Blind Speech Separation in Time-Domain Using Block-Toeplitz Structure of Reconstructed Signal Matrices

Zbyněk Koldovský[1,2]*, Jiří Málek*[1]*, and Petr Tichavský*[1,2]

[1]Faculty of Mechatronics, Informatics and Interdisciplinary Studies,
Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic
[2]Institute of Information Theory and Automation,
P.O.Box 18, 182 08 Prague 8, Czech Republic

zbynek.koldovsky@tul.cz, jiri.malek@tul.cz, tichavsk@utia.cas.cz

Abstract

Methods for Blind Source Separation (BSS) aim at recovering signals from their mixture without prior knowledge about the signals and the mixing system. Among others, they provide tools for enhancing speech signals when they are disturbed by unknown noise or other interfering signals in the mixture. This paper considers a recent time-domain BSS method that is based on a complete decomposition of a signal subspace into components that should be independent. The components are used to reconstruct images of original signals using an ad hoc weighting, which influences the final performance of the method markedly. We propose a novel weighting scheme that utilizes block-Toeplitz structure of signal matrices and relies thus on an established property. We provide experiments with blind speech separation and speech recognition that prove the better performance of the modified BSS method.

Index Terms: blind speech separation, independent component analysis, time-domain method, weighting

1. Introduction

Blind Source Separation (BSS) of convolutive mixtures of original sources is important for speech enhancement and similar applications [1]. We consider the situation when several acoustical sources sound simultaneously in a room and are recorded by an array of microphones. Let there be m microphones. Natural mixing conditions are underpinned by the convolutive model, so a signal observed by the ith microphone is equal to

$$x_i(n) = \sum_{k=1}^{d} \sum_{\tau=0}^{M_{ik}-1} h_{ik}(\tau) s_k(n-\tau), \quad i = 1, \ldots, m, \tag{1}$$

where $s_1(n), \ldots, s_d(n)$ are unknown original signals, and h_{ik}'s are source-microphone room impulse responses, each of length M_{ik}. The goal is to find separating MIMO filters whose outputs give the individual microphone responses (images) of sources

$$s_k^i(n) = \sum_{\tau=0}^{M_{ik}-1} h_{ik}(\tau) s_k(n-\tau), \tag{2}$$

i.e., the separated signals.

Many BSS methods assume the independence of original signals. Independent Component Analysis (ICA) is then used as the hearth of separation. Since ICA algorithms primarily work with instantaneous mixing model, the convolutive model (1) must be transformed. A popular way is to transform signals into the frequency-domain, where the convolutive mixing model changes to a set of instantaneous models, one for each frequency. The main problem then consists in solving the permutation problem since the order of separated frequency components is random [2].

Time-domain approaches, addressed by this paper, usually search independent signals in a signal subspace spanned by rows of matrix

$$\mathbf{X} = \begin{bmatrix} x_1(N_1) & \ldots & \ldots & x_1(N_2) \\ x_1(N_1-1) & \ldots & \ldots & x_1(N_2-1) \\ \vdots & \vdots & \vdots & \vdots \\ x_1(N_1-L+1) & \ldots & \ldots & x_1(N_2-L+1) \\ x_2(N_1) & \ldots & \ldots & x_2(N_2) \\ x_2(N_1-1) & \ldots & \ldots & x_2(N_2-1) \\ \vdots & \vdots & \vdots & \vdots \\ x_m(N_1-L+1) & \ldots & \ldots & x_m(N_2-L+1) \end{bmatrix}, \tag{3}$$

where N is the number of available samples, and $1 \leq N_1 < N_2 \leq N$. The dimension of the subspace (the number of rows of $\mathbf{X}$) is equal to mL. L is a free integer parameter corresponding to the length of separating MIMO filters.

Note that $\mathbf{X}$ can be written as $\mathbf{X} = \mathbf{S}_1 + \cdots + \mathbf{S}_d$ where each $\mathbf{S}_k$ has the same block-Toeplitz structure (where each block consists of L rows) as $\mathbf{X}$ but contains the responses (2) of the kth source only, because $x_i(n) = \sum_{k=1}^{d} s_k^i(n)$. The goal of BSS can be formulated as to find transforms $\mathbf{H}_k$ such that $\mathbf{H}_k\mathbf{X} \approx \mathbf{S}_{\pi(k)}$, where $\pi(\cdot)$ denotes an unknown permutation of separated signals[1]. Final signal estimates are then derived from $\mathbf{H}_k\mathbf{X}$.

