Acoustic and Biomedical Engineering

Convolutive Blind Signal Separation Spatial Effectiveness in Speech Intelligibility Improvement

J. Kociński*, A. Sęk and P. Libiszewski

Institute of Acoustics, Adam Mickiewicz University, Umultowska 85, 61-614 Poznań, Poland

Blind signal separation is one of the latest methods to improve the signal to noise ratio. The main objective of blind source separation is the transformation of mixtures of recorded signals to obtain each source signal at the output of the procedure, assuming that they are statistically independent. For acoustic signals it can be concluded that the correct separation is possible only if the source signals are spatially separated. That finding suggests analogies with the classical spatial filtering (beamforming). In this study we analyzed an effect of the angular separation of two source signals (i.e. speech and babble noise) to improve speech intelligibility. For this purpose, we chose the blind source separation algorithm based on the convolutive separation, based on second order statistics only. As a system of sensors a dummy head was used (one microphone inside each ear canal), which simulated two hearing aids of a hearing impaired person. The speech reception threshold, before and after the blind source separation was determined. The results have shown significant improvement in speech intelligibility after applying blind source separation (speach reception threshold fell even more than a dozen dB) in cases where the source signals were angularly separated. However, in cases where the source signals were coming from the same directions, the improvement was not observed. Moreover, the effectiveness of the blind source separation, to a large extent, depended on the relative positions of signal sources in space.

PACS: 43.72.−p, 43.60.−c, 43.60.+d

1. Introduction

There are two main groups of speech enhancement methods, namely: monosensorial and mulitisensorial. The former use spectral subtraction [1–3], Wiener filtering [4], etc. and they have some limitations: they work well for stationary disturbances only. Moreover, they often cause so-called musical noise (an unpleasant effect leading to speech being qualified as unacceptable to listen even if it is intelligible) or speech distortion [5] that leads to intelligibility degradation, even though the signal-to--noise-ratio (SNR) is improved [6]. This happens because speech signal is often modified by such kind of processing in a way that makes it more difficult to understand, e.g. some important information may be subtracted from the signal spectra.

Algorithms from the second group deal with multi--microphone array recordings. The mixture of source signals in each microphone is described by Eq. (1)

$$x_n(t) = \sum_{m=1}^{M} \sum_{k=0}^{K} s_m(t-k) a_{nm}(k)\,, \tag{1}$$

where s_m are source signals and a_{nm} are the length K mixing filters, which in a simplest approach describe the delays and reverberation.

To separate out source signals from their mixtures two main ways can be chosen. Both of them lead to the same effect, however they are based on completely different assumptions.

The first method is based on spatial and spectro--temporal properties of sources, namely the objective is to make so-called beamformer (or spatial filter) that amplifies the signal reaching from one direction while reducing signals coming from other directions [7–10]. The simplest basis of this technique is an appropriate interference of waves from each microphone. Using different delays of each wave and different weights, various spatial filters can be designed. It is obvious that the selectivity (in spatial terms) of the beamformer increases with number of microphones used in the array.

The second group of techniques, exploiting fine differences in signals at successive microphones in a microphone array, is called blind source separation (BSS). Despite the fact that the final effect should be similar to the beamformer [11], this method is based only on statistical properties of the signals coming out from separate sources that are assumed to be statistically independent.

Many approaches to solve the problem of statistical separation were introduced so far [12–19]. It was shown [20–22] that considering only simple statistics, such as decorrelation, it is possible to separate out the original signals. Moreover, such approach seems to be easier, computationally less consuming and more stable than methods using higher order statistics (HOS), which of-

* corresponding author; e-mail: jedrzej.kocinski@amu.edu.pl

ten work satisfactorily in computer simulations while performing poorly for recordings in a real environment [21] and the results are somewhat unpredictable. Opposite to beamforming method, in BSS any assumption on the signals (e.g. their spatial configurations) is not required, except the independence of signals sources. This assumption is met in most of the real acoustical cases.

2. Method

2.1. Aim

The goal of this paper was to investigate a spatial efficiency of the convolutive BSS by means of subjective speech intelligibility and in terms of beamforming. Namely, the speech intelligibility improvement (i.e. the difference in speech reception threshold, SRT, before and after the BSS was applied) was determined for different spatial configurations of target speech and disturbance. A dummy head with two microphones was used and array of sensors, which seems to be the simplest simulation of two hearing aids of a hearing impaired person.

2.2. Intelligibility test and disturbance configurations

The Polish sentence test (PST) [23, 24] was used as a speech material. The sentences of the PST were presented in a background of so-called babble noise. All 25 PST lists are phonemically balanced and contain grammatically correct and semantically neutral utterances consisting of 4 to 6 words. The power spectrum of the babble noise signal optimally matches the power spectra of test sentences as the noise was produced by summing up all of the test sentences with random time shifting and reversing [23].

