Sparse, Hierarchical and Semi-Supervised Base Learning for Monaural Enhancement of Conversational Speech

Felix Weninger[1], *Martin Wöllmer and Björn Schuller*

Institute for Human-Machine Communication, Technische Universität München, Germany
Email: {weninger,woellmer,schuller}@tum.de
Web: www.mmk.ei.tum.de

Abstract

We address the learning of noise bases in a monaural speaker-independent speech enhancement framework based on non-negative matrix factorization. Bases are estimated from training data in batch processing by means of hierarchical and non-hierarchical sparse coding, or determined during the speech enhancement process based on the divergence of the observed noisy speech signal and the speech base. In extensive test runs on the Buckeye corpus of highly spontaneous speech and the CHiME corpus of non-stationary real-life noise, we observe that semi-supervised learning of noise bases leads to overall best results while a-priori learning of noise bases is useful to speed up computation.

1 Introduction

Automatic speech recognition (ASR) in many realistic scenarios, including hands-free natural human-computer interaction and multimedia retrieval, has to deal with interfering sources as well as large variability of spontaneous speech. Furthermore, in many situations, such as analysis of on-line videos, only one audio channel is available. To enhance robustness of ASR with a monaural front-end, effective techniques for speech source separation based on non-negative matrix factorization (NMF) have been proposed, whose applicability has been demonstrated in cross-talk separation [1] and noise suppression [2].

Such NMF-based methods for monaural speech and noise separation use a factorization of the observed spectrogram into the product of speech and noise dictionaries with non-negative activations. Thus, optimizing the process of learning these dictionaries from training data is crucial for the separation results. In contrast to the above-mentioned studies which consider rather artificial tasks, particularly read and/or small vocabulary speech, noise with limited variability, and speaker- and noise-dependent dictionaries, we aim towards realistic use cases by addressing spontaneous conversational speech from the Buckeye database in highly variable noise from a domestic environment as featured in the 2011 PASCAL CHiME Challenge [3]. To cope with the large variation of speech as well as noise, we adapt the NMF methodology from our previous study on speaker dependent small vocabulary speech enhancement in CHiME noise [4] by considering speaker independent phoneme models, and investigating various methods to learn dictionaries from large amounts of variable, non-stationary noise. As a baseline method corresponding to our previous study [4] we employ data reduction by random noise subsampling and NMF. We then extend this method by introducing sparsity constraints. Further, random subsampling is replaced by a hierarchical strategy to take into account all available data. Finally, adaptive semi-supervised learning for NMF is evaluated, which has been introduced in [5] and evaluated in a study on speech and music separation in [2]. Thus, the main contribution of this study is to demonstrate the effectiveness of NMF based speech separation in a challenging task and to provide a large-scale comparative evaluation of noise base learning algorithms in a well defined experimental setup using publicly available databases and NMF implementations. Details of the experimental setup including the evaluation database are given in Section 3. Evaluation of speech enhancement is performed in Section 4 before concluding in Section 5.

2 Base Learning in NMF-based Speech Enhancement

NMF-based techniques for monaural speech enhancement as used in this study are based on the assumption that the wanted speech signal can be approximated as linear combinations of speech and noise dictionaries $\mathbf{W}^{(s)}$ and $\mathbf{W}^{(n)}$ with non-negative activation coefficients $\mathbf{H}^{(s)}$ and $\mathbf{H}^{(n)}$:

$$\mathbf{V} \approx \mathbf{\Lambda} = \mathbf{\Lambda}^{(s)} + \mathbf{\Lambda}^{(n)} := \mathbf{W}^{(s)}\mathbf{H}^{(s)} + \mathbf{W}^{(n)}\mathbf{H}^{(n)},$$

where $\mathbf{V}$ is an observed magnitude spectrogram of speech overlaid by interfering noise, and $\mathbf{\Lambda}^{(s)}$ and $\mathbf{\Lambda}^{(n)}$ denote approximations of the speech and noise parts, respectively. In the remainder of this paper, we assume that the speech basis $\mathbf{W}^{(s)}$ is fixed after estimation from training data as detailed in Section 3.2. Then, the following cost function is minimized by an iterative multiplicative update algorithm [6].

