ARCHIVES OF ACOUSTICS
33, 4, 455–474 (2008)

LOGATOM ARTICULATION INDEX EVALUATION OF SPEECH ENHANCED BY BLIND SOURCE SEPARATION AND SINGLE-CHANNEL NOISE REDUCTION

Szymon DRGAS, Jędrzej KOCIŃSKI, Aleksander P. SĘK

Adam Mickiewicz University
Faculty of Physics
Institute of Acoustics
85 Umultowska Str., 61-614 Poznan, Poland
e-mail: Szymon.Drgas@amu.edu.pl

(*received January 24, 2008; accepted July 28, 2008*)

The subjective logatom articulation index of speech signals enhanced by means of various digital signal processing methods has been measured. To improve intelligibility, the convolutive blind source separation (BSS) algorithm by PARRA and SPENCE [1] has been used in combination with classical denoising algorithms. The efficiency of these algorithms has been investigated for speech material recorded in two spatial configurations. It has been shown that the BSS algorithm can highly improve speech recognition. Moreover, a combination of the BSS with single-microphone denoising methods can additionally increase the logatom articulation index.

Keywords: speech enhancement, logatom articulation index, blind source separation.

1. Introduction

The acoustic signal reaching our ears comprises sound waves from multiple sources as well as their reflections from surfaces in the environment. One of the most remarkable achievements of human perception is the ability to resolve a chosen source from a mixture of different signals. However, for people with impaired hearing this ability is reduced. Hearing aids that merely amplify the signal do not give satisfactory results for people with sensorineural hearing loss. This is because the signal-to-interference ratio (SIR) is not improved with the use of such devices. The way to help people with impaired hearing is to apply speech enhancement algorithms in hearing aids. There have been many studies in the area of speech enhancement algorithms used in hearing aids and cochlear implants, e.g. [2–10], however combining BSS and denoising algorithms is a quite new idea.

The vast family of speech enhancement algorithms may be classified into two broad categories: single- and multiple-microphone methods. The advantage of single microphone methods is low cost and small size, while multiple microphone methods result in much higher performance. The principle of operation of single microphone speech enhancement algorithms is estimation of noise or speech signal. The multiple microphone algorithms exploit spatial diversity.

The perception of a speech signal is usually measured in terms of its quality (naturalness and ease of listening) and intelligibility (percentage of words/sentences that can be correctly identified by listeners) or articulation index (percentage of words without meaning that can be correctly identified by listeners). Most of speech enhancement techniques improve speech quality; although some of them even reduce the intelligibility. This is because speech enhancement algorithms can distort the target speech signal. While the signal-to-noise ratio (SNR) is usually accepted as an objective measure of efficiency [1, 11]; it must be emphasized that it does not reflect intelligibility. Hence, speech intelligibility measurements or articulation index are still the best methods to test the efficiency of speech enhancement methods [12, 13].

No speech enhancement algorithms can reject noise completely. In order to additionally inhibit noise, a combination of single and multiple microphone methods can be used. This is, to date, still an unexplored issue. In this work, the logatom articulation index is measured for degraded speech processed by BSS and postprocessed by well known single-microphone speech denoising algorithms.

The main purpose of this work is to investigate the effects of combining single and multiple microphone methods on improving the speech articulation index. The experiments were carried out in an anechoic chamber. Only an additive noise from point sources was considered. Thus, the goal is to test if the gain in signal-to-interference ratio, i.e. SIR enhancement is greater than the loss in speech signal distortion.

1.1. Blind source separation (BSS)

Recently, much attention has been paid to the development of BSS algorithms. The main goal of convolutive BSS is to filter the signals from a microphone array to extract the original sources, while reducing interfering signals. Given independent sources (e.g. target speech and maskers) $s_m(t)$, $m = 1, 2, ..., M$, where t denotes time, the real mixing process (including delays) can be described as:

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

where M is the number of the independent sources s_m and a_{nm} are the length K mixing filters, describing the delays and acoustical environment. In such a case, the recovery of the independent signals $u_i(t)$ can be described as:

$$u_i(t) = \sum_{n=1}^{N} \sum_{k=0}^{K} x_n(t-k) h_{in}(k), \tag{2}$$

where h_{in} are unmixing filters to be estimated. As can be seen, Eqs. (1) and (2) consist of a convolution of signals. Thus, the estimation of separating filters is a time consuming process. To solve this problem, it is possible to apply an appropriate Fourier transform to Eq. (2). As a result, the time series are converted into polynomials and the convolution is transformed to element-wise multiplications [1, 14–16]:

$$U_n(f) = \sum_{m=1}^{M} X_m(f) H_{nm}(f), \tag{3}$$

where $U_n(f)$ is the f-th spectral coefficient of the n-th signal estimate, $X_m(f)$ denotes the spectrum of the signal recorded by the m-th microphone while $H_{nm}(f)$ is the frequency response of the filter corresponding to the n-th signal estimate and m-th sensor.

Four fundamental approaches have been developed to solve this equation [17, 18]. In the first group there are algorithms that use a statistical measure of independence, namely nongaussianity or sparseness. Higher order statistics is essential to solve this problem. Another approach is based on the spectro-temporal features of signals. This approach leads to the concept of time frequency component analyzer (TFCA) [19].

The third approach exploits the temporal structures of sound sources. It assumes that each source has a non-vanishing temporal correlation. In such a case less restrictive conditions than statistical independence can be used such as decorrelation obtained by second-order statistics (SOS) analysis. Several approaches are based on this assumption [20–24]. The last fundamental approach is based on the non-stationarity properties and second order statistics (SOS). The non-stationarity was first taken into account by MATSUOKA *et al.* [25]. More general solutions based on this approach were proposed by PARRA and SPENCE [1] and PHAM *et al.* [26].

As mentioned above, in the frequency domain this problem becomes much easier and can be solved using statistical methods. The main goal of the BSS method is to invert the mixing process and find an unmixing matrix, that give estimates of source signal with accuracy to scale and permutation. Although the frequency domain BSS can be solved quickly, there is a price to be paid for such a solution. The joint diagonalization provides matrices $\mathbf{H}(f)$ only up to a scale and permutation. There are several ways to cope with this problem. The first is to constrain the separating FIR filters length in the time domain. This solution is motivated by the fact that permutations induce filters' impulse responses with very long tails [1]. It may not be easy to handle, as for long responses the inverse filter is usually even longer. There are also algorithms exploiting the assumption that the signals coming from one source are temporally correlated in the adjacent frequency [26]. Another idea is based on source localization and subsequent grouping filters which attenuate the same jammers. The localization approach is robust since a misalignment at one frequency does not affect other frequencies. Unlike localization, the correlation approach is not robust since misalignment at one frequency affects the results of other frequencies and may cause consecutive misalignments [27].

Moreover, the localization approach is not precise for some frequencies. This happens mainly for low frequencies, where the phase difference caused by the sensor spacing is very small. In contrast, the correlation approach is precise as long as signals are well separated by independent component analysis (ICA), since the measurement is based on the separated signals [27].

