A Prior Knowledge-based Noise Reduction Method with Dual Microphones

Hao Chen, Chang-chun Bao and Bing-yin Xia

Speech and Audio Signal Processing Lab, School of Electronic Information and Control Engineering, Beijing University of Technology, Beijing 100124, China
E-mail: chenhaozzj@emails.bjut.edu.cn, baochch@bjut.edu.cn, xby-abc@emails.bjut.edu.cn

Abstract—In this paper, a noise reduction method with dual microphones, based on the prior knowledge, is proposed to reduce the residual noise especially in the period of target speech absence (TSA). First, two cases, i.e. target speech presence and target speech absence were modeled by Gaussian mixture model (GMM), respectively. Then, we calculated the frame-based target speech present probability (TSPP) using Bayesian classification. Finally, a mask filter was presented by modifying the gain function of the improved phase-error based filter (IPBF) method using TSPP. Simulation results show that the proposed method outperforms the reference methods and could reduce noise effectively, particularly in the period of TSA.

I. INTRODUCTION

Noise reduction plays an important role in many applications related with speech communication. To suppress the noises, various speech enhancement techniques have been investigated in the past four decades. Now, speech enhancement with one microphone is often used for mobile communication. The most important limitation of single microphone method, such as Optimally-Modified Log-Spectral Amplitude (OM-LSA)[1], is a lack of ability to distinguish the interfering speech from the target speech at the same time. On the other hand, many speech enhancement methods with microphone array have been developed to remove the interfering speech from target speech by spatial filtering. One of the most well-known methods is the generalized sidelobe canceller (GSC) algorithm [2], which has achieved an outstanding performance in directional noise reduction, while there is lots of musical noise. In our early work, an improved phase-error based filter (IPBF) [3] method has been proposed, which could reduce the musical noise caused by the PBF algorithm and keep the performance of the directional noise reduction at the same time. But, there are still some residual noises in our previous method.

In mobile situation, the direction of the target speech is often fixed, for instance, the people who are in a noisy room or in a vehicle often speak to cellphone’s microphone directly. This prior knowledge related to fixed direction could be used in our noise reduction procedure and the conventional methods do not make use of this knowledge to improve the enhancement performance.

In this paper, a novel noise reduction method with dual microphones, based on the fixed target speech direction, is proposed to modify the gain function of the IPBF. First, the cases of target speech present and target speech absent are modeled by GMM, respectively. Then, the frame-based TSPP is estimated based on the Bayesian classifier. Finally, the gain function of the IPBF is modified using the TSPP.

The rest of this paper is organized as follows. The proposed method is presented in detail in section 2. The simulation results are shown in section 3. The conclusions are given in section 4.

II. PROPOSED DUAL MICROPHONE METHOD

The proposed method, consisting of training process and noise reduction process, is shown in Fig. 1. In the training process, the sub-band phase error (SBPE) is adopted as the training features extracted from a large speech corpus. Then, two GMMs are trained to represent two classes of features for target speech presence ($\lambda_1$) and target speech absence ($\lambda_0$), respectively. For noise reduction, the frame-based TSPP is calculated according to the Bayesian classifier. Finally, a mask filter is derived from the modification of the gain function of the IPBF using TSPP.

Feature extraction: Sub-band phase error

Training Procedure

Noise Reduction

Fig. 1 Block diagram of the proposed method

A. Feature extraction

The sub-band phase errors (21 dimensions), which could distinguish the target speech from noise source effectively,
are used as a classification feature in this paper. The phase error (PE) between two microphones is defined as [3, 4]:

$$\theta(l, k) = \angle Y_1(l, k) - \angle Y_2(l, k)$$

where $Y_1(l, k)$ and $Y_2(l, k)$ are the FFT spectrum of the signals collected by two microphones, respectively, $l$ and $k$ are the frame index and frequency index, respectively. The time delay of arrival (TDOA) is known in this paper.