1.1. T-ABCD method

To find $\mathbf{H}_k$, it was recently shown in [3] that it is advantageous to apply ICA so that the whole $\mathbf{X}$ is first decomposed into mL independent components (ICs) $\mathbf{C} = \mathbf{W}\mathbf{X}$, where $\mathbf{W}$ is the ICA decomposing square matrix. In fact, the components cannot be all independent, but can be organized into d groups forming independent subspaces corresponding to individual sources.

The grouping can be done by clustering ICs according to their similarity. The founded clusters (groups) are represented

[0]This work was supported by Ministry of Education, Youth and Sports of the Czech Republic through the project 1M0572 and by Grant Agency of the Czech Republic through the project P103/11/1947.

[1] The permutation uncertainty must be taken into account due to the inherent indeterminacy of ICA. For ease of exposition, we can assume that $\pi(k) = k$ for all k.

by diagonal matrices $\mathbf{\Lambda}_k$, $k = 1, \dots, d$, whose diagonal elements determine how much the components belong to the kth cluster. The estimate of $\mathbf{S}_k$ is then given by

$$\widehat{\mathbf{S}}_k = \mathbf{W}^{-1}\mathbf{\Lambda}_k\mathbf{C} = \mathbf{W}^{-1}\mathbf{\Lambda}_k\mathbf{W}\mathbf{X}, \tag{4}$$

which corresponds to the choice $\mathbf{H}_k = \mathbf{W}^{-1}\mathbf{\Lambda}_k\mathbf{W}$. This approach was in [3] named T-ABCD and will be described in details in Section 3. Similar idea was used in, e.g., [1, 4].

The diagonal elements of $\mathbf{\Lambda}_k$, from here referred to as weights, can be chosen in several ad hoc ways; see e.g. [5]. However, experimental comparisons in [3] with weights that are optimal in mean square error sense (computed when the target matrices $\mathbf{S}_k$ are known) have shown significant difference in performance. In other words, the separation performance of T-ABCD might be significantly improved through the weights.

In this paper, we propose weights whose computation is not ad hoc but comes from the Block-Toeplitz structure of the matrices $\mathbf{S}_k$. A criterion of "block-toeplitzity", which is a quadratic form subject to the weights, is derived. It is optimized through finding eigenvectors of the quadratic form under constraints respecting groups (clusters) of components. Experiments with real-world recordings demonstrate a better performance of the resulting weights.

The computation of weights is derived in the following section. The separation method using the weights is summarized by Section 3. In Section 4, several experiments of blind speech separation are presented. Section 5 concludes the paper.

2. Weights Adjustment

Let the diagonal elements of $\mathbf{\Lambda}_k$ be denoted by $\lambda_1^k, \dots, \lambda_{mL}^k$. Let K_k be the set of indices of ICs that were assigned to the kth cluster. A binary weighting is defined as

$$\lambda_\ell^k = \begin{cases} 1 & \text{for } \ell \in K_k \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

This weighting is consistent in the sense that if all ICs contain no residual interference (each component contains one source only) and the clustering is correct then $\mathbf{H}_k\mathbf{X}$ must be exactly equal to $\mathbf{S}_k$ since the original signals are independent.

Since in practice the residual interference remains present in ICs, it is meaningful to select weights that reflect it. For instance, in [5] the weights are selected according to a fuzzy clustering, which allows making a trade-off between the achieved signal-to-interference and signal-to-distortion ratios of separated signals.

2.1. Block-Toeplitzity Criterion

Now we propose a way how to define weights respecting the block-Toeplitz structure of $\mathbf{S}_k$. We define a criterion that measures this property of $\widehat{\mathbf{S}}_k = \mathbf{H}_k\mathbf{X}$ as

$$\mathcal{G}(\widehat{\mathbf{S}}_k) = \sum_{r=1}^{m}\sum_{p=1}^{L}\sum_{n=N_1}^{N_2-L+1}\Bigg[(\widehat{\mathbf{S}}_k)_{(r-1)L+p,n+p-1} - \frac{1}{L}\sum_{q=1}^{L}(\widehat{\mathbf{S}}_k)_{(r-1)L+q,n+q-1}\Bigg]^2. \tag{6}$$

Easily can be verified that if $\widehat{\mathbf{S}}_k$ has the block-Toeplitz structure with blocks of the length L, both terms in brackets are the same, and $\mathcal{G}(\widehat{\mathbf{S}}_k) = 0$. For example, if $\widehat{\mathbf{S}}_k = \mathbf{S}_k$, then $(\widehat{\mathbf{S}}_k)_{(r-1)L+p,n+p-1} = s_k^r(n)$, which is independent of p.