In the present study, the non-on-line algorithm introduced by PARRA and SPENCE [1] and implemented by HARMELING [28] is used. Previous research using this algorithm showed very promising results [29, 30].

It must be emphasized, that there also exist also some algorithms of the convolutive BSS that are able to separate out the signals on-line, e.g. [31, 32].

1.2. Single-microphone denoising algorithms

After the BSS separation, single microphone speech enhancement methods can be used. Since their efficiency depends on SIR, it is better to use them after BSS. In this work logatom articulation index processed by four denoising algorithms [33–36] separately was measured. These algorithms take into account additive noise. In general, a recorded noisy signal model can be expressed as follows:

$$y\left(t\right) = x\left(t\right) + d\left(t\right), \tag{4}$$

where $y(t)$ is a noisy speech signal, composed of the speech signal $x(t)$ and noise $d(t)$. Single microphone speech enhancement is a statistical estimation problem of the speech signal, [$x(t)$], from the noisy speech signal, [$y(t)$]. The algorithms used in this study work in the frequency domain, where $X_t(f_k)$ and $Y_t(f_k)$ are the Fourier coefficients of the speech signal and noisy speech, respectively, at time frame t and frequency f_k. These conventional speech enhancement methods are based on filtering the signal with a gain function [36]:

$$\left|G_t\left(f_k\right)\right| = \frac{\left|\hat{A}_t\left(f_k\right)\right|}{\left|Y_t\left(f_k\right)\right|}, \tag{5}$$

where $\left|\hat{A}_t\left(f_k\right)\right|$ is an estimate of the speech amplitude and $|Y_t(f_k)|$ is the observed amplitude in a given frame t at frequency f_k.

In the spectral subtraction algorithm by BOLL [33], the $\hat{A}_t\left(f_k\right)$ estimator is obtained by subtracting an estimate of the noise spectrum from the noisy speech spectrum. Spectral information required to describe the noise spectrum is obtained from the signal measured during nonspeech activity. The estimation error is corrected in Boll's algorithm by means of several techniques. The first is a magnitude averaging; this method exploits a symmetric distribution of noise distortion, however, it brings the risk of temporal smearing. The second step of minimizing noise error is rectification. When the spectral magnitude of the noisy signal is less than the average noise spectral magnitude then the output is set to zero. The advantage is the noise floor reduction. However, there is a risk of speech and noise frames to be floored. Thus the

logatom articulation index may decrease in this technique. After these steps there is still residual noise which has a magnitude between zero and the maximum value measured during nonspeech activity. The audible effects of the residual noise can be reduced by taking advantage of its frame-to-frame randomness, thus it can be suppressed by replacing its current value with its minimum value chosen from the adjacent frames analysed.

It should be noted that the short-time power spectrum of white noise still displays peaks and valleys. Frequency locations of these peaks and valleys are random and they vary in frequency and amplitude from frame to frame. When the smoothed estimate is subtracted from the noise, all spectral peaks are shifted down while valleys are set to zero. Thus after such a subtraction the noise components still remain present. These components are narrower and perceived as time varying tones and are called the musical noise. BEROUTI [34] proposed a modified version of spectral subtraction. He developed a spectral flooring technique to reduce musical noise, which can be described by the following equation [34]:

$$D\left(f_k\right) = \left|Y_t\left(f_k\right)\right|^2 - \left|\lambda\left(f_k\right)\right|^2$$

$$\hat{A}\left(f_k\right)^2 = \begin{cases} D\left(f_k\right) & \text{if} \quad D\left(f_k\right) > \beta\left|\lambda\left(f_k\right)\right|^2, \\ \beta\left|\lambda\left(f_k\right)\right|^2 & \text{otherwise.} \end{cases} \tag{6}$$

$D(f_k)$ is the difference between the noise and the noisy signal; $|Y_t(f_k)|^2$ and $|\lambda(f_k)|^2$ are the noisy speech and noise power spectra, respectively. After noise estimate subtraction, the filling with the noise spectrum is made. The spectral components of $\hat{A}\left(f_k\right)$ are prevented from descending below the lower band $\beta\lambda(f)$. With the use of this technique the valleys between peaks are not as deep as for the case $\beta = 0$. This procedure reduces spectral excursions.

In the third speech enhancement algorithm (MMSE STSA – minimum mean square error short time spectral amplitude [35]), the normal distribution probability of Fourier expansion coefficients of speech signal is assumed. With this assumption the spectral amplitude estimators can be derived from noisy speech. The filtering function $G_{<>}$ can be obtained from the relationship:

$$G_{<>} = \frac{\hat{A}_k}{R_k}, \tag{7}$$

where A_k denotes the spectral amplitude estimator and R_k is the spectral amplitude of the noisy speech.

Therefore, the function $G_{<>}$ is defined by the equation:

$$G_{<>}^{EM} = \frac{\mathrm{SNR}_{\mathrm{prio}}}{1 + \mathrm{SNR}_{\mathrm{prio}}} \exp\left\{\frac{1}{2}\int\limits_{v_k}^{\infty} \frac{e^{-t}}{t}\,\mathrm{d}t\right\}. \tag{8}$$

The parameters SNR *a priori* ($\mathrm{SNR}_{\mathrm{prio}}$) and SNR *a posteriori* ($\mathrm{SNR}_{\mathrm{post}}$) are defined as:

$$\mathrm{SNR}_{\mathrm{prio}} = \frac{|X_t(f_k)|^2}{\lambda^2} \tag{9}$$

and

$$\mathrm{SNR}_{\mathrm{post}} = \frac{|Y_t(f_k)|^2}{\lambda^2}. \tag{10}$$

In practical implementations of speech enhancement systems, these parameters are unknown in advance as the noisy speech alone is available. It has been reported that the *a priori* SNR acts is a key parameter in the reduction of speech distortions and musical noise. In the MMSE STSA method these unknown parameters are replaced with estimates of the noise power spectral density. The estimate of an *a priori* SNR is given by:

$$\overline{\mathrm{SNR}}_{\mathrm{prio}} = \alpha G^2\left(\mathrm{SNR}_{\mathrm{prio}}(n-1), \mathrm{SNR}_{\mathrm{post}}(n-1)\right)\mathrm{SNR}_{\mathrm{post}}(n-1) + (1-\alpha)\,P\left[\mathrm{SNR}_{\mathrm{post}} - 1\right], \tag{11}$$

where $P[\,]$ denotes half-wave rectification. The parameter α controls the trade-off between the noise reduction and the transient distortion introduced into the signal. This is a *decision-directed* estimator since it is updated on the basis of a previous amplitude estimate.

SCALART and FILHO [36] investigated experimentally the estimate and true values of these parameters for sinusoidal tone with additive noise. They showed that the *a posteriori* SNR estimate exhibits large standard deviation. They proposed to include the *a priori* concept in the classical speech enhancement schemes (like Wiener), spectral subtraction or maximum likelihood. This can be done by considering $E\{\mathrm{SNR}_{\mathrm{post}}(f_k)\} = 1 + \mathrm{SNR}_{\mathrm{prio}}(f_k)$. The fourth algorithm used in the current study is the Wiener filter with the *a priori* SNR derived from the decision directed method concept.

