Journal of Signal Processing, Vol. 15, No. 4, pp. 291-294, July 2011

SELECTED PAPER

Towards Intelligent Binaural Speech Enhancement by Meaningful Sound Extraction

Duc Thanh Chau[1], Junfeng Li[2] and Masato Akagi[1]

[1] Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
Phone/Fax: +81-761-51-1391/+81-761-51-1149
E-mail: duc.chau@jaist.ac.jp, akagi@jaist.ac.jp

[2] Institute of Acoustics, Chinese Academy of Sciences
21 Beisihuan Xilu, Haidian, Beijing 100190, China
Phone: +86-10-62528010-1678
E-mail: lijunfeng@hccl.ioa.ac.cn

Journal of
Signal Processing

信号処理

Journal of Signal Processing, Vol. 15, No. 4, pp. 291-294, July 2011

SELECTED PAPER

Towards Intelligent Binaural Speech Enhancement by Meaningful Sound Extraction

Duc Thanh Chau[1], Junfeng Li[2] and Masato Akagi[1]

[1] Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
Phone/Fax: +81-761-51-1391/+81-761-51-1149
E-mail: duc.chau@jaist.ac.jp, akagi@jaist.ac.jp

[2] Institute of Acoustics, Chinese Academy of Sciences
21 Beisihuan Xilu, Haidian, Beijing 100190, China
Phone: +86-10-62528010-1678
E-mail: lijunfeng@hccl.ioa.ac.cn

Abstract

Current speech enhancement applications, such as binaural hearing aids, mainly aim at suppressing interference signals and preserving the target signal with its binaural cues. However, in addition to the target signal, human beings are able to catch other important or meaningful sounds (e.g., calls from others) in daily conversation. This attention mechanism to meaningful signals is seldom taken into consideration in state-of-the-art signal processing systems. In this paper, we propose an intelligent binaural speech enhancement model by extracting the meaningful signals as well as enhancing the target signal. In particular, the proposed model consists of two main parallel processes: binaural target signal enhancement and binaural meaningful signal extraction, and finally yields the binaural outputs. Experimental results showed that the proposed system is able to not only suppress the interfering noise signals, but also enhance the target signal and non-target meaningful signals.

1. Introduction

The main purpose of speech enhancement is to preserve only one signal, the target signal, and reduce all the undesirable signals, such as the background noise, reverberation, and non-target speech. However, in addition to the target speech, there may be other meaningful signals that usually provide important (at least useful) information. Such meaningful signals are quite popular in daily life, e.g., the ring of the telephone or a call from someone probably behind the listener. In addition, in some urgent cases, it is quite dangerous if some non-target (meaningful) signals, e.g., the sound from a car horn or fire-alarm signal, are not perceived. However, state-of-the-art speech enhancement systems do not include the function for extracting these meaningful signals, which may lead to inconvenient and/or dangerous situations for users [1]. Therefore, detecting and extracting meaningful signals should be indispensable for speech enhancement in speech communication and hearing assistant systems.

Due to the high performance in suppressing interfering signals, multi-channel speech enhancement techniques have proven to be far superior to single-channel techniques. Many multi-channel speech enhancement systems have been proposed and widely researched, such as the delay-and-sum beamformer, generalized sidelobe canceller (GSC) beamformer [2], transfer function GSC [3], GSC with post-filtering, multi-channel Wiener filter [4], and blind source separation (BSS) [5]. However, these systems normally require a large array of spatially distributed microphones to achieve a high spatial selectivity and yield single-channel monaural output, which suffers from a high level of complexity and loss of the binaural cues in the output.

Consequently, binaural speech enhancement using two-inputs and two-outputs has been studied because of its smaller physical size and lower computational cost. Dorbecker *et al.* proposed a two-input two-output spectral subtraction approach [6], Kollmeier *et al.* introduced a binaural noise reduction scheme based on the interaural phase difference (IPD) and interaural level difference (ILD) in the frequency domain [7]. Lotter *et al.* proposed a dual-channel speech enhancement approach based on superdirective beamforming [8]. These methods are usually based on some strict assumptions that might not be satisfied in practical environments, e.g., a zero correlation between the noise signals, a diffused noise field, etc. More recently, Li *et al.* proposed a two-stage binaural speech enhancement (TS-BASE) algorithm, which was confirmed effective in dealing with non-stationary multiple-source interference signals and preserving binaural cues [9]. However, in the original TS-BASE algorithm, no meaningful signals (other than the target signal) were taken into account and preserved at the outputs [9].

