ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Speech Enhancement Using Sparse Convolutive Non-negative Matrix Factorization with Basis Adaptation*

Michael A. Carlin[1], *Nicolas Malyska*[2], *Thomas F. Quatieri*[2]

[1]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD USA
[2]MIT Lincoln Laboratory, Lexington, MA USA

macarlin@jhu.edu, {nmalyska,quatieri}@ll.mit.edu

10.21437/Interspeech.2012-178

Abstract

We introduce a framework for speech enhancement based on *convolutive non-negative matrix factorization* that leverages available speech data to enhance arbitrary noisy utterances with no *a priori* knowledge of the speakers or noise types present. Previous approaches have shown the utility of a *sparse* reconstruction of the speech-only components of an observed noisy utterance. We demonstrate that an underlying speech representation which, in addition to applying sparsity, also *adapts* to the noisy acoustics improves overall enhancement quality. The proposed system performs comparably to a traditional Wiener filtering approach, and the results suggest that the proposed framework is most useful in moderate- to low-SNR scenarios.
Index Terms: speech enhancement, convolutive non-negative matrix factorization, basis adaptation, sparsity

1. Introduction

Speech in mobile telephone environments is often corrupted by additive, wideband, and non-stationary interference. Accordingly, there is a need to denoise these signals not only for human listeners but also for automated downstream tasks like speaker identification and speech recognition. In particular, non-stationary noises such as passing vehicles and interfering speakers present many challenges to traditional enhancement algorithms, motivating exploration of alternatives.

Data-driven approaches to speech enhancement based on non-negative matrix factorization (NMF) have shown particular promise in recent years. In general, NMF finds a low-rank decomposition of a non-negative matrix X as the product of two non-negative matrices such that $X \approx WH$ [1]. The columns of W are interpreted as a *basis* or *codebook* for representing the columns of X, and the rows of H specify the corresponding time-varying *activation* of the bases. NMF yields a decomposition that is purely additive with a codebook that is readily interpretable as the constituent parts of the observed data. When X is a spectrogram, such a representation is attractive for modeling speech corrupted by additive interference.

When modeling acoustic signals, explicitly accounting for temporal context, which NMF does not, is important for a rich characterization of time-varying acoustic scenes. This was addressed by a generalization of NMF termed *convolutive* non-negative matrix factorization (CNMF). This extension models an observed spectrogram as a shifted sum of time-varying codebook entries [2]. When applied to speech, the learned basis is a collection of phone-like units and necessarily forms a richer description of the observed acoustics.

For speech enhancement, formulations of NMF can benefit from regularization that constrains the structure of the activations, depending on the degree of *a priori* knowledge of the speakers or noise types present in an observed signal. Wilson *et al.* [3] introduced a supervised framework that constrained the covariance structure of the activations given the availability of training data. In contrast, using no training data, de Frein and Rickard [4] described a CNMF framework that exploited the sparse nature of speech spectra to enhance utterances corrupted by wideband noise by suitably constraining the activations.

Although in many cases the specific speakers to be encountered are unknown, a large amount of speech is often available to reliably train a CNMF basis for representing arbitrary speakers. Such available data could be in-domain, i.e., representing similar speakers in a particular channel, or it could be out-of-domain, i.e., not reflective of the observed speakers or channel conditions. It is therefore of interest to study how available data can be used to adapt a CNMF basis to arbitrary noisy utterances.

In this paper, we introduce a CNMF-based enhancement framework for scenarios where *a priori* knowledge about the speakers and noise types is unavailable. Our contribution is a method to adapt a previously trained CNMF codebook to represent speech in arbitrary noisy utterances while simultaneously exploiting the sparse nature of speech spectra. We explore basic aspects of the framework applied to speech corrupted by additive, non-stationary noise, and we compare our results to a standard Wiener filtering-based approach, demonstrating comparable performance in moderate- to low-SNR scenarios.

