A comparative study on sparsity penalties for NMF-based speech separation: beyond LP-norms

Cyril Joder, Felix Weninger, David Virette, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Joder, Cyril, Felix Weninger, David Virette, and Björn Schuller. 2013. "A comparative study on sparsity penalties for NMF-based speech separation: beyond LP-norms." In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 26-31 May 2013, Vancouver, BC, Canada*, edited by Rabab Ward, Li Deng, Michael Adams, and Vicky Zhao, 858–62. Piscataway, NJ: IEEE. https://doi.org/10.1109/icassp.2013.6637770.

A COMPARATIVE STUDY ON SPARSITY PENALTIES FOR NMF-BASED SPEECH SEPARATION: BEYOND LP-NORMS

Cyril Joder[1], *Felix Weninger*[1], *David Virette*[2], *Björn Schuller*[1]

[1] Machine Intelligence & Signal Processing Group, MMK, Technische Universität München, Germany
[2] Huawei European Research Center, Germany

ABSTRACT

In this work, we study the usefulness of several types of sparsity penalties in the task of speech separation using supervised and semi-supervised Nonnegative Matrix Factorization (NMF). We compare different criteria from the literature to two novel penalty functions based on Wiener Entropy, in a large-scale evaluation on spontaneous speech overlaid by realistic domestic noise, as well as music and stationary environmental noise corpora. The results show that enforcing the sparsity constraint in the separation phase does not improve the perceptual quality. In the learning phase however, it yields a better estimation of the base spectra, especially in the case of supervised NMF, where the proposed criteria delivered the best results.

Index Terms— Source separation, single-channel speech enhancement, noise cancellation

1. INTRODUCTION

Isolating speech from environmental noise remains a challenging problem, especially in the presence of highly non-stationary noise. On the other hand, a large variety of applications would benefit from a robust separation of speech. Among them are the reduction of acoustic noise in speech communications [1] or hearing aids [2]. Other possible applications comprise the automatic recognition of words [3], speaker [4] or emotion [5] in speech.

One of the most popular approaches for single-channel source separation is Nonnegative Matrix Factorization (NMF) [6], which is based on a decomposition of the spectrogram of the input mixture into a nonnegative combination of several spectral bases. This method has already been successfully applied to speech separation [1, 7, 8]. However, in the standard NMF method, the estimation of the dictionary of spectral bases often suffers from some inaccuracies and results in components representing several sources at the same time. Hence, several modifications of the standard NMF method have been proposed in order to limit this problem by integrating some structural constraints in the decomposition [9, 10, 11, 12].

Among the most widely used constraints is the activation sparsity property [13], which relates to the fact that the proportion of non-zero component activations (or, more generally, of non-negligible values) in the decomposition is very small. Several criteria have been proposed for enforcing sparsity [14, 15, 16]. However, to the authors' knowledge, there has not been any study on the relative advantages of these criteria. In the present paper, we compare the usefulness of several sparsity penalty functions from the literature on the task of speech separation using supervised and semi-supervised NMF. Besides, we propose the use of two novel criteria based on the Wiener entropy function, which significantly outperform the other method in the case of supervised NMF.

After presenting the NMF separation methods in Section 2, we detail the considered sparsity criteria in Section 3. Section 4 describes the experiments conducted, before some conclusions are drawn and the relation to prior work is discussed.

The research leading to these results has received funding from the HUAWEI Innovation Research Program (GLASS project).

2. SPEECH SEPARATION WITH NMF

2.1. Nonnegative Matrix Factorization

Given a matrix of nonnegative data $\mathbf{V} \in \mathbb{R}_+^{m \times n}$, NMF aims at finding the two nonnegative matrices, $\mathbf{W} \in \mathbb{R}_+^{m \times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r \times n}$, which minimize the error $D(\mathbf{V}, \mathbf{WH})$, where D is some divergence measure. In our audio source separation application, $\mathbf{V}$ is the original magnitude spectrogram. The columns of $\mathbf{W}$ then represent characteristic spectra of the recording and $\mathbf{H}$ contains the corresponding 'activation' values of these basis spectra.

