Speech Enhancement by Classification of Noisy Signals Decomposed Using NMF and Wiener Filtering

Mahmoud Fakhry, Amir Hossein Poorjam, and Mads Græsbøll Christensen
Audio Analysis Lab, CREATE, Aalborg University

E-mails:{mfak, ahp, mgc}@create.aau.dk

***Abstract*—Supervised non-negative matrix factorization (NMF) is effective in speech enhancement through training spectral models of speech and noise signals. However, the enhancement quality reduces when the models are trained on data that is not highly relevant to a speech signal and a noise signal in a noisy observation. In this paper, we propose to train a classifier in order to overcome such poor characterization of the signals through the trained models. The main idea is to decompose the noisy observation into parts and the enhanced signal is reconstructed by combining the less-corrupted ones which are identified in the cepstral domain using the trained classifier. We apply unsupervised NMF followed by Wiener filtering for the decomposition, and use a support vector machine trained on the mel-frequency cepstral coefficients of the parts of training speech and noise signals for the classification. The results show the effectiveness of the proposed method compared with the supervised NMF.**

***Index Terms*—Speech enhancement, signal decomposition, unsupervised NMF, Wiener filtering, SVM.**

I. Introduction

The objective of speech enhancement is to reduce unwanted noise from a noisy speech signal [1], [2]. A major challenge in current enhancement techniques is to accurately estimate the noise statistics, particularly in non-stationary environments. The classical estimators are based on voice activity detectors or on tracking the non-stationarity in short-length segments of the signal [3]–[6]. However, these techniques are less-accurate when tracking highly non-stationary noise with low signal-to-noise ratio (SNR). Unlike the classical estimators, the time-frequency (TF) analysis methods such as the empirical mode decomposition (EMD), do not require estimation of the noise statistics [7]. In these methods, the decomposition is applied to the noisy speech signal, and a decision criterion identifies the less-corrupted components in order to use them to reconstruct the enhanced signal.

Most recently proposed speech enhancement methods rely on training the spectral diversity of sources, e.g., speech and noise, from relevant training data [8]–[11]. Such enhancement techniques are based on supervised non-negative matrix factorization (NMF). NMF is part-based factorization that aims to approximate the spectral power of a source, e.g. speech and noise, by spectral basis vectors and temporal activation ones, with non-negative element constraints [12]–[15]. Speech enhancement based on supervised NMF consists of two phases, namely, training and enhancement. In the training phase, spectral basis vectors, describing the speech and noise in a noisy signal, are trained independently by factorizing the spectral power of training signals. The trained spectral basis vectors are used subsequently in the enhancement phase to estimate temporal activation vectors of the speech signal and the noise signal based on the spectral power of the noisy speech signal. The enhanced signal is obtained via Wiener filtering using the trained basis and estimated activation vectors.

In some situations, speech signals of different speakers and signals of different types of noise are not known a priori, especially when the training data is available through unlabeled speech signals and unlabeled noise signals. Moreover, in some cases the data does not contain training speech signals and training noise signals that match a particular noisy signal. Joint training of spectral basis vectors on unlabeled data was presented in [16]. We show that the performance of the joint training is limited when the training set does not contain specific data that is explicitly relevant to a noisy observation.

To tackle this problem, we propose to train a classifier using a support vector machine (SVM) [17]. To perform classification with a good performance, however, enough amount of data is required for the training and the testing. In order to provide the classifier with enough data and to simplify its task, inspired by the EMD, we propose to decompose speech and noise signals for the training and a noisy speech signal for the testing into non-overlapping parts by using NMF followed by Wiener filtering. These multiple parts have different patterns and they sum up to the signal to be decomposed. The enhanced signal is then reconstructed by combining the less-corrupted parts of the noisy signal. These parts are identified in the cepstral domain [18] using the SVM that is trained on the parts of training speech and noise signals.

