Reduction of Broadband Noise In Speech Signals by Multilinear Subspace Analysis

Yusuke Sato[1], *Tetsuya Hoya*[1,2], *Hova Bakardjian*[2], *and Andrzej Cichocki*[2]

[1] Department of Mathematics, College of Science and Technology,
Nihon University, 1-8-14, Kanda-Surugadai, Chiyoda-Ku, Tokyo 101-8308, Japan
[2] Laboratory for Advanced Brain Signal Processing,
BSI RIKEN, 2-1, Hirosawa, Wakoh-City, Saitama 351-0198, Japan
ysato@grad.math.cst.nihon-u.ac.jp, hoya@math.cst.nihon-u.ac.jp
hova@brain.riken.jp, cia@brain.riken.jp

Abstract

A new noise reduction method for speech signals is proposed in this paper. The method is based upon the N-mode singular value decomposition algorithm, which exploits the multilinear subspace analysis of given speech data. Simulation results using both synthetically generated and real broadband noise components show that the enhancement quality obtained by the multilinear subspace analysis method in terms of both segmental gain and cepstral distance, as well as informal listening tests, is superior to that by a conventional nonlinear spectral subtraction method and the previously proposed approach based upon sliding subspace projection.

Index Terms: speech enhancement, broadband noise reduction, multilinear subspace analysis

1. Introduction

Removal of broadband noise has remained an important and central topic in speech enhancement. For this purpose, nonlinear spectral subtraction (NSS) [1] has been a well-known method for this purpose. The method, however, introduces annoying artefacts known as "musical tone" in the enhanced speech, due to the block-based processing in the spectral domain. It also normally requires an elaborate tuning of many parameters.

On the other hand, the approaches based upon subspace analysis [2]-[6] have recently attracted a growing interest in the speech enhancement area of study. Related to this, in a recent work [7], a tensor decomposion method based upon an N-mode orthogonal iteration algorithm is proposed for the effective dimensionality reduction of image data.

In this paper, we propose a novel approach for noise reduction in speech signals by applying a multi-linear (i.e. tensor-based) subspace analysis method. We then consider that more effective separation of signal and noise subspaces can be achieved, since the 3D data space, represented as tensors, contains the signal information spanning across multiple time frames. For the application side, it should also be noted that the multilinear subspace analysis proposed in this paper does not suffer from the aforementioned problems as in NSS.

2. The Multilinear Subspace Analysis

Without loss of generality, we here consider an ordinary two-channel environment and the observed speech signals $x_i(k)$ $(i = 1, 2)$ are expressed as

$$\begin{aligned} x_1(k) &= a \cdot s_1(k) + n_1(k) \,, \\ x_2(k) &= a \cdot s_2(k) + n_2(k) \,, \end{aligned} \tag{1}$$

where $s_i(k)$ and $n_i(k)$ denote the noise-clean speech signals and the corresponding additive noise components (i.e. assumed with zero-mean and uncorrelated with the speech signals), respectively, and the constant 'a' controls the input SNR. Then, the objective is to remove only the noise components $n_i(k)$ from $x_i(k)$.

2.1. Construction of Tensors from Noisy Speech Data

In multilinear algebra, a higher-order generalisation of a vector and a matrix is generally called a *tensor*. First, we construct a 3^{rd}-order tensor $\mathcal{D}$ ($\in \Re^{M \times L \times P}$) (hereafter, the time index k is omitted, where appropriate) from the observed data as illustrated in Fig. 1. (In this paper, since the observation is considered as two-channel signals, we set $M = 2$.) The element of $\mathcal{D}$ is given as

$$\begin{aligned} d_{i_1 i_2 i_3} &= x_{i_1}(k - L(P - i_3 + 1) + i_2) \\ &\quad (1 \le i_1 \le 2, 1 \le i_2 \le L, 1 \le i_3 \le P) \,. \end{aligned} \tag{2}$$

