Ludeña-Choez, J. & Gallardo-Antolín, A. (2012). Speech Denoising Using Non-negative Matrix Factorization with Kullback-Leibler Divergence and Sparseness Constraints. In Torre Toledano, D., et al. (eds.) *Advances in Speech and Language Technologies for Iberian Languages: IberSPEECH 2012 Conference, Madrid, Spain, November 21-23, 2012. Proceedings.* (pp. 207-216). (Communications in Computer and Information Science; 328). Springer Berlin Heidelberg. DOI: http://dx.doi.org/10.1007/978-3-642-35292-8_22

Speech Denoising Using Non-negative Matrix Factorization with Kullback-Leibler Divergence and Sparseness Constraints

Jimmy Ludeña-Choez and Ascensión Gallardo-Antolín

Dept. of Signal Theory and Communications, Universidad Carlos III de Madrid,
Avda. de la Universidad 30, 28911 - Leganés (Madrid), Spain
{jimmy,gallardo}@tsc.uc3m.es

Abstract. A speech denoising method based on Non-Negative Matrix Factorization (NMF) is presented in this paper. With respect to previous related works, this paper makes two contributions. First, our method does not assume a priori knowledge about the nature of the noise. Second, it combines the use of the Kullback-Leibler divergence with sparseness constraints on the activation matrix, improving the performance of similar techniques that minimize the Euclidean distance and/or do not consider any sparsification. We evaluate the proposed method for both, speech enhancement and automatic speech recognitions tasks, and compare it to conventional spectral subtraction, showing improvements in speech quality and recognition accuracy, respectively, for different noisy conditions.

Keywords: Non-Negative Matrix Factorization, Kullback-Leibler Divergence, Sparseness Constraints, Speech Denoising, Speech Enhancement, Automatic Speech Recognition.

1 Introduction

The quality of speech is degraded in the presence of noise. Noisy speech signals are a common problem in many applications, e.g. Automatic Speech Recognition (ASR), landline and mobile phone communications, etc. In ASR systems, the problem is harder because machine understanding is still far from humans and speech enhancement is sometimes performed as a preprocessing stage for those systems. In this paper, we have concentrated our efforts on enhancing speech for both, human consumption and ASR. Several methods for reducing the influence of noise have been proposed. Among them, it is worth mentioning the Wiener filtering technique [1] and Spectral Subtraction (SS) [2], which consists of subtracting an estimate of the noise spectrum from the noisy speech spectrum. Both of them produce a more intelligible signal but generate the so called musical noise as a side effect.

Recently, Non-Negative Matrix Factorization (NMF) has been successfully used in areas related to speech processing, including speech denoising [3], sound separation [4], speaker separation [5] and feature extraction [6]. NMF provides a

way of decomposing a signal into a convex combination of nonnegative building blocks (also called basis vectors) by minimizing a cost function. Typical cost functions are the Euclidean distance and the Kullback-Leibler (KL) divergence. Therefore, NMF is capable of separating sound sources when their corresponding building blocks are sufficiently distinct, as is the case of speech and noise.

In this paper, we propose a NMF-based method for speech denoising which is very close to the one developed in [3] for speech enhancement tasks. The technique in [3] is based on a prior model of speech and noise, and therefore it assumes a priori knowledge of the type of noise which contaminates speech. In contrast, our method does not use this explicit information about noise, because it works with the only-noise segments of the current utterance to be denoised, after being detected with a Voice Activity Detector (VAD). Besides, we report results for both, speech enhancement and automatic speech recognition. On the other hand, several studies point out that it may be useful to explicit control the degree of sparsity in NMF decompositions for sound and speaker separation tasks. In this sense, the method for speaker separation proposed in [5] introduces a penalty term in the NMF with Euclidean distance that allows to control the sparsity of the solution. However, recent NMF-based techniques in speech processing report better results by using NMF with KL divergence [6], [4]. For this reason, in this paper, we propose a NMF-based method for speech denoising which combines the use of the KL divergence with sparseness constraints following the procedure described in [7].

This paper is organized as follows: Section 2 introduces the mathematical background of NMF; in Section 3 we present the speech denoising process using NMF. In Sections 4 and 5 we describe the application of the method to speech enhancement and automatic speech recognition, respectively, and end with some conclusions in Section 6.