II. Problem Formulation

The observation model in the time domain is given by

$$x(t) = s(t) + n(t), \tag{1}$$

where $x(t)$ indicates a noisy speech signal, $s(t)$ a clean speech signal, $n(t)$ a noise signal, and t a time-index. The observation

This work was funded by Independent Research Fund Denmark: DFF 4184-00056

$x(t)$ can be represented in the time domain by the sum of K_x parts as follows

$$x(t) = \sum_{k_x=1}^{K_x} x_{k_x}(t). \tag{2}$$

The speech signal $s(t)$ and the noise signal $n(t)$ can also be represented by the sum of parts. Let $\mathbf{X}_{k_x}$ be the matrix of the complex STFT coefficients of $x_{k_x}(t)$. Due to the linearity of the STFT, the proposed signal model in the time-frequency domain is given by

$$\mathbf{X} = \sum_{k_x=1}^{K_x} \mathbf{X}_{k_x}. \tag{3}$$

We assume that the parts of the speech signal and the parts of the noise signal in the noisy speech signal $\mathbf{X}$ are less-overlapping. In this sense, the part $x_{k_x}(t)$ can be classified as a part of either $s(t)$ or $n(t)$. The parts of the noisy speech signal classified as speech are then combined together in order to reconstruct an enhanced signal. To this aim, we propose to use NMF followed by Wiener filtering in order to decompose the noisy speech signal into K_x parts, and a trained SVM in order to classify them.

III. Proposed Method

The proposed method consists of two phases, namely training and enhancement. In the training phase, a classifier is trained using the parts of speech and noise signals. Later, the parts of a noisy speech signal are classified using the classifier and the ones of speech are linearly combined to reconstruct an enhanced signal in the enhancement phase. This explicitly requires to decompose the noisy speech signal into less-overlapping and easily separable parts. Such decomposition can be achieved by exploiting certain signal diversity, e.g., spectral or temporal diversity. The noisy signal can, for example, be passed through a bank of spectral filters, or can be broken down into short-length temporal segments. However, signal separability is not always guaranteed by doing so.

It is widely known that audio signals in the time-frequency domain are sparse in their nature. This property can be exploited to apply spectro-temporal filtering in order to decompose a noisy speech signal into less-overlapping parts. The main challenge, however, is to find a way to compute spectro-temporal filters. In this paper, we propose to use NMF to compute a filter-bank of spectro-temporal Wiener gains, which is used later to decompose the noisy speech signal into parts in the time-frequency domain. In this sense, we exploit the sparsity of audio signals and of the nonnegativity constraint imposed on the factorization. The parts are finally obtained in the time domain via the inverse short time Fourier transform (ISTFT) of their STFT complex coefficients.

The proposed method also requires a means to identify the less-corrupted signal parts. In [19], [20], it is demonstrated that the distribution of mel-frequency cepstral coefficients (MFCCs) is predictably modified by additive noise and the amount of change is related to the noise level in a noisy speech signal. Considering this property, we propose to use the MFCCs as features and a support vector machine (SVM) as the classifier to detect these less-corrupted parts.

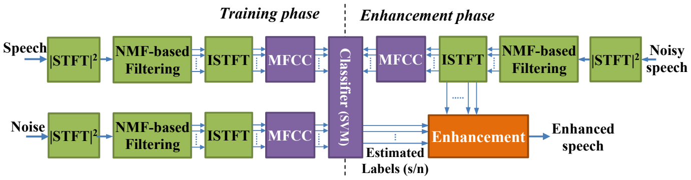


Fig. 1. Block diagram of the proposed method in training and enhancement phases.

Fig.1 illustrates the block diagram of the proposed method. In both the training and enhancement phases, the signals in the time domain are represented by parts through the STFT, the NMF-based filtering, and the ISTFT. The MFCCs of the parts of the speech signals and the noise signals are computed and used for training the classifier in the training phase. In the enhancement phase, the MFCCs of the parts of the noisy speech signal are extracted, and the speech and noise parts are identified using the trained classifier. The parts classified as speech are then combined together in order to reconstruct the enhanced signal in the time domain.

