JAIST Repository
https://dspace.jaist.ac.jp/

Title	Towards an intelligent binaural speech enhancement system by integrating meaningful signal extraction
Author(s)	Chau, Duc Thanh; Li, Junfeng; Akagi, Masato
Citation	2011 International Workshop on Nonlinear Circuits, Communication and Signal Processing (NCSP'11): 344-347
Issue Date	2011-03-03
Type	Conference Paper
Text version	publisher
URL	http://hdl.handle.net/10119/9978
Rights	This material is posted here with permission of the Research Institute of Signal Processing Japan. Duc Thanh Chau, Junfeng Li and Masato Akagi, 2011 International Workshop on Nonlinear Circuits, Communication and Signal Processing (NCSP'11), 2011, pp.344-347.
Description	

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

2011 International Workshop on Nonlinear Circuits,
Communication and Signal Processing
NCSP'11, Tianjin, China, March 1-3, 2011

Towards an intelligent binaural speech enhancement system by integrating meaningful signal 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
[(2)] Institute of Acoustics, Chinese Academy of Sciences, 21, Beisihuan Xilu, Haidian, Beijing, China
duc.chau@jaist.ac.jp, akagi@jaist.ac.jp, lijunfeng@hccl.ioa.ac.cn

***Abstract*—Current speech enhancement applications, such as binaural hearing aids, mainly aim to suppress interference signals and enhance the target signal with preservation of binaural cues. However, in addition to the target signal, human beings are able to pay attention to other important or meaningful sounds (e.g., the call from others) in daily conversation. This attention mechanism to meaningful signals is seldom considered in the 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 the target signal. Specifically, the proposed model consists of two main parallel processes: binaural target signal enhancement and binaural meaningful signal extraction, finally yielding the binaural outputs. Experimental result showed that the proposed system is able to not only suppress interfering noise signals, but also enhance target signal and meaningful signals.**

I. Introduction

The main purpose of speech enhancement is to preserve only one signal which is considered as the target signal and reduce all undesired signals such as background noise, reverberation and non-target speech. However, in addition to the target speech, there may be other meaningful signals which usually provide important (at least useful) information. Such meaningful signals are quite popular in daily life, e.g., the ring of telephone and the call from someone probably behind the listener. In some urgent cases, furthermore, it is quite dangerous if some non-target (meaningful) signals e.g., the sound from car hooter and fire-alarm signal, are not perceived. However, state-of-the-art speech enhancement systems do not involve the function of extracting these meaningful signals, which may lead to inconvenient and/or dangerous 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 technique has shown great superiority to single-channel technique. So far, many multi-channel speech enhancement systems have been proposed and widely researched, such as, 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 high spatial selectivity and yield single-channel monaural output, which suffers from the high complexity and loss of binaural cues at the output.

Consequently, binaural speech enhancement with two-input two-output has been studied for small physical size and low 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 interaural phase difference (IPD) and interaural level difference (ILD) in 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 which might not be satisfied in practical environments, e.g. zero correlation between noise signals, diffuse 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) are taken into account and preserved at the outputs [9].

For further development of binaural hearing systems, we aim to a smart speech enhancement system which not only enhance desired signal but also detect and present meaningful sound for user at the same time. Motivated by this idea and the advantage of 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 through two parallel processes. The first process is to enhance target signal from a given direction by using the traditional TS-BASE. The second process will detect and extract meaningful signal other than target. Finally the enhanced target signal and the extracted signal are combined to generate the final outputs with preserving binaural cues for sound directions. Experimental results showed that the iTS-BASE approach maintains the good performance in enhancing target signal and preserving the binaural cues, and is successful in extracting the meaningful signals.

II. Original TS-BASE

Two-stage binaural speech enhancement (TS-BASE) was firstly proposed by Li *et al* [10] and consequently improved in [9]. Basically, the TS-BASE exploits Equalization-Cancellation (EC) model and Wiener Filter to enhance target signal from a given direction through two stages (Fig. 1).