In our method, a 512-point FFT is computed, and a 257-dimension PE is available for training, but the high dimension features would increase the complexity of training and enhancement procedures. Besides, the single dimensional feature, which is the summation over the 257 dimensions of PE feature, is also not appropriate in our method due to the low frequency resolution and poor training robustness. Then a trade-off solution is presented: the 257-dimension PE could be divided into 21 sub-bands like [5] and next, the SBPE is obtained by calculating the mean in each sub-band, which is expressed as:

$$\theta_{sub}(l, b) = \frac{1}{w_b(l) - w_h(l)} \sum_{k=w_h(l)}^{w_b(l)} |\theta(l, k)|$$

where $w_b(l)$ and $w_h(l)$ are the lowest and highest frequency index in each sub-band, $b$ is the sub-band index.

In order to verify the effectiveness of SBPE for classification, a simple experiment is adopted. For each of the two classes, there are 200 SBPE feature vectors (21-dimension) which are extracted from noisy speech material (200 frames). The direction of target speech is fixed and the direction of noise source is selected arbitrarily. The dimensionality of SBPE feature vectors is reduced to 3 from 21 by PCA [6] which is a kind of effective method for dimensionality reduction. The scatter diagrams of the 3-dimensional features are given in Figure 2. Although there are some overlaps, these 3-dimensional features could almost distinguish the presence and absence of target speech.

In principle, the training data should include all the possible directions of noise source except the direction of target speech. Actually, it is not necessary to use small steps (e.g. 1° showed in Fig. 3(a)) for determining the direction of noise sources. The number of Gaussian mixtures $K$ is 32.

$$\theta_{sub}(l, b) = \frac{1}{w_b(l) - w_h(l)} \sum_{k=w_h(l)}^{w_b(l)} |\theta(l, k)|$$

where $w_b(l)$ and $w_h(l)$ are the lowest and highest frequency index in each sub-band, $b$ is the sub-band index.

In order to verify the effectiveness of SBPE for classification, a simple experiment is adopted. For each of the two classes, there are 200 SBPE feature vectors (21-dimension) which are extracted from noisy speech material (200 frames). The direction of target speech is fixed and the direction of noise source is selected arbitrarily. The dimensionality of SBPE feature vectors is reduced to 3 from 21 by PCA [6] which is a kind of effective method for dimensionality reduction. The scatter diagrams of the 3-dimensional features are given in Figure 2. Although there are some overlaps, these 3-dimensional features could almost distinguish the presence and absence of target speech.

In order to verify the effectiveness of SBPE for classification, a simple experiment is adopted. For each of the two classes, there are 200 SBPE feature vectors (21-dimension) which are extracted from noisy speech material (200 frames). The direction of target speech is fixed and the direction of noise source is selected arbitrarily. The dimensionality of SBPE feature vectors is reduced to 3 from 21 by PCA [6] which is a kind of effective method for dimensionality reduction. The scatter diagrams of the 3-dimensional features are given in Figure 2. Although there are some overlaps, these 3-dimensional features could almost distinguish the presence and absence of target speech.
posteriori probabilities as the inputs. The frame-based TSPP estimation with sigmoid function is defined as follows:

$$TSPPL(l) = \frac{1}{1 + \exp\left(-\left[p(\lambda_l | x) - p(\lambda_l | x)\right]\right)}$$  \hspace{1cm} (6)

Considering the influence of adjacent frames, the smoothed TSPP is obtained by first-order recursive averaging.

The IPBF method, taking the advantages of PBF and OM-LSA, could eliminate the musical noise (for a stationary background noise condition) caused by PBF effectively. However, when the target speech is absent, the residual noise is obvious especially in babble noise condition or when the interference speech is present. In order to reduce this residual noise, the gain function of the IPBF is modified by TSPP, which could be defined as:

$$G_{IPBF}(l,k) = TSPP(l) \times G_{IPBF}(l,k)$$ \hspace{1cm} (7)

where $$G_{IPBF}(l,k)$$ is the gain function of IPBF [4], which is defined as:

$$G_{IPBF}(l,k) = \max(\min(G_{omLSA}(l,k)),\delta)$$ \hspace{1cm} (8)

where $$G_{omLSA}(l,k)$$ and $$G_{IPBF}(l,k)$$ are the gain functions of PBF[3] and OM-LSA[1], respectively. $$\delta$$ is the minimum gain allowed.