Second, we obtain the three matrices $\boldsymbol{D}_{(1)}$ $(2 \times LP)$, $\boldsymbol{D}_{(2)}$ $(L \times 2P)$, $\boldsymbol{D}_{(3)}$ $(P \times 2L)$ by flattening the tensor $\mathcal{D}$:

$$\boldsymbol{D}_{(1)} = \begin{bmatrix} x_1(k) & x_1(k-1) & \cdots & x_1(k-LP+1) \\ x_2(k) & x_2(k-1) & \cdots & x_2(k-LP+1) \end{bmatrix} , \tag{3}$$

$$\boldsymbol{D}_{(2)} = [\boldsymbol{D}^1_{(2)} \,;\, \boldsymbol{D}^2_{(2)}] \,, \tag{4}$$

$$\boldsymbol{D}^{i_1}_{(2)} = \begin{bmatrix} x_{i_1}(k-LP+1) & x_{i_1}(k-L(P-1)+1) & \cdots & x_{i_1}(k-L+1) \\ x_{i_1}(k-LP+2) & x_{i_1}(k-L(P-1)+2) & \cdots & x_{i_1}(k-L+2) \\ \vdots & \vdots & & \vdots \\ x_{i_1}(k-LP+L) & x_{i_1}(k-L(P-1)+L) & \cdots & x_{i_1}(k) \end{bmatrix} ,$$

$$\boldsymbol{D}_{(3)} = [\boldsymbol{D}_{(3)}(1) \,;\, \boldsymbol{D}_{(3)}(2) \,;\, \cdots \,;\, \boldsymbol{D}_{(3)}(L)] \,, \tag{5}$$

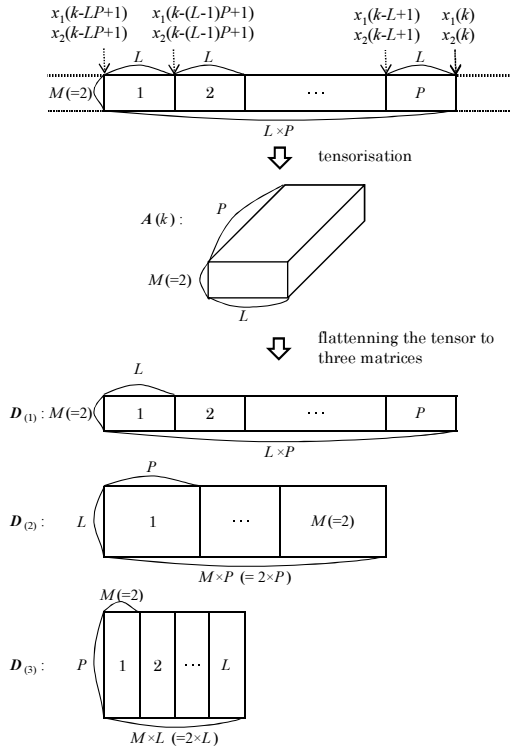


Figure 1: *Flattening the data tensor $\mathcal{D}(k)$ to obtain $\boldsymbol{D}_{(1)}$, $\boldsymbol{D}_{(2)}$, and $\boldsymbol{D}_{(2)}$*

$$\boldsymbol{D}_{(3)}^T(i_2) = \begin{bmatrix} x_1(k-LP+i_2) & x_1(k-L(P-1)+i_2) \\ x_2(k-LP+i_2) & x_2(k-L(P-1)+i_2) \\ \cdots & x_1(k-L+i_2) \\ \cdots & x_2(k-L+i_2) \end{bmatrix} .$$

2.2. Noise Reduction by the N-mode Orthogonal Iteration Algorithm

Vasilescu and Terzopoulous [7] developed an iterative algorithm for multilinear subspace analysis using the N-mode singular value decomposition (SVD) algorithm, which can be regarded as a natural extension to higher-order tensors of the conventional matrix SVD, and successfully applied to dimensionality reduction in static facial images.

Using the tensor $\mathcal{D}$ (in (2)) and its three flattened matrices $\boldsymbol{D}_{(1)}$, $\boldsymbol{D}_{(2)}$, and $\boldsymbol{D}_{(3)}$ (in (3)-(5)), we apply the following N-mode orthogonal iteration algorithm (cf. [7]):

1) Obtain the matrices $\boldsymbol{U}_i$ ($i = 1, 2, 3$) by computing the SVD of each flattened matrix $\boldsymbol{D}_{(i)}$ (given by (3)-(5)) and setting $\boldsymbol{U}_i$ to be the left matrix of the SVD; truncate $\boldsymbol{U}_i$ to R_i columns and set them as initial mode matrices $\boldsymbol{U}_i^0$.