2. Background

To begin, let $X \in \mathbb{R}_{\geq 0}^{F \times n}$ denote an $F \times n$ matrix whose elements are non-negative. When X is a magnitude spectrogram, standard NMF does not explicitly account for temporal context. To address this issue, Smaragdis [2] introduced CNMF, which solves the following optimization problem:

$$\arg\min_{W(\tau),H} D(X||\widehat{X}) \text{ subject to } W(\tau), H \geq 0 \quad \forall \tau.$$

In general, $D(X||\widehat{X})$ is a divergence function that measures the error between the observed data X and the CNMF reconstruction $\widehat{X}$, where

$$\widehat{X} = \sum_{\tau=0}^{T-1} W(\tau) \overrightarrow{H}^{\tau}. \quad (1)$$

Here, $\{W(\tau)\}$ is a collection of time-varying bases, $W(\tau) \in \mathbb{R}_{\geq 0}^{F \times K}$, $\overrightarrow{H} \in \mathbb{R}_{\geq 0}^{K \times n}$, and $\tau = 0, 1, \cdots, T-1$. The nota-

*This work is sponsored by the United States Air Force Research Laboratory under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

tion $\overrightarrow{H}^{\tau}$ is understood as a zero-padding of H with τ columns of zeros to the left and truncated at the right to maintain correct dimensionality; an analogous shift-and-truncate operation is defined for $\overleftarrow{H}^{\tau}$. Moreover, T denotes the temporal extent (in frames) of the CNMF bases. In this work, we consider $D(X||\widehat{X}) = ||X - \widehat{X}||_F$, where $||\cdot||_F$ is the Frobenius norm (i.e., the square root of summed squared matrix entries). As in the original NMF algorithm, optimization of the CNMF problem is efficiently carried out by a series of alternating multiplicative updates to the bases and activations [5]. Note that CNMF for $T = 1$ corresponds to the baseline NMF case.

In general, formulations of NMF for speech enhancement can benefit from the use of regularization to incorporate heuristics that constrain the structure of the bases and activations. Regularization amounts to penalizing the NMF divergence by terms that quantify the desired constraints on the bases and activations. Cichocki *et al.* [6] have previously described a general framework for regularized NMF-based source separation, and we elaborate on how to apply their results for CNMF-based enhancement in the next section.

Finally, given a noisy signal, we assume that the magnitude spectra of the speech and noise are additive, i.e., $X = S + N$. While this assumption is untrue, previous source separation studies have shown that the additivity assumption yields suitable results [2, 4]. Thus, the goal in this paper is to recover an estimate of the speech spectrogram $\widehat{S}$ from the CNMF reconstruction $\widehat{X}$, from which a denoised signal $\hat{s}(t)$ can be synthesized. In the following section, we describe a regularized CNMF-based framework for this purpose.

3. Proposed Enhancement Framework

To incorporate prior knowledge about a representation for speech, one can simply initialize a CNMF codebook using available speech corpora. However, since this data may not reflect the observed speaker or channel conditions, it is necessary to allow the initial speech codebook to adapt according to the observed acoustics. Also, as previously observed [4], when CNMF is applied to clean speech spectrograms, the bases tend to be sparsely activated over time. Thus, speech corrupted by wideband noise would tend to excite more bases than necessary to account for non-speech components and hence preserving sparsity of the speech activations is desirable for enhancement.

The proposed enhancement procedure is as follows. First, we observe a noisy speech spectrogram $X \in \mathbb{R}_{\geq 0}^{F \times n}$. Next, we assume that we have available speech data with which to train clean CNMF bases $W_S(\tau) \in \mathbb{R}_{\geq 0}^{F \times K_S}$ which we augment with randomly initialized noise bases $W_N(\tau) \in \mathbb{R}_{\geq 0}^{F \times K_N}$. Here, K_S and K_N are the number of speech and noise codebook entries, respectively. Moreover, we denote $\widetilde{W}_S(\tau)$ as an *adapted* speech codebook as CNMF learning proceeds. We next randomly initialize speech and noise activation matrices $H_S \in \mathbb{R}_{\geq 0}^{K_S \times n}$ and $H_N \in \mathbb{R}_{\geq 0}^{K_N \times n}$, respectively. We then seek to minimize the following objective function with respect to the bases and activations:

$$J_T := \frac{1}{2}||X - \widehat{X}||_F^2 + \alpha \cdot J_W(\widetilde{W}_S) + \beta \cdot J_H(H_S), \quad (2)$$

subject to $\widetilde{W}_S(\tau), W_N(\tau), H_S, H_N \geq 0$ for all τ, where

$$J_W(\widetilde{W}_S) := \frac{1}{2}\sum_{\tau=0}^{T-1} ||\widetilde{W}_S(\tau) - W_S(\tau)||_F^2 \quad (3)$$