2 Non-negative Matrix Factorization (NMF)

Given a matrix $V \in \mathbb{R}_+^{F \times T}$, where each column corresponds to a data vector, non-negative matrix factorization (NMF) approximates it as a product of two matrices of nonnegative low rank W and H, such that

$$V \approx WH \tag{1}$$

where $W \in \mathbb{R}_+^{F \times K}$ and $H \in \mathbb{R}_+^{K \times T}$ and normally $K \leq min\,(F, T)$. This way, each column of V can be written as a linear combination of the K basis vectors (columns of W), weighted with the coefficients of activation or gain located in the corresponding row of H. NMF can be seen as a dimensionality reduction of data vectors from an $F-$dimensional space to the $K-$dimensional space. This is possible if the columns of W uncover the latent structure in the data [8]. The factorization is achieved by an iterative minimization of a given cost function as, for example, the Euclidean distance or the generalized Kullbak Leibler (KL) divergence,

$$D_{\mathrm{KL}}\left(V\|WH\right) = \sum_{ij}\left(V_{ij}log\frac{V_{ij}}{(WH)_{ij}} - (V-WH)_{ij}\right) \tag{2}$$

In this work, we consider the KL divergence because it has been recently used with good results in speech processing tasks, such as sound source separation [4], speech enhancement [3] or feature extraction [6]. In order to find a local optimum value for the KL divergence between V and (WH), an iterative scheme with multiplicative update rules can be used as proposed in [8] and stated in (3)

$$W \leftarrow W \otimes \frac{\frac{V}{WH}H^T}{1H^T} \qquad\qquad H \leftarrow H \otimes \frac{W^T\frac{V}{WH}}{W^T1} \tag{3}$$

where 1 is a matrix of size V, whose elements are all ones and the multiplications $\otimes$ and divisions are component wise operations.

The NMF algorithm does not assume any sparsity or mutual statistical independence between columns of W. However, NMF usually provides sparse decomposition [8]. There are several ways to achieve some control of the sparsity. In this paper, we follow the approach proposed in [7] and [9] for KL cost functions, in which the NMF is regularized using non-linear projections based on (3). Applying this procedure, the regularized learning rules are the following,

$$W \leftarrow \left[W \otimes \frac{\left[\frac{V}{WH}H^T\right]^{\omega}}{1H^T}\right]^{(1+\alpha_w)} \qquad H \leftarrow \left[H \otimes \frac{\left[W^T\frac{V}{WH}\right]^{\omega}}{W^T1}\right]^{(1+\alpha_h)} \tag{4}$$

where α_w and α_h are the regularization parameters or sparse factors and ω is a relaxation parameter which also controls the sparsity and, in addition, speeds up the algorithm convergence. Note that with the sparse factors, the exponent of the learning rules are greater than one, which implies that the small values in the non-negative matrix tend to zero as the number of iterations increase [9]. In this paper, we only consider sparsification on the matrix H.

3 Speech Denoising Using NMF

NMF-based methods perform speech denoising under the hypothesis that noisy speech signals are the additive mixture of two sufficiently different sources: speech and noise. NMF is applied to magnitude spectra as it is assumed that the short-time magnitude spectra of a noisy signal, $|V_{\mathrm{mix}}|$ can be expressed as a linear combination of several distinct components, those representing only-speech spectra (W_{speech}) and those representing only-noise spectra (W_{noise}). These components are called Spectral Basis Vectors (SBV). The NMF representation of a noisy signal is shown in Fig. 1, wherein the speech SBVs (W_{speech}) and their corresponding speech activation coefficients (H_{speech}) can be used to reconstruct the clean speech signal ($|V_{\mathrm{speech}}| \approx W_{\mathrm{speech}}H_{\mathrm{speech}}$), while the noise SBVs (W_{noise}) and their corresponding noise activation coefficients (H_{noise}) can also be used to reconstruct the noise signal ($|V_{\mathrm{noise}}| \approx W_{\mathrm{noise}}H_{\mathrm{noise}}$) if required.

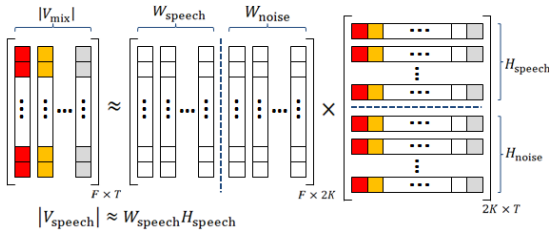

Fig. 1. NMF representation of noisy speech signals

The speech enhancement process consists of two different stages, training and denoising itself, as detailed below.

