Integrating noise estimation and factorization-based speech separation: a novel hybrid approach

Cyril Joder, Felix Weninger, David Virette, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Joder, Cyril, Felix Weninger, David Virette, and Björn Schuller. 2013. "Integrating noise estimation and factorization-based speech separation: a novel hybrid approach." 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, 131–35. Piscataway, NJ: IEEE. https://doi.org/10.1109/icassp.2013.6637623.

INTEGRATING NOISE ESTIMATION AND FACTORIZATION-BASED SPEECH SEPARATION: A NOVEL HYBRID APPROACH

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

We present a novel method to integrate noise estimates by unsupervised speech enhancement algorithms into a semi-supervised non-negative matrix factorization framework. A multiplicative update algorithm is derived to estimate a non-negative noise dictionary given a time-varying background noise estimate with a stationarity constraint. A large-scale, speaker-independent evaluation is carried out on spontaneous speech overlaid with the official CHiME 2011 Challenge corpus of realistic domestic noise, as well as music and stationary environmental noise corpora. In the result, the proposed method delivers higher signal-distortion ratio and objective perceptual measure than standard semi-supervised NMF or spectral subtraction based on the same noise estimation algorithm, and further gains can be expected by speaker adaptation.

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

1. INTRODUCTION

The present paper deals with the separation of a speech signal from a single-channel noisy recording. This task, often known as speech enhancement, finds applications in the reduction of acoustic noise especially in telephonic communications [1] or in hearing aids [2]. Automatic speech recognition [3], speaker recognition [4] or emotion recognition [5]. Numerous methods have been proposed in the speech enhancement literature [6]. Most of the approaches rely on an estimation of the background noise, which is then "removed" from the signal. However, they assume that the background noise is stationary, i.e. changes slowly over time, which is not always verified in real-life noisy environments.

Other approaches such as codebook-based methods [7, 8, 9] have been proposed to overcome this limitation by using models for speech and noise. In particular, Non-negative Matrix Factorization (NMF) has been recently applied to this problem [10, 11, 12]. This method is based on a decomposition of the spectrogram of the mixture into a non-negative combination of several spectral bases, belonging to either the speech or the interfering noise.

However, we observed that conventional noise estimators are often superior to NMF for capturing the stationary background noise, with a smaller computational complexity. Hence, in the present paper we introduce a novel hybrid approach which incorporates a stationarity constraints as well as time-varying noise estimates from denoising techniques into an NMF model. The stationary part of the background noise is estimated by a standard noise estimation algorithm, while the non-stationary components are computed by a standard NMF. A large-scale evaluation show that the proposed approach can outperform both standard NMF and spectral subtraction.

In the rest of this paper, the hybrid NMF model is introduced in Section 2. Then, Section 3 describes the experimental evaluation. Finally, some conclusions are drawn before 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. METHODOLOGY

2.1. NMF Framework

The proposed monaural speech separation method is based on an extension of NMF in the time-frequency domain. The assumption is that the magnitude spectrogram $\mathbf{V} \in \mathbb{R}_+^{M \times N}$ of the noisy speech signal (with observations in columns) can be modeled as the sum $\mathbf{V} = \mathbf{V}^{(s)} + \mathbf{V}^{(n)}$, where $\mathbf{V}^{(s)}$ and $\mathbf{V}^{(n)}$ are the spectrograms of the speech signal and of the noise, respectively. Conventionally, it is assumed that both, the speech and noise spectrograms can be approximated as linear combinations of non-negative speech and noise dictionaries $\mathbf{W}^{(s)}$ and $\mathbf{W}^{(n)}$ with corresponding non-negative activation coefficients $\mathbf{H}^{(s)}$ and $\mathbf{H}^{(n)}$.

$$\mathbf{V} \approx \mathbf{\Lambda} = \mathbf{\Lambda}^{(s)} + \mathbf{\Lambda}^{(n)} := \mathbf{W}^{(s)}\mathbf{H}^{(s)} + \mathbf{W}^{(n)}\mathbf{H}^{(n)}, \quad (1)$$

where $\mathbf{\Lambda}$, $\mathbf{\Lambda}^{(s)}$ and $\mathbf{\Lambda}^{(n)}$ denote approximations of $\mathbf{V}$, $\mathbf{V}^{(s)}$ and $\mathbf{V}^{(n)}$, respectively.

