-MM, the Bi-MM-PDSI method achieves higher PESQ scores. This indicates that the Bi-MM helps preserve speech quality.

Table 1
PESQ scores of different decorrelation methods.

Method	Bi-MM-PDSI	Mo-MM-PDSI	Mo-MM-WNI	SPM	NLT
PESQ in the low-frequency band	4.2	4.0	3.9	4.3	3.7
PESQ in the whole spectrum	4.3	4.0	3.9	4.4	3.9

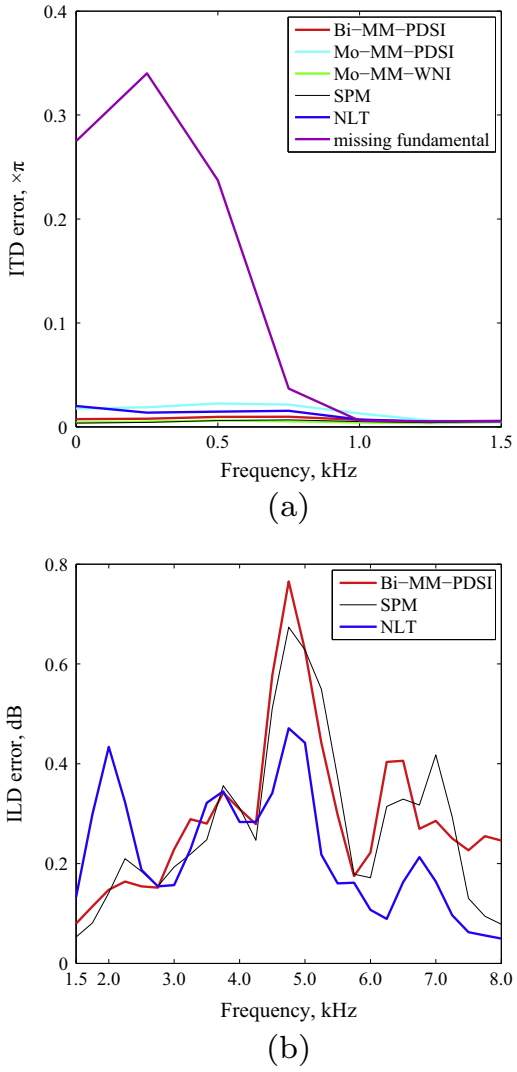

Fig. 11. Evaluation of stereophonic perception. (a) ITD error in the low-frequency band [0,1.5] kHz. (b) ILD error in the high-frequency band [1.5,8] kHz.

4.4. ITD/ILD

With regard to the performance of stereophonic perception, low-frequency band ITD and high-frequency band ILD are evaluated. In order to obtain the binaural signals, speech segments are filtered through the room impulse responses and the MIT HRTFs. The ITD/ILD errors [43] between the decorrelated and the original speeches are worked out and depicted in Fig. 11.

The low-frequency band ITD errors are plotted in Fig. 11(a). In order to better show the performance of the Bi-MM-PDSI method, the missing-fundamental method [6] is also taken into consideration. As can be seen from Fig. 11(a), the Bi-MM-PDSI as well as the Mo-MM-WNI and the SPM methods are the best at preserving the stereophonic perception in the low-frequency band, since their ITD errors in this frequency band are the lowest. Note that this should owe to the Bi-MM rather than the PDSI, since there is no ITD-error improvement when comparing the Mo-MM-PDSI method with the Mo-MM-WNI method. Another thing to be explained is that the ITD error of the missing-fundamental method is much larger than that of the other methods, which is predictable because the notch filter used in [6] removes lots of extra components in addition to the speech fundamental.

The high-frequency band ILD error is depicted in Fig. 11(b). As can be seen from this figure, all of these methods have a comparable stereophonic performance in the high-frequency band. Note that only objective test results are presented in this paper for the following two considerations. First, these objective measurements have already been proved that they are highly correlated with subjective test results. Second, our informal listening test results also confirm that the proposed algorithm outperforms the competitive algorithms in some aspects, which are consistent with our objective test results.