2.3. Apparatus

All recordings were carried out in an anechoic chamber. Five horizontal angles of target speech source were used: 0° (in front of the dummy head), 30°, 60°, 90°, 180°. For each of the target speech positions eight different angles of the masker source were investigated, namely 0°, 15°, 30°, 45°, 60°, 75°, 90°, 180°, clockwise. The notation used in this paper is as follows: S_x stands for speech signal source placed at the angle of x degrees, N_y stands for disturbance signal source placed at the angle of y degrees, e.g. $S_{30}N_{45}$ describes the configuration in which speech source was placed at 30° and noise source was placed at 45° (clockwise). At the first stage of the experiment, the recordings were carried out in an anechoic chamber using custom PC software implemented in Matlab 6.5: the signals (sentences) were sent via ADAT interface from PC to the Yamaha 01 V digital console used as a D/A converter. Then, their level was adjusted using a Pioneer A-505R amplifier and delivered to a Tonsil Altus 300 loudspeaker placed in an anechoic chamber. The signals from the loudspeaker were recorded using the Neumann KU100 dummy head and additionally by a small (1/2 inch) reference microphone (Svantek SV01A with pre-amp Svantek SV08A) placed just above the dummy head. The signal from this microphone was used to adjust a proper SNR in listening sessions. Then, the signals from the left and the right ear and from the reference microphone were fed to the Yamaha 01V console, converted into digital form and delivered via ADAT to a PC where they were finally stored on a hard drive. A dummy head was placed at a professional swivel table with an angular scale on it, thus the different angular target-masker configurations were possible to obtain.

An adaptive procedure [25] was used during listening sessions, thus speech and noise were recorded separately: first only the PST was recorded for all angles and next the masking signal was recorded analogously. Such a procedure allowed to obtain different SNRs and angular target-masker configurations just by means of mixing up the speech and noise in the PC. In order to get signals at different SNRs in the binaural listening (before BSS) case, the following procedure was used. The root mean square (RMS) of the signals (target and masker) recorded via reference microphone was calculated. According to this value the RMS of the signals from the left and the right ear were adjusted to get an appropriate SNR (the signals from both microphones of the dummy head were multiplied by the same value related to the desirable SNR). For the speech signal, the RMS was calculated for whole sentences. After this procedure the speech and masker were mixed up (separately for left and right channels), D/A converted (TDT RP2) and fed to the headphone buffer (TDT HB7). The signals were presented binaurally to the subjects via Sennheiser HD 580 headphones at the level of 70 dB SPL. When the BSS was applied, the same mixtures were used as input signals to the BSS. The Parra and Spence algorithm [21] implemented by Harmeling [26] was used in the experiment. As the algorithm does not work on-line, the separating filters were calculated previously.

The procedure of filter estimation was as follows. A random sentence was chosen and the procedure of SNR adjustment was proceeded. Then the separating filters were calculated using the mixtures from the left and the right ear. Finally, the set of filters was stored on a hard drive. This procedure was carried out for all the combinations of target-masker spatial configurations and 51 SNRs (between 0 and −50 dB). During the listening session in the "after BSS" paradigm, after SNR adjustment, there was an additional step of separation using previously estimated filters. Then, the separated target signal was presented monaurally to the subject using the same experimental setup as in the binaural case. It must be emphasized that before BSS was applied all binaural cues could be exploited by the subject. However, after the BSS (with two input channels) one of the output signals contained target speech while the other one contained the masker. Therefore, binaural listening was pointless: the only way to present the target signal binaurally was to deliver the same signal (containing the speech) to both ears. This procedure, however, should not change speech

intelligibility. During the listening sessions standard psychoacoustical equipment was used: Tucker–Davis Technologies (TDT) System 3 (real-time processor RP2 and headphone buffer HB7).

The subjects were seated in a double-walled, acoustically-insulated booth. SRTs for each target--masker configuration was determined for 6 subjects. The listeners were paid for participation in the measurements and all of them were Polish native speakers aged between 21 and 28 with no history of hearing disorders.

3. Results

The results of the experiment are shown in Fig. 1. Successive curves show the gathered data for different position of the speech signal: S_0: 0° — filled squares with solid line; S_{30}: 30° — empty circles with dotted line; S_{60}: 60° — filled triangles with dashed line; S_{90}: 90° — filled asterisks with dash-dot line; S_{180}: 180° — crosses with dot-dash line.