In this study, we assume that the NMF speech dictionary $\mathbf{W}^{(s)}$ is estimated a-priori from training data. This is done by applying unsupervised NMF as follows. Let $\mathbf{T}^{(s)}$ be the concatenation of training speech spectrograms. Then, in the training phase, we compute $\mathbf{W}^{(s)}$ and $\mathbf{H}^{(s)}$ by minimizing

$$d(\mathbf{T}^{(s)}|\mathbf{W}^{(s)}\mathbf{H}^{(s)}) + \lambda \sum_{j=1}^{N} \frac{\prod_{i=1}^{R} \mathbf{H}_{i,j}^{(s)}}{\frac{1}{R}\sum_{i=1}^{R} \mathbf{H}_{i,j}^{(s)}}^{1/R} \quad (2)$$

where $d(\cdot|\cdot)$ is the Kullback-Leibler divergence, R is the rank of the decomposition (number of atoms in the resulting dictionary), and λ is a free parameter controlling the weighting of the reconstruction error and the enforcement of sparse activations. The algorithm applied for the training, which uses the 'flatness' of the columns of $\mathbf{H}^{(s)}$ as a sparsity criterion, is described in details in [20].

In the speech enhancement phase, the noise dictionary $\mathbf{W}^{(n)}$ as well as all the NMF activations $\mathbf{H} := \begin{matrix}\mathbf{H}^{(s)} \\ \mathbf{H}^{(n)}\end{matrix}$ are initialized randomly and then estimated by an iterative multiplicative update algorithm minimizing the Kullback-Leibler divergence $d(\mathbf{V}|\mathbf{\Lambda})$, yielding a semi-supervised NMF speech separation algorithm as in [13, 14].

After determining the free parameters of the model by means of NMF, an estimate $\widehat{\mathbf{V}^{(s)}}$ of the clean speech magnitude spectrogram

is calculated by a 'soft masking' approach as

$$\widehat{\mathbf{V}^{(s)}} = \frac{\mathbf{\Lambda}^{(s)}}{\mathbf{\Lambda}} \otimes \mathbf{V}. \quad (3)$$

The result is transformed to the time domain by means of inverse short-time Fourier transformation and overlap-add.

2.2. Integration of Noise Estimators

The proposed method extends the noise modeling to a hybrid approach with an NMF component $\mathbf{W}^{(n)}\mathbf{H}^{(n)}$ and an additional noise estimate $\mathbf{B}$ computed by 'traditional' unsupervised methods such as minimum statistics [15]. Formally, the noise spectrogram $\mathbf{V}^{(n)}$ is represented as

$$\mathbf{\Lambda}^{(n)} = \mathbf{W}^{(n)}\mathbf{H}^{(n)} + (\mathbf{1} \cdot \mathbf{h}^{(b)}) \otimes \mathbf{B} \quad (4)$$

where $\mathbf{1}$ is an all-one column vector of dimension M and $\mathbf{h}^{(b)}$ is a row vector of dimension N that allows scaling of the noise estimate per frame. Intuitively, $\mathbf{B}$ corresponds to a 'background', quasi-stationary noise floor as estimated by traditional de-noising algorithms, whereas $\mathbf{W}^{(n)}\mathbf{H}^{(n)}$ focuses on modeling the 'foreground', i. e., non-stationary sounds. The noise estimate $\mathbf{B}$ can be regarded as an NMF component that varies from frame to frame, instead of just adding a constant column to the $\mathbf{W}^{(n)}$ matrix. Incorporating a stationarity constraint, the vector $\mathbf{h}^{(b)}$ is set a-priori to an all-one vector. Alternatively, it can be updated iteratively in the multiplicative update NMF framework in order to mitigate the influence of errors in $\mathbf{B}$. We call this approach 'adaptive noise scaling'. The proposed iterative algorithm minimizes the Kullback-Leibler divergence $d(\mathbf{V}|\mathbf{\Lambda})$ (cf. above) by means of