Single microphone denoising methods applied in this study differ in the statistical models of speech and enhanced speech distortion measure. Evaluations of these algorithms by logatom articulation index measurements can show how denoising algorithms used after BSS could be useful in future application in hearing aids.

2. Aim

The main purpose of this work was to assess the logatom articulation index improvement after using the convolutive BSS procedure combined with single microphone denoising algorithms applied to the signal after BSS. Speech signals were initially recorded in different noisy conditions after which BSS was applied [1]. Next, the signal after BSS characterised by a higher SNR was additionally transformed by each of the above described denoising algorithms [33–36] separately. Therefore, the efficiency of BSS only and BSS with each of the single microphone denoising algorithms expressed

by means of logatom articulation index was estimated and compared. It should be emphasized that only additive sources were considered. The reflections from surfaces and background noise were neglible as recordings were carried out in an anechoic chamber.

The efficiency of the combined algorithms (i.e. BSS and BSS with denoising algorithms) was assessed using two different spatial configurations. The purpose of using different spatial configurations was to show the efficiency of the dependence of the speech enhancement method on the spatial localization of the target sound source and disturbing sound sources.

3. Stimuli

3.1. Recordings

The Polish nonsense word list (logatoms) [37–39] was used as speech material. The logatoms were initially recorded in an anechoic chamber in the presence of masking disturbing sound sources. The background noise and reverberation were neglible. Two spatial configurations shown in Fig. 1 and Fig. 2 were used. The target speech source was always placed in front of the microphone array of four unidirectional AKG 1000S microphones. The distance between the microphones was 8 cm.

In the first spatial configuration there were two masking sources: the concurrent speech (azimuth angle -45° clockwise) and music (azimuth angle 90°). The speech that was used as a masker signal was uttered while reading a popular book by a professional male lector. This signal was processed in order to remove pauses. Each logatom was mixed with different parts of concurrent speech. In the second scenario there was a third masking source generating white noise (azimuth angle -135°). The level of concurrent speech was 68 dB SPL, music – 65 dB SPL and white noise 65 dB SPL. The target signal level (TSL) was changed between 59 and 70 dB SPL.

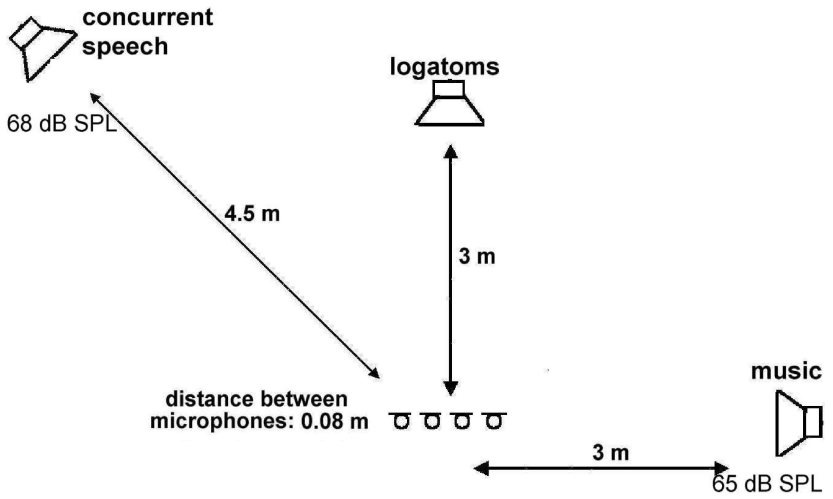

Fig. 1. Spatial configuration of sources with two maskers.

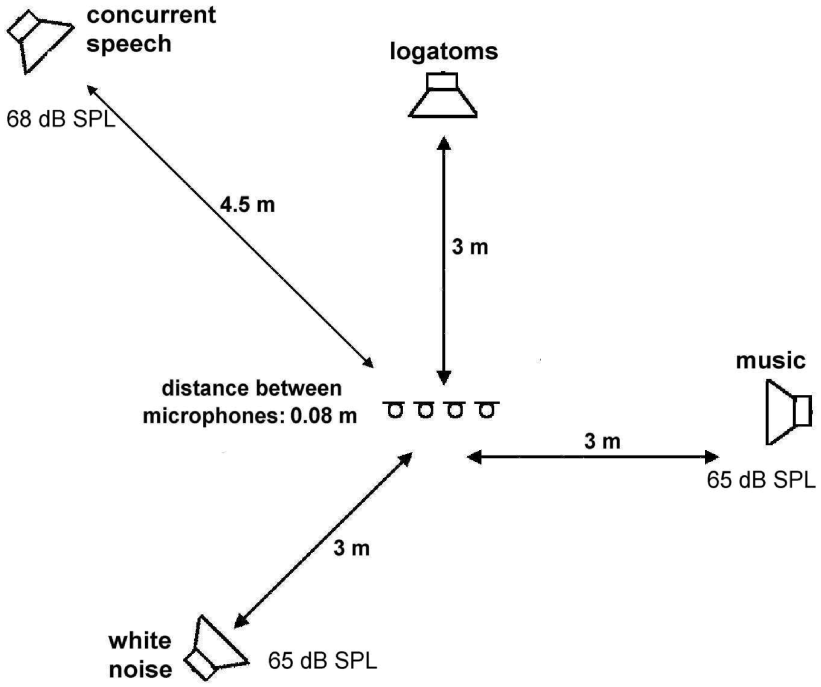

Fig. 2. Spatial configuration of sources with three maskers.

Target speech and masking signals were played using CD players except for white noise that was generated by a PC. The sounds were amplified using two amplifiers (Pioneer A-505R) while the target speech was amplified using a Sony STR-DE475 amplifier. Then, the signals were delivered to three-way loudspeakers Tonsil ZG-60 (one for each signal) placed in the anechoic chamber.

All signals were recorded through the array of four microphones AKG 1000S and preamplified using a Soundcraft M8 console. Next, the signals were fed to the digital recorder (Fostex D824) and stored on the hard drive at a sampling rate of 44100 Hz.

3.2. Signal transformations

The signals recorded were processed by the *convbss* algorithm by PARRA and SPENCE [1] implemented in MATLAB by HARMELING [28]. The length of the FFT was set to 512 samples. The length of the filter in the time domain was 128 samples. Five matrices were diagonalized to calculate the separation filters. The maximum number of algorithmic iterations was set to 1000 and the learning rate was 1.0. These parameters were chosen on the basis of the results obtained by the authors of the algorithm and results of an earlier study by KOCIŃSKI [29]. Thus they seem to be optimal as they cover the delays between consecutive microphones. In terms of beamforming (the equations in both methods, i.e. beamforming and BSS are the same, but the way of separation

is different) small values of the filter length might be insufficient to create an optimal (sharp) spatial filter thus the separation is poor. In contrast very high filter length increases computational time and does not lead to better separation.

