Single-channel Dynamic Exemplar-based Speech Enhancement

Nasser Mohammadiha, Simon Doclo

Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4all
University of Oldenburg, Germany

Abstract

This paper proposes an exemplar-based speech enhancement method based on high-resolution STFT magnitude spectrograms, where a selection of the nonnegative training data is used as the dictionary to provide a holistic nonnegative representation of the test data. We discuss how this exemplar-based model ensures that the enhanced speech signal falls on the speech manifold, which improves the quality of the enhanced speech signal. To exploit the temporal continuity, a vector autoregressive model is used to model the activations where the model parameters are learned using a new NMF-based approach. Results from several supervised and semi-supervised speech enhancement experiments indicate that the proposed exemplar-based method outperforms the considered supervised and unsupervised denoising algorithms in terms of both segmental SNR and PESQ at different input SNRs.

Index Terms: nonnegative matrix factorization, exemplar-based noise reduction, overcomplete dictionary

1. Introduction

Most of the state-of-the-art noise reduction algorithms estimate the clean speech signal from a noisy signal in an unsupervised fashion [1]. When some additional information about the acoustic environment or speaker identity is available, supervised speech enhancement methods, e.g., [2, 3, 4, 5, 6, 7, 8], can be used to obtain a higher-quality enhanced speech signal. In the supervised scenario, it is assumed that training samples corresponding to the targeted noise environment and/or the speaker are available. The required information about the speaker identity and noise type can be obtained using speaker recognition and acoustic environment classification algorithms, or are readily available in some special applications, e.g., in pilot communications [4].

In this paper, we propose a dynamic exemplar-based speech enhancement algorithm that operates in the high-resolution short-time discrete Fourier transform (STFT) magnitude spectrogram domain. The proposed technique is closely related to nonnegative matrix factorization (NMF) [9], where instead of learning a low-dimensional dictionary, an overcomplete dictionary is obtained by sampling from the training data. By doing so, the source dictionaries become more representative of the underlying classes and the obtained source estimates will lie on the manifolds of the original sources (cf. Section 2). Other exemplar-based techniques have been proposed in the past for the purpose of speech recognition and source separation [10, 11]. The approach in [10] was designed in the low-resolution Mel-scale magnitude spectrogram domain and was used for both sparse classification and feature enhancement and improved the speech recognition results. In [11], a sparse exemplar-based single-channel source separation method was proposed that significantly outperformed the competing algorithm.

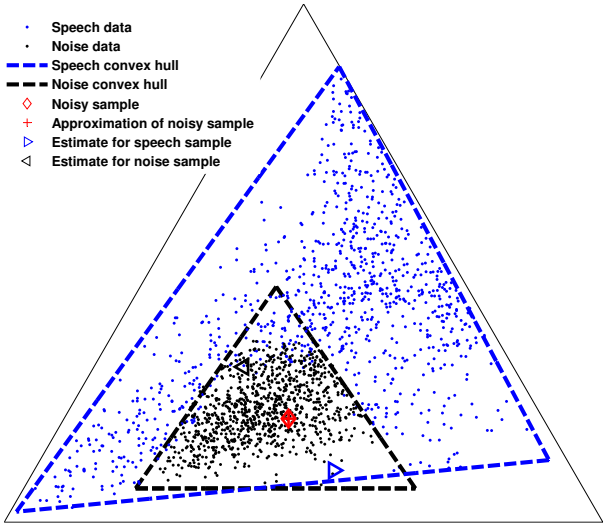

Figure 1: Visualization of exemplar- and NMF-based representations. Three Mel bands are chosen from the speech and babble noise Mel-scale magnitude spectrograms and are normalized to sum to one and are plotted in a two-dimensional simplex. An enhancement example is also shown, where the NMF-based speech estimate lies outside the speech manifold.

We use a recently proposed nonnegative vector autoregressive model (N-VAR) to efficiently exploit the temporal continuity of the speech and noise signals, where the nonnegative model parameters are obtained using a new NMF-based scheme (cf. Section 3). Our proposed method is causal and does not depend on the future short-time frames, where the sliding window approach in [10] admits a delay around 200 ms to produce the best results.