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T \cdot \frac{\mathbf{V}}{\mathbf{\Lambda}}}{\mathbf{W}^T \cdot \mathbf{1}}; \quad \mathbf{W}^{(n)} \leftarrow \mathbf{W}^{(n)} \otimes \frac{\frac{\mathbf{V}}{\mathbf{\Lambda}} \cdot \mathbf{H}^{(n)T}}{\mathbf{1} \cdot \mathbf{H}^{(n)T}} \quad (5)$$

where $\mathbf{W} := [\mathbf{W}^{(s)}\mathbf{W}^{(n)}]$, and

$$\mathbf{h}^{(b)} \leftarrow \mathbf{h}^{(b)} \otimes \frac{\mathbf{1} \cdot \left(\mathbf{B} \otimes \frac{\mathbf{V}}{\mathbf{\Lambda}}\right)}{\mathbf{1} \cdot \mathbf{B}}. \quad (6)$$

The update rules in (5) correspond to the standard update rules from [16] adapted to integrate the noise estimate (4). Note that no sparsity penalty is applied at the speech enhancement phase, as found beneficial in [20]. However, the presence of the constant background noise component implicitly enforces sparsity on the other components, due to the non-negativity constraint.

3. EXPERIMENTAL SETUP

Our method is evaluated on mixtures of speech and noise from publicly available corpora. We use spontaneous speech from the Buckeye corpus [17] to reflect use cases such as multimedia retrieval in web videos. Furthermore, to simulate the influence of various noise types, we consider (i) the CHiME 2011 Challenge [18] 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 [19] for mostly stationary, environmental noise. The MTV collection consists of 200 songs covering the years from 1981 to 2000 as well as various genres, and featuring male as well as female singers. All data are converted to 16 kHz sampling rate, monophonic audio.

As evaluation data, we use 60 test sentences from the Buckeye corpus from 30 speakers (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 180 test files. The remaining 10 speakers of the Buckeye corpus are used for speaker-independent training of NMF speech dictionaries.

The parameters are chosen after a previous study [1]. NMF is applied to magnitude spectrograms computed using Hann windows of 32 ms length with 50 % overlap. NMF speech dictionaries are learnt from the 10 training speakers by concatenating several utterances (about one minute per speaker) and applying NMF with 25 components ($R = 25$). For reference purposes, we also consider speaker-dependent NMF speech dictionaries which are learnt from a set of (clean) utterances that is disjoint from the set of test utterances. In this base learning procedure, the flatness weight λ (cf. (2)) is set to 100 [20]. The number of atoms of the noise dictionary computed during the enhancement step is set to 8. All atoms of the NMF dictionaries are normalized to unity Euclidean norm.

We compare the proposed hybrid method (Section 2.2) to conventional NMF (Section 2.1) and magnitude domain spectral subtraction. Furthermore, we consider naïve cascading of spectral subtraction and conventional NMF, in either order. For spectral subtraction, an over-subtraction factor a as introduced by [21] is taken into account. The algorithm subtracts a times the noise estimate from the noisy recording. This over-subtraction factor is supposed to reduce musical noise, by removing not only the constant part of the noise, but also some of the small peaks which appear randomly and which cause the musical noise. We consider $a = 1$ (no over-subtraction) and $a = 3$ as put forth by [21]. The noise estimation algorithm used is a publicly available[1] implementation of the procedure proposed by [15].

The hybrid NMF-denoising system considers the noise matrix $\mathbf{B}$ estimated by the same algorithm [15]. The corresponding scaling vector $\mathbf{h}^{(b)}$ is initialized to $a \cdot \mathbf{1}$. In the baseline setting, $\mathbf{h}^{(b)}$ is fixed, whereas in the *adaptive noise scaling* version, it is updated for each frame, using (6). However, $\mathbf{h}^{(b)}$ is constrained to the range $]0, a]$. For the estimation steps according to (5, 6), we perform a fixed number K of iterations, where K is chosen from $\{1, 2, 4, \ldots, 128\}$.

It must be noted that the hybrid system does not exploit the spectral subtraction concept, because the reconstruction phase uses a Wiener-like soft mask (3), as in the basic NMF approach. Hence, it is not based on any specific 'de-noising' algorithm. The only algorithm borrowed from the traditional speech enhancement framework is the noise estimator. Note also that the tested version is an off-line algorithm. However, it may be used in an on-line context with a sliding-window approach [1, 12].