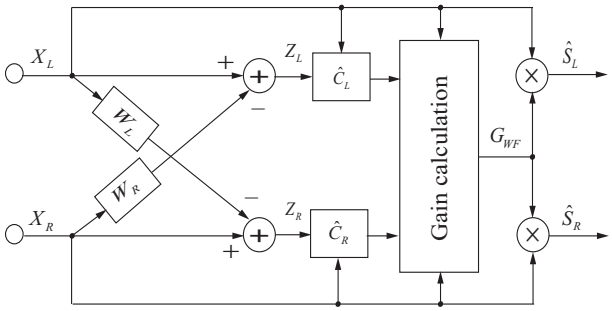

Fig. 1. *Block diagram of TS-BASE.*

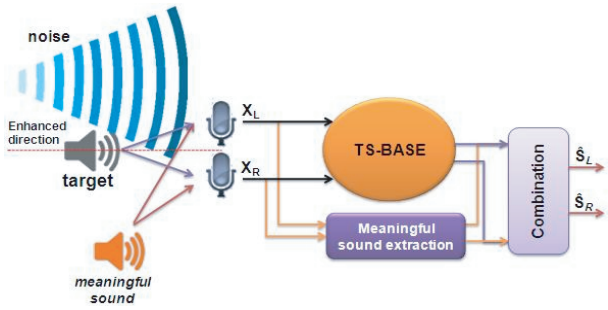

Fig. 2. *The conceptual model of the proposed iTS-BASE.*

1) *Estimation of interference signals.* The EC model is applied to estimate interference signals in which the equalization process is performed in training process to construct two equalizers (left and right) and the cancellation process applies the two equalizers to cancel the target signal in each channel. A compensation process is further performed to make the remaining signal equivalent to interference signals based on Wiener theory. As a result, the remaining signal contains only 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 speech enhancer which is shared in both channels for binaural cues preservation. Finally, the gain function is applied to the original binaural input to get the enhanced signal.

III. The Proposed Intelligent TS-BASE

A. Principle of iTS-BASE

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

The key factor in this research is to detect and extract the meaningful sounds which were never 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., telephone ring, sound of car hooter, sound of fire alarm). In principle, however, it is an extremely difficult to determine which sound is meaningful among a vast of mixture sounds because it is highly dependent upon the situations where human perceives sounds. Though meaningful signals have diverse characteristics that attract human's perceptual attention, in this paper, the meaningful signals were limited to the sounds with the following physical characteristics for simplicity:

- *Strong energy*: The meaningful signals that human beings are interested in are normally strong enough in intensity. This is because that the weak sounds will be masked by other stronger sounds in practical environments.
- *Enough temporal duration*. The meaningful sounds are normally long enough for human to perceive. The too short sound in duration is difficult to be recognized by human.
- *Sudden occurrence*: Some meaningful sounds (e.g., telephone ring) occur with sudden increase in energy, which easily attracts the attention of human in daily-life conditions.

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

B. 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. For the meaningful signal, we define meaningful signal as a signal (other than target signal) which satisfies two conditions: (1) its energy is strong enough (e.g., larger than a specific threshold); (2) its duration is long enough (e.g., last for a certain 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.

It is noticed that to enhance signal from a given direction, the TS-BASE aims to preserve signal from that direction and suppress all signals from other directions. This means

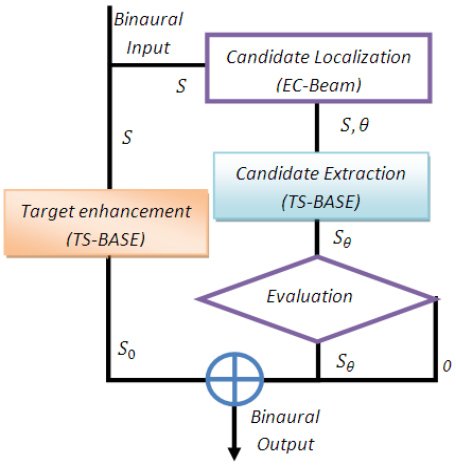

Fig. 3. *The implementation flowchart of the proposed iTS-BASE.*

that the TS-BASE can be used to extract meaningful signal if its direction is determined. Therefore, the TS-BASE is employed again in the second process as follows: a sound source localization task is carried out to estimate the direction of arrival (DOA) of candidate of meaningful signal, followed by the candidate meaningful signal extraction by TS-BASE, an evaluation process is performed to judge whether the extracted signal is meaningful (satisfies two conditions) and eventually outputting the binaural target and meaningful signals by combining the output signals from two processes. The implementation flowchart of the proposed iTS-BASE is shown in Fig. 3.