$$c(\mathbf{W}^{(n)}, \mathbf{H}) = c_r(\mathbf{W}^{(n)}, \mathbf{H}) + \lambda\, c_s(\mathbf{H}), \qquad (1)$$

where c_r corresponds to the reconstruction error (Kullback-Leibler divergence) and c_s is a *sparsity constraint* penalizing the L1 norm of $\mathbf{H}$. Informally, c_s enforces that only a few basis vectors can be active at a given time, which is reasonable if the basis vectors correspond to, e. g., phonemes, or spectra originating from different noise sources. Depending on whether the noise basis $\mathbf{W}^{(n)}$ is estimated a priori from training data and subsequently kept constant in the minimization of (1)—this is called *supervised NMF* in the ongoing—, or treated as a free parameter during speech enhancement, various base learning algorithms can be derived, which will be presented in detail below. A fixed number of NMF iterations K is applied starting from a (Gaussian) random solution. Generally, more iterations result in more precise modeling of the noisy original spectrogram in terms of the cost function, and often improve the noise suppression, but at the same time may lead to overadaptation (e. g., 'mis-use' of noise bases to model speech and vice versa), and hence introduce separation artifacts [7]. We will evaluate the influence of K in Section 4.

[1] The research leading to these results has been partly funded by the German Research Foundation (DFG) through grant no. SCHU 2502/2-1.

After the NMF iterations, an estimate of the clean speech spectrogram, $\widehat{\mathbf{V}}^{(s)}$, is obtained by filtering the observed spectrogram $\mathbf{V}$:

$$\widehat{\mathbf{V}}^{(s)} = \frac{\mathbf{\Lambda}^{(s)}}{\mathbf{\Lambda}} \otimes \mathbf{V}. \qquad (2)$$

All experiments for this paper are based on the NMF implementations found in our open-source toolkit openBliSSART [8] to enforce reproducibility of our results.

The remainder of this paper investigates different approaches to learn $\mathbf{W}^{(n)}$ by means of NMF. Thereby the dictionary size (number of columns in $\mathbf{W}^{(n)}$) is a crucial factor since the computational complexity of NMF is linear in the dictionary size. Thus, several studies [1, 9] propose to build these dictionaries by applying data reduction methods to a set of spectrograms from training data in batch processing. This data reduction can be carried out by clustering or NMF itself; the latter was found superior in [10]. If NMF is used for the data reduction step, it has to be taken into account that the basic algorithm requires to keep all training data in memory; hence, it cannot be directly applied to large amounts of training data. To cope with this, we propose random subsampling as in [1] as well as a hierarchical method.

Random Subsampling and NMF

For the CHiME task, we have proposed in [4] to concatenate a randomly selected subset of the training noise spectrograms into a matrix $\mathbf{T}^{(n)}$, then reduce it to a noise dictionary $\mathbf{W}^{(n)}$ of fixed size by means of NMF:

$$\mathbf{T}^{(n)} \rightsquigarrow \mathbf{W}^{(n)}\mathbf{H}^{(n)}.$$

1 000 noise exemplars are extracted and 100 NMF iterations are used for the reduction, minimizing $c(\mathbf{W}^{(n)}, \mathbf{H}^{(n)})$ (cf. Eqn. 1) by alternating updates of $\mathbf{W}^{(n)}$ and $\mathbf{H}^{(n)}$ as proposed in [6]. To provide a baseline in analogy to the experiments in [4], no sparsity constraint is used therein (i. e., $\lambda = 0$).

Sparse Coding by NMF

If the above-mentioned NMF process is applied without enforcing a sparsity constraint, we often observe that the resulting noise bases correspond to 'building blocks': Actual noise events in the signal are modeled by additive combinations of spectra which often only extend to certain frequency bands. In contrast, using $\lambda > 0$ forces $\mathbf{W}^{(n)}$ to model sparsely occurring events corresponding to actual noise sources, since modeling by additive combinations of dictionary atoms is penalized. These observations are in accordance with the ones made by [11] in the context of image processing. Thus, it can be conjectured that the atoms in a noise dictionary learnt by sparse coding are harder to 'mis-use' for modeling speech; this undesired behavior is one of the major drawbacks of basic NMF-based monaural speech separation and leads to separation artifacts: The part of the speech that is modeled by noise atoms will be suppressed in the filtering process (see Eqn. 2).

Hierarchical Decomposition

To cope with the computational complexity of reducing training noise to an incomplete dictionary, the above-mentioned methods introduce a rather ad-hoc strategy of random subsampling, thus neglecting large portions of the training noise. Thus, we now propose a hierarchical learning algorithm to take into account all available training noise while keeping the computational complexity low. This algorithm processes the training noise in B blocks and estimates a basis $\mathbf{W}^{(n,b)}$ from the spectrogram of block b, $\mathbf{T}^{(n,b)}$, $b = 1, \dots, B$ by means of NMF. 100 iterations are performed as in the above. Then, in a second step, the concatenation of the block-wise bases, $[\mathbf{W}^{(n,1)} \cdots \mathbf{W}^{(n,B)}]$ is reduced to the final base $\mathbf{W}^{(n)}$ by means of NMF. The dimensionality of the factorization in the first step is determined by a reduction factor ρ_h, which is the ratio of the block size in signal frames (number of columns of $\mathbf{T}^{(n,b)}$) and the size of the block-wise dictionaries (number of columns of $\mathbf{W}^{(n,b)}$). The second step NMF is used to eliminate redundancies in the dictionaries which could be present if the noise events in the blocks overlap (as will usually be the case in real-life recordings taken over a period of time, such as the CHiME noise corpus).