To further development of binaural hearing systems, we aim to create a smart speech enhancement system that not only enhances the desired signal but also detects and presents meaningful sounds to users at the same time. Motivated by this idea and the advantages of using TS-BASE, we propose an intelligent speech enhancement approach for binaural hearing applications, namely intelligent TS-BASE (iTS-BASE). In principle, the proposed model is performed in two parallel processes. The first process enhances the target signal from a given direction using the traditional TS-BASE. The second process detects and extracts the other meaningful signals besides the target. Finally, the enhanced target signal and the extracted signals are combined to generate the final outputs while preserving the binaural cues for the sound directions. The experimental results showed that the iTS-BASE approach maintains the good performance of TS-BASE in enhancing the target signal and preserving the binaural cues, and is successful in extracting meaningful signals.

2. Original TS-BASE

Two-stage binaural speech enhancement (TS-BASE) is a speech enhancement method proposed by Li *et al.* [9]. Basically, the TS-BASE exploits the Equalization-Cancellation (EC) model and Wiener Filter to enhance target signal from a given direction in two stages (Figure 1).

1. *Estimation of interference signals.* The EC model is ap-

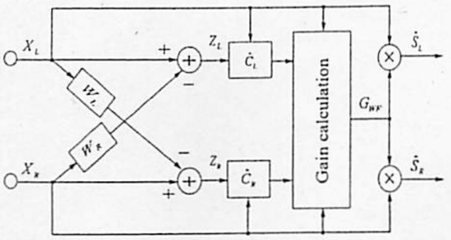

Figure 1: Block diagram of TS-BASE algorithm

plied to estimate the interference signals in which the equalization process is performed in the training process to construct two equalizers (left and right), and the cancellation process uses the two equalizers to cancel the target signal in each channel. A compensation process is conducted to make the remaining signal equivalent to the interference signals based on the Wiener theory. As a result, the remaining signal contains only the interference signals received in each microphone.

2. *Enhancement of target signal.* The estimated interference signals in the first stage are used to construct the gain function of the speech enhancer which is shared in both channels for the preservation of the binaural cues. Finally, the gain function is applied to the original binaural input to obtain the enhanced signal.

3. Proposed Intelligent TS-BASE

3.1 Principle of iTS-BASE

The conceptual model is proposed as shown in Figure 2 to construct an intelligent TS-BASE algorithm, including the two main parallel processes: (1) the first process implements the original TS-BASE algorithm to enhance the target signal from a specific direction. The result from this process is expected to be only the signal from the target direction, and thus, the signals from the other directions should be suppressed. (2) The second process attempts to detect and extract the meaningful signals considered important to the user. This process must strictly be concurrently performed, and share the same input with the first process. Moreover, the meaningful signals from the non-target direction are also binaural signals with binaural cues, which are very important in some serious cases. One typical example is someone should be able to judge where a car is when they hear the car horn.

The key factor in this research is to detect and extract the meaningful sounds that were not considered by the state-of-the-art speech enhancement systems. In real-world environments, there are a huge number of meaningful sounds, including speech (e.g., a call from someone) and non-speech (e.g., a telephone ring, the sound of a car horn, or the sound of a fire alarm system). In principle, however, it is extremely difficult to determine which sound is meaningful from among a vast mixture of sounds because it is highly dependent upon the situations where a human can perceives the presented sounds. Although meaningful signals have diverse characteristics that attract a human's perceptual attention, in this paper, the meaningful signals were limited to the sounds with the following physical characteristics:

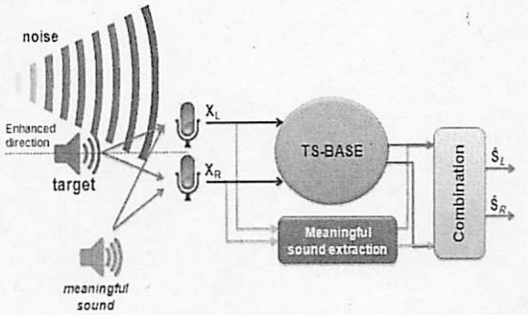

Figure 2: Conceptual model of the proposed iTS-BASE algorithm

- *Strong energy*: The meaningful signals that human beings are typically interested in are normally strong in intensity. This is because weak sounds are masked by other stronger sounds in practical environments.
- *Sufficient temporal duration*. Meaningful sounds are normally long enough for a human to perceive. A sound that is too short in duration is not easily recognizable by humans.
- *Sudden occurrence*: Some meaningful sounds (e.g., telephone ring) occur with a sudden increase in energy, which easily attracts the attention of humans in daily-life.