III. EXPERIMENTAL RESULTS AND DISCUSSION

The experiments were performed in a 7.1m×5.1m×3m conference room with a reverberation time (RT) of approximately 150ms as shown in Fig.4. Point A is the target speech source which is selected from NTT database (6 males and 6 females). The noise contains two types: babble noise taken from Noisex92 [9] and the mixed noise. The mixed noise, recorded in a real conference room, contains white noise and an interfering speech. Then, the received signals of two microphones for target speech and noise source are generated using image method [8], respectively. To acquire the enhanced performance of the inclusive and exclusive training data, the direction of noise source is divided into two cases (case A and case B). For case A, the directions of the noise sources, denoted as p1 to p4, are the same with the training data. For case B, the directions of the noise source are not included in the training data, and are denoted as q1 to q4, which is showed in fig.4. The noisy signals are mixed using the target speech and noise sources with different directions and the input SNRs conditions are 0dB, 3dB, 6dB and 9dB, respectively.

![Fig 4 Simulation configuration in a room](image)

To evaluate the performance of speech enhancement methods, three objective speech quality measures are used, i.e. Segmental Signal to Noise Ratio Improvement (SegSNRI) [10], Log-spectral distance (LSD) [11] and Perceptual Evaluation of Speech Quality (PESQ) [12]. The performance of the proposed method (IPBF+TSPP) is investigated by comparing with the three methods including OM-LSA [1], PBF [3] and IPBF [4].

![Fig 5 Spectrograms](image)

Fig.5 Spectrograms of (a) clean target signal, (b) corrupted signal by babble noise, SNR = 3dB, noise direction: q3, (c) enhanced signal by OM-LSA[10], (d) enhanced signal by PBF[3], (e) enhanced signal by IPBF[4], (f) enhanced signal by IPBF+TSPP (the proposed method).

Fig. 5 shows the spectrogram comparison among four methods. For the three reference methods, there are lots of residual noises existed while target speech is absent. Whereas, the proposed method could remove residual noise effectively while target speech is absent.

The results of the objective evaluation are shown in Fig.6. The blue solid line and black dotted line represent the mixed noise and babble noise environments, respectively. From Fig. 6(a) and Fig. 6(b), we can see that the proposed method outperforms other three methods with respect to SegSNRI, LSD and PESQ in case A. For case B shown in Fig.6(c) and Fig. 6(d), i.e. the directions of noise sources are out of the training data, the SegSNRI and LSD of the proposed method are better than other three methods. For the PESQ test, the proposed method is better than the others expect IPBF which gets a slightly higher PESQ score than the proposed method.

For case B shown in Fig.6(c) and Fig. 6(d), i.e. the directions of noise sources are out of the training data, the SegSNRI and LSD of the proposed method are better than other three methods. For the PESQ test, the proposed method is better than the others expect IPBF which gets a slightly higher PESQ score than the proposed method. When the directions of noise sources are not included in the training data, there would be some estimation error for TSPP and a slight distortion would be introduced into the enhanced speech. In comparison with IPBF method, the proposed method has the better noise reduction result and lower PESQ score in case B.
In this paper, we propose a dual-microphone noise reduction method based on the prior knowledge. First, target speech presence and absence are represented by GMMs respectively by using sub-band phase error as the classified feature with an off-line training. Then, we present a soft frame based TSPP estimation method based on Bayesian classification. Finally, the TSPP is adopted to modify the gain function of the IPBF to improve its enhanced performance. Simulation results show that the proposed method outperforms the reference methods and it could reduce noise effectively when target speech is absent.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 61072089), Beijing Natural Science Foundation Program and Scientific Research Key Program of Beijing Municipal Commission of Education (No.KZ201110005005).

REFERENCES