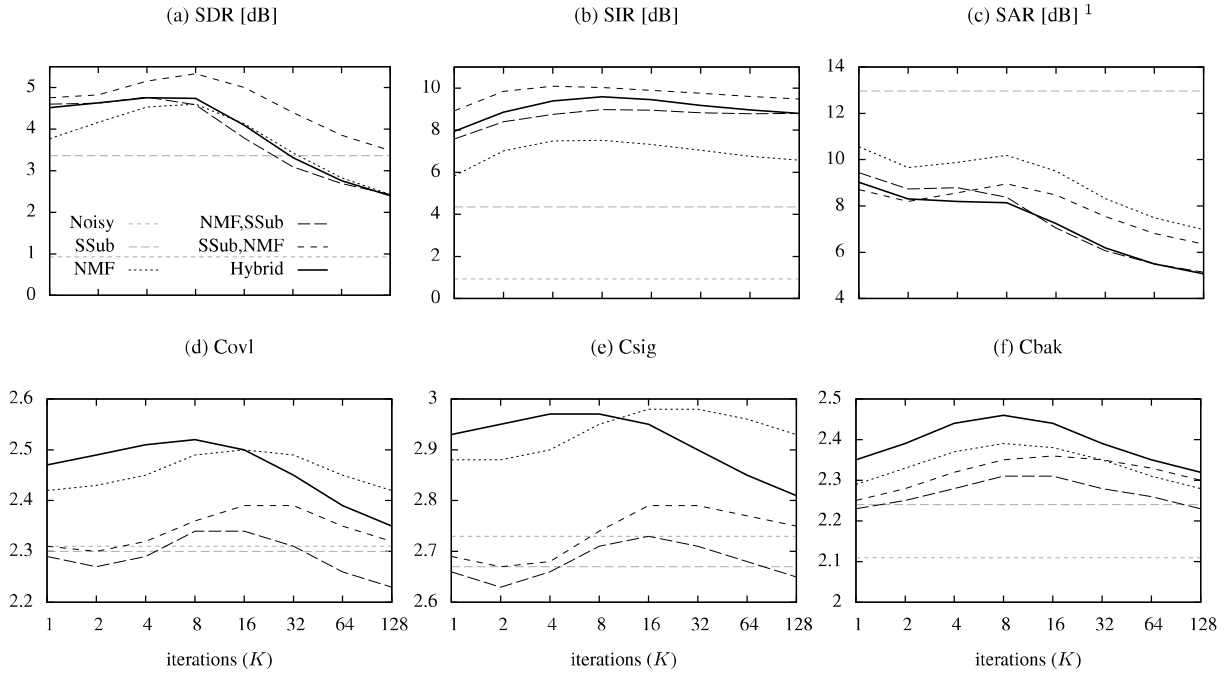
We measure the performance of speech enhancement in terms of energy-based measures—source-distortion ratio (SDR), source-interference ratio (SIR) and source-artifact ratio (SAR) [22]—and the Covl, Csig and Cbak measures [23] representing mean opinion scores (MOS) of overall perceptual quality, perceived quality of the wanted signal and perceived quality of the interference signal, on a scale from 1–5.

4. RESULTS

Figure 1 shows the chosen evaluation measures in speaker-independent speech separation across all test utterances, and for different numbers of NMF iterations (K). In terms of SDR (Figure 1a), the proposed hybrid NMF method significantly outperforms standard spectral subtraction. At $K = 8$, we have $p \ll .001$ and a 95 % confidence interval of the true SDR difference = $[0.98, 1.82]$ according to a two-tailed paired t-test, but the gain over standard semi-supervised

[1] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html#enhance

Fig. 1: Energy-based measures (SDR, SIR, SAR) and objective perceptual measures on Buckeye test set for varying numbers of iterations of the NMF-based algorithms. Average results across CHiME, MTV and NOISEX noise types. Noisy and spectral subtraction (SSub) baselines are shown. ',' indicates naïve cascading of algorithms. [1] For noisy baseline (no processing), SAR $\rightarrow +\infty$.