1) DOA estimation of the meaningful signal: Concerning DOA estimation of the meaningful signal, the algorithm based on EC theory and beamforming scanning techniques, namely EC-Beam, that we previously proposed was exploited [11]. The EC-Beam algorithm was shown effective in high-accurately estimating the DOA of sound source in the presence of HRTF effects. Another advantage of utilization of EC-Beam for DOA estimation of the meaningful signal is that both TS-BASE and EC-Beam are based upon EC-theory, so they can share the same equalizers in 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 through EC-based beamforming.

2) Extraction of meaningful signal: After the DOA of candidate of meaningful signal is estimated, the candidate signal will be extracted using the TS-BASE algorithm. [9]. Then, the extracted candidate signal is evaluated whether it is meaningful or not. Specifically, the candidate is only considered as meaningful signal if its energy is stronger than a pre-defined threshold and last longer than a pre-defined duration. In the implementation, these thresholds were experimentally set: the threshold in intensity was at 0.5 of average energy of the whole signal, and that in duration was 0.2 second. The output of the meaningful signal extraction will be the extracted candidate signal if it satisfied all criteria; and zero in otherwise.

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.

IV. Experiments and Results

A. Experimental configuration

In the experiments, a situation is simulated, in which the target speaker is localized in the front of the listener and another guy calls the listener from behind (i.e., the meaningful signal) .

The target signal is the utterance selected from ATR database [12] and the meaningful signal is a recorded sound of speech "hello". To obtain the binaural sounds, the HRTF database from MIT Media lab [13] was used. The speech data were first up-sampled to 44.1 kHz and convolved with the HRTF, then down-sampled to 8 kHz. Binaural background noise was recorded at cafeteria using two microphones at the two ears of a dummy head. 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 signals was controlled to make the ratio of the meaningful signal to the target signal (MTR) in average amplitude be 0.5 and 1.0, respectively. The mixture of the target and meaningful signals was then considered as the *clean signal* to be estimated. The noisy signal was generated by adding the recorded cafeteria noise into the mixture of the target and meaningful signal at SNRs of $0, 5, 10, 15$ dB. In DOA estimation of the meaningful signal by EC-Beam, the direction from $[-10^o, 10^o]$ was considered as the target direction and was ignored for scanning meaningful signal.

B. Experimental results and discussions

The performance of the proposed iTS-BASE algorithm was evaluated in terms of two measures, namely, perceptual evaluation of speech quality (PESQ) score [14] and log-spectral distance (LSD). The evaluation results of PESQ are shown in Fig. 4. In general, the PESQ of the iTS-BSE algorithm is higher than that of the TS-BASE algorithm, which indicates the performance of the iTS-BASE algorithm is better than the original TS-BASE algorithm in improving speech quality. Both TS-BASE and iTS-BASE algorithms provide much higher PESQ improvements compared with the unprocessed noisy inputs. In the case MTR = 1.0, it can be observed that the PESQ of iTS-BASE is steady above the other PESQs. In this case, when SNR becomes high (or the noise becomes low), the performance of TS-BASE gets worse. The reason 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 is just able to enhance signal from only one direction (target) and tends to reduce signal from other direction, including meaningful signal. When the noise becomes low, the energy of the non-target signal is mainly from meaningful signal. Since the TS-BASE algorithm removed the meaningful sound, its PESQ value becomes lower

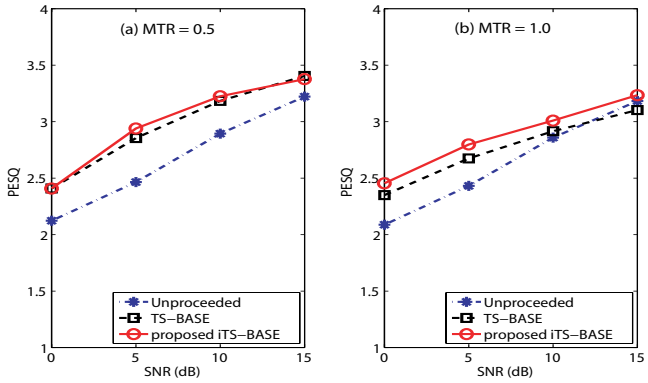

Fig. 4. *Experimental results in terms of perceptual evaluation of speech quality (PESQ)of the noisy signal, the signals enhanced by the TS-BASE algorithm and the iTS-BASE algorithm.*