This paper focuses on the speech enhancement application and several experiments are carried out in order to evaluate the performance of the exemplar-based denoising algorithms. Our supervised and semi-supervised (where a general speech model is used and a single noise model is constructed using the training samples of the four targeted noise types) experiments show that the proposed method significantly outperforms the competing approaches ([10], [12], and [13]) in terms of both segmental SNR and PESQ.

2. Exemplar- versus NMF-based Representations

NMF is a popular representation method and has been successfully used in different applications. In a source separation or speech enhancement problem, it is common to learn a dictionary for each source in advance using source-specific training

data. The dictionary elements of each source define a convex hull that surrounds the training data of that source [11]. In the following, we use speech and noise signals to describe and visualize why this may lead to unsatisfactory enhancement performance.

Speech data is obtained using 54 sentences from the TIMIT database [14], where as the noise data, the babble noise signal from the NOISEX-92 database [15] is considered. Both speech and noise signals were transformed to a time-frequency domain by applying the STFT using a frame length of 64 ms with 75% overlapping Hann windows. A subset of the resulting high-resolution magnitude spectrograms were chosen and transformed into the Mel-domain, with 26 Mel-scale frequency bands. Figure 1 shows the scatter plot of the resulting speech and noise data marked with blue and black dots, respectively, where only three Mel bands are chosen and plotted in a two-dimensional simplex. The convex hulls of the speech and noise data are also plotted using dashed lines in Figure 1. The three vertices of the convex hulls correspond to the learned dictionary elements.

As can be seen in the figure, the convex hulls of the speech and noise data are highly overlapped, which causes an ambiguity in signal separation. In other words, each convex hull surrounds not only the samples from the corresponding source but also the samples associated with the other source. Therefore, the speech dictionary can also be used to approximate the points that do not belong to the speech class. Figure 1 additionally shows a noisy sample, a possible NMF approximation of the noisy sample and corresponding estimates of the underlying speech and noise samples. As can be seen, the speech estimate lies outside the speech manifold in this example. If only a few points (ideally one point) from the speech data were used to obtain an estimate for the speech sample, the resulting speech estimate would have lied on the speech manifold.

Although in the preceding discussion we used three-dimensional data points for the purpose of visualization, the reasoning is valid for higher dimensions as well. Using training data (STFT magnitudes of the training signals) as the dictionary elements for one source provides a richer model that can better explain an unseen magnitude spectra from that source. In the enhancement phase, a sparse combination of the speech training data provides an estimate for the underlying speech STFT magnitude spectra that lies on the manifold of the speech magnitude spectra, which results in a higher-quality enhanced speech signal, as also verified by our experiments (cf. Figure 2).

3. Proposed Method

We present our new exemplar-based speech enhancement approach in this section. The signal model and speech estimation method are presented in Subsection 3.1, where we assume that the speech and noise dictionaries are given. In Subsection 3.2, we explain how the speech and noise dictionaries are constructed from the training data.

3.1. High-resolution Exemplar-based Processing

We assume an additive noise model where speech and noise signals are added to obtain the noisy signal. All the signals are transformed into a time-frequency domain by applying the STFT. The magnitude spectrogram matrices (the magnitude of the STFT) of the noisy, speech, and noise signals are respectively denoted by $\mathbf{Y}$, $\mathbf{S}$ and $\mathbf{N}$. Let $\mathbf{y}_t = [y_{1t} \dots y_{Kt}]^T$ denote the noisy spectral vector at time t, where K is the number of STFT frequency bins, and T denotes the vector transpose. The speech and noise spectral vectors, $\mathbf{s}_t$ and $\mathbf{n}_t$, are defined similarly. In the following, the noisy spectral vector is approximated by adding the speech and noise spectral vectors [5], and each source's spectral vector is approximated by a linear combination of the associated dictionary elements:

$$\begin{aligned} \mathbf{y}_t &\approx \mathbf{s}_t + \mathbf{n}_t \\ &\approx \mathbf{W}^s \mathbf{h}_t^s + \mathbf{W}^n \mathbf{h}_t^n = \mathbf{W}\mathbf{h}_t, \end{aligned} \tag{1}$$

where $\mathbf{W}^s$ and $\mathbf{W}^n$ are the speech and noise exemplar-based dictionaries (cf. Subsection 3.2), and $\mathbf{h}_t^s$ and $\mathbf{h}_t^n$ are the corresponding activation vectors. Eq. (1) is basically an NMF representation of the input vector $\mathbf{y}_t$. There is however an important remark that the dictionary elements in (1) are samples from the training data and therefore the representation is holistic, rather than a parts-based representation usually learned by NMF [9]. In contrast to [10], Eq. (1) models the noisy signal in the high-resolution STFT domain rather than the reduced-resolution Mel domain.

To model the temporal modulations of the speech and noise signals, we model the activation vectors using a first-order non-negative vector autoregressive (N-VAR) model, where the activation vector at time t is approximated by multiplying the non-negative coefficient matrix $\mathbf{A}$ and the activation vector at time $t-1$:

$$\mathbf{h}_t^s \approx \mathbf{A}^s \mathbf{h}_{t-1}^s, \quad \mathbf{h}_t^n \approx \mathbf{A}^n \mathbf{h}_{t-1}^n. \tag{2}$$

Since (2) can be seen as an NMF approximation of $\mathbf{h}_t^s$ and $\mathbf{h}_t^n$, we propose an NMF-based approach to learn the source-dependent matrices $\mathbf{A}^s$ and $\mathbf{A}^n$ from the training data in Subsection 3.2.

Eq. (1) and (2) describe our signal model. Given a noisy spectral vector $\mathbf{y}_t$ and the estimated activation vector at time $t-1$ ($\hat{\mathbf{h}}_{t-1}$), we would like to estimate the enhanced speech spectral vector $\hat{\mathbf{s}}_t$. For this purpose, we first use a two-step algorithm to obtain $\hat{\mathbf{h}}_t$. In the first step, (2) is used to predict the activation vector as

$$\tilde{\mathbf{h}}_t = \mathbf{A}\hat{\mathbf{h}}_{t-1}, \tag{3}$$

where $\mathbf{A}$ is the diagonal concatenation of $\mathbf{A}^s$ and $\mathbf{A}^n$. In the second step, we apply probabilistic latent component analysis (PLCA)[1] [16] on $\mathbf{y}_t$ to obtain $\overline{\mathbf{h}}_t$ such that $\mathbf{W}\overline{\mathbf{h}}_\mathbf{t}$ best approximates the input $\mathbf{y}_t$. Following the approach proposed in [13], $\hat{\mathbf{h}}_t$ is now obtained as

$$\hat{\mathbf{h}}_t = \frac{\overline{\mathbf{h}}_t \odot (\tilde{\mathbf{h}}_t)^{\boldsymbol{\beta}}}{\sum \overline{\mathbf{h}}_t \odot (\tilde{\mathbf{h}}_t)^{\boldsymbol{\beta}}}, \tag{4}$$

where $\odot$, $\div$, and $(\cdot)^\beta$ denote the element-wise multiplication, division, and power operators, respectively, and $\boldsymbol{\beta}$ is a prior weight vector. We use two different scalar weights (β^s, β^n) for speech and noise activations. A sparse PLCA approach [11] can also be used to obtain $\overline{\mathbf{h}}_t$; however, as the resulting estimate using (4) is usually sparse (sparser than $\overline{\mathbf{h}}_t$), the sparse PLCA did not substantially improve the results over the basic PLCA in our experiments, and it is not considered in our evaluations.