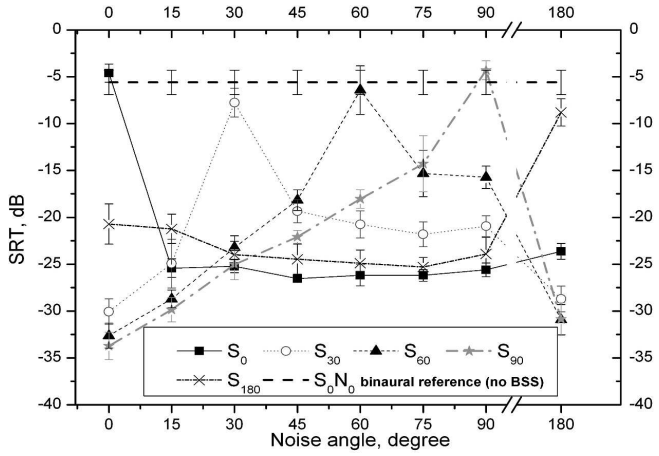

Fig. 1. Juxtaposition of BSS spatial efficiency patterns for five different angles of target speech source (S_0: 0° — filled squares with solid line; S_{30}: 30° — empty circles with dotted line; S_{60}: 60° — filled triangles with dashed line, S_{90}: 90° — filled asterisks with dash-dot line, S_{180}: 180° — crosses with dot-dash line). Additionally, reference SRT (for binaural listening in S_0N_0 configuration with no BSS is plotted with thick dashed line at −5.6 dB and SD of 1.3 dB). The efficiency of BSS is depicted in terms of SRT (in dB) as a function of disturbing noise horizontal angle.

To obtain a reference value the SRT for the S_0N_0 configuration with no signal processing was determined. In such a case the binaural signal was presented to the listener and the same adaptive procedure was used and we obtained a mean value of SRT = −5.6 dB (SD = 1.3 dB) which is in line with the reference value for monaural listening with no signal processing in normal hearing people is −6.1 dB (SD = 1 dB) [23].

The gathered data was subjected to a within-subject analysis of variance (ANOVA) that showed highly significant differences between spatial configurations. However, as can be seen from the figure the pattern of the data for all configurations is homogeneous and very similar. The SRT is the highest (the poorest intelligibility) when the angular position of the speech source coincides with that of the noise and is equal to the SRT characterizing the PST itself. However, when the angular positions of the speech source and that of the noise are different a substantial increase in speech intelligibility is observed. The pattern of the data is consistent with the basic assumption of BSS: when differences (in terms of intensity and/or phase) between the noise reaching two microphones are the same as the differences of speech signal the BSS is unable to separate out the mixture. However, once there are some differences between mixture reaching two sensors, then BSS algorithm is highly efficient. In general, it may be stated that the increase in spatial separation of the sources in the horizontal plane brought about a substantial increase in speech intelligibility. It must be emphasized that the result of the beamformer can be related to the beamforming: when both sources are placed at the same beam there is no possibility to separate them. Only spatial (angular) separation of sources can provide speech intelligibility improvement, as for these cases a target source signal can be separated and the interfering noise signals can be attenuated.

By analogy with psychophysical tuning curves [27], it can be stated that the BSS algorithm is characterized by very steep and asymmetric "spatial tuning": when the sources are at the same place, the BSS efficiency is poor, however for small angular separation the efficiency is significantly better and leads to a high speech intelligibility improvement. It seems, however, that the "steepness" is asymmetric: it is higher on the left hand side of the maximum. The asymmetric pattern of these curves is probably a consequence of the head-related transfer function (HRTF) as it modifies the signals reaching separate ears in different ways, while the signals are not presented directly in front or at the back of the head. It is also caused by the different angular positions of the sources: when we consider the result of BSS in terms of beamforming, different efficiencies should be obtained depending on mutual angular positions of the sources and positions of the sensor matrix as the performance of the beamformer depends on the angle the main beam is formed at.

Based on the presented data it is difficult to say how sharp the observed tuning is (or to introduce a parameter as for example Q3 dB or Q10 dB) because the data were gathered for discrete position of the sources with a resolution of 15 degrees which was not too fine. However, it has been shown [28] that shifting a signal in the time domain by one sample only, which is equivalent to (approximately) 2 degrees, brought about a significant speech intelligibility improvement. Moreover, the parameters of the BSS used here have been fixed at, somehow, arbitrary values and recordings were done in an anechoic chamber using two well separated microphones at a distance of about 23 cm. Therefore it seems that to describe the

spatial resolution of the BSS procedure in detail, it is necessary to carry out more systematic study. However, the presented results are very promising for the future usage of BSS, although they were collected in very specific conditions. They simply show that using the BSS procedure speech intelligibility improvement of more than 10 dB may be easily reached if noise and speech source are separated by 15 degrees.