Many algorithms for performing this optimization rely on multiplicative update rules, in order to maintain the nonnegativity of the matrices $\mathbf{W}$ and $\mathbf{H}$. The cost function used is the generalized Kullback-Leibler divergence, as it showed good results in previous experiments [17]. It is defined as:

$$D_{\mathrm{KL}}(\mathbf{X}, \mathbf{Y}) = \sum_{i,j} x_{i,j} \log \frac{x_{i,j}}{y_{i,j}} - x_{i,j} + y_{i,j}. \qquad (1)$$

The corresponding update rules proposed by [6, 18] are as follows:

$$\mathbf{W} \leftarrow \mathbf{W} \cdot \frac{\frac{\mathbf{V}}{\mathbf{WH}}\mathbf{H}^T}{\mathbf{1}\mathbf{H}^T} \qquad (2) \qquad\qquad \mathbf{H} \leftarrow \mathbf{H} \cdot \frac{\mathbf{W}^T\frac{\mathbf{V}}{\mathbf{WH}}}{\mathbf{W}^T\mathbf{1}} \qquad (3)$$

where $\mathbf{X} \cdot \mathbf{Y}$ and $\frac{\mathbf{X}}{\mathbf{Y}}$ denote element-wise operations and $\mathbf{1}$ is a matrix of ones.

Assuming that each source is described by a set of columns of $\mathbf{W}$ with corresponding rows in $\mathbf{H}$, separated signals can then be reconstructed as follows. Let $\mathbf{W}^{(k)}$ be the sub-matrix containing the columns of $\mathbf{W}$ corresponding to a source k, and let $\mathbf{H}^{(k)}$ be the according activation sub-matrix. The magnitude spectrogram of the isolated source $\mathbf{V}^{(k)}$ is obtained by the Wiener-like formula: $\mathbf{V}^{(k)} = \mathbf{V} \cdot \frac{\mathbf{W^{(k)}H^{(k)}}}{\mathbf{WH}}$. This spectrogram is then inverted using the phase of the original mixture and the time-domain signal is obtained by the overlap-add procedure.

2.2. Adding Activation Sparsity Penalty

The standard NMF method described in the previous subsections aims to minimize the reconstruction error between the original input and the decomposition, without taking into account the structure of

the individual signals. Hence, the estimated bases can capture some unstructured "building blocks" which can be used to reconstruct several sources, whereas the goal is to match each basis to a specific source. Hence, activation sparsity is often enforced by adding a penalty term to the cost function, which becomes:

$$C(\mathbf{W},\mathbf{H};\mathbf{V}) = D_{\mathrm{KL}}(\mathbf{V},\mathbf{WH}) + \lambda g(\mathbf{H}), \tag{4}$$

where $g(\mathbf{H})$ is a sparsity criterion and λ is a parameter controlling the weight of the sparsity penalty. This formulation is called sparsity-constrained NMF.

For the optimization of this cost function, we adopt the multiplicative update heuristic used in [15, 14]. In comparison to the previous subsection, only the update of $\mathbf{H}$ (3) is modified. It becomes

$$\mathbf{H} \leftarrow \mathbf{H} \cdot \frac{\mathbf{W}^T \frac{\mathbf{V}}{\mathbf{WH}} + \lambda \nabla g^-(\mathbf{H})}{\mathbf{W}^T\mathbf{1} + \lambda \nabla g^+(\mathbf{H})}, \tag{5}$$

where the gradient $\nabla g(\mathbf{H})$ of the sparsity criterion computed at $\mathbf{H}$ is written as a subtraction of two element-wise nonnegative terms $\nabla g(\mathbf{H}) = \nabla \mathbf{g}^+(\mathbf{H}) - \nabla \mathbf{g}^-(\mathbf{H})$. At each iteration, the columns of the base matrix $\mathbf{W}$ are normalized to have unity Euclidean norm.