and

$$J_H(H_S) := \sum_{m=1}^{n} ||\mathbf{h}_m^S||_1, \quad (4)$$

with $||\mathbf{h}_m^S||_1$ denoting the L_1-norm of the m-th column of H_S. Observe that the adaptation parameter α controls how "close" the adapted speech bases remain to those learned from the training data. Moreover, the sparsity parameter β controls how many bases are activated to reconstruct the observed speech spectrogram by the definition of the L_1-norm. Thus, large values of α mean that the speech bases adapt little from training whereas large values of β favor sparse activation of the speech bases.

Following [6], we compute the element-wise gradient of Eq. 2 with respect to the speech bases, and can express the multiplicative updates to the bases as

$$\widetilde{W}_S(\tau) \leftarrow \widetilde{W}_S(\tau) \circ \frac{\left\{X \overrightarrow{H}_S^{\tau T} - \alpha \cdot [\widetilde{W}_S(\tau) - W_S(\tau)]\right\}_\epsilon}{\widehat{X} \overrightarrow{H}_S^{\tau T}} \quad (5)$$

and

$$W_N(\tau) \leftarrow W_N(\tau) \circ \frac{X \overrightarrow{H}_N^{\tau T}}{\widehat{X} \overrightarrow{H}_N^{\tau T}}, \quad (6)$$

where $\circ$ denotes the Hadamard (element-wise) product, division is understood to be element-wise, and $\{\cdot\}_\epsilon := \max(\cdot, \epsilon)$ is applied element-wise to ensure $\widetilde{W}_S(\tau)$ remains non-negative (we set $\epsilon = 10^{-16}$). As is typically done, we normalize each of the speech and noise bases to have unit Frobenius norm after each update [4]. Similarly, we can express the multiplicative updates to the activation matrices as

$$H_S \leftarrow \left\langle H_S \circ \frac{\left\{\widetilde{W}_S^T(\tau) \overleftarrow{X}^{\tau} - \beta \cdot \mathbf{1}_{K_S \times n}\right\}_\epsilon}{\widetilde{W}_S^T(\tau) \overleftarrow{\widehat{X}}^{\tau}} \right\rangle \quad (7)$$

and

$$H_N \leftarrow \left\langle H_N \circ \frac{W_N^T(\tau) \overleftarrow{X}^{\tau}}{W_N^T(\tau) \overleftarrow{\widehat{X}}^{\tau}} \right\rangle, \quad (8)$$

where $\mathbf{1}_{K_S \times n}$ is a matrix of all ones and the average $\langle\cdot\rangle$ is computed over all τ. We alternate between updates to the bases and activation matrices, ceasing when the relative change in J_T is less than a threshold of 0.1% or we exceed 500 iterations, whichever occurs first. Upon convergence, we obtain

$$\widehat{X} = \sum_{\tau=0}^{T-1} [\widetilde{W}_S(\tau)\, W_N(\tau)] \begin{bmatrix} \overrightarrow{H}_S^{\tau} \\ \overrightarrow{H}_N^{\tau} \end{bmatrix} = \widehat{S} + \widehat{N}, \quad (9)$$

which yields the enhanced waveform $\hat{s}(t)$ from $\widehat{S}$ by overlap-add synthesis using the phase $\angle X$ from the observed noisy utterance. In the next section, we explore some general characteristics of the proposed method, and evaluate enhancement performance.

4. Experiments and Results

For the discussion that follows, all audio was sampled at 8 kHz and we considered magnitude spectrograms computed from the short-time Fourier transform of a standardized acoustic waveform using 32 ms Hamming-weighted windows overlapped by 50%. In the figures, p-values for assessing statistical significance using a two-tailed Wilcoxon rank-sum test at the 5% level are indicated by asterisks ($* : p < 0.05$; $** : p < 0.005$; $***: p \ll 0.005$) or labeled as not significant (n.s.).