5. Discussion and conclusion

A novel interchannel decorrelation method has been proposed for stereo acoustic echo cancellation by using a simplified binaural masking model. The binaural masking model based pitch-driven sinusoidal injection is applied to the low-frequency band ($\leqslant$1.5 kHz), and a sinusoidal phase modulation scheme is applied to higher frequencies. Evaluations have been performed in terms of decorrelation, filter misalignment, speech distortion and stereophonic perception. Results have verified the effectiveness of the proposed method. In a word, there are mainly three advantages of the proposed method. First, the binaural masking model helps retain good speech quality as well as stereophonic perception, since it takes into account the contribution of spatial cues in binaural listening, which is the common case in stereo acoustic echo cancellation systems. Second, higher decorrelation are obtained in the low-frequency band comparing with the masked-noise injection method, thanks to the concentration of injection energy by using the pitch-driven sinusoidal injection technique. Third, high decorrelation together with low filter misalignment are achieved over the whole spectrum, due to the combination of the sinusoidal phase modulation in the high-frequency band.

Because the proposed algorithm needs to estimate the fundamental frequency of the far-end speech signal, it would somewhat increase the computational load. However, we can use a very efficient approach to estimate the fundamental frequency in practice to solve this problem. Another problem is that the proposed algorithm can only be applied to speech signals. In other words, if the far-end signal is not a speech signal, the proposed algorithm cannot be used directly and some extensions are needed, which can be considered for future work. For speech applications, the proposed algorithm is robust to the double-talk situations due to that the adaptive filter coefficients will stop updating automatically with the help of some double-talk detection schemes.

Acknowledgement

This work was supported by National Science Fund of China Under Grant No. 61571435 and No. 61302168.