2.3. Supervised and Semi-Supervised NMF

In our work, we perform speech separation by using supervised and semi-supervised NMF. In the supervised case, the spectral base matrix $\mathbf{W}$ is learned *a priori* from training data. This learning consists in applying unsupervised NMF to two different training sequences containing isolated speech and noise respectively. The matrix $\mathbf{W}$ is built by concatenating the two resulting basis matrices. Then, this matrix is kept constant during the separation phase, and only the matrix $\mathbf{H}$ is updated, according to (3) or (5).

In the semi-supervised case, only the columns of $\mathbf{W}$ corresponding to speech are learned and kept fixed for the separation. The second part of the matrix is randomly initialized and updated on each recording according to (2).

3. SPARSITY CRITERIA

We consider several different criteria for enforcing the sparsity of the matrix $\mathbf{H}$, whose characteristics are summed up in Table 1.

3.1. Norm-Based Criteria

The "natural" way of measuring sparsity is to count the number of non-zero components. However the resulting metric, known as the L-0 norm, is not differentiable and thus it often leads to intractable (NP-hard [19]) optimization problems. Thus, most of the sparsity measures used in the literature approximate this function by other norms.

The first function, already used in [14], is the sum of the L1-norm of the columns of $\mathbf{H}$. This boils down to the sum of all the values of the matrix. This function is denoted by L1.

The second criterion, which we call Row-Normalized L1-norm (RNL1), was proposed by [15]. It is equal to the sum of the L1-norms of the rows of $\mathbf{H}$, which are normalized by their Euclidean norm.

The Column-Normalized L1-norm (CNL1), used in [16], is similar to the previous criterion. However, the columns of $\mathbf{H}$ are normalized to have unity Euclidean norm.

For a sparsity criterion whose behavior is closer to the L-0 norm, we also exploit the L1/2 quasi-norm which is given by the definition of a p-norm with $p = \frac{1}{2}$. Although this function is not convex, it is differentiable and can be used with our iterative optimization approach.

3.2. Wiener Entropy Criteria

The Wiener Entropy, also called *spectral flatness*, of a set of nonnegative values is the ratio between the geometric mean and the arithmetic mean of the values. It is always between zero and one, and is maximal when all the values in the set are equal. Intuitively, a large value of the Wiener entropy corresponds to a "flat" plot and a small value corresponds to a "peaky" plot. Hence, to enforce the sparsity property of the NMF decomposition, it has to be ensured that the Wiener entropy of each column is small. Thus we use the sum of the Wiener entropy of the columns of $\mathbf{H}$ as sparsity measure. Note that in practice, a small positive number is added to the values of $\mathbf{H}$, in order to ensure that the penalty term is positive and that sparsity is enforced even when one component is equal to zero.

The value of the Wiener entropy is scale-independent, since it is the ratio of two means. However, the Kullback-Leibler divergence used as reconstruction error is dependent on the scale and for a nonnegative real number α, we have

$$D_{\mathrm{KL}}(\alpha\mathbf{V},\alpha\mathbf{WH}) = \alpha D_{\mathrm{KL}}(\mathbf{V},\mathbf{WH}). \tag{6}$$

Thus, in order to make the relative orders of magnitude of the sparsity term and the reconstruction error approximately constant, one can weight the Wiener entropies of the columns of $\mathbf{H}$ by their arithmetic mean. This reduces to using the geometric mean as sparsity criterion.