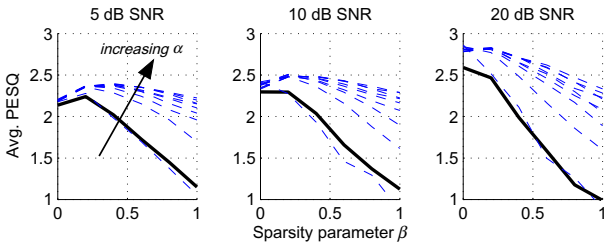

Figure 1: *Varying the adaptation and sparsity parameters and their effect on average PESQ for AWGN-AM interference at various SNR ($T = 3$, $K_S = 100$, $K_N = 50$).*

4.1. Database and Evaluation Criterion

To evaluate the proposed enhancement framework, we assembled a small corpus using utterances from the TIMIT database. For training the CNMF bases, we used approximately three minutes of clean speech from the TIMIT `train` subset; utterances were chosen at random, half of which were male and half female. For testing, we selected a single utterance from 10 male and 10 female speakers, all chosen at random from the TIMIT `test` subset. Each test utterance was corrupted with segments of interference at 20, 10 and 5 dB overall SNR. We considered randomly selected segments of *babble* and *street* noise from the NOISEX92 and AURORA corpora, respectively, as well as a white Gaussian noise signal whose envelope was modulated at 5 Hz (*AWGN-AM*).

We assessed enhancement quality using the ITU standard Perceptual Evaluation of Speech Quality (PESQ) measure [7]. Despite being originally developed to objectively measure the perceptual quality of coded speech over telephone networks, it is commonly used to measure the quality of speech enhancement algorithms. PESQ ranges between 1 and 5 and correlates well with listener-reported mean opinion scores of perceptual quality, with higher scores indicating higher quality.

4.2. Parameter Selection

To study the utility of *both* basis adaptation and sparsity regularization on enhancement quality, we calculated PESQ for $\alpha \in [0, 200]$ and $\beta \in [0, 1]$, computing the average across all test utterances as a function of α and β. Results are shown in Fig. 1 for test utterances corrupted by *AWGN-AM* interference for $K_S = 100$, $K_N = 50$, and $T = 3$. For comparison, we also considered a system where we held the speech codebook used for enhancement *fixed* to that used for training, updating only the noise bases, the speech and noise activations, and varying only the sparsity parameter. We observe that allowing the speech bases to adapt during enhancement (various α, dashed lines) yields larger average PESQ values than if the speech bases had been held fixed (solid lines). Moreover, favoring a sparse reconstruction of the speech spectrogram also improves enhancement quality, before performance drops off, with the effect seemingly more pronounced at lower SNRs.

To explore the dependence of the adaptation and sparsity parameters on SNR, we considered the distributions of α and β that maximized PESQ for each test utterance; this is akin to a user tweaking the system by hand to yield a high-quality enhancement. These results are shown in Fig. 2, again for *AWGN-AM* interference, with $K_S = 100$, $K_N = 50$, and $T = 3$. In this figure, black horizontal lines indicate population median, thick vertical lines indicate interquartile range, thin vertical lines indicate population extrema, and grey circles correspond to outliers. In panel (a), it is clear that low-SNR conditions favor larger values of α, i.e., *less* deviation from the training speech codebook, compared to high-SNR conditions where the smaller values of α are favored, i.e., *more* deviation from the training speech codebook. Furthermore, in panel (b) we confirm that high-quality enhancements in low-SNR conditions generally warrant reconstructions of the speech spectrograms that are more sparse than in high-SNR conditions.

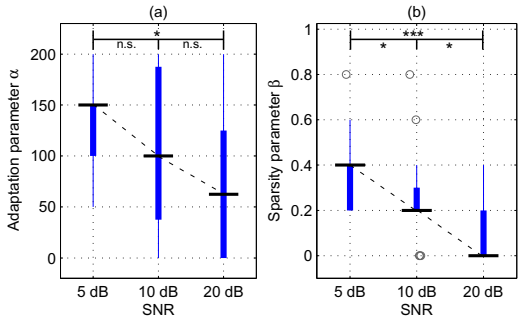

Figure 2: *Distribution of optimal (a) adaptation and (b) sparsity parameters for AWGN-AM interference at various SNRs ($T = 3$, $K_S = 100$, $K_N = 50$).*