A. Signal Decomposition using NMF and Wiener Filtering

NMF has been applied for low-rank modeling of signals. In the context of modeling audio signals, the spectral power of a source is approximately represented as linear combinations of spectral basis vectors using temporal activation vectors. A combination through the outer product of a basis vector and its corresponding activation one results in a rank-1 matrix, which is a part of the low-rank approximation of the spectral power. The parts obtained by different combinations are supposed to be overlapped to a lesser extent and they can be used to compute a filter-bank of spectro-temporal Wiener gains.

1) NMF and Wiener Filtering: NMF aims at approximating source spectral power $\mathbf{V}$ of size $F \times L$ by a spectral basis matrix $\mathbf{W} = [\mathbf{w}_1 \mathbf{w}_2 \ \dots \ \mathbf{w}_K]$ and a temporal activation matrix $\mathbf{H} = [\mathbf{h}_1 \mathbf{h}_2 \ \dots \ \mathbf{h}_K]^T$, so that $\mathbf{V} \approx \tilde{\mathbf{V}} = \mathbf{WH}$. $\mathbf{w}_k$ is a spectral basis vector of length F and $\mathbf{h}_k$ is a temporal activation vector of length L, and both have non-negative entries. Here, the decomposition is performed by minimizing the error measured by the Kullback-Leibler (KL) divergence [14]. $\mathbf{W}$ and $\mathbf{H}$ are iteratively updated to minimize the divergence by alternating the following multiplicative update algorithm [15]

$$\mathbf{W} \leftarrow \mathbf{W} \odot \frac{[\mathbf{V} \odot (\mathbf{WH})^{\cdot -1}]\mathbf{H}^T}{\mathbf{1} \ \mathbf{H}^T}, \tag{4}$$

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\mathbf{W}^T[\mathbf{V} \odot (\mathbf{WH})^{\cdot -1}]}{\mathbf{W}^T \ \mathbf{1}}, \tag{5}$$

where $\odot$ denotes the element-wise product, and the power and the division are also element-wise. The matrices are often initialized by random positive numbers. $\mathbf{1}$ is a matrix of ones of size $F \times L$ and ^{T} indicates the matrix transposition. After

ISBN 978-90-827970-1-5 © EURASIP 2018

each update of $\mathbf{W}$, its columns are normalized using the l_1-norm and the rows of $\mathbf{H}$ are scaled, accordingly, in order to avoid the scaling indeterminacy. Given the matrices $\mathbf{W}$ and $\mathbf{H}$, spectro-temporal Wiener gains can be calculated through the outer product of the spectral basis vector $\mathbf{w}_k$ and its corresponding temporal activation one $\mathbf{h}_k$, as follows

$$\mathbf{G}_k = \frac{\mathbf{w}_k \mathbf{h}_k^T}{\mathbf{WH}}, \tag{6}$$

where the division is element-wise, $\mathbf{w}_k \mathbf{h}_k$ is a rank-1 matrix and $\mathbf{WH}$ is a rank-K matrix. The filter-bank of the Wiener gains is then obtained for all the K vectors, i.e., $\mathbf{G} = [\mathbf{G}_1 \;..\; \mathbf{G}_k \;..\; \mathbf{G}_K]$.