4. EXPERIMENTS

4.1. Evaluation Databases

The methods are evaluated on mixtures of speech and noise from publicly available corpora. We use spontaneous speech from the Buckeye corpus [20] to reflect use cases such as speech enhancement in wideband telephone channels or multimedia retrieval in web videos. Furthermore, to simulate the influence of various noise types, we consider (i) the CHiME 2011 Challenge [21] background noise corpus as an example for realistic noise recorded in a domestic environment that contains both stationary and non-stationary noises; (ii) the 'Twenty Years on MTV' collection of popular music as non-stationary 'noise'; and (iii) the NOISEX database [22] for mostly stationary, environmental noise. The MTV collection consists of 200 songs covering the years from 1981 to 2000 as well as various genres from hip-hop to country music, and featuring male as well as female singers. Other types of noise are gathered to build a 17 min training sequence, composed of noise recordings from the SiSEC 2010 noisy speech database[1] as well as the SPIB noise database[2] and street noise from the *soundcities* website[3]. All data are converted to 16 kHz sampling rate, monophonic audio.

As evaluation data, we use 80 test sentences from the 40 speakers of the Buckeye corpus (two from each speaker). These are mixed with random recordings of each of the three noise databases at SNRs between -9 dB and 12 dB. This results in 240 test files.

[1] http://sisec2010.wiki.irisa.fr/tiki-index.php?page=Source+separation+in+the+presence+of+real-world+background+noise

[2] http://spib.rice.edu/spib/select_noise.html

[3] http://www.soundcities.com

Acronym	Name	$g(\mathbf{H})$	$\left[\nabla g^{+}(\mathbf{H})\right]_{i,j}$	$\left[\nabla g^{-}(\mathbf{H})\right]_{i,j}$
L1	L1-norm	$\sum_{j=1}^{n}\sum_{i=1}^{r} h_{i,j}$	1	0
RNL1	Row-Normalized L1-norm	$\sum_{i=1}^{r}\frac{\sum_{j=1}^{n} h_{i,j}}{\sqrt{\sum_{k=1}^{n} h_{i,k}^2}}$	$\frac{1}{\sqrt{\frac{1}{n}\sum_{k=1}^{n} h_{i,k}^2}}$	$\frac{\sqrt{N}h_{i,j}\sum_{k=1}^{n} h_{i,k}}{\left(\sum_{k=1}^{n} h_{i,k}^2\right)^{3/2}}$
CNL1	Column-Normalized L1-norm	$\sum_{j=1}^{n}\frac{\sum_{i=1}^{r} h_{i,j}}{\sqrt{\sum_{k=1}^{r} h_{k,j}^2}}$	$\frac{1}{\sqrt{\frac{1}{n}\sum_{k=1}^{r} h_{k,j}^2}}$	$\frac{\sqrt{N}h_{i,j}\sum_{k=1}^{r} h_{k,j}}{\left(\sum_{k=1}^{r} h_{k,j}^2\right)^{3/2}}$
L1/2	L1/2 quasi-norm	$\sum_{j=1}^{n}\left(\sum_{i=1}^{r}\sqrt{h_{i,j}}\right)^2$	$\frac{\sum_{k=1}^{r}\sqrt{h_{k,j}}}{\sqrt{h_{i,j}}}$	0
WE	Wiener Entropy	$\sum_{j=1}^{n}\frac{\left(\prod_{i=1}^{r} h_{i,j}\right)^{\frac{1}{r}}}{\frac{1}{r}\sum_{i=1}^{r} h_{i,j}}$	$\frac{\left(\prod_{k=1}^{r} h_{k,j}\right)^{\frac{1}{r}}}{h_{i,j}\sum_{k=1}^{r} h_{k,j}}$	$\frac{r\left(\prod_{k=1}^{r} h_{k,j}\right)^{\frac{1}{r}}}{\left(\sum_{k=1}^{r} h_{k,j}\right)^2}$
GM	Geometric Mean	$\sum_{j=1}^{n}\left(\prod_{i=1}^{r} h_{i,j}\right)^{\frac{1}{r}}$	$\frac{\left(\prod_{k=1}^{r} h_{k,j}\right)^{\frac{1}{r}}}{h_{i,j}}$	0

Table 1. Sparsity criteria and nonnegative decomposition of the gradient used.

4.2. Experimental Setup

NMF is applied to magnitude spectrograms computed using Hamming windows of 32 ms length with 50 % overlap.

