# Real-Time Speech Separation by Semi-supervised Nonnegative Matrix Factorization

Cyril Joder[1], Felix Weninger[1], Florian Eyben[1],
David Virette[2], and Björn Schuller[1,⋆]

[1] Institute for Human-Machine Communication,
Technische Universität München, 80333 München, Germany
[2] HUAWEI Technologies Düsseldorf GmbH, Germany
cyril.joder@tum.de

**Abstract.** In this paper, we present an on-line semi-supervised algorithm for real-time separation of speech and background noise. The proposed system is based on Nonnegative Matrix Factorization (NMF), where fixed speech bases are learned from training data whereas the noise components are estimated in real-time on the recent past.

Experiments with spontaneous conversational speech and real-life non-stationary noise show that this system performs as well as a supervised NMF algorithm exploiting noise components learned from the same noise environment as the test sample. Furthermore, it outperforms a supervised system trained on different noise conditions.

## 1 Introduction

Isolating speech from environmental noise remains a challenging problem, especially in the presence of highly non-stationary noise such as background speech or music. On the other hand, a great variety of applications could benefit from a robust separation of speech, such as telephony, automatic speech recognition or hearing aids. In the case of telephony, additional constraints have to be taken into account, since usually only one microphone is available and the separation has to be performed in a real-time, on-line framework with a very small latency between audio input and output, in order to preserve natural communication.

One of the most popular approaches for single-channel source separation is Nonnegative Matrix Factorization (NMF) [3]. It has been shown efficient for speech separation [14,13], when both speech and noise models where learned prior to the separation. In [9], a variant of this algorithm is used, in which only one source is learned, the other being estimated from the mixture. However, this estimation requires off-line processing, where the whole signal is known.

Some studies have considered adapting the NMF algorithm to an incremental, on-line framework. In [11], pattern learning from large amounts of audio data using an on-line version of (convolutive) NMF is discussed. In [1], NMF is used

⋆ The research leading to these results has received funding from the HUAWEI Innovation Research Program 2010 (project GLASS).

F. Theis et al. (Eds.): LVA/ICA 2012, LNCS 7191, pp. 322–329, 2012.

to decompose a sequence formed by the new observation and the basis vectors, which are supposed to encompass the past observations. The approach of [12] and [6] first optimizes the activations for each coming observation, with fixed basis vectors, and then updates the bases based on the past activations. Still, we are not aware of a study on speech separation using on-line NMF algorithms.

In this work, we exploit a simple sliding window approach, where a classic NMF decomposition is performed on the recent past and the noise components are adapted in real-time to the current conditions. We test this semi-supervised on-line NMF method on a speech separation task with realistic data. Results show that the obtained system performs as well as a supervised NMF trained on the same noise environment, with a setting allowing for real-time capabilities.

After presenting the general NMF method in Section 2, we outline the proposed on-line NMF algorithms in Section 3. Then, experiments are detailed in Section 4, before drawing some conclusions.

## 2 Nonnegative Matrix Factorization (NMF) for Source Separation

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 $\mathbf{H}$. For example, with the generalized Kullback-Leibler divergence:

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

the update rules proposed by [5] are as follows:

$$\mathbf{W} \leftarrow \mathbf{W} \cdot \frac{\frac{\mathbf{V}}{\mathbf{WH}}\mathbf{H}^T}{\mathbf{1}\mathbf{H}^T} \tag{2}$$

$$\mathbf{H} \leftarrow \mathbf{H} \cdot \frac{\mathbf{W}^T \frac{\mathbf{V}}{\mathbf{WH}}}{\mathbf{W}^T}, \tag{3}$$

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 equation:

$$\mathbf{V_k} = \mathbf{V} \cdot \frac{\mathbf{W_k}\mathbf{H_k}}{\mathbf{WH}}. \tag{4}$$

This spectrogram is then inverted using the phase of the original mixture.