Adaptive Noise Learning

Finally, we use an *adaptive noise learning* algorithm that estimates a dictionary $\mathbf{W}^{(n)}$ during the speech enhancement process according to (1), starting from a random dictionary. In other words, no a priori information about the noise characteristics is used, but an optimal dictionary to model the noise in the utterance is estimated only based on the mismatch between the fixed speech dictionary and the observed spectrogram—as measured by the NMF cost function (Eqn. 1). This is a *semi-supervised* NMF approach as proposed, e. g., in [2, 5].

3 Experimental Setup

3.1 Evaluation Database

We used the Buckeye corpus [12] recorded in clean conditions and mixed with the CHiME noise corpus [3] to simulate spontaneous speech encountered in a noisy domestic environment at controlled noise levels. The Buckeye corpus contains recordings of interviews with 40 speakers. 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 segment of more than 0.5 s length. Only the subjects' speech is used, amounting to a total length of 26 hours. We use a speaker-independent subdivision into a training set (13 557 utterances from 32 speakers), development set (1 631 utterances from four speakers), and test set (1 985 utterances from four speakers), stratified by speaker age and gender.

The additive noise considered in this study is taken from the corpus of the 2011 PASCAL CHiME Challenge [3]. This corpus contains genuine recordings of highly non-stationary noise from a domestic environment obtained over a period of several weeks. To create the noisy version of our evaluation database, we followed the protocol which was used to create the CHiME Challenge ASR task [3]: In the development and test set, we employ six signal-to-noise ratios (SNRs) ranging from 9 dB down to -6 dB in steps of 3 dB by selecting matching noise segments from the CHiME development/test noise. As proposed in [3], the noisy utterances are not constructed by artificial scaling of the speech or noise amplitudes, but by choosing noise segments as they were recorded in a real life situation. This means that noisy utterances at low SNRs co-occur with noise that naturally has high energy, such as broad band impact noises. The SNRs are measured on first order differences of speech and noise signals. For the experiments reported in this paper, all signals are down-mixed to mono by averaging channels.

3.2 NMF Parameterization

To apply NMF, spectrograms of the signals are calculated by short-time Fourier Transform using Hann windows of 25 ms length at 10 ms frame shift. A shorter window size and frame shift than in our previous study on the small vocabulary CHiME Challenge ASR task [4] have been chosen to cope with higher variability of spontaneous conversational speech. To build speaker-independent speech bases for NMF, for each phoneme, the corresponding spectrograms are extracted from the Buckeye training set according to a forced alignment. These concatenated phoneme spectrograms are reduced to a single dictionary atom by a 1-component NMF, and the column-wise concatenation of these atoms constitutes the matrix $\mathbf{W}^{(s)}$. Thus, the number of speech components is equivalent to the number of phonemes (39). The advantage of such phoneme-dependent speech bases over unsupervisedly learnt ones has been shown in [1]. Noise bases are estimated from the CHiME training background noise [3]. The same set of randomly sampled noise spectrograms is used as in [4]. For hierarchical base learning, reduction factors of $\rho_h \in \{50, 500\}$ are tested, and blocks simply correspond to the noise audio files (each spanning 5 minutes). The sparsity weight is set to $\lambda = 0.1$ in sparse base learning and in all speech enhancement processes; for the latter, sparsity is used to guide the NMF algorithm towards a solution where only few phonemes and noise atoms can be activated at a time. This weight has been found optimal on the development set. The size of the noise dictionary is 40 except for adaptive base learning, where four components were found sufficient to model the noises actually occurring in the signal, based on preliminary experiments on the development set. The most influential parameter for the overall separation quality turned out to be the number of iterations K; an optimal value is selected from $\{1, 2, 4, 8, 16, 32\}$ based on the development set as will be described in the following section.