Two types of separation methods are considered. The first one consists in supervised NMF, which is well-suited for on-line processing [23, 24]. Speaker-dependent speech dictionaries are learned using sparsity-penalized NMF from a 1-minute set of (clean) utterances that is disjoint from the set of test utterances. The noise bases are learned from the training noise sequence. The number of components for each source is set to $r = 25$. In the second experiment, the separation is done by semi-supervised NMF, using the same speech dictionaries. The noise bases are estimated by the NMF separation algorithm, where the number of noise components is set to 8.

In the separation step, we perform a fixed number K of update iterations, where K is chosen from $\{1, 2, 4, \ldots, 128\}$. For each sparsity criterion, several values of the parameter λ (4) are tested.

We measure the performance of speech enhancement in terms of energy-based measures—source-distortion ratio (SDR), source-interference ratio (SIR) and source-artifact ratio (SAR) [25]—and the Covl, Csig and Cbak measures [26] estimating mean opinion scores (MOS) of overall perceptual quality, perceived quality of the wanted signal and perceived quality of the interference signal, on a scale from 1–5. The obtained scores are compared to the ones corresponding to the original noisy signal and the baseline system, constituted by a standard unconstrained NMF system. Statistical significance tests are performed by paired-sample t-tests.

4.3. Results

In a first experiment, we compare two NMF approaches for the separation. In the "constrained separation" approach, the separation is performed with the same sparsity-constrained NMF algorithm as was used for the learning of the bases. In the "unconstrained separation" approach, the bases are also learned with sparsity penalty, but the separation uses the standard NMF algorithm as outlined in Section 2.1.

Fig. 1 displays the average overall perceptual measures (Covl) obtained by supervised NMF with each of the tested sparsity criteria after $K = 32$ iterations. This parameter has been found to be the optimal number of iterations on this experiment. For every setting, the unconstrained approach delivers better Covl scores. This does not depend on the number of NMF iterations. Even if the constrained separation leads to some small SDR increases (up to 0.4 dB) using the semi-supervised NMF with some specific settings, the perceptual score is not improved in any of the tested systems.

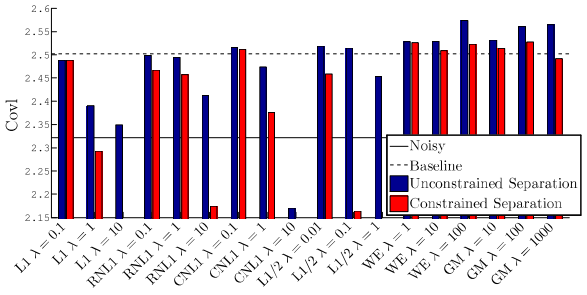

Fig. 1. Average Covl measures obtained with unconstrained and constrained supervised NMF, after $K = 32$ iterations. The non-represented bars correspond to scores lower than 2.15.

These results indicate that, in the case of incomplete dictionaries as considered in this study, applying a sparsity penalty term in the separation phase does not really improve the speech separation quality. On the contrary, it can interfere with the accurate modeling of the signal and deteriorate the quality of the resulting signals. Hence, the rest of the experiments will only focus on the unconstrained separation approach.

Fig. 1 and 2 displays the average Covl and SDR measures obtained with supervised NMF after $K = 32$ iterations. It can be observed that the L1 and RNL1 sparsity criteria actually degrades the separation results compared to the baseline unconstrained NMF, in terms of both used metrics. The other tested systems significantly improve the SDR (except for the CNL1 with $\lambda = 10$, where $p > 0.001$). According to this metric, the best setting is the L1/2 criterion with $\lambda = 1$ (5.7 dB against 4.2 dB for the baseline). However, the corresponding Covl is lower than the baseline (2.45 against 2.50). Indeed,

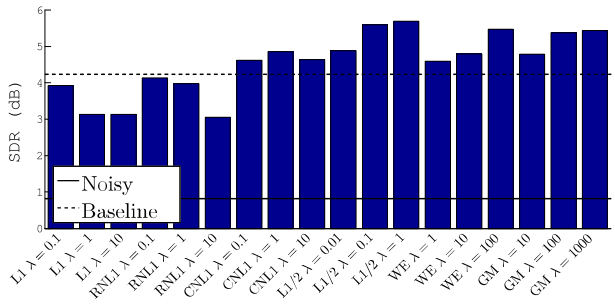

Fig. 2. Average SDR obtained with supervised NMF after $K = 32$ iterations.