Actually, in addition to the above-mentioned basic characteristics, there are a lot of other characteristics for diverse meaningful signals that generally depend on the perceptual attention of the listeners in different environments. Although the dominant factors for determining meaningful signals greatly vary under different circumstances, generally speaking, the above characteristics are common features for most meaningful signals in the real world. In the current implementation of our iTS-BASE algorithm, only the first two characteristics are considered.

3.2 Implementation of iTS-BASE

The proposed iTS-BASE approach consists of the original TS-BASE for enhancing the target signal in the first process, and the meaningful signal extraction in the second process, which will be detailed in this section. We define a meaningful signal for extraction as a signal (other than target signal) that satisfies two conditions: (1) its energy is sufficiently strong (e.g., larger than a specific threshold) and (2) its duration is long enough (e.g., lasts for a given duration). In this research, only one meaningful signal is considered at a time. As a result, it is the biggest signal other than target signal.

We noticed that to enhance a signal from a given direction, the TS-BASE algorithm aims to preserve the signal from that direction and suppress all the signals from the other directions. This means that the TS-BASE algorithm can be used to extract meaningful signals if their direction can be determined. Therefore, this algorithm is exploited again in the second process as follows: a sound source localization task is carried out to estimate the direction of arrival (DOA) of a candidate of a meaningful signal, followed by the candidate meaningful signal extraction by the TS-BASE algorithm, an evaluation process is performed to judge whether the extracted signal is meaningful

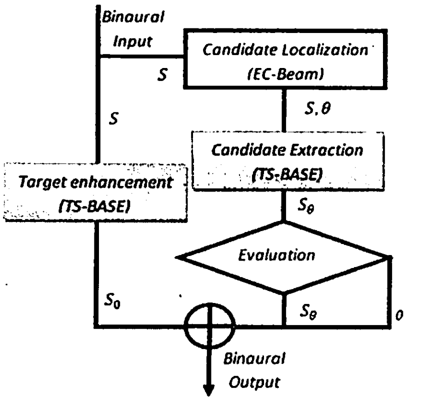

Figure 3: Implementation flowchart of the proposed iTS-BASE algorithm

(satisfies two conditions) and eventually outputting the binaural target and meaningful signals by combining the output signals from the two processes. The implementation flowchart of the proposed iTS-BASE is shown in Figure 3.

3.2.1 DOA estimation of a meaningful signal

Concerning the DOA estimation of a meaningful signal, the algorithm based on EC theory and beamforming scanning techniques, namely EC-Beam, which we previously proposed, was exploited [10]. The EC-Beam algorithm was proven effective in highly-accurate estimation of the DOA of a sound source in the presence of HRTF effects. Another advantage of the utilization of EC-Beam for DOA estimation of a meaningful signal is that both the TS-BASE and EC-Beam algorithms are based upon the EC-theory, so they can share the same equalizers in the cancellation stage. Since the meaningful signal is different from the target signal, in the current implementation, the DOA of the meaningful signals is determined by scanning the non-target directions using EC-based beamforming.

3.2.2 Extraction of meaningful signal

After the DOA of meaningful signal candidate is estimated, the candidate signal will be extracted using the TS-BASE algorithm [9]. Then, the extracted candidate signal is evaluated for whether it is meaningful. In particular, the candidate is only considered meaningful if its energy is stronger than a predefined threshold and lasts longer than a pre-defined duration. In the implementation, these thresholds were experimentally set: the threshold in intensity was set at 0.5 of the average energy of the whole signal, and that in duration was set at 0.2 second. The output of the meaningful signal extraction will be the extracted candidate signal if it satisfies all the criteria; and zero otherwise.

3.2.3 Enhancement of target and meaningful signals

The output of the proposed iTS-BASE algorithm is finally generated by combining the output of the original TS-BASE algorithm (the enhanced target signal), and the output of the meaningful signal extraction.