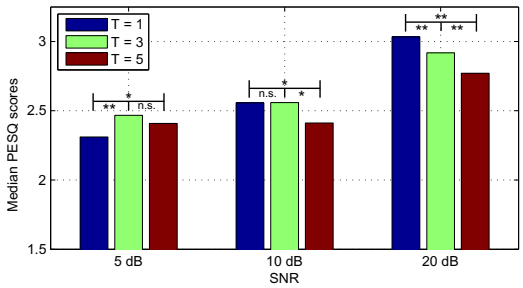

Figure 3: *Median optimal PESQ values for different T using optimal (α, β) for AWGN-AM interference at various SNRs ($K_S = 100$, $K_N = 50$).*

We also considered the effect of codebook duration T on enhancement quality. Shown in Fig. 3 are the median values of optimal PESQ scores for short to long codebook durations for low- to high-SNR conditions. At 20 dB SNR, enhancement based on standard NMF ($T = 1$) yields the largest median PESQ compared to $T = \{3, 5\}$ frames. However, at 5-dB SNR we observe that an increased codebook length of $T = 3$ frames yields a significantly higher median PESQ compared to $T = 1$. Thus, overall analysis of the proposed framework suggests that in low-noise conditions, higher-quality enhanced speech is obtained using a simpler model (short T), whose speech codebook varies relatively freely (small α) with little-to-no requirement of a sparse reconstruction of the speech spectrogram (small β). However, in noisy environments, the system performs best when the speech codebook extends over multiple frames ($T = 3$) and stays close to that seen from available training data (large α), with sparsity of the speech codebook activations playing an important role (large β).

4.3. Comparison to Baseline Wiener Filter

Finally, we compared the performance of the proposed framework to an implementation of a traditional Wiener filtering algorithm that estimates *a priori* SNR [8]. While we have applied the algorithm "out-of-the-box" and not optimized its various operation parameters (e.g., voice activity detector thresh-

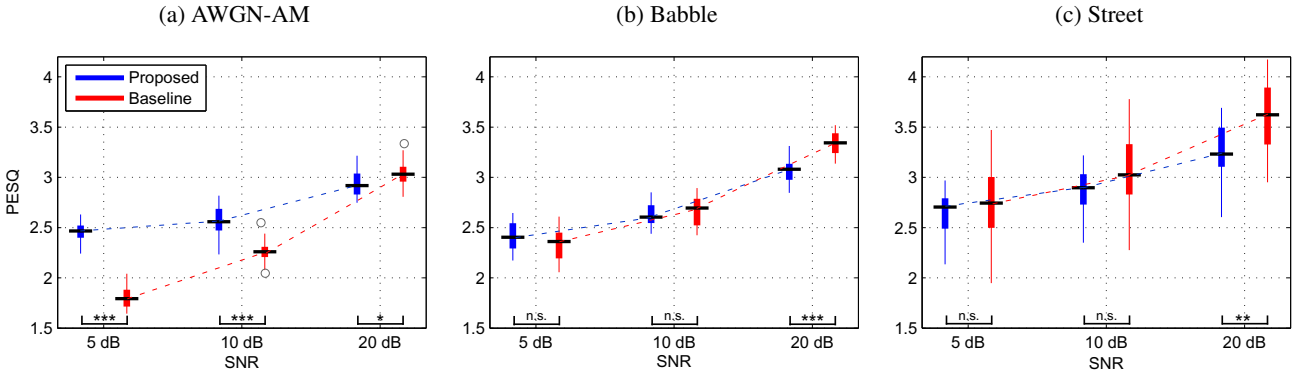

Figure 4: *Enhancement results for speech corrupted by (a) AWGN-AM (at 5 Hz), (b) babble, and (c) street interference at various SNR conditions. The proposed algorithm (blue) is is compared to a baseline approach based on a traditional Wiener algorithm (red).*

old), it nevertheless provides a useful baseline for interpreting PESQ scores obtained from the proposed algorithm. Results for test utterances corrupted by *AWGN-AM*, *babble*, and *street* interference are shown in Fig. 4. For the proposed system, we set $K_S = 100$, $K_N = 50$, and $T = 3$, and we consider values of α and β that maximize PESQ for each test utterance.

First, in panel (a) we observe that for *AWGN-AM* interference in low-SNR conditions (10 and 5 dB) the proposed method significantly outperforms the baseline. This is of interest since it has been suggested that systematic degradation of the temporal modulations in the 4 Hz range—roughly the syllabic rate of speech—result in a corresponding reduction in speech intelligibility [9]. Thus, despite the non-stationary nature of the interference, the proposed method yields an enhanced speech signal that may help preserve intelligibility.