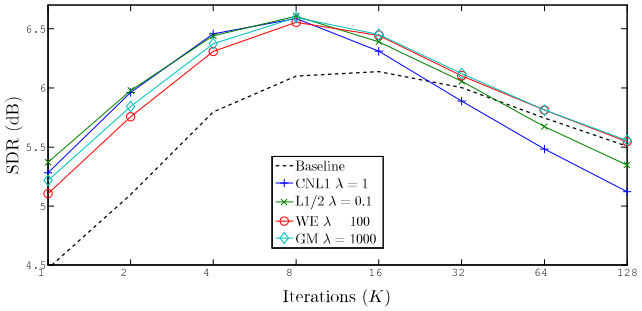

Fig. 3. Average SDR obtained with semi-supervised NMF as a function of the number of iterations. Sparsity constraint used in speech base learning only. Average SDR of the noisy samples: 0.8 dB.

because of the very high level of sparsity imposed during the training, the learned base is composed of "broadband spectra", which do not discriminate well speech and noise.

However, the Wiener entropy-based criteria allow for a significant increase of perceptual score (p<0.001). The best average Covl of 2.57 is obtained by the WE system with $\lambda = 100$ which also exhibits a high SDR (5.5 dB). This shows that the proposed criteria allow for an efficient learning of the speech and noise dictionaries.

With both supervised and semi-supervised NMF, we observed that the L1 and RNL1 criteria do not significantly improve the overall quality of the speech separation. These criteria also favors "temporal sparsity", according to which the proportion of the frames where a component is active is small, and this property may not be well suited to speech signals.

For each of the other sparsity criteria, we selected the value of λ which yielded the best results. The average SDR and Covl scores obtained are displayed in Figs. 3 and 4. We can first observe that the separation quality is consistently better than previously, thanks to a better estimation of the noise dictionaries. Besides, the optimal number of iterations is modified. For supervised NMF, it was found to be $K = 32$ independently of the setting, whereas in the semi-supervised case the best SDR is obtained after only 8 iterations.

Another observation is that the curves corresponding to the two metrics do not exhibit exactly the same behavior: The SDR obtains its declines after 8 iterations, while the Covl continues to increase until about 32 iterations. The decrease of SDR is due to the addition of artifacts (the SAR decreases, because of an overfitting phenomenon), which is not balanced by the suppression of noise. However, these artifacts become perceptually disturbing only after some further iterations. An interesting property of the considered sparsity criteria, is

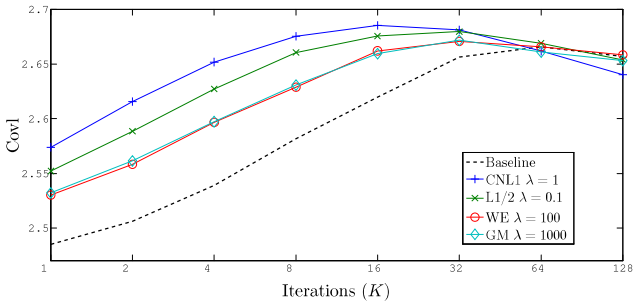

Fig. 4. Average Covl obtained with semi-supervised NMF as a function of the number of iterations. The average score of the noisy recordings is 2.32.