NMF is not significant ($p > .05$). Furthermore, the overall best SDR of 5.3 dB is achieved by simply applying spectral subtraction after standard semi-supervised NMF. In terms of interference reduction measured by SIR (Figure 1b), the gain by integrating the additional noise estimate, instead of simple semi-supervised NMF, is much more pronounced (95 % confidence interval: $[1.75, 2.38]$ dB SIR). However, looking at SAR (Figure 1c) the energy of the artifacts seems to increase whenever noise estimates and NMF are combined; the best SAR is achieved by spectral subtraction, followed by standard semi-supervised NMF, with SAR decreasing with the number of NMF iterations for any of the NMF-based methods.

Regarding the objective perceptual measures (Figures 1d–1f), we observe a different picture. The proposed method outperforms standard semi-supervised NMF in terms of Covl, and achieves comparable Csig for fewer iterations. In terms of Cbak, the proposed hybrid method is superior to standard NMF (95 % confidence interval: $[.055, .083]$), reflecting the results for SIR. Spectral subtraction by itself actually worsens the scores for overall quality and quality of the wanted signal (Covl, Csig) compared to the noisy baseline. In terms of all three objective perceptual measures, both of the naïve cascades of spectral subtraction and NMF are inferior to NMF itself, and the proposed integration of the noise estimator into NMF.

Next, we investigate the behavior of the different methods depending on the noise type. Results are displayed in Figure 2. The overall highest Covl is obtained on domestic noise from the CHiME database, when the proposed hybrid method is used, while spectral subtraction or its naïve combination with NMF decreases the score. Additive music is most challenging for all the methods considered; here, the overall best result is achieved by standard NMF, probably due to problems of the noise estimator with this highly non-stationary type of noise. Finally, for the NOISEX database, we observe a significant gain by spectral subtraction—this is somehow expected due to the mostly stationary noise. Still, NMF-based methods can add another gain on top of that, with the hybrid method yielding the overall best result of 2.65 for $K = 8$.

Table 1: Influence of the SNR on separation scores, with $K = 8$. The best result in either scenario is highlighted. SNR: *low* $< -1.2\,\text{dB} \leq$ *med* $< 3\,\text{dB} \leq$ *high*

	SDR [dB]			Covl		
SNR	*low*	*med*	*high*	*low*	*med*	*high*
Noisy	-3.79	0.85	5.64	1.86	2.41	2.65
SSub	0.04	3.05	**6.90**	1.90	2.40	2.58
NMF	2.27	5.14	6.39	2.17	2.59	2.71
NMF,SSub	**3.71**	**5.70**	6.59	2.10	2.45	2.52
SSub,NMF	3.15	4.93	5.66	2.06	2.43	2.52
Hybrid	3.12	5.04	6.05	**2.23**	**2.61**	**2.72**

We further investigate the influence of the original signal-to-noise ratio (SNR). We split the database into three subsets of equal size, according to whether the SNR is *low* (lower than -1.2 dB), medium (*med*, between -1.2 dB and 3 dB) or *high* (above 3 dB). The obtained SDR and Covl after $K = 8$ iterations are displayed in Table 1. The relative performance of the NMF methods are consistent with low and medium SNRs. In the case of high SNRs, the spectral subtraction obtains comparatively better results since it delivers the best SDR (6.90 dB). However, in terms of objective perceptual scores, the hybrid method consistently outperforms all others.

Besides comparing the speaker-independent NMF algorithms to the speaker-independent spectral subtraction, it is of interest to compare the NMF algorithms in a speaker-dependent setup. Results are shown in Table 2. As expected, the overall results are now vastly

Fig. 2: Overall composite mean opinion score (Covl) for speech separation from different types of background noise.