4. Experiments and Results

4.1 Experimental Configuration

A situation is simulated in which the target speaker is located in front of the listener and another guy calls for the listener from behind (i.e., a meaningful signal) in the experiments. The target signal is the utterance selected from the ATR database [11] and the meaningful signal is a recorded sound of speech, "hello". The HRTF database from the MIT Media lab [12] was used to obtain the binaural sounds. The speech data were first upsampled to 44.1 kHz and convolved with the HRTF, then downsampled to 8 kHz. Binaural background noise was recorded at a cafeteria using two microphones placed on the head of a dummy where the ears should be. The target signal was assumed from the front of the listener (i.e., 0^o), while the direction of the meaningful signal was set to 60^o. The amplitude of the meaningful signal was controlled to make the ratio of average amplitude between the meaningful signal and the target signal (MTR) be 0.5 and 1.0, respectively. The mixture of the target and meaningful signals was then considered as a clean signal to be estimated. The noisy signal was generated by adding the recorded cafeteria noise into the mixture of the target and meaningful signals at SNRs of 0, 5, 10, and 15 dB. In the DOA estimation of the meaningful signal using the EC-Beam algorithm, the direction from $[-10^o, 10^o]$ was considered the target direction and was ignored in order to scan for the meaningful signal.

4.2 Experimental Results and Discussions

The performance of the proposed iTS-BASE algorithm was evaluated in terms of two measures, namely, the perceptual evaluation of speech quality (PESQ) score [13] and log-spectral distance (LSD). The PESQ evaluation results are shown in Figure 4. In general, the PESQ of the iTS-BASE algorithm was higher than that of the TS-BASE one, which indicates the performance of the iTS-BASE algorithm is better than that of the original TS-BASE algorithm in improving the speech quality. Both the TS-BASE and iTS-BASE algorithms provide a much higher PESQ improvement compared with the non-processed noisy inputs. In the case of MTR = 1.0, we observed that the PESQ of iTS-BASE is steadily above the other PESQs. In this case, when the SNR becomes high (or the noise becomes low), the TS-BASE performance worsens. The reason for this is that the clean signal contains signals from two separate directions (the target signal is from 0^o and the meaningful signal from 60^o). However, the TS-BASE algorithm is just able to enhance the signal from only one direction (target) and tends to reduce the strength of the signals from the other directions, including the meaningful signal. When the noise becomes low, the energy of the non-target signal is mainly from the meaningful signal. Since the TS-BASE algorithm removed the meaningful sound, its PESQ value lowers to an even lower level than the non-processed signal. In contrast, by enhancing the target signal and extracting the meaningful signal at the same time, the iTS-BASE algorithm performs well and is stable for almost all the SNR levels.

The LSD results plotted in Figure 5 show that the performance of the TS-BASE algorithm worsens when the SNR increases for both MTR = 0.5 and MTR = 1.0. This is also explained by the fact that the TS-BASE algorithm removes all the

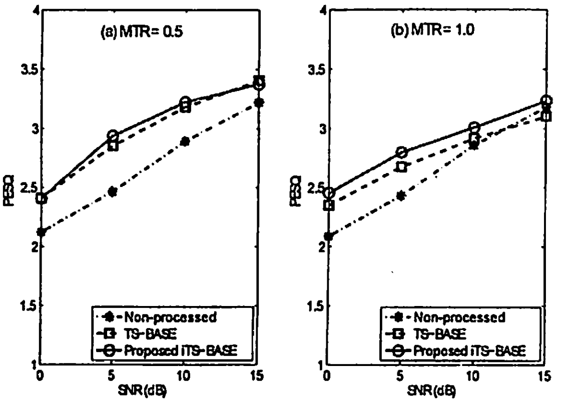

Figure 4: Experimental results in terms of PESQ of noisy signal, signals enhanced by TS-BASE and iTS-BASE algorithms

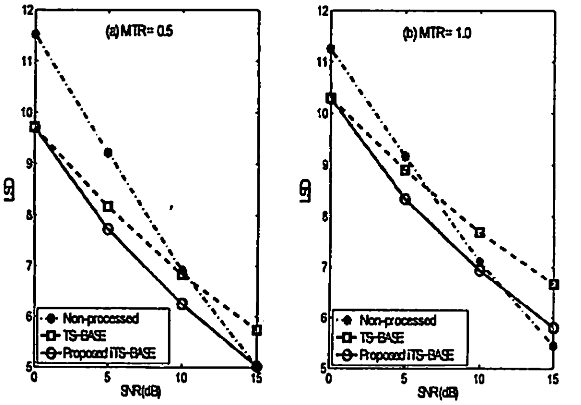

Figure 5: Experimental results in terms of LSD of noisy signal, signals enhanced by TS-BASE and iTSBASE algorithms