System	Noisy	Baseline	CNL1	L1/2	WE	GM
λ	–	–	1	0.1	100	1000
K	–	64	16	32	32	32
SDR (dB)	0.8	5.7	6.3	6.1	6.1	6.2
SIR (dB)	0.8	9.4	10.8	10.9	10.3	10.4
SAR (dB)	∞	10.4	10.3	9.8	10.2	10.3
Covl	2.32	2.67	2.69	2.68	2.67	2.67
Csig	2.75	3.20	3.22	3.22	3.20	3.20
Cbak	2.10	2.50	2.55	2.55	2.54	2.54

Table 2. Average scores obtained with semi-supervised NMF.

that they "accelerate" the separation process compared to the baseline NMF, which attains the maximum SDR after 16 iterations and the maximum Covl after 64 iterations. Thus, a similar (or even better) separation quality can be obtained with half the number of iterations.

The scores obtained with the selected systems after the optimal number of iterations (according to the Covl metric) are summarized in Table 2. All the systems using sparsity constraints obtain higher SDRs and SIRs as the baseline setting, while the loss in SAR is not statistically significant (p>0.001), with the exception of L1/2. They also allow for an improvement of the perceptual measures, although the increase in Covl is not significant ($p > 0.001$).

5. CONCLUSION

We have studied the influence of several common forms of sparsity penalties in the context of speaker-dependent speech separation by NMF. We have also introduced two novel criteria based on the Wiener entropy function. The results of an evaluation on spontaneous speech corrupted by a wide range of noise showed that the use of a sparsity penalty in the separation phase was not useful. However, spectral bases learned using the proposed criteria delivered the highest separation quality in the case of supervised NMF. In semi-supervised NMF experiments, several sparsity criteria provided slightly better results than the baseline NMF, with the advantage of a faster convergence to the optimal quality.

6. RELATION TO PRIOR WORK

Several sparsity criteria have been proposed in the literature for source separation using NMF [15, 14, 16]. However, to the authors' knowledge, no systematic analysis on the usefulness of these penalties have been conducted. We compare several common approaches, along with two novel criteria, on a large scale evaluation of speaker-dependent speech separation.

7. REFERENCES

[1] Cyril Joder, Felix Weninger, Florian Eyben, David Virette, and Björn Schuller, "Real-time speech separation by semi-supervised nonnegative matrix factorization," in *Proc. of Inter. Conf Latent Variable Analysis and Signal Separation (LVA/ICA)*, Tel Aviv, Israel, Mar. 2012.

[2] Nilesh Madhu, Ann Spriet, Sofie Jansen, Raphael Koning, and Jan Wouters, "The potential for speech intelligibility improvement using the ideal binary mask and the ideal Wiener filter in single channel noise reduction systems: Application to auditory prostheses," *IEEE Trans. Audio, Speech, Language Processing*, vol. 21, no. 1, pp. 63–72, Jan. 2013.

[3] Ron W. Weiss and Dan P. W. Ellis, "Speech separation using speaker-adapted eigenvoice speech models," vol. 24, no. 1, pp. 16–29, Jan. 2010, Speech Separation and Recognition Challenge.

[4] Ji Ming, Timothy J. Hazen, James R. Glass, and Douglas A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. Audio, Speech, Language Processing*, vol. 15, no. 5, pp. 1711–1723, June 2007.

[5] Felix Weninger, Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi, "Recognition of nonprototypical emotions in reverberated and noisy speech by nonnegative matrix factorization," *Journal on Advances in Signal Processing, Special Issue on Emotion and Mental State Recognition from Speech*, vol. 2011, 2011, Article ID 838790, 16 pages.

[6] Daniel D. Lee and H. Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[7] Kevin W. Wilson, Bhiksha Raj, Paris Smaragdis, and Ajay Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. of IEEE Inter. Conf. Acoust. Speech, Signal Processing (ICASSP)*, Las Vegas, NV, USA, Apr. 2008, pp. 4029 –4032.

[8] Zhiyao Duan, Gautham J. Mysore, and Paris Smaragdis, "Speech enhancement by online non-negative spectrogram decomposition in non-stationary noise environments," in *Proc. of Interspeech*, Portland, OR, USA, Sept. 2012.