References

[1] Benesty J, Morgan DR, Sondhi MM. A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation. IEEE Trans Speech, Audio Process 1998;6:156–65.
[2] Morgan DR, Hall JL, Benesty J. Investigation of several types of nonlinearities for use in stereo acoustic echo cancellation. IEEE Trans Speech Audio Process 2001;9:686–96.
[3] Gänsler T, Benesty J. New insights into the stereophonic acoustic echo cancellation problem and an adaptive nonlinearity solution. IEEE Trans Speech Audio Process 2002;10:257–67.
[4] Wu M, Lin Z, Qiu X. A novel nonlinearity for stereo acoustic echo cancellation. IEICE Trans Fundam 2005;E88-A:1757–9.
[5] Wu S, Qiu X, Wu M. Stereo acoustic echo cancellation employing frequency-domain preprocessing and adaptive filter. IEEE Trans Audio, Speech, Lang Process 2011;19:614–23.
[6] Cecchi S, Romoli L, Peretti P, Piazza F. A combined psychoacoustic approach for stereo acoustic echo cancellation. IEEE Trans Audio, Speech, Lang Process 2011;19:1530–9.
[7] Yang H, Zheng C, Li X. A spectral dominance decorrelation method for stereo acoustic echo cancellation. In: Int Congr on Sound Vibration (ICSV), Beijing, China; June 2013.
[8] Ali M. Stereophonic acoustic echo cancellation system using time varying all-pass filtering for signal decorrelation. In: Proc IEEE Int Conf on Acoust, Speech, Signal Process (ICASSP), Seattle, USA; May 1998.
[9] Nguyen DQ, Gan WS, Khong AWH. Time-reversal approach to the stereophonic acoustic echo cancellation problem. IEEE Trans Audio, Speech, Lang Process 2011;19:385–95.
[10] Herre J, Buchner H, Kellermann W. Acoustic echo cancellation for surround sound using perceptually motivated convergence enhancement. In: Proc IEEE Int Conf on Acoust, Speech, Signal Process (ICASSP), Honolulu, USA; April 2007.
[11] Benesty J, Morgan DR, Hall JL, Sondhi MM. Stereophonic acoustic echo cancellation using nonlinear transformations and comb filtering. In: Proc IEEE Int Conf on Acoust, Speech, Signal Process (ICASSP), Seattle, USA; May 1998.
[12] Romoli L, Cecchi S, Peretti P, Piazza F. A mixed decorrelation approach for stereo acoustic echo cancellation based on the estimation of the fundamental frequency. IEEE Trans Audio, Speech, Lang Process 2012;20:690–8.
[13] Valin JM. Perceptually-motivated nonlinear channel decorrelation for stereo acoustic echo cancellation. In: IEEE hands-free speech communication and microphone arrays (HSCMA), Trento, Italy; May 2008.
[14] Gilloire A, Turbin V. Using auditory properties to improve the behaviour of stereophonic acoustic echo cancellers. In: Proc IEEE Int Conf on Acoust, Speech, Signal Process (ICASSP), Seattle, USA; May 1998.
[15] Li J, Sakamoto S, Hongo S, Akagi M, Suzuki Y. Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication. Speech Commun 2011;53:677–89.
[16] Kim YI, Kil RM, Lee SY. A zero-crossing-based binaural masking model for speech recognition in noisy and reverberant conditions. In: 9th China-India–Japan–Korea Joint Workshop on Neurobiology and Neuroinformatics, Korea; July 2007.
[17] Zheng C, Schwarz A, Kellermann W, Li X. Binaural coherent-to-diffuse-ratio estimation for dereverberation using an ITD model. In: 2015 Eur Signal Process Conf (EUSIPCO2015), Nice, France; August 2015.
[18] Brothánek M, Jandák V, Jiříček O, Švec P. Monaural and binaural parameters of Rudolfinum concert halls in Prague. Appl Acoust 2012;73:1201–8.
[19] Aoki S, Toba M, Tsujita N. Sound localization of stereo reproduction with parametric loudspeakers. Appl Acoust 2012;73:1289–95.
[20] Durlach NI. Equalization and cancellation theory of binaural masking-level difference. J Acoust Soc Am 1963;35:1206–18.
[21] Kopčo N, Cunningham BGS. Spatial unmasking of nearby pure-tone targets in a simulated anechoic environment. J Acoust Soc Am 2003;114:2856–70.
[22] Blauert J. Spatial hearing: the psychophysics of human sound localization. Cambridge: MIT Press; 1997. p. 257–71.
[23] Ahrens J. Analytic methods of sound field synthesis. Berlin: Springer; 2012. p. 4–7.
[24] Shimamura T, Kobayashi H. Weighted autocorrelation for pitch extraction of noisy speech. IEEE Trans Speech Audio Process 2001;9:727–30.
[25] Zheng C, Li X. Detection of multiple sinusoids in unknown colored noise using truncated cepstrum thresholding and local signal-to-noise-ratio. Appl Acoust 2012;73:809–16.
[26] Bosi M, Goldberg RE. Introduction to digital audio coding and standards. Berlin: Springer; 2003. p. 151–91.
[27] Brungart DS, Simpson BD. The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal. J Acoust Soc Am 2002;112:664–76.
[28] Gardner B, Martin K. HRTF measurements of a KEMAR dummy-head microphone. <http://sound.media.mit.edu/resources/KEMAR/>; 2011.
[29] Gilkey RH, Robinson DE, Hanna TE. Effects of masker waveform and signal-to-masker phase relation on diotic and dichotic masking by reproducible noise. J Acoust Soc Am 1985;78:1207–19.
[30] Hafter ER, Carrier SC. Masking-level difference obtained with a pulsed tonal masker. J Acoust Soc Am 1970;47:1041–7.
[31] Robinson DE, Langford TL, Yost WA. Masking of tones by tones and of noise by noise. Percept Psychophys 1974;15:159–67.
[32] Grantham DW, Robinson DE. Role of dynamic cues in monaural and binaural signal detection. J Acoust Soc Am 1977;61:542–51.
[33] Jefftess LA, Blodgett HC, Sandel TT, Wood CL. Masking of tonal signals. J Acoust Soc Am 1956;28:416–26.
[34] Wightman FL. Binaural masking with sine-wave maskers. J Acoust Soc Am 1969;45:72–8.
[35] Wightman FL. Detection of binaural tones as a function of masker bandwidth. J Acoust Soc Am 1971;50:623–36.
[36] Yost WA. Tone-on-tone masking for three binaural listening conditions. J Acoust Soc Am 1972;52:1234–7.
[37] Hu X, Wang S, Zheng C, Li X. A cepstrum-based preprocessing and postprocessing for speech enhancement in adverse environments. Appl Acoust 2013;74:1458–62.
[38] Cecchi S, Romoli L, Carini A, Piazza F. A multichannel and multiple position adaptive room response equalizer in warped domain: realtime implementation and performance evaluation. Appl Acoust 2014;82:28–37.
[39] Allen JB, Berkley DA. Image method for efficiently simulating small room acoustics. J Acoust Soc Am 1979;65:943–50.
[40] Boroujeny BF. Adaptive filters: theory and applications. West Sussex: John Wiley & Sons; 1998. p. 172–5.
[41] Waters G. Sound quality assessment material—recordings for subjective tests: user's handbook for the EBUSQAM compact disk. Eur Broadcasting Union, Tech Rep; 1988.
[42] ITU-T-REC-P.862. Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. Geneva: ITU-T; 2001.
[43] Cornelis B, Doclo S, Bogaert TVD, Moonen M, Wouters J. Theoretical analysis of binaural multimicrophone noise reduction techniques. IEEE Trans Audio, Speech, Lang Process 2010;18:342–55.