Next, the denoising algorithms were used to additionally enhance the best signal selected after BSS in the separation stage. It must be emphasized, that at the output of the BSS procedure four signals were obtained as four microphones were used. Each of these signals was the best estimate of the signals produced by separate sources. The best target signal selected for the denoising procedure was determined based on subjective impression of the experimenter. It is worth mentioning that BSS markedly increased the SIR, thus, it was easier for the denoising algorithm to subtract the noise. Four different denosing algorithms were applied to the best target signal after BSS, separately. The time window in each algorithm was chosen as suggested by the author of the implementations to 25 ms with 40% overlapping.

3.3. Logatom articulation index evaluation

The enhanced speech was presented in a double-walled, acoustically isolated chamber to normally hearing subjects. The signals stored on the hard drive were fed to the TDT-RP2 (Tucker-Davis Technologies, System 3) processor and then amplified in the headphone buffer, TDT-HB7, to the overall level of 75 dB SPL at the tympanic membrane. Next, the signals were delivered to the Sennheiser HDA580 headphones and presented diotically to the subjects. All the recordings and presentations were carried out using MatLab 6.5 computing language (MathWorks Inc.).

To find out the advantages of using BSS combined with the denoising algorithms the logatom articulation index of unprocessed speech, speech processed by BSS only and that of the speech processed by a combination of BSS and one of following algorithms: spectral subtraction [33, 34], and statistical methods [35, 36] was measured.

4. Subjects

Three subjects aged 23–25 with audiologically normal hearing were asked to listen to the speech signals and write down all the understood logatoms on a special form. In all figures presented below the subjects are depicted as JK (the author), LS and AK. All subjects were instructed and took part in a short training session (about 2 hours) to be familiarized with the task.

5. Results

The data gathered in this experiment, i.e. the logatom articulation index as a function of the Target Source Level (TSL) of the target signal (or alternatively signal-to-interference ratio, SIR of the recorded signal), for all subjects and spatial configuration with two masking sources are depicted in Fig. 3. Analogous results for the case with

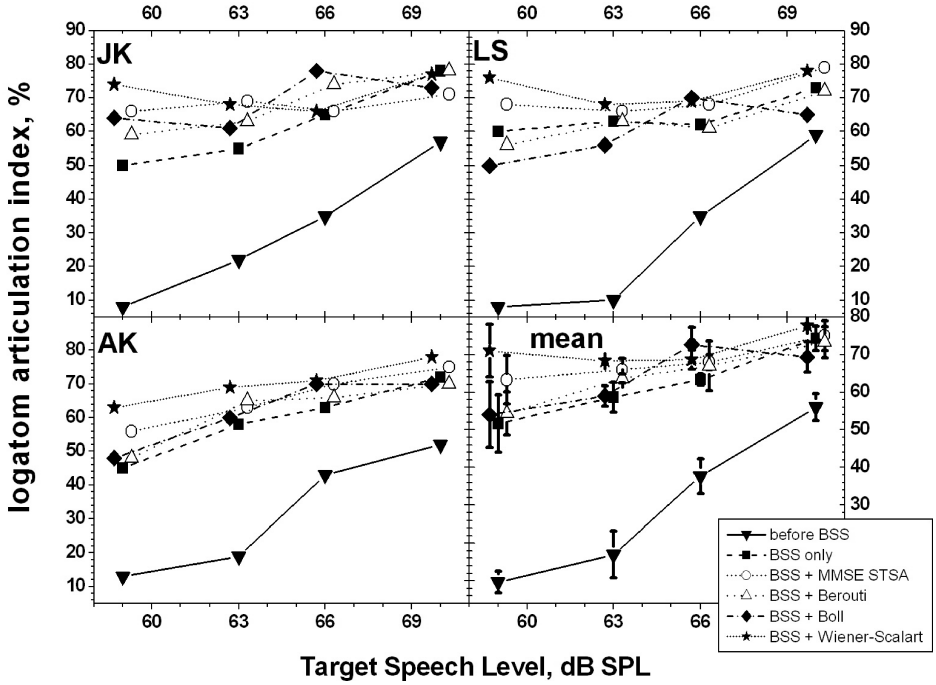

Fig. 3. Results for spatial configuration with two masking sources.

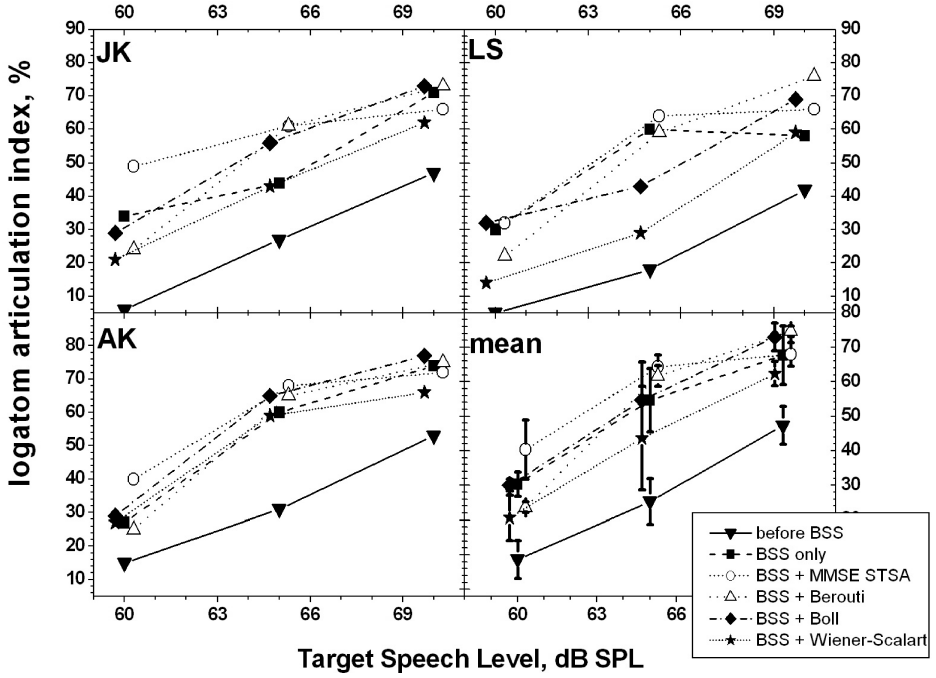

Fig. 4. Results for spatial configuration with three masking sources.

three masking sources are shown in Fig. 4. The filled triangles show the results for the signals before the BSS was applied while the filled squares show the results after BSS . The empty circles depict the results after BSS and denoising algorithm using the minimum mean-square error log-spectral amplitude estimator algorithm (*MMSE STSA*) [35, 40]. The filled asterisks show the results after BSS and a denoising algorithm based on the Wiener-Scalart idea [36, 41]. Filled diamonds and empty triangles indicate the results after BSS and the BOLL'S [33] or BEROUTI'S *et al.* [34] spectral subtraction algorithms respectively. For better visual clarity the symbols for each SNR were slightly shifted in the figures.

5.1. Spatial configuration with two masking sources

For the spatial configuration with two masking sources the use of BSS increased the logatom articulation index by about 40 percentage points for the lowest TSL and by about 20 percentage points for the highest TSL. Therefore, one may say that the efficiency of BSS is inversely related to SIR.