Training Stage. In the training stage, the SBVs representing speech and noise signals are determined. This is done by separately performing NMF on clean speech and noise data. First, the spectrum magnitude of both, clean speech ($|V_{\text{speech}}|$) and noise ($|V_{\text{noise}}|$) is computed. Afterwards, the KL divergence between the magnitude spectra and their corresponding factored matrices (($W_{\text{speech}}H_{\text{speech}}$) and ($W_{\text{noise}}H_{\text{noise}}$)) is minimized using the learning rules in (3). Since it is an iterative algorithm, it is important to perform a proper initialization of the matrices. Note that the spectral basis vectors contained in W_{speech} and W_{noise} are used in the next stage as speech and noise models.

For building the speech model, it is assumed that enough clean speech data is available. For the noise model, we have explored two different alternatives:

- *Offline Noise Data (OND).* In this approach, a priori knowledge about the type of the noise is assumed as in [3]. Therefore, a separate noise model for each of the noise types considered is trained using some offline available noise data. This approach will provide an upper limit of the proposed NMF-based denoising method performance.
- *Voice Activity Detector Noise Data (VADND).* In this approach, a VAD is used in order to explicit detect the only-noise segments of the utterance to be denoised. Afterwards, the noise model is built using these noise frames. Therefore, one noise model is trained for each utterance to be enhanced. This approach is more computational costly, but it avoids the need of the a priori knowledge about the type of noise, which it is not always possible.

Denoising Stage. As W_{speech} and W_{noise} are assumed to be good spectral basis functions to describe speech and noise, in the denoising stage they are kept fixed and are concatenated to form a single set of SBVs called W_{all}. Given the magnitude spectrum of the noisy speech signal ($|V_{\text{mix}}|$), we compute its factorization $|V_{\text{mix}}| \approx W_{\text{all}}H_{\text{all}}$ by minimizing the KL divergence between $|V_{\text{mix}}|$ and ($W_{\text{all}}H_{\text{all}}$), updating only the activation matrix H_{all} with the learning rules in (4). In order to control the sparseness of H_{all}, appropriate values for the regularization parameters (ω and α_h) need to be chosen (see subsection 4.2).

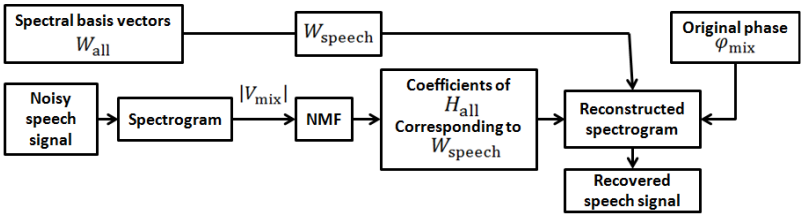

Fig. 2. Block diagram of the speech denoising process using NMF

The magnitude spectrum of the denoised speech is estimated as $|V_{\text{speech}}| \approx W_{\text{speech}}H_{\text{speech}}$, being H_{speech} the rows of H_{all} corresponding to the activation coefficients of W_{speech}. Finally, the spectrogram is recovered using the phase spectrum of the original noisy signal and the denoised speech signal is transformed into the time domain by means of a conventional overlap-add method. The whole process of speech denoising is shown in the block diagram of Fig. 2.

4 Application to Speech Enhancement

In this section, the evaluation of the proposed methods (OND and VADND) on a speech enhancement task is presented.

4.1 Database and Experimental Setup

The evaluation of speech enhancement was conducted on the AURORA-2 database [10], which is based on the TIDIgits database and it contains the recordings of 52 male and 52 females US-American adults pronouncing sequences of digits. Originally the database was recorded in clean conditions and subsequently contaminated with several types of noises at different SNRs. The sampling frequency is 8KHz. The database was end-pointed using the G.729 VAD.

For training the speech SBVs we used around 420 clean files belonging to the clean training set of the AURORA-2 database. In the OND method, the specific noise models were trained using the corresponding noise signals included in the database. In the VADND approach, the noise model for each utterance was trained using the initial only-noise frames detected by the VAD. In order to perform the study in subsection 4.2 we used 1,200 files from the test set A, which correspond to different noisy versions of 200 arbitrarily selected files with car noise added at SNRs from -5dB to 20dB with 5dB step. Finally, experiments in subsection 4.3 were conducted over 4,800 files from the test set A containing speech contaminated with subway, babble, car and exhibition hall noises at the SNRs previously mentioned. These files are noisy versions of 200 arbitrarily selected speech signals different from the ones used in subsection 4.2.