After estimating $\hat{\mathbf{h}}_t$, a real-valued gain function is computed as

$$\mathbf{g} = \frac{\mathbf{W}^s \hat{\mathbf{h}}_t^s}{\mathbf{W}\hat{\mathbf{h}}_t}, \tag{5}$$

[1] PLCA is a probabilistic NMF approach that minimizes a weighted Kullback-Leibler divergence between the input and its approximation.

after which the enhanced speech signal is obtained using $\mathbf{g} \odot \mathbf{y}_t$, the inverse DFT, and the overlap-add method.

We additionally suggest a modified Mel-scale counterpart of the proposed method and evaluate its performance in Section 4. For this purpose, all the spectral vectors in (1) are transformed into the Mel spectral domain by summing the adjacent elements using overlapping triangular filters. The resulting vectors are denoted by $\mathbf{y}_t^{\text{mel}}$, $\mathbf{s}_t^{\text{mel}}$, and $\mathbf{n}_t^{\text{mel}}$. To use the robustness of the Mel-scale representation with respect to the pitch changes, and also to use the high resolution of the original STFT representation, in addition to $\mathbf{W}^s$ and $\mathbf{W}^n$, we construct Mel-domain dictionaries for each source that are identically aligned with $\mathbf{W}^s$ and $\mathbf{W}^n$. In the enhancement phase, the activation vector $\hat{\mathbf{h}}_t$ is estimated using the noisy spectral vector $\mathbf{y}_t^{\text{mel}}$ and the Mel-domain speech and noise dictionaries. The real-valued gain function $\mathbf{g}$ is however calculated using the computed $\hat{\mathbf{h}}_t$, $\mathbf{W}^s$ and $\mathbf{W}^n$, as explained before. Pilot tests indicated that this hybrid approach yields better results than the pure Mel-domain enhancement method, and hence only this hybrid scheme is considered in our experiments in Section 4.

3.2. Dictionary Construction

In the following, we explain how the speech and noise dictionaries $\mathbf{W}^s$ and $\mathbf{W}^n$ and the N-VAR coefficient matrices $\mathbf{A}^s$ and $\mathbf{A}^n$ are obtained using speech and noise training data that are denoted by $\mathbf{S}^{\text{tr}}$, $\mathbf{N}^{\text{tr}}$, respectively, where the columns of $\mathbf{S}^{\text{tr}}$ and $\mathbf{N}^{\text{tr}}$are normalized to sum to one. Here, we only explain the procedure for the speech signal, as the same algorithm is also used for the noise signal.

As the training signal might be too long and redundant, it is sufficient to only select a subset of $\mathbf{S}^{\text{tr}}$ to create $\mathbf{W}^s$, which reduces the memory requirements and the computational complexity. We used two selection approaches for this purpose. The first approach is adopted from [10], where we use a uniformly-distributed random frame-shift of 4 to J samples to construct $\mathbf{W}^s$ from $\mathbf{S}^{\text{tr}}$. J is chosen such that the resulting $\mathbf{W}^s$ has approximately I^s elements, where I^s is our desired number of dictionary elements. If required, additional samples are taken from $\mathbf{S}^{\text{tr}}$ such that $\mathbf{W}^s$ has exactly I^s columns. As the second approach, we implemented the manifold-preserving quantization approach from [17]. The performance of both approaches is compared in Section 4.

To learn the coefficient matrix $\mathbf{A}^s$, we propose an NMF-based algorithm. To do so, we first obtain the nonnegative approximation of the entire speech training data using the constructed speech dictionary as:

$$\mathbf{S}^{\text{tr}} \approx \mathbf{W}^s \mathbf{H}^{s,\text{tr}}, \tag{6}$$

where we use PLCA approach to estimate $\mathbf{H}^{s,\text{tr}}$. Let us define the matrix $\mathbf{V}^s$ such that its t-th column is the $(t-1)$-th column of $\mathbf{H}^{s,\text{tr}}$. Eq. (2) can now be written as:

$$\mathbf{H}^{s,\text{tr}} \approx \mathbf{A}^s \mathbf{V}^s, \tag{7}$$

which provides an NMF approximation of $\mathbf{H}^{s,\text{tr}}$ in terms of the dictionary $\mathbf{A}^s$ and the activation matrix $\mathbf{V}^s$. We learn $\mathbf{A}^s$ by applying the PLCA approach on $\mathbf{H}^{s,\text{tr}}$, while $\mathbf{V}^s$ is held fixed.