The use of a single microphone denoising algorithm, in general, additionally increased the logatom articulation index. The best performance was obtained for the Wiener-Scalart algorithm. Similarly to the BSS procedure, the highest increase in performance can be noticed for the lowest TSLs used in the experiment. The efficiency of this algorithm however, was much poorer at higher TSL. For TSL = 60 dB there was no difference in logatom articulation index for signal after BSS and after BSS combined with the Wiener-Scalart algorithm. Good performance can be also noticed for the MMSE STSA algorithm (10 percentage points increase in the logatom articulation index). It is worth adding that the above mentioned algorithms are similar in a sense, as they are based on the assumption that the signals' frequency components have Gaussian distributions.

The data gathered in this experiment were analysed using a within-subject analysis of variance (ANOVA) and are collected in Table 1. All factors used in the experiment (target speech level, speech enhancement method) were proved to be statistically significant, $F(3,\ 48) = 66.63$, ($p < 0.05$) for the target speech level and $F(3,\ 48) = 127.6$, ($p < 0.05$) for the method respectively. The F denotes Fisher's distribution critical value while values in semicolons correspond to degrees of freedom for factor and error respectively. This analysis of variance fully confirmed the above presented conclusion as the BSS procedure markedly enhanced logatom articulation index. The interaction

Table 1. Results of the analyses of variance for logatom articulation indexes obtained before and after the BSS for the spatial configuration 1 (where F denotes Fisher's F-distribution).

Factor	Before BSS	BSS
Target speech level	$F(3,\ 48) = 66.63$, ($p < 0.05$)	$F(4,\ 40) = 29.36$, ($p < 0.001$)
Speech enhancement method	$F(3,\ 48) = 127.6$, ($p < 0.05$)	$F(3,\ 40) = 7.95$, ($p < 0.001$)
Interaction	$F(15,\ 48) = 6.78$, ($p < 0.05$)	$F(12,\ 40) = 2.22$, ($p < 0.03$)

between these two factors was also significant $F(15,\ 48) = 6.78$, ($p < 0.05$) and it proves that the efficiency of BSS strongly depends on SIR (or alternatively the TSL).

As can be seen from Fig. 3 denoising algorithms, on average, improved the logatom articulation index. However, on the basis of this figure it is difficult to asses the efficiency of these algorithms as well as differences between them. Therefore, a separate analysis of variance was conducted for the data on the logatom articulation index after the above mentioned algorithms were used. The logatom articulation index data obtained after BSS had been applied were also included in this analysis. As expected, the effect of the TSL was statistically significant $F(4,\ 40) = 29.36$, $p < 0.001$. Also the influence of denoising algorithm was highly significant $F(3,\ 40) = 7.95$ ($p < 0.001$), which proves that the efficiency of these algorithms is quite different. The intelligibilities obtained for the lowest TSL was the highest (76%) for the Wiener-Scalart algorithm and the lowest (52%) for Boll's algorithm which was very close to the logatom articulation index after BSS only. The interaction between the algorithm and the TSL was also statistically significant $F(12,\ 40) = 2.22$, ($p < 0.03$). This confirms that the efficiency of these algorithms strongly depends on SIR (or TSL) of the input signals, being the highest for the lowest TSL. As can be seen from Fig. 3. the use of additional denoising algorithm brought about an increase in the logatom articulation index. Indeed, if one can ignore the logatom articulation index data obtained for the raw recorded signals (before any speech enhancement method was applied) the worst performance can be noticed for BSS.

This happens especially for the lowest TSLs. For the highest TSLs the differences between data collected for different algorithms (including BSS) are quite small and it is difficult to asses the strict relation between them.

The data presented in Fig. 3 suggests that the use of additional denoising algorithms may be quite useful especially in highly noisy conditions. However, the use of these algorithms is SIR dependent.

The analyses of variance applied to each of the denoising algorithms and to the logatom articulation index after the BSS, were performed to assess the significance of the efficiency obtained for of each of the considered denoising algorithms. The results of these analyses are shown in Table 2. The F-values were calculated for one factor, i.e. additional denoising method. The statistically significant improvement was obtained for the MMSE STSA and Wiener-Scalart algorithms. In the case of spectral subtraction algorithms by Boll and Berouti, statistically significant changes were not obtained.

Table 2. Results of the analyses of variance that tested logatom articulation index improvement after use of single-microphone denoising methods for the spatial configuration 1.

Algorithm	Fisher's F-distribution
Ephraim	$F(1,\ 16) = 11.27$, ($p < 0.05$)
Berouti	$F(1,\ 16) = 1.77$, ($p = 0.2$)
Boll	$F(1,\ 16) = 0.71$, ($p = 0.41$)
Scalart	$F(1,\ 16) = 29.7$, ($p < 0.05$)

5.2. Spatial configuration with three masking sources

In the second spatial configuration (see Fig. 2) there was an additional masking source generating a white noise. Therefore, one can say that the stationary signal was added to the masking mixture. The logatom articulation index data gathered in this case for three subjects (including averages across subjects) are presented in Fig. 4. As can be seen from this figure the use of the BSS algorithm, in general, brought about a logatom articulation index improvement. The smallest improvement was 21 percentage points (for the lowest and the highest TSL) and the highest one was about 28 percentage points (for medium TSL). The logatom articulation index improvement is much smaller then that observed in the case of the first spatial configuration (see Sec. 5.1). It is also less dependent on TSL in comparison to the two masking sources case. The increase in logatom articulation index in most of the observed cases is a monotonic function of the TSL, like in the first spatial configuration. However, the logatom articulation index improvement after BSS is much smaller then in the previous case. The use of an additional denoising algorithm after BSS brought about, in general, further increase in the logatom articulation index. However, the efficiency of the algorithms applied was different. The best logatom articulation index improvement was observed for the MMSE STSA algorithm (by about 20 percentage points for the lowest TSL). The worst performance was observed for the Wiener-Scalart algorithm, for which the logatom articulation index was decreased in comparison with BSS only (by about 10 percentage points for TSL of 65 dB). However, the logatom articulation index improvement caused by the use of any of the denoising algorithms did not depend on TSL.

Some deterioration in speech enhancement efficiency can be also noticed for Berouti's *et al.* algorithm but only for the lowest TSL while for higher TSLs the use of this algorithm is beneficial (increase in the logatom articulation index by about 7 percentage points). The use of Boll's spectral subtraction method did not give a noticeable change in logatom articulation index.

The data collected in this part of the experiment was subjected to several separate analyses of variance. In the first part of this analysis TSL and the denoising method were tested. This analysis showed that the TSL was a statistically significant factor $F(2,\ 36) = 163.28$, ($p < 0.001$). For speech enhancement method factor ANOVA showed also statistical significance $F(5,\ 36) = 24.71$, ($p < 0.001$). However, an interaction between the speech enhancement method and TSL was not statistically significant $F(10,\ 36) = 1.67$, ($p = 0.13$), contrary to the earlier considered case with two masking sources. The results of this analysis fully confirmed that TSL and the denoising method markedly increased the logatom articulation index. These results also confirmed that the logatom articulation index improvement was nearly independent of TSL as well.