[9] Paris Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, Language Processing*, vol. 15, no. 1, pp. 1–12, Jan. 2007.

[10] Tuomas Virtanen, A. Taylan Cemgil, and Simon Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proc. of IEEE Inter. Conf. Acoust. Speech, Signal Processing (ICASSP)*, Las Vegas, NV, USA, Apr. 2008, pp. 1825–1828.

[11] Madhusudana V.S. Shashanka and Paris Smaragdis, "Probabilistic latent variable models as non-negative factorizations," *Special issue on Advances in Non-negative Matrix and Tensor Factorization, Computational Intelligence and Neuroscience Journal*, vol. 2008, May 2008, Article ID 947438, 8 pages.

[12] Jean-Louis Durrieu, Bertrand David, and Gaël Richard, "A musically motivated mid-level representation for pitch estimation and musical audio source separation," *IEEE J. Select. Topics Signal Processing*, vol. 5, no. 6, pp. 1180 –1191, Oct. 2011.

[13] Julian Eggert and Edgar Korner, "Sparse coding and NMF," in *Proc. IEEE International Joint Conference on Neural Networks*, July 2004, vol. 4, pp. 2529–2533.

[14] Andrzej Cichocki, Rafal Zdunek, Anh Huy Phan, and Shun-Ichi Amari, *Nonnegative Matrix and Tensor Factorizations*, Wiley & Sons, 2009.

[15] Tuomas Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Language Processing*, vol. 15, no. 3, pp. 1066 –1074, march 2007.

[16] Patrik O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, Dec. 2004.

[17] Felix Weninger and Björn Schuller, "Optimization and parallelization of monaural source separation algorithms in the openBliSSART toolkit," *Journal of Signal Processing Systems*, vol. 69, no. 3, pp. 267–277, Dec. 2012.

[18] Daniel D. Lee and H. Sebastian Seung, "Algorithms for non-negative matrix factorization," in *Proc. of NIPS Conf.*, Denver, CO, USA, Oct. 2000, pp. 556–562, MIT Press.

[19] Morten Mørup, Kristoffer Hougaard Madsen, and Lars Kai Hansen, "Approximate l0 constrained non-negative matrix and tensor factorization," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, Seattle, WA, USA, May 2008, pp. 1328–1331.

[20] Mark A. Pitt, Laura Dilley, Keith Johnson, Scott Kiesling, and William Raymond, *Buckeye Corpus of Conversational Speech (2nd release)*, Department of Psychology, Ohio State University (Distributor), Columbus, OH, USA, 2007, [www.buckeyecorpus.osu.edu].

[21] Heidi Christensen, Jon Barker, Ning Ma, and Phil Green, "The CHiME corpus: a resource and a challenge for computational hearing in multisource environments," in *Proc. of Interspeech*, Makuhari, Japan, 2010, pp. 1918–1921.

[22] Andrew Varga and Herman J.M. Steeneken, "Assessment for automatic speech recognition: II. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, July 1993.

[23] Nasser Mohammadiha, Timo Gerkmann, and Arne Leijon, "A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization," in *Proc. of IEEE Workshop Applicat. of Signal Processing to Audio and Acoust (WASPAA)*, New Paltz, NY, USA, Oct. 2011, pp. 45–48.

[24] Nasser Mohammadiha, Jalil Taghia, and Arne Leijon, "Single channel speech enhancement using bayesian NMF with recursive temporal updates of prior distributions," in *Proc. of IEEE Inter. Conf. Acoust. Speech, Signal Processing (ICASSP)*, Kyoto, Japan, Mar. 2012, pp. 4561–4564.

[25] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Processing*, vol. 14, no. 4, pp. 1462 –1469, July 2006.

[26] Yi Hu and Philipos C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Processing*, vol. 16, no. 1, pp. 229–238, Jan. 2008.