## 3 On-Line NMF

In this paper, on-line NMF refers to a sliding window method which decomposes the spectrum of the recent past into matrices $\mathbf{W}$ and $\mathbf{H}$ as detailed above. The sliding window contains the recent past spectra of the signal. Once a new frame is received, the sliding window is shifted by one frame. The activation matrix $\mathbf{H}$ is also shifted and the new column is initialized randomly. The matrices are then updated using a fixed number of NMF iterations. By using a sliding window approach, context information (in particular, the activations of the older frames) is available for this update so that a low number of iterations is sufficient.

### 3.1 On-Line Supervised NMF

In order to exploit the NMF decomposition for a practical source separation task, one needs to determine the source corresponding to each part of the decomposition. For our supervised algorithm, we assume that the sources which are to be separated are known in advance. This can correspond, for example, to the case of a teleconference, where several known people can talk simultaneously. We can then perform a learning of the characteristics of each isolated source. Hence, a spectral basis matrix is created for each considered source, using the (unsupervised) NMF decomposition of the learning data, as in [8].

For the separation phase, the $\mathbf{W}$ matrix is built by concatenating the basis matrices of the isolated sources. This matrix is kept constant and only the activation matrix $\mathbf{H}$ is updated using eq. (2). This particular case is straightforward to implement in an on-line system. This is because the update rule for each column $\mathbf{h}_{:,t}$ of $\mathbf{H}$ can be rewritten as:

$$\mathbf{h}_{:,t} \leftarrow \mathbf{h}_{:,t} \cdot \frac{\mathbf{W}^T \frac{\mathbf{v}_{:,t}}{\mathbf{W}\mathbf{h}_{:,t}}}{\mathbf{W}^T \mathbf{1}_{:,t}}. \tag{5}$$

Thus, each column of the activation matrix can be updated independently of the others, using only the current observation spectrum $\mathbf{v}_{:,t}$. The obtained factorization is then equivalent to an off-line version of supervised NMF.

### 3.2 On-Line Semi-supervised NMF

In the semi-supervised version of the on-line NMF algorithm, we consider that one source is unknown (modeling for example noise, or a new speaker). Thus, the spectral basis matrix $\mathbf{W}$ is no longer fully determined in advance. In the separation phase, the columns corresponding to the unknown source are initialized randomly, and updated with each new frame, following eq. (2). The other columns are kept constant.

With this semi-supervised algorithm, it is no longer possible to process each frame independently of the others, since the two matrices $\mathbf{W}$ and $\mathbf{H}$ depend on each other. Thus, the length of the sliding window — and thus of the amount of context information considered — does have an impact on the decomposition. Intuitively, a meaningful estimation of the 'noise spectral basis', i.e. the non-constant part of $\mathbf{W}$ , requires a whole sequence of observations.

### 3.3 Real-Time Implementation

For a real-time implementation of the on-line semi-supervised NMF, another important parameter is the *delay* parameter. It denotes the position in the sliding window of the frame to be output. If this parameter is equal to 0, only the past context is used for the NMF decomposition. By increasing this value, later observations can be considered. Moreover, the precision of the activations depends not only on the estimation of the matrix $\mathbf{W}$ (which can be controlled by the length of the sliding window), but also on the number of iterations that have been used for computation, which increases with the delay parameter.

The total latency $L$ introduced by the system (neglecting computation time) is then determined by the frame size $s$ and the delay parameter $d$, thanks to the relation: $L = (d + 1)s$. Note that the delay parameter is not relevant for the supervised algorithm, since the sliding window can be limited to the single current frame. Our implementation of the systems exploits the openSMILE [4] framework, which allows for an efficient incremental processing of audio data.

## 4 Experimental Evaluation

### 4.1 Experimental Settings

We evaluate the system on speech that was artificially mixed with real-life noise. Speech was taken from the Buckeye database [7], which contains recordings of interviews. The speech is highly spontaneous and contains a variety of non-linguistic vocalizations. Thus, we believe that this corpus is better suited for evaluation of speech separation in real-life conditions than, e. g., the popular TIMIT corpus of read speech, which is characterized by lower variation. We subdivided the Buckeye recording sessions, each of which is approximately 10 min long, into turns by cutting whenever the subject's speech was interrupted by the interviewer, or by a silence of more than 0.5 s length. Only the subject's speech is used. In these experiments, we only exploit turns of at least 3 s.