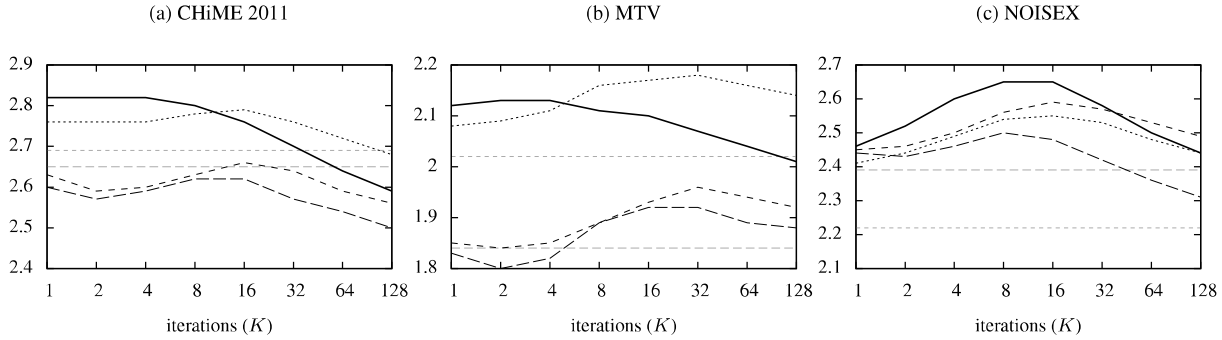


Table 2: Comparison of NMF-based algorithms in speaker-independent and speaker-dependent scenarios. The best result in either scenario is highlighted. Buckeye test set, $K = 8$, average across all noise types.

	SDR	SIR [dB]	SAR	Covl	Csig	Cbak
Noisy	0.93	0.93	∞	2.31	2.73	2.11
Speaker-independent NMF speech dictionary						
NMF	4.60	7.51	**10.19**	2.49	2.95	2.39
NMF,SSub	**5.33**	**10.02**	8.95	2.36	2.74	2.35
Hybrid	4.74	9.58	8.14	**2.52**	**2.97**	**2.46**
Speaker-dependent NMF speech dictionary						
NMF	6.44	9.83	**11.36**	2.61	3.11	2.53
NMF,SSub	**7.00**	**12.25**	10.06	2.49	2.90	2.49
Hybrid	6.79	12.17	10.57	**2.68**	**3.17**	**2.60**

superior to speaker-independent separation; for example, the absolute increase in Covl going from the noisy baseline to the best speaker-independent algorithm is .21 (95 % confidence interval: $[.17, .26]$), and another gain of .16 ($[.14, .17]$) can be obtained with a speaker dependent base. Due to this improvement of the speech dictionaries, the NMF methods outperform spectral subtraction even on high SNRs. For example, the "SSub,NMF" method obtains an SNR of 8.51 dB, against 6.90 dB for spectral subtraction. Furthermore, we observe the same ranking of the different algorithms as before with respect to the six evaluation measures considered: The hybrid NMF delivers the highest objective perceptual scores while the best SDR is achieved by cascading NMF and spectral subtraction. Finally, we found that the behavior of the algorithms with regard to the number of iterations did not change when switching to a speaker-dependent NMF base.

To conclude our evaluation, in Table 3 we present the results obtained with over-subtraction as well as adaptive scaling of the noise estimate. For spectral subtraction itself, a noticeable gain of .46 dB SDR (95 % confidence interval: $[.36, .56]$) can be obtained by considering over-subtraction with $a = 3$. However, if spectral subtraction is applied as post-processing to standard semi-supervised NMF, the SDR gain (.01 dB) is not significant ($p \gg .05$). For the hybrid approach, setting $a = 3$ actually decreases the SDR. Among the hybrid approaches, a combination of noise scaling and 'over-subtraction' seems to deliver best results (6.83 dB average SDR), but this is only insignificantly ($p > .05$) above the default hybrid approach (6.79 dB). The Covl measure cannot be improved by either noise scaling or over-subtraction. We conclude that including the stationarity constraint (constant $\mathbf{h}_b = \mathbf{1}$) into NMF is meaningful.

Table 3: Influence of over-subtraction on spectral subtraction and hybrid NMF, and adaptive noise scaling (iterative update of $\mathbf{h}_b$) for hybrid NMF. Speaker-dependent NMF base, $K = 8$, Buckeye test set, average across all noise types.