4. Experimental Results

We used speech signals from the TIMIT database [14], babble and factory noise signals from the NOISEX-92 database [15], and highway traffic and city traffic noise signals from the Sound Ideas database [18]. All the signals were down-sampled to 16-kHz and the STFT analysis was performed using a frame length of 1024 samples (64 ms) with 75% overlapping Hann windows. The Mel-scale spectrograms were obtained using 26 overlapping triangular filters.

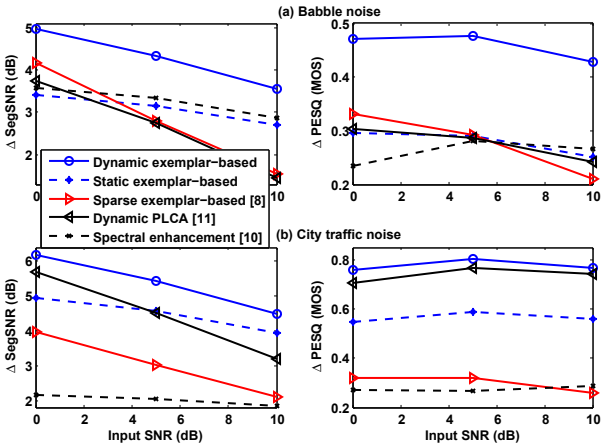

Figure 2: Results of supervised speech enhancement for babble (top) and city traffic (bottom) noise types. Speaker- and noise-dependent dictionaries are used for all the methods except the spectral enhancement approach.

The performance of the proposed method is studied with ($\beta^s = 0.4, \beta^n = 0.1$) and without ($\beta^s = \beta^n = 0$) the temporal continuity and are referred to as "dynamic exemplar-based" and "static exemplar-based", respectively. We also consider the proposed hybrid (STFT+Mel) scheme, which is referred to as "dynamic exemplar-based, hybrid". We additionally compare the denoising performance of the proposed approaches with that of [10] (referred to as "sparse exemplar-based" where the sparsity weight parameters were experimentally set to obtain the best denoising performance and $T^{[8]}$ was set to 12 frames that results to a delay around 180 ms for the output signal), [13] (referred to as "dynamic PLCA"), and the speech spectral enhancement using generalized Gamma priors [12] (referred to as "spectral enhancement", with $\nu^{[10]} = \gamma^{[10]} = 1$ and noise power spectral density estimated using [19]). We use the segmental SNR (SegSNR) and PESQ implemented by Loizou [1] to evaluate the speech enhancement performance.

4.1. Supervised Speech Enhancement

We present the results of our supervised speech enhancement experiments in this subsection. We constructed noise-dependent and speaker-dependent dictionaries, each consisting of 800 randomly-sampled spectral vectors. Speech signals from 14 speakers were considered in our experiments, where the speech dictionary was constructed using 9 out of 10 available speech sentences from each speaker and the other speech sentence was used for testing. The first 75% of each noise signal was used for noise dictionary construction and the last 25% was used for testing.

Figure 2 shows the SegSNR and PESQ improvements at different input SNRs for the babble noise (top panel) and the city traffic noise (bottom panel). The results show that the best performance is obtained using our "dynamic exemplar-based" approach, which is significantly better than the other methods