The test signals were then corrupted using noise recordings from the official corpus provided for the 2011 PASCAL CHiME Challenge [2]. These contain genuine recordings from a domestic environment obtained over a period of several weeks in a house with two small children. The noise is highly instationary due to abrupt changes such as appliances being turned on/off, impact noises such as banging doors, and interfering speakers [2]. All these data are publicly available[1]. The noise mixed with the speech was randomly drawn from the six hours of noise recordings in the database. We intentionally do not scale speech or noise to attain a distribution of noise levels corresponding to a real-life environment.

The sampling rate of the recordings is 16 kHz and the tested systems employ 32 ms analysis frames, with a 50 % overlap. In our experiments, we used 12 randomly chosen segments of speech, between 3 s and 20 s long. For each speech sample, a training sequence is created by concatenating 20 other speech segments from the same speaker, yielding lengths between 1.5 min and 5.5 min.

[1] `http://spandh.dcs.shef.ac.uk/projects/chime/PCC/datasets.html`

We constructed two different noise training sequences for supervised NMF. The first was created by concatenating 1024 short segments (0.5 s) drawn from diverse locations in the CHiME noise recordings. Hence, this training sequence contains most of the noise sources that can be found in the database. In order to assess the generalization property of the system to different types of noise, we also constructed another 17 min training sequence, composed of noise recordings from the SiSEC 2010 noisy speech database[2] as well as some extract of the SPIB noise database[3] and some street noise from the *soundcities* website[4]. These sequences are referred to as *matched* and *mismatched* training noise.

Several speech separation systems are tested here. All of them exploit constant basis components for speech, previously learned from the training sequence. The first two systems exploit the on-line supervised NMF algorithm presented in subsection 3.1, with noise components learned respectively from the matched and mismatched training noise. For these systems, the numbers of NMF components for speech and noise are equal to $c_s = c_n = 50$, which has been empirically found satisfactory for the speaker separation task. All the training processes use 256 iterations. The other system uses the on-line semi-supervised NMF algorithm of subsection 3.2, with $c_s = 50$ speech components. The tested values of the different parameters are displayed in Table 1. This values were chosen to maintain a limited computational complexity.

**Table 1.** Tested values of the parameters for the on-line semi-supervised NMF system

| | Parameter | Tested Values |
|---|---|---|
| $c_s$ | number of speech components | {50} |
| $c_n$ | number of noise components | {1,2,4,8,12,16} |
| $\ell$ | sliding window length | {2,4,6,8,12,16,20,25,30} |
| $d$ | delay | {0,1,2,3,4,5,6,7} |
| $n$ | number of optimization iterations | {1,2,4,8,16,32,64} |

Several evaluation criteria were computed from the separated speech: the Source to Distortion Ratio (SDR), the Source to Interference Ratio (SIR) and the Source to Artifact Ratio (SAR) [10]. For comparison of the on-line approach, we consider an 'optimal' off-line version of the semi-supervised NMF algorithm, which outperforms supervised NMF on our test data. For this system, 256 iterations are used and the number of noise components was chosen from the set $c_n \in \{1, 2, 4, 8, 12, 16, 20, 25, 30, 35, 40, 45, 50\}$. The value $c_n = 30$ is selected, maximizing the average SDR in the test database. This optimal SDR is equal to 5.2 dB, which represents the best result that can be achieved with basic NMF speech separation algorithms on our test data.

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

[3] `http://spib.rice.edu/spib/select_noise.html`

[4] `http://www.soundcities.com`

### 4.2 Results

The results obtained by the supervised NMF systems are displayed in Fig. 1. It can be observed that for both systems, the SIR increases with the number of iterations. However, this reduction of the interferences is at the cost of more artifacts, since the SAR concurrently decreases. The optimal trade-off is here realized for a single iteration, yielding a 4.2 dB SDR, against 0.6 dB for the original corrupted speech. Although the optimal number of iterations may be dependent on the data; this shows that a very small number of iterations is sufficient for a satisfactory separation. Thus, the obtained complexity is very low, achieving a real-time factor of 2 % on a 3.4 GHz, 64 bits CPU.