To evaluate the performance of the proposed methods, we use the so-called *Perceptual Evaluation of Speech Quality (PESQ)*, which is a measure recommended by the ITU-T for end-to-end speech quality assessment. The PESQ

score is able to predict subjective quality with good correlation in a very wide range of conditions (noise, filtering, coding distortions, etc.) [11] and uses a 5-point scale with 1 the worst and 5 the best values. PESQ values were computed using the code available in [12] and considering the clean speech signal as the reference. Results are presented using the following relative measure,

$$Ef_{\text{rel}} = \frac{PESQ_{\text{denoised}} - PESQ_{\text{noisy}}}{PESQ_{\text{noisy}}} \times 100\% \tag{5}$$

where $PESQ_{\text{noisy}}$ and $PESQ_{\text{denoised}}$ are the PESQ scores before and after applying the speech enhancement process, respectively. Increments imply a quality improvement and decrements a degradation with respect to the noisy signal.

4.2 Study on the Influence of the NMF Parameters

This set of experiments was performed in order to study the impact of several NMF parameters on the quality of the enhanced speech. The considered parameters were the analysis window length and the frame shift used for spectrograms computation, the number of spectral basis vectors and the values of the regularization factors, ω and α_h. In all cases, NMF was initialized by running 10 times the Alternating Least Squares NMF (ALS NMF) algorithm [9], in such a way that the factorization with the smallest euclidean distance between V and (WH) was chosen for initialization. Then, these initial matrices were refined by minimizing the KL divergence with sparseness constraints as indicated en Section 2. Preliminary experiments considering the Euclidean distance as cost function instead of the KL divergence produced worse results in terms of PESQ. The main experiments and results are summarized in next paragraphs:

- The window length was varied from 10ms to 45ms with 5ms step. From this set of experiments, it was observed that PESQ scores decreased with the window length, obtaining the best results in the range from 25ms to 45ms.
- The frameshift was studied in the range from 1ms to 10ms. In this case, the speech quality improved as the frameshift became smaller. Best PESQ scores were found in the range from 1ms to 5ms.
- The number of SBVs was varied from 10 to 80 with 10 step. Results showed that the quality of the denoised speech degraded when using a small number of SBVs (below 30), whereas best PESQ scores were obtained in the range from 40 to 80 SBVs. This result indicates that for an adequate representation of speech signals in NMF, it seems necessary to consider more than 30 SBVs.
- With respect to the regularization parameters, several experiments were performed varying α_h from 0 to 1.2 (fixing $\omega = 1$) and varying ω from 1 to 2.5 (fixing $\alpha_h = 0$). Results for the OND approach are shown in Fig. 3a and Fig. 3b, respectively. Similar trends were observed for the VADND method. As it can be observed, PESQ scores degrade when no regularization is used (this case corresponds to $\alpha_h = 0$ in Fig. 3a and $\omega = 1$ in Fig. 3b). However, when the values of the regularization parameters increase, the speech quality improves, being the best performance found for the combination of

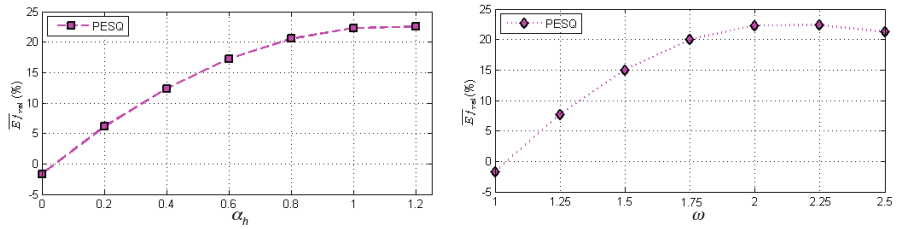

Fig. 3. Relative PESQ measure for the OND approach and a) $\omega = 1$ with different values of α_h and b) $\alpha = 0$ with different values of ω

α_h around 1 and $\omega = 1$ or the combination of ω around 2 and $\alpha_h = 0$). Other combinations of these parameters were tried, not obtaining significant improvements with respect to these PESQ values.