2) Repeat the following, until the condition $\|\boldsymbol{U}_i^{j+1T} \cdot \boldsymbol{U}_i^j\|^2 < (1-\epsilon)R_i$ ($j = 0, 1, \ldots$) is met:

   2.1) Set $\tilde{\mathcal{U}}_1^{j+1} = \mathcal{D} \times_2 \boldsymbol{U}_2^{j\,T} \times_3 \boldsymbol{U}_3^{j\,T}$; mode-1 flatten the tensor $\tilde{\mathcal{U}}_1^{j+1}$ to obtain the matrix $\tilde{\boldsymbol{U}}_1^{j+1}$; set the columns of $\boldsymbol{U}_1^{j+1}$ to an orthonormal basis for the R_1-dimensional dominant subspace of $\tilde{\boldsymbol{U}}_1^{j+1}$. The dominant subspace can be actually computed via the left matrix obtained by the SVD of $\tilde{\boldsymbol{U}}_1^{j+1}$. (So are the matrices $\boldsymbol{U}_2^{j+1}$ and $\boldsymbol{U}_3^{j+1}$, as computed in the following sub-steps.)

   2.2) Set $\tilde{\mathcal{U}}_2^{j+1} = \mathcal{D} \times_1 \boldsymbol{U}_1^{j\,T} \times_3 \boldsymbol{U}_3^{j\,T}$; mode-2 flatten the tensor $\tilde{\mathcal{U}}_2^{j+1}$ to obtain the matrix $\tilde{\boldsymbol{U}}_2^{j+1}$; set the columns of $\boldsymbol{U}_2^{j+1}$ to an orthonormal basis for the R_2-dimensional dominant subspace of $\tilde{\boldsymbol{U}}_2^{j+1}$.

   2.3) Set $\tilde{\mathcal{U}}_3^{j+1} = \mathcal{D} \times_1 \boldsymbol{U}_1^{j\,T} \times_2 \boldsymbol{U}_2^{j\,T}$; mode-3 flatten the tensor $\tilde{\mathcal{U}}_3^{j+1}$ to obtain the matrix $\tilde{\boldsymbol{U}}_3^{j+1}$; set the columns of $\boldsymbol{U}_3^{j+1}$ to an orthonormal basis for the R_3-dimensional dominant subspace of $\tilde{\boldsymbol{U}}_3^{j+1}$.

3) Set the converged matrices to $\hat{\boldsymbol{U}}_1, \hat{\boldsymbol{U}}_2, \hat{\boldsymbol{U}}_3$. Using these three matrices, compute the core tensor

$$\hat{\mathcal{Z}} = \tilde{\mathcal{U}}_3^{j+1} \times_3 \hat{\boldsymbol{U}}_3^T . \tag{6}$$

The rank-reduced approximation of $\mathcal{D}$ is eventually given by

$$\hat{\mathcal{D}} = \hat{\mathcal{Z}} \times_1 \hat{\boldsymbol{U}}_1 \times_2 \hat{\boldsymbol{U}}_2 \times_3 \hat{\boldsymbol{U}}_3 . \tag{7}$$

For the application of the algorithm in the above to noise reduction in speech, we follow the same principle as in [6] that the signal space can be decomposed into signal-plus-noise (i.e. this corresponds to the dominant subspace in Step 2.1) and noise subspace by an SVD operation. Hence, we take only the one-dimensional dominant subspace in mode-1 matrix $\boldsymbol{U}_1^j$ during the iteration, i.e. $R_1 = 1$, for performing the noise reduction. (Thus, the degrees of freedom of the proposed method yields essentially three, i.e. the selection of the parameters L,$R_1 (\leq L)$, and P.)