non-target signals including the meaningful signals. When the noise level decreases, the meaningful signal becomes the main focus from the non-target signals and removing it makes the results from the TS-BASE algorithm differ from that of the clean signal. In contrast to the TS-BASE algorithm, the iTS-BASE algorithm generally performs well and is more stable. One thing should be noted, in both cases, the LSD value of the TS-BASE and iTS-BASE algorithms is the same when SNR = 0. This is because at this SNR, the noise is much bigger than meaningful sound, and therefore, the extracted signal is not considerable compared to the remaining noise. However, under high SNR conditions, the iTS-BASE algorithm becomes increasingly better than the TS-BASE algorithm. This confirms the effectiveness of the proposed iTS-BASE algorithm in extracting meaningful signals.

5. Conclusion

Many binaural speech enhancement methods have been proposed for binaural hearing applications. However, the problem of preserving the non-target meaningful signals has not yet been taken into consideration. This may lead to inconvenient or dangerous situations for the user in some practical situations. In this research, we proposed an intelligent binaural speech enhancement system based on the TS-BASE algorithm, namely the iTS-BASE algorithm, which not only enhances the target signal but also captures and presents the non-target meaningful sounds. The iTS-BASE algorithm basically includes two main processes: the first process uses the TS-BASE algorithm to enhance the target signal, and the second process detects, captures, and represents the meaningful signal with the target one. In the experiments, we took into consideration the criteria for simple alarm sounds, such as the signal's energy and the signal's duration. The experimental results showed that the proposed iTS-BASE algorithm performs as well as the TS-BASE one, but can also deal with some of the more simple meaningful sounds.

References

[1] M. Brandstein and D. Ward: Microphone Arrays, Digital Signal Processing, Springer, ISBN 3-540-41953-5, 2001.

[2] J. Griffiths: An alternative approach to linearly constrained adaptive beamforming, IEEE Trans. Antennnas Propagat., Vol. 30, pp. 27-34, 1982.

[3] S. Gannot, D. Burshtein and E. Weinstein: Signal enhancement using beamforming and nonstationarity with applications to speech, IEEE Trans. Signal Processing, Vol. 49, No. 8, pp. 1614-1626, 2001.

[4] S. Doclo, A. Spriet, J. Wouters and M. Moonen: Frequency-domain criterion for the speech distortion weighted multichannel Wiener filter for robust noise reduction, Speech Communication, Vol. 49, No. 7-8, pp. 636-656, 2007.

[5] R. Aichner, H. Buchner, M. Zourub and W. Kellermann: Multi-channel source separation preserving spatial information, Proc. ICASSP2007, pp. I5-8, 2007.

[6] M. Dorbecker and S. Ernst:Combination of two-channel spectral subtraction and adaptive Wiener post-filtering for noise reduction and dereverberation, Proc. EUSIPCO1996, pp.995-998, 1996.

[7] B. Kollmeier, J. Peissig and V. Hohmann: Binaural noise-reduction hearing aid scheme with real-time processing in the frequency domain, Scand. Audio. Suppl., Vol. 38, pp. 28-38, 1993.

[8] T. Lotter, B. Sauert and P. Vary: A stereo input-output superdirective beamformer for dual channel noise reduction, Proc., Eurospeech2005, pp. 2285-2288, 2005.

[9] J. Li, S. Sakamoto, S. Hongo, M. Akagi and Y. Suzuki: A two-stage binaural speech enhancement with Wiener filter for high-quality speech communication, Speech Communication, 2010.

[10] D. Chau, J. Li and M. Akagi: A DOA estimation algorithm based on Equalization-Cancellation Theory, Proc. Interspeech2010, pp. 2770-2773, 2010.

[11] A. Kurematsu, K. Takeda, H. Kuwabara, K. Shikano, Y. Sagisaka and S. Katagiri: ATR Japanese speech database as a tool of speech recognition and synthesis, Speech Communication, Vol. 9, No.4, pp.357-363, 1990.

[12] B. Gardner and K. Martin: HRTF measurements of a KEMAR dummy head microphone, http://sound.media.mit.edu/KEMAR.html, Accessed April, 2010.

[13] ITU-T P.862: Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, ITU-T Recommendation P.862, 2000.