From (4), the ijth element of $\widehat{\mathbf{S}}_k$ can be expressed as

$$(\widehat{\mathbf{S}}_k)_{ij} = \sum_{\ell=1}^{mL}(\mathbf{W}^{-1})_{i,\ell}(\mathbf{\Lambda}_k)_{\ell,\ell}\mathbf{C}_{\ell,j} = \sum_{\ell=1}^{mL}\lambda_\ell^k(\mathbf{W}^{-1})_{i,\ell}\mathbf{C}_{\ell,j}.$$

It follows that $\mathcal{G}(\widehat{\mathbf{S}}_k)$ is purely quadratic subject to the weights $\lambda_1^k, \dots, \lambda_{mL}^k$ and can be represented by a symmetric matrix $\mathbf{G}$ such that

$$\mathcal{G}(\widehat{\mathbf{S}}_k) = \boldsymbol{\lambda}_k^T\mathbf{G}\boldsymbol{\lambda}_k, \tag{7}$$

where $\boldsymbol{\lambda}_k = [\lambda_1^k, \dots, \lambda_{mL}^k]^T$. The derivation of $\mathbf{G}$ is straightforward and is not shown here due to lack of space.

2.2. Finding Weights by Minimization of $\mathcal{G}(\widehat{\mathbf{S}}_k)$

Unfortunately, $\boldsymbol{\lambda}_k$ cannot be searched directly by minimizing $\mathcal{G}(\widehat{\mathbf{S}}_k)$. The reason is two-fold. First, a minimum is achieved for $\boldsymbol{\lambda}_k = [1, \dots, 1]^T$, because then $\widehat{\mathbf{S}}_k = \mathbf{X}$, which means no separation. This is because $\mathbf{X}$ has the exact block-Toeplitz structure by its definition (3). Second, there are other local minima of $\mathcal{G}(\widehat{\mathbf{S}}_k)$ that might correspond to unwanted solutions such as $\widehat{\mathbf{S}}_k \approx \mathbf{S}_{i_1} + \dots + \mathbf{S}_{i_s}$, $2 \le s \le d$, because all these matrices have the block-Toeplitz structure as well.

Therefore, we propose to constrain the minimization of $\mathbf{G}$ by optimizing selected elements of $\boldsymbol{\lambda}_k$ only, while the other elements are set equal to zero. In view of (5), it is natural to optimize those elements whose indices are in K_k, which correspond to ICs that were assigned to the kth cluster.

The solution of the constrained problem is obtained by the following steps.

1. Put $\mathbf{G}_k$ equal to the submatrix of $\mathbf{G}$ with only those rows and columns whose indices are in K_k.
2. Find the eigenvector $\mathbf{v}_k$ of $\mathbf{G}_k$ corresponding to the minimum eigenvalue of $\mathbf{G}_k$.
3. Put elements of $\boldsymbol{\lambda}_k$ in K_k equal to $\mathbf{v}_k$ (keep the same order) and the other put equal to zero.

2.3. Penalty term

We will show by experiments that signals separated by use of the weighting derived above often have good signal-to-interference ratio (SIR) but poor signal-to-distortion ratio (SDR). It happens when the eigenvector $\mathbf{v}_k$ is sparse, i.e., has many elements close to zero, which means that only few ICs from the kth cluster are used to reconstruct $\widehat{\mathbf{S}}_k$.