2) Signal Decomposition: The spectral power matrices of the training speech signal $s(t)$, the training noise signal $n(t)$, and the testing noisy speech signal $x(t)$ are factorized using NMF, i.e., $\mathbf{V}_s = |\mathbf{S}| \odot |\mathbf{S}| \approx \mathbf{W}_s\mathbf{H}_s$, $\mathbf{V}_n = |\mathbf{N}| \odot |\mathbf{N}| \approx \mathbf{W}_n\mathbf{H}_n$, and $\mathbf{V}_x = |\mathbf{X}| \odot |\mathbf{X}| \approx \mathbf{W}_x\mathbf{H}_x$, where $\mathbf{S}$ and $\mathbf{N}$ are the complex STFT coefficients of $s(t)$ and $n(t)$, respectively, and $|.|$ indicates the absolute value. Subsequently, as explained in Section III-A1, filter-banks of spectro-temporal Wiener gains for each signal is obtained using the factorization of its corresponding spectral power matrix, as in (6), namely, $\mathbf{G}^s = [\mathbf{G}_1^s \;..\; \mathbf{G}_{k_s}^s \;..\; \mathbf{G}_{K_s}^s]$, $\mathbf{G}^n = [\mathbf{G}_1^n \;..\; \mathbf{G}_{k_n}^n \;..\; \mathbf{G}_{K_n}^n]$, and $\mathbf{G}^x = [\mathbf{G}_1^x \;..\; \mathbf{G}_{k_x}^x \;..\; \mathbf{G}_{K_x}^x]$. K_s, K_n, and K_x are the ranks of the approximations of the matrices $\mathbf{V}_s$, $\mathbf{V}_n$, and $\mathbf{V}_x$, respectively. Each signal is then decomposed into non-overlapping parts using its corresponding filter-bank, namely

$$\mathbf{S}_{k_s} = \mathbf{G}_{k_s}^s \odot \mathbf{S}, \quad \mathbf{N}_{k_n} = \mathbf{G}_{k_n}^n \odot \mathbf{N}, \quad \text{and} \quad \mathbf{X}_{k_x} = \mathbf{G}_{k_x}^x \odot \mathbf{X}.$$

The k th parts of the training speech signal $s_{k_s}(t)$, the training noise signal $n_{k_n}(t)$, and the testing noisy speech signal $x_{k_x}(t)$ are obtained in the time domain via the ISTFT of $\mathbf{S}_{k_s}$, $\mathbf{N}_{k_n}$, and $\mathbf{X}_{k_x}$, respectively. The parts $s_{k_s}(t)$ with k_s from 1 to K_s, and $n_{k_n}(t)$ with k_n from 1 to K_n, are used to train the SVM. The parts $x_{k_x}(t)$ with k_x from 1 to K_x are classified afterwards using the trained SVM.

Fig.2 shows examples of using NMF and Wiener filtering to decompose a clean speech signal, a bird noise signal, and their linear mixture signal into 5 parts. By focusing our attention on the last row and comparing the signal parts to the ones in the upper two rows, we can easily observe that the proposed method is capable of producing separable parts of the noisy speech signal.

B. Training the SVM Classifier

The SVM [17] is a discriminative classifier that attempts to model the boundary between two classes of data by finding the maximum margin separation hyper-plane such that it generalizes well to the test data. The SVM is trained using the MFCCs extracted from the parts of the training speech signal $s_{k_s}(t)$ with k_s from 1 to K_s, and the parts of the training noise signal $n_{k_n}(t)$ with k_n from 1 to K_n. To compute the MFCCs, each part is first segmented into short-length frames and the spectral power of each frame is passed through a set of filters, linearly spaced on the mel-frequency scale. The MFCCs are the amplitudes of the discrete cosine transform (DCT) taken from the output energy of the filters on a logarithmic scale [18]. The MFCCs extracted from each signal part are averaged over the frames in order to form one vector.

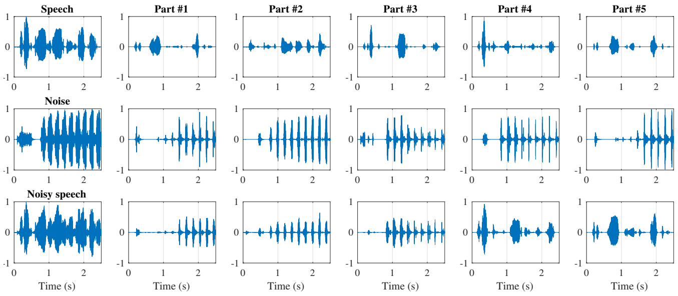


Fig. 2. Decomposition of a speech signal, a noise signal, and their linear mixture.