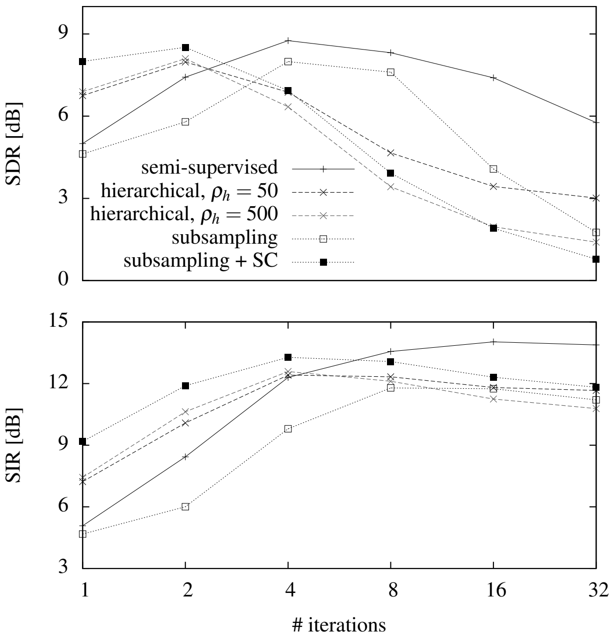
4 Results

Results are evaluated with the BSS_Eval [13] toolkit in terms of signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and signal-to-artifact ratio (SAR).

4.1 Development Set

Firstly, we evaluate the influence of the number of iterations K on the development set for the various base learning algorithms listed in Section 2. In order to reduce the computational effort for parameter tuning, a subset of the development set is used for these measurements, consisting of ten randomly selected utterances of each of the four speakers, with lengths between 5 s and 10 s, at six different SNRs (240 utterances in total). The average SDR, SIR, and SAR are computed across SNRs, and the value of K is optimized on SDR. In terms of average SDR (Figure 1, top), we observe different optima for different bases. The overall best quality (SDR = 8.8 dB) is obtained by adaptive base learning, using $K = 4$. However, 8.5 dB SDR can be achieved by the supervised NMF method with a sparsely learnt noise basis and $K = 2$ iterations updating only the activations, i. e., at significantly lower computational effort. It can be seen that the sparsely learnt noise basis helps to guide the factorization into the 'right direction' faster than the baseline noise basis, which achieves its optimum SDR for $K = 4$ iterations, at 8.0 dB. When performing more iterations, however, it is surprising that the noise bases estimated with sparsity introduce more artifacts (lower SDR) than the baseline: One would expect lower discrimination capabilities of the latter due to its genericity, as explained in Section 2. Generally, the performance of the hierarchical learning methods is disappointing, as they cannot outperform sparse coding from subsampled data except for high numbers of iterations where a general performance drop occurs. This might be due to the fact that the number of distinct noise sources in the CHiME dataset is rather small and thus spectrograms of most sources can be captured by subsampling. In contrast, adaptive noise learning by semi-supervised NMF seems to be most robust against over-adaptation (as it avoids the large drop in SDR for $K > 4$), and provides best interference reduction (SIR, cf. Figure 1, bottom) for $K > 4$; particularly, the optimum SIR is 14.0 dB at $K = 16$. These observations provide evidence that the method is able to estimate a noise base 'on-the-fly' which provides good discrimination from speech in the spectral domain. Again, the subsampling + SC method provides good interference reduction at much lower computational complexity (SIR = 13.3 dB at $K = 4$).

Figure 1: Buckeye development set: Average SDR (top) and SIR (bottom) with different noise base learning algorithms, across SNRs from -6 to 9 dB. SC = sparse coding ($\lambda = 0.1$).



4.2 Test Set

In Table 1, show the performance of the methods on the test set with the optimal number of iterations K as determined on the development set, in terms of SDR, SIR, and SAR. Best results in terms of SDR, SIR, and SAR are

Table 1: Evaluation on the noisy Buckeye test set. [1]Without separation, SIR = SDR and SAR → ∞.

		[dB]	SNR						
Base learning method	K		-6	-3	0	3	6	9	avg
— (no separation)	0	SDR[1]	0.7	2.5	4.3	5.7	7.0	8.2	4.7
subsampling	4	SDR	2.8	4.8	6.7	8.4	9.9	11.2	7.3
		SIR	3.9	6.1	8.0	9.8	11.5	12.8	8.7
		SAR	15.0	15.9	16.9	17.5	18.2	18.9	17.1
subsampling + SC	2	SDR	4.3	6.0	7.5	8.9	10.1	11.1	8.0
		SIR	6.2	8.1	9.8	11.5	13.1	14.5	10.5
		SAR	12.0	12.8	13.7	14.2	14.8	15.2	13.8
hierarchical, $\rho_h = 500$	2	SDR	4.2	5.7	7.0	8.2	9.2	10.1	7.4
		SIR	5.9	7.5	8.9	10.4	11.7	12.9	9.5
		SAR	11.6	12.5	13.3	13.7	14.2	14.5	13.3
semi-supervised	4	SDR	5.3	6.6	7.8	8.9	10.1	11.0	8.3
		SIR	8.1	9.3	10.5	11.8	13.2	14.5	11.2
		SAR	11.2	12.6	14.0	14.9	15.5	16.1	14.0