In order to balance this effect, we propose to add a penalty term to the criterion (6) that forces the elements of $\boldsymbol{\lambda}_k$ to be closer to the binary weighting (5). To this end, we define a term that penalizes differences between weights and their average value. The new criterion is therefore defined as

$$\widetilde{\mathcal{G}}(\widehat{\mathbf{S}}_k) = \boldsymbol{\lambda}_k^T\mathbf{G}\boldsymbol{\lambda}_k + \alpha\sum_{\ell=1}^{mL}\Big[(\boldsymbol{\lambda}_k)_\ell - \frac{1}{mL}\boldsymbol{\lambda}_k^T\mathbf{1}_{mL\times 1}\Big]^2 \tag{8}$$

$$= \boldsymbol{\lambda}_k^T\Big(\mathbf{G} + \alpha\big(\mathbf{I} - \frac{1}{mL}\mathbf{1}_{mL\times mL}\big)\Big)\boldsymbol{\lambda}_k, \tag{9}$$

where α is a free non-negative parameter, $\mathbf{I}$ stands for the identity matrix, and $\mathbf{1}$ is the matrix of ones of given dimensions. As follows from (9), the modified criterion remains purely quadratic. Therefore, the computation of weights is the same as described in the previous subsection but with the modified matrix

$$\widetilde{\mathbf{G}} \leftarrow \mathbf{G} + \alpha\big(\mathbf{I} - \frac{1}{mL}\mathbf{1}_{mL\times mL}\big). \tag{10}$$

2.4. Phase and scale correction

It is desired that the resulting weights are non-negative so that the phase and scale of reconstructed signals is not changed. Since an eigenvector is uniquely determined up to its scale and sign[2] (provided that it is only one eigenvector corresponding to its eigenvalue), we normalize $\boldsymbol{\lambda}_k$, $k = 1, \dots, d$, by taking absolute values of its elements. To preserve the scale, $\boldsymbol{\lambda}_k$ is then divided by its maximum element so that its largest element is equal to one.

2.5. Choice of K_k

The sets K_k, $k = 1, \dots, d$, are normally determined by the clusters founded by a clustering algorithm. A hard clustering results in disjoint sets K_k. When using a fuzzy clustering, the sets need not be disjoint and can share components that usually contain high residual interference. The advantage of non-binary weightings is then that such ICs are reasonably outweighted so as not to decrease SIR of separated signals.

We therefore use the relational fuzzy C-means clustering algorithm (RFCM) as in [5] and define K_k as the set of indices of ICs whose degree of affiliation to the kth cluster is above a given threshold τ.

3. Summary of the BSS Algorithm

Now we summarize our modification of the time-domain method T-ABCD from [3] which is endowed by the novel weighting approach. Also an improved similarity of components is proposed here; details are given in the next subsection.

The input of the method are the signals from microphones $x_i(n)$, $i = 1, \dots, m$, $n = N_1, \dots, N_2$, and the parameters L, α and τ. Then, the steps are as follows.

1. Construct $\mathbf{X}$ according to (3). Find the decomposing matrix $\mathbf{W}$ by an ICA algorithm (we use the BGSEP algorithm from [6]) giving mL independent components of $\mathbf{X}$, $\mathbf{C} = \mathbf{W}\mathbf{X}$.
2. Group the ICs (rows of $\mathbf{C}$) into clusters so that each cluster contains ICs corresponding to the same original source based on the similarity measure defined in subsection 3.1. Do the clustering by means of the RFCM algorithm using known number of clusters (sources) d.
3. Define the sets K_k, $k = 1, \dots, d$, as described in Subsection 2.5. Then, evaluate $\widetilde{\mathbf{G}}$ defined by (10), compute the weighting vectors $\boldsymbol{\lambda}_k$ following the steps in Subsection 2.2., and put $\widehat{\mathbf{S}}_k = \mathbf{W}^{-1}\boldsymbol{\Lambda}_k\mathbf{C}$, $k = 1, \dots, d$, where $\boldsymbol{\Lambda}_k = \text{diag}[\boldsymbol{\lambda}_k]$.
4. Compute estimates of the responses (2) of separated signals as

$$\widehat{s}_k^i(n) = \frac{1}{L}\sum_{p=1}^{L}(\widehat{\mathbf{S}}_k)_{(i-1)L+p,n+p-1} \tag{11}$$

(i.e. take averages over time-shifted rows within the blocks of $\widehat{\mathbf{S}}_k$).

3.1. Similarity of ICs

The clustering is done based on a similarity measure between the components. Since components corresponding to the same source should, ideally, be the same up to an unknown FIR filter, projection of one component to a subspace spanned by delayed copies of the other components were used in [3] as the measure of their similarity. Here we propose to use a coherence measure, which yields comparable results as the projection, but is computationally simpler.