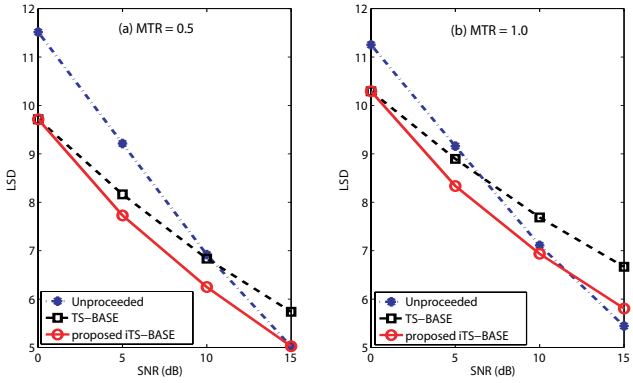

Fig. 5. *Experimental results in terms of log-spectral distance (LSD) of the noisy signal, the signals enhanced by the TS-BASE algorithm and the iTS-BASE algorithm.*

even compared to the un-proceeded signal. In contrast, by enhancing the target signal and extracting the meaningful signal at the same time, the iTS-BASE performs well and stable for almost all SNR level.

The results of LSD plotted in Fig. 5 show that the performance of the TS-BASE algorithm becomes worse when the SNR increases in both cases MTR = 0.5 and MTR = 1.0. This is also explained by the fact that TS-BASE removes all non-target signals including meaningful signals. When the noise decreases, the meaningful signal will become the main part in non-target signals and removing it makes the result from the TS-BASE algorithm more different to the clean signal. In contrast to the TS-BASE algorithm, the iTS-BASE algorithm generally performs well and more stable. There is one notice that, in both cases, the LSD value of TS-BASE and iTS-BASE is the same when SNR = 0. It is because at this SNR, the noise is much bigger than meaningful sound, so that the extracted signal is not considerable compare to the remaining noise. However, in high SNR conditions, the iTS-BASE algorithm becomes better than the TS-BASE algorithm more and more. This confirms the effectiveness of the proposed iTS-BASE algorithm in extracting meaningful signals.

V. Conclusion

Many binaural speech enhancement methods have been proposed for binaural hearing applications. However, the problem of preserving non-target meaningful signal has not been considered. This may lead to inconvenient or dangerous for user in some practical situations. In this research, we proposed an intelligent binaural speech enhancement system based on TS-BASE, namely iTS-BASE, which not only enhance target signal but also capture and present non-target meaningful sound. Essentially, the iTS-BASE includes two main processes: the first process is TS-BASE to enhance target signal; the second process detect, capture and represent meaningful signal with target. In the experiment, we have considered the criteria for simple alarm sounds such as the signal's energy, the signal's duration. Experimental result showed that the proposed iTS-BASE remains good performance of TS-BASE and can deal with some simple meaningful sounds.

References

[1] M. Brandstein and D. Ward, Microphone Arrays, Digital Signal Processing, Springer, ISBN 3-540-41953-5, pp.157-201, 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. On 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, W. Kellermann, "Multi-channel source separation preserving spatial information," in Proc. ICASSP2007, pp. I5-8, 2007.

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

[7] B. Kollmeier, J. Peissig, 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, In Proc., Eurospeech2005, pp. 2285-2288, 2005.

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

[10] J. Li, S. Sakamoto, S. Hongo, M. Akagi and Y. Suzuki, "A speech enhancement approach for binaural hearing aids," in Proc. the 22nd Signal Processing Symposium, pp. 263-268, Sendai, Japan, November, 2007.

[11] D. Chau, J. Li, M. Akagi, "A DOA Estimation Algorithm based on Equalization-Cancellation Theory", In *Proc. Interspeech2010*, Tokyo, 2010. (In Press)

[12] A. Kurematsu, K. Takeda, H. Kuwabara, K. Shikano, Y. Sagisaka, S. Katagiri, "ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis," Speech Communication, vol. 9, no.4, pp.357-363, 1990.

[13] B. Gardner and K. Martin, "HRTF Measurements of a KEMAR Dummy Head Microphone", Available at http://sound.media.mit.edu/KEMAR.html, Accessed April, 2010.

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