4.3 Experimental Results

In this subsection, we compare the performance of the two NMF-based denoising approaches (OND and VADND) with the conventional Spectral Subtraction (SS) in terms of the relative PESQ measure. According to the results achieved in the previous study, for the NMF-based methods, we used a window length of 40ms, a frameshift of 2.5ms, 50 SBVs, $\omega = 1$ and $\alpha_h = 1$. For a fair comparison, in SS we considered the same values for the window length and the frameshift.

Fig. 4 shows the relative PESQ measure with respect to the noisy signal for the four types of noise considered at several SNRs. For subway, babble and exhibition hall noises, the NMF-based methods clearly outperform SS at low and medium SNRs (from -5 dB to 10 dB). For SNR = 15 dB, results obtained with OND, VADND and SS are rather similar. However, at higher SNR (20 dB), SS produces better results than the NMF-based techniques. For the car noise, OND is better than SS at low and medium SNRs (-5 dB, 0 dB and 5 dB). Nevertheless, SS outperforms OND for higher SNRs. VADND produces worse results than SS at all SNRs, being more noticeable the differences for SNRs over 15 dB. In general, results show that OND and VADND are more suitable than SS for low and medium ranges of SNR.

With respect to the comparison between OND and VADND, it can be observed that the quality of the enhanced signal is better with OND in all cases. This result is expectable because OND uses more information than VADND in the denoising process. In fact, it needs to know the type of noise (not the SNR) presented in the noisy utterances. Nevertheless, VADND is capable of effectively denoise the speech signal using only the information contained in the only-noise segments of each utterance.

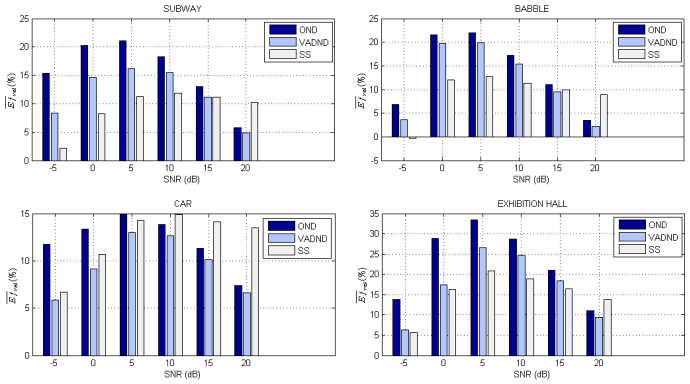

Fig. 4. Relative PESQ measure for SS, OND and VADND techniques

5 Application to Automatic Speech Recognition (ASR)

In this section, we present the evaluation of the proposed techniques on an ASR task. In this case, firstly noisy signals are denoised using the NMF-based techniques (OND or VADND) and then, these enhanced signals are fed into the ASR system.

5.1 Database and Experimental Setup

The experiments were conducted over the AURORA-2 database [10] as for the speech enhancement task. The recognizer was based on HTK (Hidden Markov Model Toolkit) software package with the configuration included in the standard experimental protocol of the database. Acoustic models were obtained from the clean training set of the database, whereas test files correspond to the complete test set A. Results are shown in terms of the recognition accuracy.

Acoustic features consist of the conventional Mel-Frequency Cepstrum Coefficients (MFCC). In particular, twelve MFCCs were computed every 10 ms using a Hamming analysis window of 25 ms long and 23 mel-spaced spectral bands. The log-energy of each frame and the corresponding delta and acceleration coefficients were also computed and added, yielding feature vectors of 39 components. Finally, cepstral mean normalization (CMN) was applied.

5.2 Experimental Results

Fig. 5 shows the recognition results achieved by the different NMF-based denoising techniques as well as for Spectral Subtraction (SS) and the baseline system (without denoising). For SS, OND and VADND, the same configuration parameters as in the case of speech enhancement were used, except for the regularization parameters, that were set to $\omega = 1.25$ and $\alpha_h = 0.2$, after a preliminary experimentation.