Let $\mathbf{c}_i$ and $\mathbf{c}_j$ be the ith and jth component (row of $\mathbf{C}$), respectively. Let $c_i(\omega_k, n)$ denote the short-time Fourier Transform of $\mathbf{c}_i$ where $n = 1, \dots, Q$ is the index of the time-window of a length P, and ω_k is the kth frequency. The coherence of the ith and jth component is defined as

$$\text{coh}[\mathbf{c}_i, \mathbf{c}_j] = \frac{1}{P}\sum_{k=1}^{P}\frac{\big|\sum_{n=1}^{Q} c_i(\omega_k, n)\overline{c_j(\omega_k, n)}\big|^2}{(\sum_{n=1}^{Q}|c_i(\omega_k, n)|^2)(\sum_{n=1}^{Q}|c_j(\omega_k, n)|^2)}. \tag{12}$$

In our experiments, we select $P = 128$ and use no overlap of the time-windows.

4. Experimental Evaluation

4.1. Separation of SiSEC 2010 data

We test the proposed BSS method on data from the SiSEC 2010 evaluation campaign[3]. The data consist of two-microphone recordings of two sources played over loudspeakers. There are six couples of signals (a male or female speech with another jammer source) each recorded at seven different positions in room #1 and five positions in room #2. In summary, there are 12 different scenarios and 72 recordings altogether. The sampling rate is 16 kHz. Each recording lasts for three seconds, but the sources are active in first two seconds only. T-ABCD was set to compute ICA using the first second only.

method	SIR [dB]	SDR [dB]	SAR [dB]
original	8.7	4.4	9.6
proposed	9.0	4.9	10.1

Table 1: Achieved SIR, SDR and SAR on SiSEC data averaged over all scenarios.

We separate the recorded signals by the T-ABCD method with $L = 20$ using the original weighting from [3] and the one proposed here to compare their performances. The separation quality is evaluated using the BSS_EVAL toolbox [7]. Three criteria are used: the Signal-to-Interference ratio (SIR), the Signal-to-Distortion ratio (SDR), and the Signal-to-Artifacts ratio (SAR). SIR evaluates the ratio of energy of the target to jammer signal in the separated signal, while SDR and SAR measure deformations of the target signal; see [7] for exact definitions. The parameters of the compared approaches were tuned to optimum values making a good trade-off between SIR and SDR, namely, $\alpha = 0.1$ and $\tau = 0.4$ in the proposed method, and $\alpha = 1$ in the original method.

Table 1 summarizes achieved results averaged over both separated responses of both separated signals and over all scenarios. The performance of the proposed weighting is signif-

[2] It is worth noting that the elements of $\boldsymbol{\lambda}_k$ are expected to have, prior to the normalization, all the same sign. Although this is not guaranteed in theory, our observation is that it happens very often.

[3] The task "Robust blind linear/non-linear separation of short two-sources-two-microphones recordings" in the "Audio source separation" category; online `http://www.irisa.fr/metiss/SiSEC10/robot/database2010.zip`

icantly better in terms of all the criteria than that of the original weighting. Detailed results for each scenario are shown in Fig. 1, where it is seen that the improvement is up to by 1.5 dB.

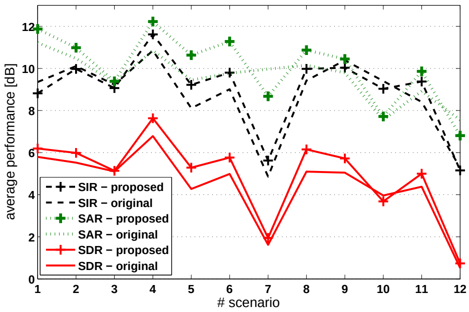

Figure 1: Detailed performance of compared approaches in all scenarios of SiSEC 2010 data.

4.2. Performance versus α

The parameter α controls the influence of the penalty term in (10). A typical behavior of the separation performance in dependence on α is shown by Fig. 2. For α close to zero, SIR is better at the expense of lower SDR and SAR. By contrast, SDR and SAR are better for higher α since the weighting is closer to the binary one defined in (5).

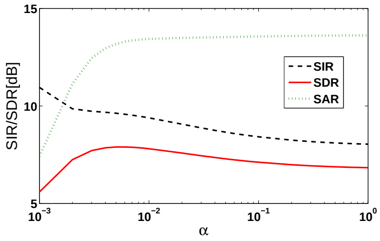

Figure 2: SIR, SDR and SAR of a separated source as functions of α.