C. Enhancement of the Noisy Speech Signal

The enhancement step is performed by identifying the less-corrupted parts of the noisy speech signal and then combining them together to obtain an enhanced speech signal. Such decision is taken by the trained SVM using the vector of the MFCCs extracted from x_{k_x}. Similarly to the training phase, the MFCCs extracted from each signal part of the noisy speech signal are averaged over the frames in order to form one vector. Once all the K_x parts are classified, the ones detected as the speech parts are summed up in the time domain to reconstruct the enhanced signal.

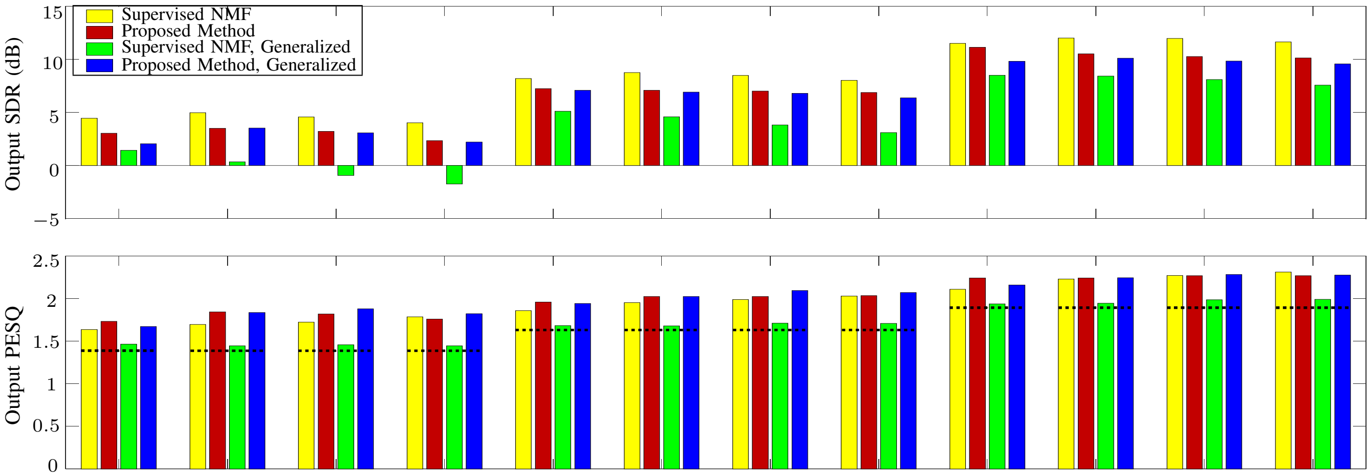
IV. Experimental Evaluation

The experiments were carried out using 25 utterances uttered by 5 English speakers, namely, 3 males and 2 females, from the NOIZEUS dataset [21], with an average length of 2.5 s. We used 4 signals of 4 different types of non-stationary noise, namely, train, bird, restaurant, and keyboard.

In order to investigate the enhancement performance of the proposed method, we considered two different scenarios:

1) The whole data were used for the training and the testing. In this case, the 25 utterances of the 5 different speakers and the 4 signals of the 4 different types of noise were used for the training and the testing. This resulted in 100 noisy utterances for the testing.
2) The 5 utterances of a target speaker and the signal of target noise in a noisy signal under testing were excluded from the training data. In this case, 20 utterances of 4 different speakers and 3 signals of 3 different types of noise were used for the training, and 5 utterances of the target speaker and one signal of the target noise were used for the testing. This procedure is repeated for each speaker and each type of noise, which resulted in a total of 100 noisy utterances for the testing.

The noisy observations $x(t)$ were generated by linearly mixing testing utterances and testing noise signals at 3 different input SNR, namely -5, 0 and 5 dB. The training is done by



K	10	30	50	70	10	30	50	70	10	30	50	70
SNR	-5 dB	-5 dB	-5 dB	-5 dB	0 dB	0 dB	0 dB	0 dB	5 dB	5 dB	5 dB	5 dB

Fig. 3. The enhancement results averaged over 100 noisy utterances in terms of K and the input SNR. The supervised NMF and the proposed methods when the testing data is excluded from the training data are labeled as “Generalized”. The dotted-line in the bottom plot indicates the input PESQ.

concatenation of the training utterances in one speech signal, $s(t)$, and the training noise signals in one noise signal, $n(t)$. In the proposed method, the concatenated signals, $s(t)$ and $n(t)$, are decomposed and the MFCCs extracted from each part are used to train the SVM classifier. The enhanced signal is then obtained by combining the less-corrupted parts of $x(t)$ classified as speech.