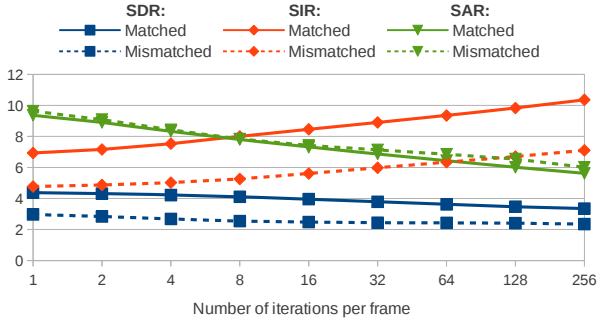


**Fig. 1.** Average source separation criteria (dB) for the supervised NMF systems, trained on the matched and mismatched noise

Our results also show the importance of an adequate noise model for the separation. Indeed, the supervised NMF systems are outperformed by the off-line semi-supervised algorithm, whose noise spectra seem to fit the observations even better, probably since they are estimated directly on each test sample. Furthermore, whereas the SARs of both supervised systems are roughly equivalent, the 'matched' noise training induces significantly higher SIRs (by over 2 dB) and thus a better separation quality.

Fig. 2 to 4 present a few of the numerous results of the on-line semi-supervised NMF system. The best SDR is equal to 4.4 dB that is slightly better than the result obtained with the supervised NMF, even with the 'matched' noise training. This shows the efficiency of the proposed method to adapt the noise model to the environment in an on-line framework.

The best score is obtained with the parameters $c_n = 8$, $\ell = 20$, $d = 0$ and $i = 1$ (see Table 1). Contrarily to the supervised case, Fig. 2 shows a degradation of the SIR when the number of iterations increases. This can be due to an 'overfitting' phenomenon, where the updated components tend to model speech as well as noise. One can see in Fig. 3 that, with a larger sliding window, the SIR decreases while the SAR is improved. This can be explained by the fact that the adaptation to the environment is then a bit less precise, but it is more robust to

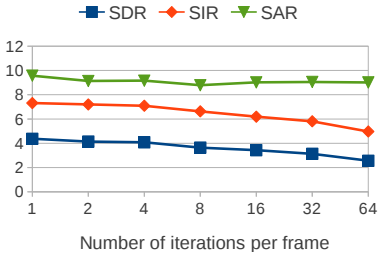


**Fig. 2.** Criteria (dB) as a function of $i$ for constant $c_n = 8$, $\ell = 20$ and $d = 0$

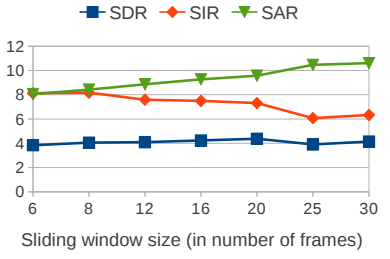


**Fig. 3.** Criteria (dB) as a function of $\ell$ for constant $c_n = 8$, $d = 0$ and $i = 1$

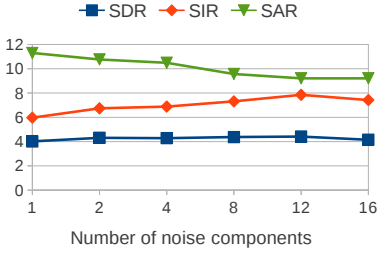


**Fig. 4.** Criteria (dB) as a function of $c_n$ for constant $l = 20$, $d = 0$ and $i = 1$

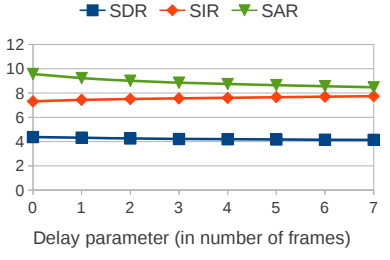


**Fig. 5.** Criteria (dB) as a function of $d$ for constant $c_n = 8$, $\ell = 20$ and $i = 1$

the overfitting phenomenon. The number of noise components seems to have the opposite influence (Fig. 4). Hence, the values $c_n = 8$ and $\ell = 20$, corresponding to a sliding window of 336 ms, here constitute a reasonable trade-off.