achieved by the semi-supervised method. The SDR difference to the second best result, achieved by the supervised method with a sparsely learnt noise basis, is significant ($p \ll .001$, 95 % confidence interval: $[0.22, 0.30]$, sample size $6 \times 1985 = 11\,910$) according to a two-sided paired t-test. Conversely, and mirroring the results on the development set, sparse coding significantly improves the performance of supervised NMF in terms of SDR (95 % confidence interval: $[0.67, 0.72]$) while non-sparse noise bases seem to introduce less artifacts. Overall, the SDR and SIR gain by semi-supervised over supervised NMF seems to be mostly due to improvements on the lower SNR end, corroborating the hypothesis put forth in [2] in a larger scale study. While—in our parameterization—semi-supervised NMF introduces more artifacts on the lower SNR end, interestingly, the speech quality loss at high SNR levels is lower than for supervised NMF. Finally, we highlight that the 95 % confidence interval for the SDR improvement in dB by semi-supervised NMF over the noisy baseline is $[3.4, 3.6]$ while it is $[6.4, 6.6]$ for the SIR improvement. We further note that all the methods are real-time capable in the sense of a real-time factor (RTF) $\ll 1$. For example, the semi-supervised NMF method, using the openBliSSART [8] implementation, factorizes the 11 910 test utterances with a total length of 15 hours in roughly two hours on a standard quad-core PC.

5 Conclusions and Future Work

We have shown effective and efficient methods to reduce non-stationary background noise in highly spontaneous speech. In particular, we have demonstrated the applicability of semi-supervised NMF in a challenging speech separation task and have shown that sparse base learning from training noise can be used to speed up the separation process significantly. Future work should address the close integration into an automatic speech recognition system, including techniques to mitigate separation artifacts and parameter tuning towards ASR accuracy, evaluation of the proposed methods in a truly on-line framework [7], and semi-supervised speaker adaptation.

References

[1] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. of Interspeech*, (Pittsburgh, PA, USA), 2006.

[2] F. Weninger, J. Feliu, and B. Schuller, "Supervised and Semi-Supervised Supression of Background Music in Monaural Speech Recordings," in *Proc. of ICASSP*, (Kyoto, Japan), pp. 61–64, 2012.

[3] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME corpus: a resource and a challenge for Computational Hearing in Multisource Environments," in *Proc. of Interspeech*, (Makuhari, Japan), pp. 1918–1921, 2010.

[4] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich 2011 CHiME Challenge Contribution: NMF-BLSTM Speech Enhancement and Recognition for Reverberated Multisource Environments," in *Proc. of CHiME Workshop*, (Florence, Italy), pp. 24–29, 2011.

[5] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. of ICA*, (London, UK), pp. 414–421, Springer, 2007.

[6] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. of NIPS*, (Vancouver, Canada), pp. 556–562, 2001.

[7] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller, "Real-time speech separation by semi-supervised nonnegative matrix factorization," in *Proc. LVA ICA, Special Session "Real-world constraints and opportunities in audio source separation"*, (Tel Aviv, Israel), pp. 322–329, Springer, 2012.

[8] F. Weninger, A. Lehmann, and B. Schuller, "openBliSSART: Design and Evaluation of a Research Toolkit for Blind Source Separation in Audio Recognition Tasks," in *Proc. of ICASSP*, (Prague, Czech Republic), pp. 1625–1628, 2011.

[9] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 1–14, 2007.

[10] X. Jaureguiberry, P. Leveau, S. Maller, and J. Burred, "Adaptation of source-specific dictionaries in non-negative matrix factorization for source separation," in *Proc. of ICASSP*, (Prague, Czech Republic), pp. 5–8, 2011.

[11] J. Eggert and E. Körner, "Sparse coding and NMF," in *Proc. of Neural Networks*, vol. 4, (Dalian, China), pp. 2529–2533, 2004.

[12] M. A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, *Buckeye Corpus of Conversational Speech (2nd release)*. Columbus, OH, USA: Department of Psychology, Ohio State University (Distributor), 2007. [www.buckeyecorpus.osu.edu].

[13] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.