The proposed method was compared to speech enhancement based on supervised NMF using the KL divergence [15]. In the supervised NMF, the matrices $\mathbf{W}_s$ and $\mathbf{W}_n$ are trained independently by factorizing the spectral power matrices $\mathbf{V}_s$ and $\mathbf{V}_n$ of $s(t)$ and $n(t)$, respectively, as in (4) and (5). The trained matrices are used later to estimate temporal activation matrices $\mathbf{H}_s$ and $\mathbf{H}_n$ of speech and noise signals based on the spectral power matrix $\mathbf{V}_x$ of $x(t)$ for the enhancement as in (5). The enhanced speech signal is obtained by means of Wiener filtering using the trained and estimated matrices. The enhancement performance was evaluated using the signal-to-distortion ratio (SDR) in decibels (dBs) [22] and the perceptual evaluation of speech quality (PESQ) [23] on a scale from 0 to 4.5. The higher the PESQ value, the better the signal quality.

A. Implementation Details

The complex STFT coefficients $\mathbf{S}$, $\mathbf{N}$, and $\mathbf{X}$ of $s(t)$, $n(t)$, and $x(t)$, respectively, are obtained using a hamming analysis window with length of 32 ms (i.e., 512 samples at 16 kHz) and shift 16 ms. The spectral power matrices $\mathbf{V}_s$, $\mathbf{V}_n$ and $\mathbf{V}_x$ for computing the sets of the filter-banks $\mathbf{G}^s$, $\mathbf{G}^n$, and $\mathbf{G}^x$ were factorized by iterating the KL factorization algorithm in (4) and (5) for 300 times, with $K_s = K_n = K_x = K$. For the supervised NMF, the matrices $\mathbf{W}_s$ and $\mathbf{W}_n$ were trained independently by using $\mathbf{V}_s$ and $\mathbf{V}_n$, respectively, with the number of spectral basis vectors equals K.

For the classification, the signal parts $s_{k_s}(t)$, $n_{k_n}(t)$, and $x_{k_x}(t)$ are segmented into short frames of 30 ms long and 10 ms overlap using a hamming window. Then, for each frame, 12 MFCCs together with the log-energy are extracted to produce a 13 dimensional feature vector. To have one vector per each signal part, the MFCCs are averaged over the frames.

B. Results and Discussion

Fig.3 shows the enhancement results averaged over all the speech and noise signals in terms of K and the input SNR. In this figure, the results of the second scenario, described in Section IV, are labeled as “Generalized”.

We can observe that the supervised NMF provides better performance compared with the proposed method in terms of the output SDR when the utterances of a target speaker and the signal of target noise are included in the training data. However, the proposed method provides an enhanced speech signal with comparable or slightly better quality according to the output PESQ values.

In the second scenario, in which the utterances of a target speaker and the signal of target noise are excluded from the training data, the enhancement performance of the supervised NMF is degraded. This is probably because the training data could not provide spectral basis vectors describing well the speech and noise signals in the noisy signal for the good reconstruction of an enhanced speech signal. However we observe that the proposed method can better generalize the problem. This is due to the fact that different noise types with similar characteristics have similar impact on the distribution of the MFCCs of a speech signal. That is, the mean and the covariance of the MFCCs extracted from a speech signal are similarly modified if the speech signal is corrupted by different noise signals with similar characteristics [20]. For this reason, we have achieved a comparable performance when the proposed method is applied for the enhancement in both scenarios. It suggests that the classifier can generalize for unseen noise types if they have similar characteristics (and not necessarily the same noise signal) to those present in the training data.