After the convergence, we eventually obtain a chunk of LP data points of the enhanced speech signals $\hat{s}_i(k)$, $\hat{s}_i(k-1)$, $\ldots$, $\hat{s}_i(k-LP+1)$ by simply unfolding the rank-reduced version of tensor $\hat{\mathcal{D}}$ in (7) (i.e. via an inverse operation from the tensor cube to the signal vectors; as illustrated in the upper part of Fig. 1):

$$\hat{s}_{i_1}(k - L(P - i_3 + 1) + i_2) = \hat{d}_{i_1 i_2 i_3} . \tag{8}$$

3. Simulation Study

In the simulation study, a total of 45 speech signals were used for the two-channel speech components s_i (as given in (1)). In the recording session of the speech signals, three non-trained Japanese native speakers (one female and two male) participated and uttered five times the three short sentences in Japanese "*o-ha-yoh-go-za-i-ma-su*"(good morning), "*kon-ni-chi-wa*" (hello), and "*ha-ji-me-ma-shi-te*"(nice to meet you), in an ordinary office room (with a negligible level of background noise). For the additive noise components n_i, we considered the two cases: 1) i.i.d. random variables generated from normal distribution (i.e. using the MATLAB $randn(\cdot)$ function) and 2) PC line noise recorded using the built-in stereophonic microphone and sound-recorder program on Windows XP (i.e. used as the real broadband noise components). Figure 2 shows a portion of the recorded two-channel PC line noise signals. All the recorded speech as well as real noise signals were originally sampled at 48kHz and down-sampled to 8kHz.

To validate the proposed method (denoted hereafter N-

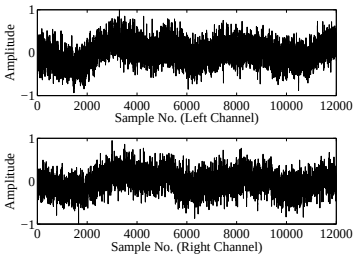


Figure 2: *Two-channel real broadband noise components used for the simulation study*

mode SVD), two existing algorithms of NSS [1] and sliding subspace projection (SSP) [6] were compared. For the NSS, since the performance is quite dependent upon the parameter choice, the three parameter combinations as used in [6] were considered for giving a fair comparison. For the N-mode SVD, the settings of $L = 32$ and $R_2 = 12$ were used for additive two i.i.d. random noise case, whilst $L = 24$ and $R_2 = 6$ for the real broadband noise case, and the same values of L were used as the data vector lengths for the SSP. Moreover, for an efficient computation of the N-mode orthogonal iteration algorithm, we constrained $R_3 = R_1 R_2$ during the simulation.

For the evaluation of enhancement, the objective measurement in terms of both the segmental gain in SNR and averaged cepstral distance was considered (see, e.g. [8]), together with the informal listening tests. The segmental gain in SNR (dB) is given by

$$G(dB) \quad = \quad \frac{1}{Mp_1} \sum_{i=1}^{M} \sum_{j=1}^{p_1} 10\, log_{10} \frac{\|\boldsymbol{n}_i\|_2^2}{\|\boldsymbol{s}_i - \hat{\boldsymbol{s}}_i\|_2^2}, \quad (9)$$

where $M = 2$ (two-channels), $\boldsymbol{s}_i = [s_i(k), s_i(k+1), \cdots, s_i(k+N_f-1)]^T$, $\hat{\boldsymbol{s}}_i = [\hat{s}_i(k), \hat{s}_i(k+1), \cdots, \hat{s}_i(k+K-1)]^T$, $\boldsymbol{n}_i = [n_i(k), n_i(k+1), \cdots, n_i(k+N_f-1)]^T$, $(k = (j-1)N_f, (j-1)N_f+1, \cdots, jN_f-1,\ j = 1, 2, ..., p_1)$ are respectively the noise-clean speech, enhanced speech, and the noise signal vector, and where $N_f (= 256)$ is the number of samples in each frame and p_1 is the number of frames. The averaged cepstral distance is defined as

$$d_{cep} \quad = \quad \frac{1}{M} \sum_{i=1}^{M} \frac{1}{p_{2,i}} \sum_{j=1}^{p_{2,i}} \sum_{k=1}^{2q} (c_{i,k}(j) - \hat{c}_{i,k}(j))^2 \quad (10)$$

where $c_{i,k}(j)$ and $\hat{c}_{i,k}(j)$ are the cepstral coefficients corresponding to the clean and the enhanced signal at the left/right channel, respectively. The parameter $q(= 8)$ is the order of the model, and $p_{2,i}$ $(i = 1, 2)$ is the number of frames where speech is present. The presence of speech was determined by manual examination using the noise-clean speech signals.