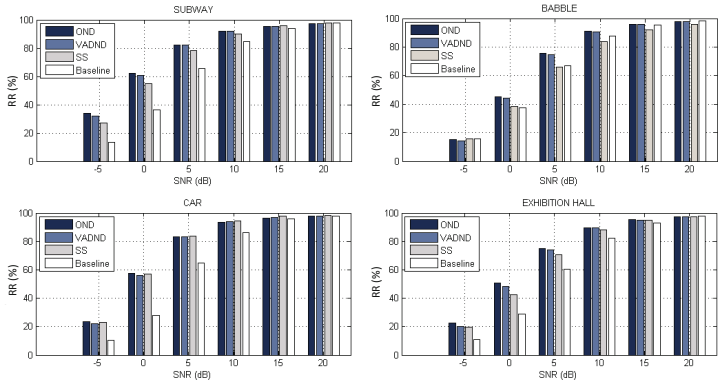

Fig. 5. Recognition Rates (%) for the baseline, SS, OND and VADND techniques

As it can be observed, for subway, babble and exhibition hall noises, both NMF-based techniques achieve better results than SS and the baseline for low and medium SNRs (from -5 dB to 10 dB). For higher SNRs, all the algorithms present a similar behaviour except for SS in the babble noise. In this case, the recognition accuracy obtained with SS is lower than the other techniques (including the baseline), probably due to the distortions introduced by SS in the denoising process. For the car noise, similar results are achieved with all techniques. On the other hand, comparing the two NMF-based methods for all noises, OND outperforms slightly VADND in most cases, being these performance differences less noticeably than in the speech enhancement task.

Table 1. Average Recognition Rates (%) for the four types of noise

Noise	OND	VADND	SS	Baseline
Subway	77.12	76.62	73.95	65.34
Babble	70.19	69.66	65.35	66.83
Car	75.29	74.94	75.72	63.86
Exhibition Hall	71.81	70.66	68.83	62.23

Table 1 shows the recognition rates averaged over all SNRs for each type of noise. It can be observed that OND and VADND outperforms SS for all noises, except for the car noise in which the results are very similar.

6 Conclusions and Future Work

In this paper we have presented a NMF-based method for speech denoising which combines the use of the Kullback-Leibler divergence with sparseness constraints on the activation matrix and it does not assume a priori knowledge about the

nature of the noise. In addition, an exhaustive study on the influence of different NMF parameters (window length, frameshift, number of spectral basis vectors and regularization parameters) on the quality of the enhanced speech has been carried out. We have compared the proposed method to conventional spectral subtraction for both, speech enhancement and automatic speech recognitions tasks, under different noisy conditions, obtaining significant improvements especially at low and medium SNRs.

For future work, we plan to experiment on real noisy signals instead of the artificially distorted ones used in this paper. Other future lines include the unsupervised learning of auditory filter banks by means of NMF and the use of the activation coefficients as acoustic parameters in ASR tasks.

Acknowledgments. This work has been partially supported by the Spanish Government grants TSI-020110-2009-103 and TEC2011-26807. Financial support from the Fundación Carolina (Jimmy Ludeña-Choez) is thankfully acknowledged.

References

1. Scalart, P., Filho, J.: Speech enhancement based on a priori signal to noise estimation. In: ICASSP 1996, pp. 629–632 (1996)
2. Berouti, M., Schwartz, R., Makhoul, J.: Enhancement of speech corrupted by acoustic noise. In: ICASSP 1979, pp. 208–211 (1979)
3. Wilson, K., Raj, B., Smaragdis, P., Divakaran, A.: Speech denoising using nonnegative matrix factorization with priors. In: ICASSP 2008, pp. 4029–4032 (2008)
4. Virtanen, T.: Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. on Audio, Speech and Language Processing 15(3), 1066–1074 (2007)
5. Schmidt, M., Olsson, R.: Single-channel speech separation using sparse nonnegative matrix factorization. In: INTERSPEECH 2006 (2006)
6. Schuller, B., Weninger, F., Wollmer, M., Sun, Y., Rigoll, G.: Non-negative matrix factorization as noise-robust feature extractor for speech recognition. In: ICASSP 2010, pp. 4562–4565 (2010)
7. Cichocki, A., Zdunek, R., Amari, S.: New algorithms for non-negative matrix factorization in applications to blind source separation. In: ICASSP 2006, pp. 621–625 (2006)
8. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)
9. Cichocki, A., Zdunek, R., Phan, A., Amari, S.: Nonnegative matrix and tensor factorizations. John Wiley and Sons, United Kingdom (2009)
10. Pearce, D., Hans, G.: The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In: ICSLP 2000 (2000)
11. Beerends, J., Hekstra, A., Rix, A., Hollier, M.: Perceptual evaluation of speech quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part II. Psychoacoustic model. Journal of the Audio Engineering Society 50(10), 765–778 (2002)
12. Hu, Y., Loizou, P.: Matlab software (2011), `http://www.utdallas.edu/~loizou/speech/software.htm`