The delay parameter has only a small influence, as shown in Fig. 5. Thus, the value $d = 0$ can be chosen so as to minimize the latency of the system. Furthermore, the best-performing setting has a relatively low complexity, since only one iteration is performed for each frame. The real-time factor then is 20 %, on the aforementioned CPU. Therefore, the system is fully real-time capable.

## 5 Conclusion

We presented a method for on-line speech separation exploiting a sliding window version of the semi-supervised Nonnegative Matrix Factorization algorithm. An extensive experimental study has been conducted, testing numerous parameter combinations. Our results show that this system performs similarly to (and even slightly better than) a supervised algorithm in which the noise components are learned from the same environment as the test samples. Furthermore, the optimal setting yields a system which is real-time capable on a recent PC.

Among the future works for further improvements of the system can be the introduction of regularization terms such as priors [14] or sparsity and continuity

constraints, in order to obtain more meaningful components in both learning and separation phases without considerably affecting the complexity. The use of a small-order Nonnegative Matrix Deconvolution algorithm [8] could also be explored, although at the cost of increased latency and computational complexity. Finally, the observed behavior depending on the number of iterations motivates introduction of relaxation [3] into the multiplicative update algorithm.

## References


1. Cao, B., Shen, D., Sun, J.T., Wang, X., Yang, Q., Chen, Z.: Detect and track latent factors with online nonnegative matrix factorization. In: Proceedings of the 20th Intern. Joint Conf. on Artifical Intelligence (IJCAI 2007), pp. 2689–2694 (2007)
2. Christensen, H., Barker, J., Ma, N., Green, P.: The CHiME corpus: a resource and a challenge for Computational Hearing in Multisource Environments. In: Proc. of Interspeech, Makuhari, Japan, pp. 1918–1921 (2010)
3. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.I.: Nonnegative Matrix and Tensor Factorizations. Wiley & Sons (2009)
4. Eyben, F., Wöllmer, M., Schuller, B.: openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. In: Proc. of ACM Multimedia, pp. 1459–1462 (2010)
5. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)
6. Lefèvre, A., Bach, F., Févotte, C.: Online algorithms for Nonnegative Matrix Factorization with the Itakura-Saito divergence. In: Proc. of IEEE Workshop on Applications of Signal Process. to Audio and Acoustics (WASPAA), pp. 313–316 (2011)
7. Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., Fosler-Lussier, E.: Buckeye Corpus of Conversational Speech (2nd release). Department of Psychology, Ohio State University (Distributor), Columbus, OH, USA (2007), `www.buckeyecorpus.osu.edu`
8. Smaragdis, P.: Convolutive speech bases and their application to supervised speech separation. IEEE Trans. Audio, Speech and Language Process. 15(1), 1–14 (2007)
9. Smaragdis, P., Raj, B., Shashanka, M.: Supervised and Semi-Supervised Separation of Sounds from Single-Channel Mixtures. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 414–421. Springer, Heidelberg (2007)
10. Vincent, E., Gribonval, R., Févotte, C.: Performance measurement in blind audio source separation. IEEE Transactions on Audio, Speech, and Language Processing 14(4), 1462–1469 (2006)
11. Wang, D., Vipperla, R., Evans, N.: Online Pattern Learning for Non-Negative Convolutive Sparse Coding. In: Proc. of Interspeech, pp. 65–68 (2011)
12. Wang, F., Li, P., König, A.C.: Efficient document clustering via online nonnegative matrix factorizations. In: SDM, pp. 908–919. SIAM / Omnipress (2011)
13. Weninger, F., Geiger, J., Wöllmer, M., Schuller, B., Rigoll, G.: The Munich 2011 CHiME Challenge Contribution: NMF-BLSTM Speech Enhancement and Recognition for Reverberated Multisource Environments. In: Proc. Internat. Workshop on Machine Listening in Multisource Environments (CHiME 2011), pp. 24–29 (2011)
14. Wilson, K.W., Raj, B., Smaragdis, P., Divakaran, A.: Speech denoising using nonnegative matrix factorization with priors. In: IEEE Intern. Conf. on Acoustics, Speech and Signal Process. pp. 4029–4032 (2008)