Due to the constraints on space, we present only a couple of the simulation results using two speech samples. Figure 4 (a) and (b) show the two different noise-clean speech signals, and the respective noisy speech (assuming the input SNR=4.8dB) are shown in (c) and (d), respectively. To generate the noisy speech, the aforementioned two i.i.d. random noise signals were used for (c), whereas the real broadband noise components for (d). In the lower part of the figure, the enhanced speech signals obtained by the three different noise reduction algorithms are compared: the enhanced speech obtained by NSS is shown in (e) and (f), that by SSP in (g) and (h), and the proposed N-mode SVD approach in (i) and (j), respectively.

It is noticeable that some parts of the enhanced speech by NSS are almost gone/corrupted (i.e. around 7000-8000 sample nos. in Fig. 4(e) and (f)). In contrast, the noise level is grossly reduced in the enhanced speech obtained by the SSP/N-mode SVD, whilst the overall shape of speech parts almost remains intact as shown in Fig. 4(g)-(j). It is also clearly seen that overall noise level in the enhanced speech obtained by the N-mode SVD is lower than that by SSP in these figures. All these observations also agreed with the informal listening tests. Moreover, for the N-mode SVD, annoying pulse-like noise as in the enhanced speech by SSP was not heard.

Figures 4(a)-(d) compare the performance of the three algorithms, NSS (with the aforementioned three different parameter settings; denoted NSS1-NSS3 in the figures), SSP, and N-mode SVD, in terms of both the segmental gain (i.e. given by (9)) and cepstral distance (i.e. given by (10)). In the figures, the plots so obtained are the averaged ones for all the 45 speech samples. As shown in Fig. 4(a), within the range of input SNR 3-12dB, the performance of N-mode SVD in terms of segmental gain is superior to the other two algorithms for the two i.i.d. random noise case, whilst a better performance in terms of the cepstral distance is obtained by the N-mode SVD for both the noise cases within the lower range of the input SNR than 7dB, as shown in Figs. 4(b) and 4(d).

4. Conclusion

In this paper, we proposed a novel approach for noise reduction in speech based upon the multilinear subspace analysis of a data tensor using the N-mode orthogonal iteration algorithm [7]. The effectiveness of the proposed approach based upon N-mode SVD has been shown, in comparison with both the conventional NSS and SSP. As in the SSP, the enhanced speech signals by N-mode SVD eventually yields monoaural (due to the operation with $R_1 = 1$). For the recovery of the stereophonic image, the post-processing by adaptive signal enhancers can therefore be effective. Moreover, unlike the SSP, the proposed approach works rather in a batch mode. Thus, for the practical utility, an approach operated in an on-line mode is desirable. We are thus currently investigating these possibilities.

5. References

[1] R. Martin, "Spectral subtraction based on minimum statistics," *Proc. EUSIPCO-94*, pp. 1182-1185, Edinburgh, 1994.

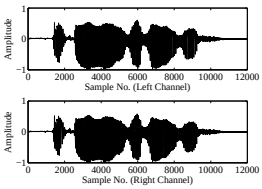
[2] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: a regenerative approach," *Speech Communication*, Vol. 10, pp. 45-57, Feb. 1991.

[3] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech, Audio Proc.*, Vol. 3, No. 4, pp. 251-266, July 1995.

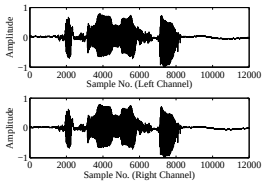
[4] S. Doclo and M. Moonen, "A novel iterative signal enhancement algorithm for noise reduction in speech," Proc. *5th Int. Conf. Spoken Language Processing*, Sydney, Australia, pp. 1435-1438, Dec. 1998, also in internal report, K. U. Leuven, Apr. 1998.

[5] F. Asano, S, Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE Trans. Speech, Audio Proc.*, Vol. 8, No. 5, pp. 497-507, Sept. 2000.