4.3. CHiME Challenge

The task of CHiME challenge, which is an evaluation campaign organized as a satellite workshop of Interspeech 2011, is to separate distant speech from noise and recognize the keywords in commands being spoken. The recordings are binaural and were obtained by a dummy head from a distance of 2 meters from speaker. Both the speaker and dummy head maintain fixed positions, but no such assumption can be made about noise sounds; see the CHiME homepage[4] or [8] for further details about the task.

We apply the T-ABCD method endowed by the proposed weighting to the CHiME development data that consist of 3600 utterances. The data were created at six different signal-to-noise ratios (SNR: -6, -3, 0, 3, 6, and 9 dB) so that there are 600 recordings for each SNR. First, we separated the recordings by T-ABCD with $L = 10$, which has the ability to capture direct-path of the target signal and isolate it from other directions. To further increase SIR, the separated signals were post-processed by a non-linear mask (masked by the separated non-target signal), and the results were sent to a baseline recognizer that was trained on noise-free commands. Table 2 compares recognition score achieved, respectively, on *untreated* and enhanced signals using the *original* and *proposed* modification of T-ABCD.

SNR [dB]	-6	-3	0	3	6	9
untreated	31.08	36.75	49.08	64.00	73.83	83.08
original	35.25	38.58	51.00	64.67	73.83	83.58
proposed	35.25	41.58	53.33	67.42	74.58	84.08

Table 2: Keyword recognition accuracies (%) for the CHiME development data sets.

The improvement of recognition accuracy is significant (1–5%) but rather small. To explain, note that CHiME is a multidisciplinary challenge that requires complex solutions to achieve better recognition score. The separation here is fully blind (up to finding the separated target signal), and no other information such as the fixed position of the target source is exploited. Nevertheless, this direct application of T-ABCD provides an illustration pointing to the applicability of the method. We prepare a more efficient and tailored solution of the CHiME task for the respective satellite workshop.

5. Conclusions

We have proposed a novel weighting of independent components in the T-ABCD method for blind separation of audio signals. Compared to previous ad hoc weighting schemes, the proposed one relies on the block-Toeplitz structure of signal matrices. The experiments show that it improves the separation performance in various scenarios, and it is demonstrated that the resulting method can be used to enhance noisy speech signal prior to the recognition.

6. References

[1] Makino, S., Lee, T.-W. and Sawada, H. (Eds.), “Blind Speech Separation”, Springer, Sept. 2007.

[2] Sawada, H., Mukai, R., Araki, S. and Makino, S., “A robust and precise method for solving the permutation problem of frequency-domain blind source separation,” IEEE Trans. Speech Audio Processing, 12(5):530–538, 2004.

[3] Koldovský, Z. and Tichavský, P., “Time-Domain Blind Separation of Audio Sources on the basis of a Complete ICA Decomposition of an Observation Space”, IEEE Trans. on Speech, Audio and Language Processing, 19(2):406–416, Feb. 2011.

[4] Jafari, M.G., Vincent, E., Abdallah, S.A., Plumbley, M.D. and Davies, M.E., “An adaptive stereo basis method for convolutive blind audio source separation”, Neurocomputing, 71:2087–2097, 2008.

[5] Málek, J., Koldovský, Z., Žďánský, J. and Nouza, J., ”Enhancement of Noisy Speech Recordings via Blind Source Separation”, Proc. of Interspeech 2008, pp. 159-162, Sept. 22-26, Brisbane, Australia, 2008.

[6] Tichavský, P. and Yeredor, A., “Fast Approximate Joint Diagonalization Incorporating Weight Matrices,” IEEE Trans. on Signal Processing, 57(3):878–891, March 2009.

[7] Vincent, E., Févotte, C. and Gribonval, R., “Performance measurement in Blind Audio Source Separation,” IEEE Trans. Audio, Speech and Language Processing, 14(4):1462–1469, 2006.

[8] Christensen, H., Barker, J., Ma, N. and Green, P. “The CHiME corpus: a resource and a challenge for Computational Hearing in Multisource Environments,” INTERSPEECH’10, Makuhari, Japan, September 2010.

[4] http://www.dcs.shef.ac.uk/spandh/chime/challenge.html