It can be seen from the data presented in Fig. 4 that the majority of the denoising algorithms brought about a logatom articulation index improvement. However, on the

basis of the data presented, it is very difficult to asses the efficiency of the algorithms applied. Therefore, several analyses of variance were performed on the logatom articulation index results obtained after BSS and after BSS postprocessed by with each of the single-microphone denoising algorithms with the factors being TSL and the denoising method. The results of ANOVA test of significance of the denoising method factor (excluding interactions and significance of TSL factor) are presented in Table 3. As can be seen in the case of one denoising method (namely MMSE STSA) the logatom articulation index was significantly different relative to that obtained after applying BSS only.

Table 3. Results of the analyses of variance that tested logatom articulation index improvement after use of single-microphone denoising methods for the spatial configuration 2.

Algorithm	Fisher's F-distribution
Ephraim	$F(1,\ 12) = 4.5$, $(p < 0.05)$
Berouti	$F(1,\ 12) = 0.88$, $(p = 0.36)$
Boll	$F(1,\ 12) = 0.24$, $(p = 0.63)$
Scalart	$F(1,\ 12) = 4.51$, $(p < 0.06)$

The data collected in this part of the study (three disturbing sources) showed much less improvement in the logatom articulation index than in the first spatial configuration (2 interfering noises). The data showed that although BSS can separate out the number of sources equal to the number of microphones, it is more effective when the number of separated sources is lower then the number of sensors.

5.3. Discusion

The results of the experiments have shown that the BSS algorithm is highly effective in subjective logatom articulation index improvement. The improvement reaches even about fifty percentage points in the logatom articulation index. However, the BSS efficiency strongly depends on several parameters. First of all, as suggested in this paper, when the number of sensors (microphones) exceeds the number of sources the efficiency is much higher. In the case when four sensors were used the efficiency of BSS was higher for three sound sources than for four sound sources. The difference in the logatom articulation index between two disturbing sources and three disturbing sources could be caused by different numbers of separation filter coefficients. In the scenario with two masking sources there are 16 filters to estimate while in the case of three interferers there are 12 filters only. Moreover, the permutation problem is easier to solve for a smaller number of sources. The results reveal also one very important feature of the BSS procedure. Namely, when four sensors were used with three sound sources a statistically significant improvement in the effectiveness of BSS was observed for the lowest TSLs. However, when four sound sources were used the effectiveness did not seem to depend on TSL.

It is also worth adding that in the case of three sound sources the two disturbances were concurrent speech signals. The third disturbing signal (4 sources case) was white noise. Significantly smaller effectiveness of BSS in the latter case is probably connected not only with the next disturbing sound source but also with the nature of white noise. White noise is a stationary signal while concurrent speech signals are nonstationary ones. Speech signals are characterized by marked changes in their amplitude envelope whose spectral maximum coincides with 4 Hz. The spectrum of the amplitude envelope of white noise is quite flat and does not show a prominent maximum. As the rate of the amplitude fluctuation of speech signal is much smaller and the fluctuations are much greater one can expect that the subjects' performance can be markedly influenced by the so-called deep-listening mechanism. When white noise is added to the disturbing sounds the deep listening does not occur nearly at all yielding a much smaller logatom articulation index improvement (see Fig. 5).

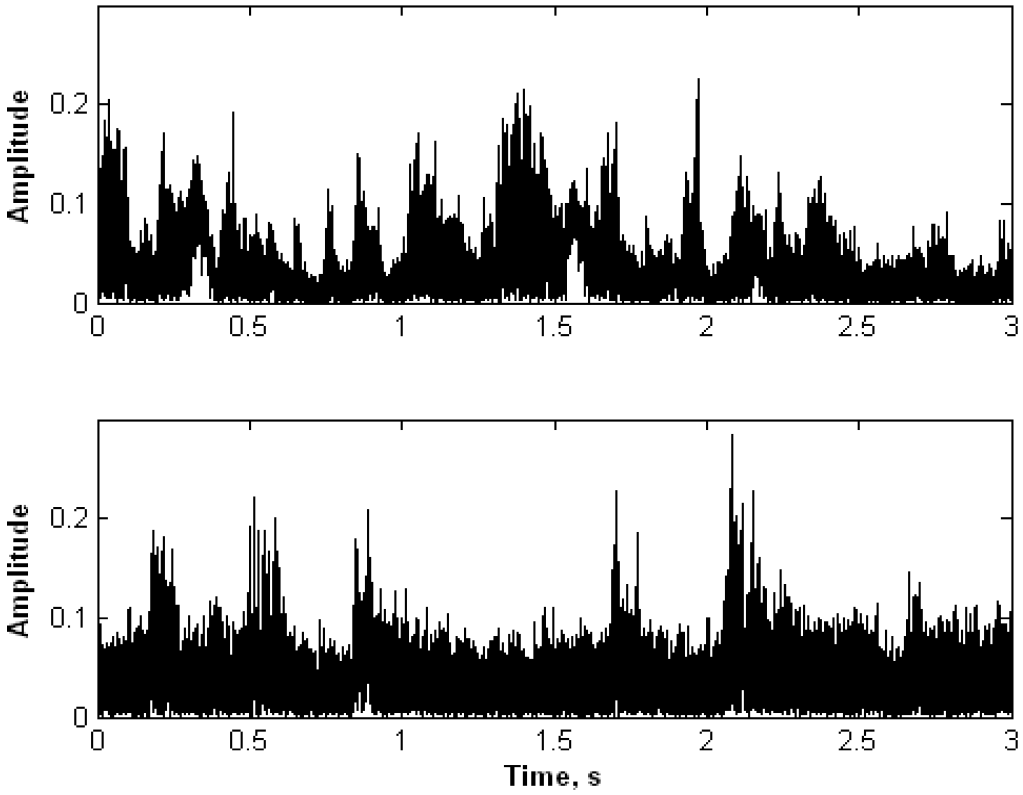

Fig. 5. Comparison of the time envelopes of the signals after BSS for the lowest SNR and for two (upper panel) and three (lower panel) interferences.

When logatoms are presented in silence, the highest logatom articulation index does not reach 100%, and it is usually close to 95% [37]. In our experiment the logatoms were presented against a background of disturbing sounds and the highest logatom ar-

ticulation index for the lowest TSL reached about 75%. Thus, one can say that the BSS algorithm failed to fully restore the speech signal. However, it is necessary to keep in mind, that BSS is a statistical procedure that gives estimates of signals instead of original source signals. The fifty percentage points in logatom articulation index improvement observed is probably the highest one that can be obtained for speech signals presented against background disturbances. This efficiency seems to be very promising for future development of hearing aids.

The results have shown that the use of the single microphone denoising methods after BSS was applied can additionally increase the logatom articulation index. However, this additional increment was much smaller than that obtained by means of BSS only. Moreover, the increment strongly depended on the type of denoising algorithm and on TSL. Similarly to the BSS procedure, the highest efficiency of the single sensor denoising algorithm was observed for the lowest TSLs. The best results were obtained for the MMSE STSA algorithm for both three and four sound sources. The high efficiency of this algorithm can be connected with an *a priori* SNR estimation or Gaussian distribution of speech model and log-MMSE criterion.