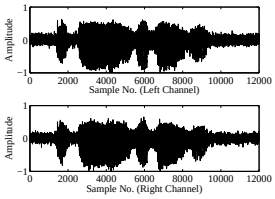
[6] T. Hoya, T. Tanaka, A. Cichocki, T. Murakami, G. Hori, and J. A. Chambers, "Stereophonic noise reduction using a combined sliding subspace projection and adaptive signal enhancement," *IEEE Trans. Speech. Audio Proc.*, Vol. 13, No. 3, pp. 309-320, May 2005.



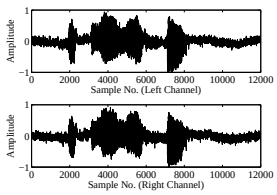
(a) *Noise clean speech (speech sample #1)*



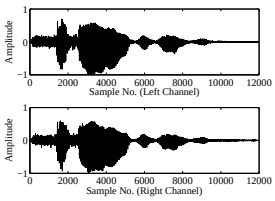
(b) *Noise clean speech (speech sample #2)*



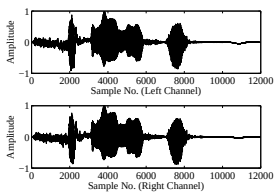
(c) *Noisy data – with additive two i.i.d. random noise (SNR=4.8dB)*



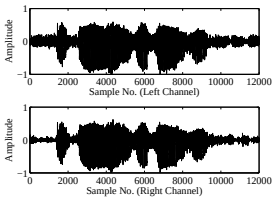
(d) *Noisy data – with additive real broadband noise (SNR=4.8dB)*



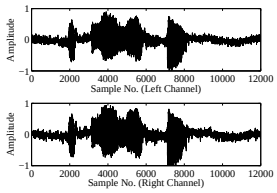
(e) *Enhanced speech by NSS*



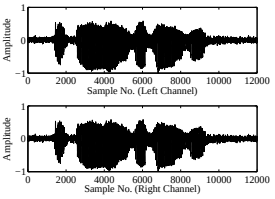
(f) *Enhanced sppech by NSS*



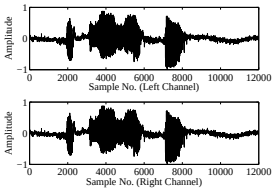
(g) *Enhanced speech by SSP*



(h) *Enhanced speech by SSP*

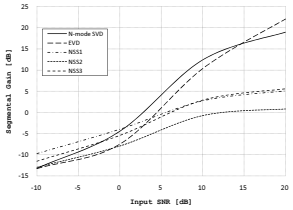


(i) *Enhanced speech by N-mode SVD*

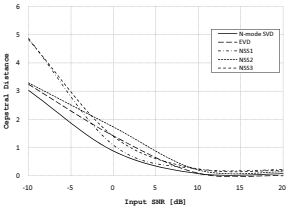


(j) *Enhanced speech by N-mode SVD*

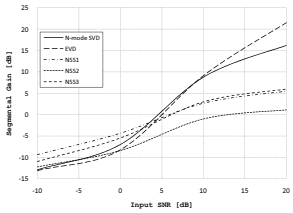
Figure 3: *Simulation results obtained using two speech samples*



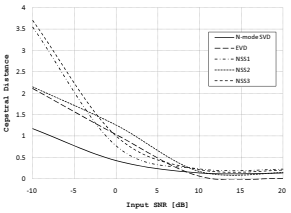
(a) *Segmental gain – two additive i.i.d. random noise case*



(b) *Cesptral distances – two additive i.i.d. random noise case*



(c) *Segmental gain – real broadband noise case*



(d) *Cesptral distances – real broadband noise case*

Figure 4: *Comparison of the segmental gain and cepstral distances*

[7] M. A. Vasilescu and D. Terzopoulous, "Multilinear subspace analysis of image ensembles," Proc. *Int. Conf. Computer Vision and Pattern Recognition*, Madison, 2003.

[8] J. R. Deller, Jr., J. G. Proakis, and J. H. L. Hansen, *Discrete-time processing of speech signals*, Macmillan Publishing Company, 1993.