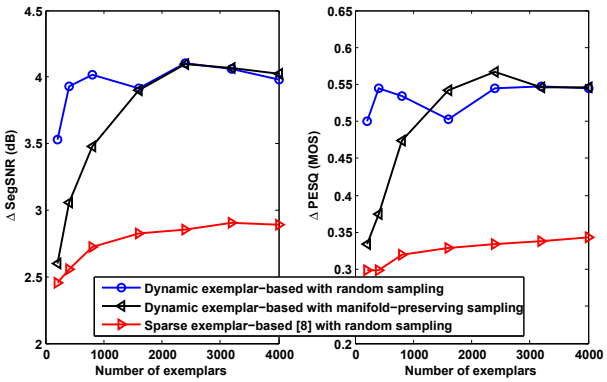

Figure 3: Semi-supervised noise reduction performance as a function of the number of exemplars used to construct the dictionary. Speech and noise dictionaries have identical number of exemplars. The input SNR is 5 dB and the results are averaged over four noise types.

including "static exemplar-based" and "sparse exemplar-based" methods, especially for the babble noise. As indicated by the results, most of the supervised algorithms fail to outperform the unsupervised spectral enhancement algorithm in the babble case. Since the speech and babble signals mostly share a similar manifold, a more constrained NMF approach can be used to improve the performance in this case [6].

As can be seen in Figure 2, all the supervised methods outperform the spectral enhancement approach for the city traffic noise. City traffic noise is a very non-stationary signal that includes different horn sounds and its NMF dictionary is distinct enough from that of the speech signal. As a result, dynamic PLCA is also able to achieve very good results in this case.

4.2. Semi-supervised Speech Enhancement

This subsection presents the results of our semi-supervised experiments, where the speech model was constructed from 216 speech sentences uttered by 24 speakers (around 41000 spectral vectors), while for the test purposes, 14 speech sentences uttered by 14 speakers (none of which were included in the training) were used. A single noise dictionary was constructed using the first 75% of the noise signals from the four considered noise types, and the last 25% of the signals were used for the testing.

We first study the effect of the number of exemplars on the speech enhancement performance. Figure 3 shows the SegSNR and PESQ improvements for different dictionary sizes, where the speech and noise dictionaries have identical number of dictionary elements. The input SNR is equal to 5 dB and the results are averaged over all noise types. As can be seen, when there are more than 1600 elements in the dictionaries the noise reduction performance does not depend on the sampling approach and that the performance does not change very much by changing the dictionary size. The large performance gap between our proposed method and that of the sparse exemplar-based approach is due to designing a high-resolution filter and having a more efficient approach to use the temporal dynamics.

Figure 4 shows the results for the babble (top) and the city traffic (bottom) noise types, where speech and noise dictionaries, each having 3200 elements, are constructed by the random sampling method. As can be seen, the proposed approach (dynamic exemplar-based) has yielded the best performance. As the dynamic PLCA approach resulted to a poor performance,

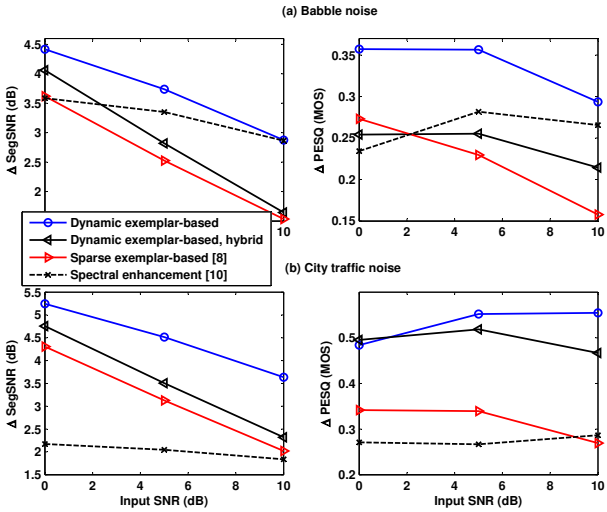

Figure 4: Semi-supervised speech enhancement results for babble (top) and city traffic (bottom) noise types. Speaker- and noise-independent dictionaries are used for all the model-based approaches.

it is omitted from the figure for brevity. We see that although the proposed hybrid method provides a good performance for the city traffic noise, it yields a poor performance for the babble noise, especially at 10dB input SNR.