V. Conclusion

A single-channel speech enhancement method for suppressing non-stationary noise in a noisy speech signal has been presented. We have tried to overcome the problem of missing training data by using a trained classifier instead of trained models based on NMF. The classifier is trained on relevant and irrelevant data. Moreover, NMF followed by Wiener filtering are applied to decompose the noisy speech signal into less-overlapping parts. These parts are then contributed in reconstructing the enhanced speech signal by linearly combining the less-corrupted ones, detected in the cepstral domain using a SVM classifier trained on MFCCs extracted from the parts of training speech signals and the parts of training noise signals. The performance of the proposed method has been evaluated using utterances of different speakers and signals of different types of noise at different input SNR levels. The experimental results showed that the proposed method can better generalize, comparing to the supervised NMF, when the signal of the target noise and the utterances of the target speaker are not included in the training data.

References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Inc., Boca Raton, FL, USA, 2nd edition, 2013.

[2] J. Benesty, M. G. Christensen, and J. R. Jensen, *Signal enhancement with variable span linear filters*, Springer, USA, 2016.

[3] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *ICASSP*, 1999.

[4] K. Manohar and P. Rao, "Speech enhancement in nonstationary noise environments using noise properties," *Speech Communication*, vol. 48, no. 1, pp. 96–109, 2006.

[5] C. Plapous, C. Marro, and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *Trans. Audio, Speech and Lang. Proc.*, vol. 14, no. 6, pp. 2098–2108, Nov. 2006.

[6] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay.," *IEEE Trans. Audio, Speech & Language Processing*, vol. 20, no. 4, pp. 1383–1393, 2012.

[7] L. Zao, R. Coelho, and P. Flandrin, "Speech enhancement with EMD and hurst-based mode selection," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 22, no. 5, pp. 897–909, 2014.

[8] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. K. Duong, "The signal separation evaluation campaign (2007-2010): achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, 2012.

[9] P. Smaragdis, B. Raj, and M. V. S. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *ICA*. 2007, vol. 4666 of *Lecture Notes in Computer Science*, pp. 414–421, Springer.

[10] H. Laurberg, M. G. Christensen, M. D. Plumbley, L. K. Hansen, and S. H. Jensen, "Theorems on positive data: on the uniqueness of NMF," *Comp. Int. and Neurosc.*, 2008.

[11] M. Fakhry, P. Svaizer, and M. Omologo, "Audio source separation in reverberant environments using β-divergence-based nonnegative factorization," *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 25, no. 7, pp. 1462–1476, 2017.

[12] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[13] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis*, Wiley, New York, NY, USA, 2009.

[14] Z. Yang, H. Zhang, Z. Yuan, and E. Oja, "Kullback-leibler divergence for nonnegative matrix factorization," in *Artificial Neural Networks and Machine Learning – ICANN 2011*, Berlin, Heidelberg, 2011, pp. 250–257.

[15] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the beta-divergence," *CoRR*, vol. abs/1010.1763, 2010.

[16] H. Laurberg, M. N. Schmidt, M. G. Christensen, and S. H. Jensen, "Structured non-negative matrix factorization with sparsity patterns," in *Asilomar Conference on Signals, Systems, and Computers*, 2008.

[17] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[18] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, New York, 2nd edition, 2000.

[19] A. H. Poorjam, J. R. Jensen, M. A. Little, and M. G. Christensen, "Dominant distortion classification for pre-processing of vowels in remote biomedical voice analysis," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 289–293.

[20] A. H. Poorjam, M. A. Little, J. R. Jensen, and M. G. Christensen, "A supervised approach to global signal-to-noise ratio estimation for whispered and pathological voices," in *ICASSP*, 2018.

[21] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7-8, pp. 588–601, 2007.

[22] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *ICA*. 2007, vol. 4666 of *Lecture Notes in Computer Science*, pp. 552–559, Springer.

[23] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, 2001, pp. 749–752.