The smallest logatom articulation index improvement (sometimes even logatom articulation index deterioration) was observed for classical spectral subtraction algorithms. In real speech, fricatives are usually low energy and broadband sounds. Therefore, spectral subtraction methods may treat these sounds as noise and, in consequence, subtract them. Missing fricatives can cause logatom articulation index degradation. Even though some single-sensor denoising procedures did not give the logatom articulation index improvement, it is worth including one of the single denoising methods in the final speech enhancement stage (after the multimicrophone denoising methods).

As follows from the results, BSS can give a significant logatom articulation index increment. In the spatial configuration with two masking sources this increment depends on SIR (or TSL). In the case with 2 masking sources BSS was the most efficient for the lowest TSL. However, in the spatial configuration with three masking sources the benefits of using BSS were TSL independent. The difference in the logatom articulation index between the results obtained for two disturbing sources and three disturbing sources could be attributed to different numbers of separation filters coefficients: 16 filters in two disturbing sources and 12 in the case of three interferers. Moreover, the permutation problem is easier to solve for a smaller number of sources.

There were some cases in which denoising algorithms decreased logatom articulation index. Figure 6 illustrates the influence of the Wiener-Scalart algorithm on speech signal. The spectrograms are obtained from speech recorded in spatial configuration with three masking sources. In the upper spectrogram, the case where signal was processed only by the BSS algorithm is presented. The lower spectrogram depicts speech after BSS and the use of the Wiener-Scalart algorithm. It can be noted that the denoising algorithm distorts the signal (see circles in the Fig. 6). These distortions (inhibition of consonants) can cause logatom articulation index decrement.

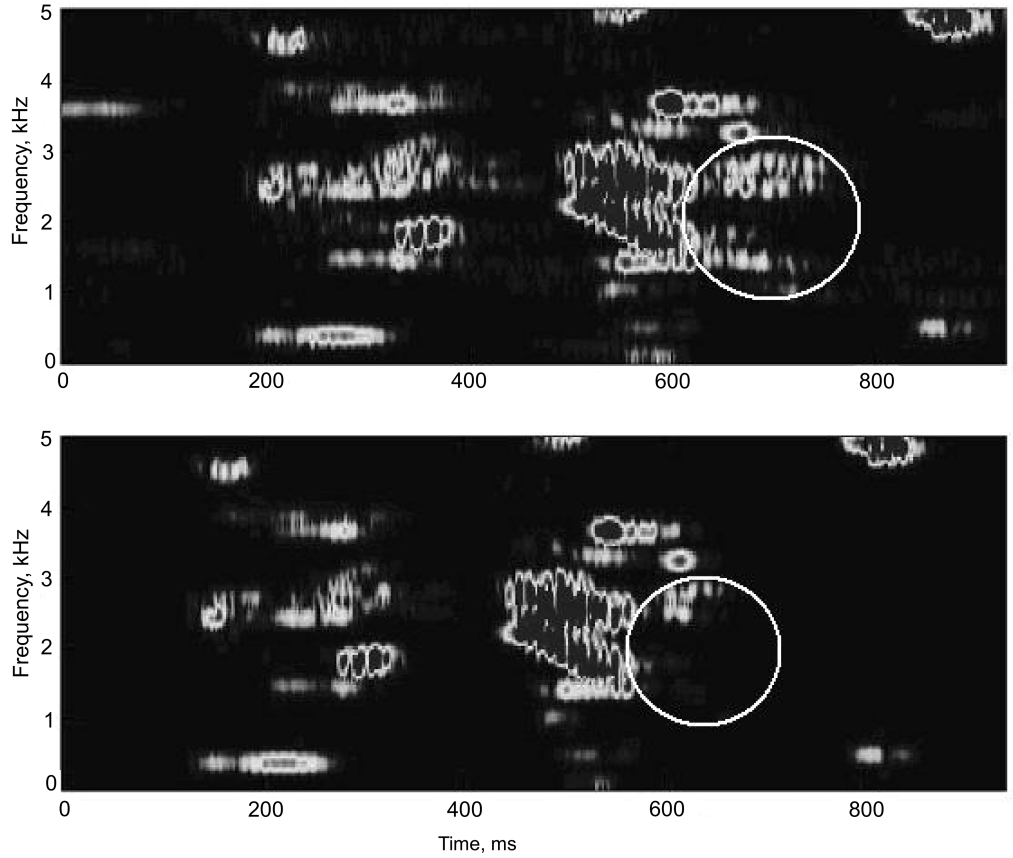

Fig. 6. Spectrograms comparison: upper spectrogram is derived from speech recorded in spatial configuration with three masking sources and processed with BSS algorithm. Lower spectrogram depicts speech after BSS and the use of the Wiener-Scalart algorithm. Degraded (by the denoising algorithm) part of the signal is marked by white circles.

6. Conclusions

The above discussed results of our experiments lead to the following conclusions:

- The use of the BSS procedure gives a marked improvement in the subjective logatom articulation index, reaching 50 percentage points.
- The efficiency of BSS seems to be much higher when the number of sound sources is lower than the number of sensors.
- BSS is more effective for lower TSLs or alternatively for lower SIRs.
- Some of the single sensor denoising algorithms applied at the postprocessing stage (i. e. after BSS) give an additional improvement in the logatom articulation index which is much smaller than that obtained with the use of BSS only.
- Denoising algorithms based on the spectral subtraction did not improve the logatom articulation index, even though the optimal parameters were used.

- Although the BSS is a relatively new idea, many different approaches to solve the problem of separating mixed signals have been introduced. Moreover, the problem of hearing loss becomes more and more commonplace. Thus, research in this area should be continued using new algorithms and solutions before they will be used in hearing aids.

Acknowledgments

This work was supported by the Ministry of Science and Higher Education, grant no. N517 028 32/4327 and partially by PR6, Hearcom. The authors would like to thank two anonymous reviewers for useful comments and remarks on the earlier version of this manuscript.

References

[1] PARRA L., SPENCE C., *Convolutive blind source separation of non-stationary sources. US Patent US6167417*, IEEE Transactions on Speech and Audio Processing., **8**, 3, 320–327 (2000).

[2] PREVES D.A., *Hearing aids and listening in noise*, Seminars in Hearing, **21**, 2, 103–122 (2000).

[3] DIGIOVONNI J.J., NELSON P.B., SCHLAUCH R.S., *A psychophysical evaluation of spectral enhancement*, Journal of Speech, Language, and Hearing Research, **48**, 5, 1121–1135 (2005).

[4] EZEKIEL S., OBLITEY W., TRIMBLE R., *Hearing aid speech enhancement: A multiresolution analysis approach*, [in:] IASTED International Conference on Internet and Multimedia Systems and Applications, pp. 533–537, EuroIMSA, 2005.