4. Conclusion

The experiment proved high efficiency of BSS procedure used in speech intelligibility improvement. Using different spatial (angular) target-masker configurations, the spatial characteristics of beamformer created using only statistical methods were obtained. It must be emphasized that these characteristics were expressed in terms of psychoacoustical measure, namely SRT, which reflects the most robust measure of the efficiency of BSS algorithm to create a spatial filter. The data gathered in this study show very effective "spatial tuning" of the algorithm: when both sources (target and masker) are placed at the same angle, the performance of BSS is poor. However small angular separation of the sources results in a very significant decrease in SRT, i.e. a substantial speech intelligibility improvement. Moreover, comparing two different listening paradigms, namely "before (binaural) BSS" (that is "natural" that takes advantage from all binaural cues) listening and "after BSS" (monaural) again it can be stated that the BSS technique brings about a very high speech intelligibility improvement even though after BSS procedure, all binaural cues are lost. However, one must keep in mind that the performance of BSS algorithms depends on the number and spacing of the microphones as well as the geometry and orientation of the microphone array. Moreover the BSS efficiency is strongly affected by acoustical properties of a room. It has been shown by [29] that to get a speech intelligibility improvement in a real room (reverberation time of 0.5 s) the unmixing filters must be much longer that made the BSS much more time consuming.

Acknowledgments

This work was supported by Polish–Norwegian Research Fund and by Polish Ministry of Science and Higher Education grant no. N N 518 502139.

References

[1] S.F. Boll, *IEEE Trans. Acoust. Speech Signal Proc.* **ASSP-27**, 113 (1979).

[2] M. Berouti, R. Schwartz, J. Makhcul, in: *IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 1979.

[3] E. Ephraim, D. Malah, *IEEE Trans. Acoust, Speech Signal Proc.* **ASSP-32**, 1109 (1984).

[4] P. Scalart, J.V. Filho, in: *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 1, 1996, p. 629.

[5] J.R. Deller, H.L. Hansen, J.G. Proakis, *Discrete-Time Processing of Speech Signals*, 2nd ed., EEE Press, New York 2000.

[6] J. Kocinski, *Speech Commun.* **50**, 29 (2008).

[7] D.H. Johnson, D.E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Prentice Hall, Englewood Cliffs, NJ 1993.

[8] M.S. Brandstein, D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*, Springer, Berlin 2001.

[9] M. Kajala, M. Hamalainen, in: *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2001.

[10] T. Yu, J.H.L. Hansen, in: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2009.

[11] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, H. Saruwatari, *EURASIP J. Appl. Signal Proc.* **11**, 1157 (2003).

[12] P. Comon, *Signal Process* **24**, 11 (1991).

[13] C. Jutten, J. Herault, *Signal Process* **24**, 1 (1991).

[14] P. Comon, *Signal Process* **36**, 287 (1994).

[15] A. Ziehe, K.R. Muller, G. Note, G.M. Macker, G. Curio, *IEEE Trans. Biomed. Eng.* **47**, 75 (2000).

[16] A. Cichocki, A. Belouchrani, in: *Third Int. Conf. on Independent Component Analysis and Signal Separation (ICA-2001), San Diego*, 2001.

[17] A. Hyvärinen, J. Karhunen, O. Erkki, *Independent Component Analysis*, Wiley, New York 2001.

[18] A. Cichocki, S. Amari, *Adaptive Blind Signal and Image Processing Learning Algorithms and Applications*, Wiley, Chichester 2003.

[19] S. Choi, A. Cichocki, H.-M. Park, S.Y. Lee, *Neural Inf. Proc.-Lett. Rev.* **6**, 1 (2005).

[20] K. Matsuoka, M. Ohya, M. Kawamoto, *Neural Networks* **8**, 411 (1995).

[21] L. Parra, C. Spence, *IEEE Trans. Speech Audio Proc.* **8**, 320 (2000).

[22] D.-T. Pham, C. Serviere, H. Boumaraf, in: *ICA 2003, Nara (Japan)*, 2003.

[23] E. Ozimek, D. Kutzner, A.P. Sęk, A. Wicher, O. Szczepaniak, *Arch. Acoust.* **31**, 431 (2006).

[24] E. Ozimek, D. Kutzner, A. Sęk, A. Wicher, *Int. J. Audiol.* **48**, 433 (2009).

[25] H. Levitt, *J. Acoust. Soc. Am.* **49**, 467 (1971).

[26] S. Harmeling, computer program, convbss Berlin, http://bme.ccny.cuny.edu/faculty/lparra/publish/, 2001.

[27] B.C.J. Moore, *An Introduction to the Psychology of Hearing*, 5th ed., Academic Press, London 2003.

[28] J. Kocinski, A. Sek, P. Libiszewski, *Speech Commun.* **53**, 390 (2011).

[29] P. Libiszewski, J. Kocinski, *Arch. Acoust.* **32**, 337 (2007).