Next, in panel (b), while the baseline algorithm outperforms the proposed method in low-noise conditions, there are no significant differences in the median PESQ scores between the proposed and Wiener filtering approaches in low-SNR conditions. Of course, *babble* interference is particularly difficult since its long-term spectral statistics somewhat resemble those of speech. As such, the performance of the proposed algorithm was not expected to be significantly different from the baseline.

Finally, in panel (c), again we observe that the baseline algorithm outperforms the proposed method in low-noise conditions, but that there are no significant differences in median PESQ values between the two approaches at low SNRs. We note, however, that the *street* noise is such that some utterances are subject to loud noises (e.g., passing vehicles, vehicle horns, etc.) interspersed with relatively long periods of silence. As such, one could expect the baseline approach—which relies heavily on the accuracy of the voice activity detector—to perform well in those high-SNR segments, but possibly less so in more noisy regions. This is perhaps reflected in the apparent higher variance of the baseline approach in 10 and 5 dB SNR conditions as compared to the proposed method.

5. Discussion and Conclusions

We have introduced a CNMF-based framework for speech enhancement that leverages available speech data to enhance noisy utterances with no prior knowledge about the speakers or noise types present. We have characterized how to configure the system as a function of SNR, demonstrated the advantages of speech basis adaptation and activation sparsity, and shown that the system performs comparably to a traditional Wiener filtering approach. The overall results suggest that the proposed framework is most useful in moderate- to low-SNR environments.

It is important to note that other heuristics and adaptation schemes have been considered in NMF-based source separation and enhancement tasks. For example, Wilson *et al.* [10] considered temporal smoothness constraints on the activations, and Grais and Erdogan [11] recently considered adaptation of NMF speech codebooks in a probabilistic setting (though here they assume the availability of a small amount of speaker-specific adaptation data). Such approaches are expected to benefit the algorithm presented in this paper. Finally, while we manually varied α and β to maximize PESQ, this measure is unavailable in operational environments. Future work should investigate methods to automatically select these parameters.

6. Acknowledgements

Many thanks to David Harwath from MIT CSAIL for feedback and support with early implementations of this work.

7. References

[1] Lee, D. and Seung, H. "Algorithms for non-negative matrix factorization." In *Proc. Neural Inf. Proc. Sys. (NIPS)*, 2000.

[2] Smaragdis, P. "Convolutive speech bases and their application to supervised speech separation." *IEEE Trans. Aud., Sp., and Lang. Proc.*, 15(1):1–12, 2007.

[3] Wilson, K. W., Raj, B., Smaragdis, P., and Divakaran, A. "Speech denoising using non-negative matrix factorization with priors." In *Proc. ICASSP*, 2008.

[4] de Frein, R. and Rickard, S. T. "Learning speech features in the presence of noise: sparse convolutive robust non-negative matrix factorization." In *Proc. IEEE Conf. on Digital Sig. Proc.*, 2009.

[5] Wang, W., Cichocki, A., and Chambers, J. A. "A multiplicative algorithm for convolutive non-negative matrix factorization based on squared Euclidean distance." *IEEE Trans. Sig. Proc.*, 57(7):2858–2864, 2009.

[6] Cichocki, A., Zdunek, R., and Amari, S. "New algorithms for non-negative matrix factorization in applications to blind source separation." In *Proc. ICASSP*, 2006.

[7] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs." Tech. Rep. ITU-T P.862, 2001.

[8] Loizou, P. *Speech Enhancement: Theory and Practice*, CRC Press, Boca Raton: FL, 2007.

[9] Drullman, R., Festen, J. M., and Plomp, R. "Effect of temporal envelope smearing on speech reception." *J. Acoust. Soc. Am.*, 95(2):1053–1064, 1994.

[10] Wilson, K. W., Raj, B., and Smaragdis, P. "Regularized non-negative matrix factorization with temporal dependencies for speech denoising." In *Proc. Interspeech*, 2008.

[11] Grais, E. M., and Erdogan, E. "Adaptation of speaker-specific bases in non-negative matrix factorization for single channel speech-music separation." In *Proc. Interspeech*, 2011.