	SDR [dB]		Covl	
	$a = 1$	$a = 3$	$a = 1$	$a = 3$
SSub	3.36	**3.82**	**2.29**	2.26
NMF,SSub	7.00	**7.01**	**2.49**	2.45
Hybrid	**6.79**	6.37	**2.68**	2.66
Hybrid ($\mathbf{h}_b$)	6.65	**6.83**	2.65	2.65

5. CONCLUSIONS

We have shown an effective and efficient way to integrate traditional noise estimation for speech de-noising into the emerging family of factorization-based speech separation algorithms. The proposed integration of stationary noise estimates and stationarity constraints into the NMF framework delivered better mean opinion scores than naïve cascading of spectral subtraction and NMF in a large scale evaluation on spontaneous speech corrupted by a wide range of noise. Using speaker dependent NMF bases greatly improved the results on top of that, motivating further research on unsupervised speaker adaptation during speech enhancement. Furthermore, since the considered noise estimator is on-line, it can be straightforwardly integrated into our real-time semi-supervised NMF framework [1]. The use of other types of noise estimators such as [24] can also be considered, in order to assess their influence on the obtained separation quality. Finally, improved phase estimation algorithms for speaker separation such as in [25] will be transferred to the speech enhancement domain.

6. RELATION TO PRIOR WORK

In the present paper, we introduce a novel NMF model integrating a stationary noise estimator for speech separation. Previous studies on NMF-based speech separation have focused on the design of priors and constraints on the factorization, in order to take into account the different behavior of the noise and speech components [11, 26]. Duan *et al.* [12] present a pure NMF solution inspired by a classic spectral subtraction framework without initialization of the speech base. The work by Mohammadiha *et al.* [27] considers a method mixing classic denoising techniques and NMF. However, it also estimates the noise by a standard NMF approach, without exploiting the stationarity property. In contrast to these works, the proposed approach exploits an independent estimation of the noise spectrum and incorporates it directly into the NMF decomposition.

7. REFERENCES

[1] 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, Mar. 2012, pp. 322–329, Springer.

[2] N. Madhu, A. Spriet, S. Jansen, R. Koning, and J. 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," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 1, pp. 63–72, Jan. 2013.

[3] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The munich 2011CHiME challenge contribution: Nmf-blstm speech enhancement and recognition for reverberated multisource environments," in *Proc. Inter. Workshop on Machine Listenning in Multisource Environments (CHiME)*, Florence, Italy, Sept. 2011, pp. 24–29.

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

[5] F. Weninger, B. Schuller, A. Batliner, S. Steidl, and D. 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] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 78, pp. 588–601, 2007.

[7] S. Srinivasan, J. Samuelsson, and W.B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 163–176, Jan. 2006.

[8] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based bayesian speech enhancement for nonstationary environments," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 441 –452, Feb. 2007.

[9] P. Mowlaee, R. Saeidi, and R. Martin, "Model-driven speech enhancement for multisource reverberant environment (signal separation evaluation campaign (sisec) 2011)," in *Proc. LVA ICA*, Tel Aviv, Israel, Mar. 2012, pp. 454–461, Springer.

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

[11] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. of ICASSP*, Las Vegas, NV, USA, Apr. 2008, pp. 4029–4032.

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

[13] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden Markov model for polyphonic audio representation and source separation," in *Proc. of WASPAA*, Mohonk, NY, USA, Oct. 2009, pp. 121–124.

[14] 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, Mar. 2012, pp. 61–64.

[15] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 9, no. 5, pp. 504–512, July 2001.

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

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

[18] 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, Sept. 2010, pp. 1918–1921.

[19] A Varga and H. 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.

[20] C. Joder, F. Weninger, D. Virette, and B. Schuller, "A comparative study on sparsity penalties for nmf-based speech separation: Beyond Lp-norms," submitted to *ICASSP 2013*.

[21] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," in *Proc. of ICASSP*, Washington DC, USA, Apr. 1979, pp. 208–211.

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

[23] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[24] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Communication*, vol. 46, no. 2, pp. 220–231, Feb. 2006.

[25] P. Mowlaee, R. Saeidi, and R. Martin, "Phase estimation for signal reconstruction in single-channel speech separation," in *Proc. of Interspeech*, Portland, OR, USA, Sept. 2012.

[26] N Mohammadiha, J. Taghia, and A. Leijon, "Single channel speech enhancement using bayesian nmf with recursive temporal updates of prior distributions," in *Proc. of ICASSP*, Kyoto, Japan, Mar. 2012, pp. 4561–4564.

[27] N. Mohammadiha, T. Gerkmann, and A. Leijon, "A new linear mmse filter for single channel speech enhancement based on nonnegative matrix factorization," in *Proc. of WASPAA*, New Paltz, NY, USA, Oct. 2011, pp. 45–48.