Comparing Figures 2 and 4, we see that all the model-based methods admit a performance loss when less supervision is used for the model training. This is more pronounced for the proposed method, which is partially due to pitch mismatch between the training and the testing speech signals. Nevertheless, our experiments indicate that the developed method leads to a better performance compared to the other competing algorithms both for the supervised and semi-supervised scenarios, which was also verified by our informal listenings. Some sound examples are available at `http://www.sigproc.uni-oldenburg.de/audio/nmoh/exemplar-based-nr/main.html`.

5. Conclusion

We proposed a causal high-resolution exemplar-based speech enhancement algorithm, where we used a vector autoregressive model to efficiently model the temporal modulations of the activations. A hybrid counterpart of the proposed method was also developed by using both the STFT and Mel-scale dictionaries. We provided a geometrical argument and discussed how an exemplar-based approach can lead to a better-quality enhanced speech signal compared to the similar NMF-based method. The results of our speech enhancement experiments indicate that the proposed high-resolution method outperforms the similar NMF-based method, an exemplar-based feature enhancement approach, and a speech spectral enhancement method with a high margin. Moreover, the results indicate that the proposed high-resolution method outperforms the hybrid counterpart.

6. Acknowledgements

This research was supported by the Cluster of Excellence 1077 "Hearing4all", funded by the German Research Foundation (DFG).

7. References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 1st ed. Boca Raton, FL: CRC Press, 2007.

[2] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 725–735, Apr. 1992.

[3] T. V. Sreenivas and P. Kirnapure, "Codebook constrained Wiener filtering for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 383–389, Sep. 1996.

[4] X. Xiao and R. Nickel, "Speech enhancement with inventory style speech resynthesis," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 18, no. 6, pp. 1243–1257, Aug. 2010.

[5] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using NMF," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[6] N. Mohammadiha and A. Leijon, "Nonnegative HMM for babble noise derived from speech HMM: Application to speech enhancement," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 5, pp. 998–1011, May 2013.

[7] S. Srinivasan, "Knowledge-based speech enhancement," Ph.D. dissertation, KTH - Royal Institute of Technology, Stockholm, Sweden, 2005. [Online]. Available: http://www.diva-portal.org/smash/get/diva2:12853/FULLTEXT01.pdf

[8] N. Mohammadiha, "Speech enhancement using nonnegative matrix factorization and hidden markov models," Ph.D. dissertation, KTH - Royal Institute of Technology, Stockholm, Sweden, 2013. [Online]. Available: http://theses.eurasip.org/theses/499/speech-enhancement-using-nonnegative-matrix/download/

[9] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[10] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based speech enhancement and its application to noise-robust automatic speech recognition," in *Proc. Int. Workshop on Machine Listening in Multisource Environments (CHiME)*, Sep. 2011, pp. 53–57.

[11] P. Smaragdis, M. Shashanka, and B. Raj, "A sparse non-parametric approach for single channel separation of known sounds," in *Advances in Neural Information Process. Systems (NIPS)*. MIT Press, 2009, pp. 1705–1713.

[12] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized Gamma priors," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.

[13] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Prediction based filtering and smoothing to exploit temporal dependencies in NMF," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, May 2013, pp. 873–877.

[14] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "TIMIT acoustic-phonetic continuous speech corpus." Philadelphia: Linguistic Data Consortium, 1993.

[15] 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, Jul. 1993.

[16] P. Smaragdis, B. Raj, and M. V. Shashanka, "A probabilistic latent variable model for acoustic modeling," in *Advances in Models for Acoustic Process. Workshop, NIPS*. MIT Press, 2006.

[17] M. Kim and P. Smaragdis, "Manifold preserving hierarchical topic models for quantization and approximation," in *Proc. Int. Conf. Machine Learning (ICML)*, 2013.

[18] B. Nimens *et al.*, "Sound ideas: sound effects collection," ser. 6000, http://www.sound-ideas.com/6000.html.

[19] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, Mar. 2010, pp. 4266–4269.