[5] CHUNG K., ZENG F.-G., ACKER K.N., *Effects of directional microphone and adaptive multichannel noise reduction algorithm on cochlear implant performance*, Journal of the Acoustical Society of America, **120**, 4, 2216–2227 (2006).

[6] GAO J., ZHANG H., HU G., *Real-time implementation of an efficient speech enhancement algorithm for digital hearing aids*, Tsinghua Science and Technology, **11**, 4, 475–480 (2006).

[7] FRANCK B.A.M., BOYMANS M., DRESCHLER W.A., *Interactive fitting of multiple algorithms implemented in the same digital hearing aid*, International Journal of Audiology, **46**, 7, 388–397 (2007).

[8] HOEGE H., *Basic parameters in speech processing. The need for evaluation*, Archives of Acoustics, **32**, 1, 67–74 (2007).

[9] ROCH M.A., HURTIG R.R., HUANG T., LIU J., ARTEAGA S.M., *Foreground auditory scene analysis for hearing aids*, Pattern Recognition Letters, **28**, 11, 1351–1359 (2007).

[10] WON J.H., SCHIMMEL S.M., DRENNAN W.R., SOUZA P.E., ATLAS L., Rubinstein J.T., *Improving performance in noise for hearing aids and cochlear implants using coherent modulation filtering*, Hearing Research, **239**, 1–2, 1–11 (2008).

[11] O'SHAUGHNESSY D., *Speech communications. Human and machine. Second edition*, Piscataway: IEEE Press (2000).

[12] HOJAN E., FASTL H., MALENDA J., HOJAN–JEZIERSKA D., *Investigations into speech intelligibility in the presence of different masking noises for hearing aids with variable attack and release times*, Archives of Acoustics, **30**, 2, 159–171 (2005).

[13] KOCIŃSKI J., SĘK A.P., *Speech intelligibility in various spatial configurations of background noise*, Archives of Acoustics, **30**, 2, 173–191 (2005).

[14] SMARAGDIS P., *Information theoretic approaches to source separation*, [in:] *MAS Department*, MSc thesis, Massachusetts Institute of Technology, Massachusetts 1997.

[15] SMARAGDIS P., *Efficient blind separation of convolved sound mixtures*, [in:] EEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 19–22, New Paltz NY 1997.

[16] HYVÄRINEN A., KARHUNEN J., ERKKI O., *Independent component analysis*, John Wiley & Sons, Inc, New York 2001.

[17] CICHOCKI A., AMARI S., *Adaptive blind signal and image processing learning algorithms and applications*, John Wiley & Sons, Ltd, Chichester / New York / Weinheim / Brisbane / Singapore / Toronto, 2003.

[18] CHOI S., CICHOCKI A., PARK H.-M., LEE S.Y., *Blind source separation and independent component analysis: a review*, Neural Information Processing – Letters and Reviews, **6**, 1, 1–57 (2005).

[19] BELOUCHRANI A., AMIN M.G., *A new approach for blind source separation using time-frequency distribution*, Proceedings SPIE, **2846**, 193–203 (1996).

[20] MOLGEDEY L., SCHUSTER H.G., *Separation of a mixture of independent signals using time delayed correlatons*, Physical Review Letters, **72**, 23, 3634–3637 (1994).

[21] BELOUCHRANI A., ABED–MERAIM K., *A blind source separation technique using second-order statistics*, IEEE Transactions on Signal Processing, **45**, 2, 434–444 (1997).

[22] CICHOCKI A., BELOUCHRANI A., *Sources separation of temporally correlated sources from noisy data using bank of band-pass filters*, Third International Conference on Independent Component Analysis and Signal Separation (ICA-2001), pp. 173–178, San Diego, USA 2001.

[23] CHOI S., CICHOCKI A., BELOUCHRANI A., *Second order nonstationary source separation*, Journal of VLSI Signal Processing, **32**, 1–2, 93–104 (2002).

[24] CHOI S., CICHOCKI A., ZHANG L.L., AMARI S., *Approximate maximum likelihood source separation using the natural gradient*, IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, **86**, 1, 198–205 (2003).

[25] MATSUOKA K., OHYA M., KAWAMOTO M., *A neural net for blind separation of nonstationary signal*, Neural Networks, **8**, 3, 411–420 (1995).

[26] PHAM D.-T., SERVIERE C., BOUMARAF H., *Blind separation of convolutive audio mixtures using nonstationarity*, [in:] ICA 2003, Nara, Japan 2003.

[27] MAKINO S., SAWADA H., MUKAI R., ARAKI S., *Blind source separation of convolutive mixtures of speech in frequency domain*, IEICE Trans. Fundamentals, **E88**, 7, 1640–1655 (2005).

[28] HARMELING S., *convbss*, FRAUNHOFER FIRST Berlin, Berlin 2001.

[29] KOCIŃSKI J., *Influence of blind source separation on speech intelligibility*, Archives of Acoustics, **30**, 4 (Supplement), 149–152 (2005).

[30] LIBISZEWSKI P., KOCIŃSKI J., *Efficiency of blind source separation in a real room*, Archives of Acoustics, **32**, 4 (Supplement), 337–342 (2007).

[31] PARRA L., SPENCE C., *On-line blind source separation of non-stationary signals*, Journal of VLSI Signal Processing, **26**, 1–2, 39–46 (2000).

[32] ASANO F., GOTO M., ITOU K., ASOH H., *Real-time sound source localization and separation system and its application to automatic speech recognition*, Eurospeech, 1013–1016, 2001.

[33] BOLL S.F., *Suppression of acoustic noise in speech using spectral substraction*, IEEE Transactions on Acoustics, Speech and Signal Processing, **ASSP-27**, 2, 113–120 (1979).

[34] BEROUTI M., SCHWARTZ R., MAKHCUL J., *Enhancement of speech corrupted by acoustic noise*, IEEE International Conference on Acoustics, Speech and Signal Processing, **4**, 208–211 (1979).

[35] EPHRAIM E., MALAH D., *Speech enhancement using a minimum mean-square error log-spectral amplitude estimator*, IEEE Transactions on Speech and Audio Processing, **ASSP-33**, 2, 443–445 (1985).

[36] SCALART P., FILHO J.V., *Speech enhancement based on a priori signal to noise estimation*, IEEE International Conference on Acoustics, Speech, and Signal Processing, **1**, 629–632 (1996).

[37] BRACHMAŃSKI S., STARONIEWICZ P., *Phonetic structure of a test material used in subjective measurements of speech quality* [in Polish], Speech and Language Technology, Poznań, 71–80 (1999).

[38] BRACHMAŃSKI S., *Effect of additive interference on speech transmission*, Archives of Acoustics, **27**, 2, 95–108 (2002).

[39] BRACHMAŃSKI S., *Estimation of logatom intelligibility with the STI method for Polish speech transmitted via communication channels*, Archives of Acoustics, **29**, 4, 555–562 (2004).

[40] ZAVAREHEI E., *MMSESTSA85.m*, http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=7655&objectType=FILE.

[41] ZAVAREHEI E., *WienerScalart96.m*, 2005, http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=7673&objectType=FILE.