

REDUCTION OF NON-STATIONARY NOISE USING A NON-NEGATIVE LATENT VARIABLE DECOMPOSITION

Mikkel N. Schmidt and Jan Larsen

Technical University of Denmark
Richard Petersens Plads, Bldg. 321
DK-2800 Kgs. Lyngby, Denmark
Email: {mns,jl}@imm.dtu.dk

ABSTRACT

We present a method for suppression of non-stationary noise in single channel recordings of speech. The method is based on a non-negative latent variable decomposition model for the speech and noise signals, learned directly from a noisy mixture. In non-speech regions an overcomplete basis is learned for the noise that is then used to jointly estimate the speech and the noise from the mixture. We compare the method to the classical spectral subtraction approach, where the noise spectrum is estimated as the average over non-speech frames. The proposed method significantly outperforms the classic approach, especially when the noise is highly non-stationary and at low signal-to-noise ratios.

1. INTRODUCTION

Reducing non-stationary background noise in single-channel recordings of speech is a challenging and important problem. One of the most successful methods, that is often used in practical applications, is spectral subtraction [1, 2] which is based on subtracting an estimate of the noise in the short-time spectral magnitude domain. When a good noise estimate is available, this approach generally leads to good results, but the problem is, of course, to find a good estimate of the noise.

One approach is to use a voice activity detector (VAD). Since we know that natural speech has many small pauses, the idea is to use the periods when there is no speech to estimate the noise. The noise spectrum is typically estimated by taking the average over a window of non-speech frames, and this estimated noise is then subtracted from the signal. This method has the drawback that since we cannot estimate the noise during speech activity, we must assume that it is stationary between each speech pause.

To construct a reliable VAD is difficult when the noise is non-stationary and at a low signal-to-noise ratio (SNR). There do exist spectral subtraction methods that do not require the use of a VAD, for example methods based on min-

imum or quantile statistics [3, 4, 5]. Here, the idea is to assume that the noise in all frequency bands is lower than the speech activity in that band, and thus the noise can be estimated by tracking for example the minimum or the median in each frequency band; however, as with the VAD-based spectral subtraction approach, these methods are unable [6] to track fast changing non-stationary noise.

Spectral subtraction systems consist of two important components: a noise estimation method and a suppression rule. In this paper we focus on the noise estimator, and propose a new method based on a non-negative latent variable decomposition. We do not discuss the suppression rule and post processing which is equally important for the final noise reduction system; however, our method can readily be combined with any existing suppression rule and post processing scheme. The method we propose makes no assumptions on the stationarity of the noise, and is related to our previous work on suppression of wind noise [7].

In the following section we present the noise estimation method. In section 3 we present experimental studies comparing the proposed method with classical spectral subtraction for a range of different types of noise at different SNRs, and finally we discuss the qualities of the proposed method.

2. METHOD

We propose a method for suppression of non-stationary noise which we apply in the log-frequency short-time Fourier transform magnitude domain. It is also possible to use linear-frequency, wavelet, perceptually weighted representations, etc., in the presented framework. The fundamental assumptions about the signal representation are: i) non-negativity; ii) approximate additivity of signal and noise components in the used representation; and iii) sufficient separation between signal and noise basis-vectors.

2.1. Non-negative latent variable model

We assume that the noisy signal is converted into a sequence of non-negative magnitude spectral vectors, $\{\mathbf{x}^{(i)}\}_{i=1}^I$, denoted frames, where i is the frame index. The noisy spectral vectors are modeled as the sum of three terms,

$$\mathbf{x}^{(i)} = \mathbf{s}^{(i)} + \mathbf{n}^{(i)} + \mathbf{r}^{(i)}, \quad (1)$$

where $\mathbf{s}^{(i)} \in \mathbb{R}_+^N$ is the speech, $\mathbf{n}^{(i)} \in \mathbb{R}_+^N$ is the noise, and $\mathbf{r}^{(i)} \in \mathbb{R}^N$ is a residual. Here, \mathbb{R}_+ denotes the non-negative real numbers, and N is the dimensionality of the spectral representation used. Since $\mathbf{x}^{(i)}$ is a non-negative spectral vector, and we assume that the speech and noise are additive in the spectral representation, the non-negativity of $\mathbf{s}^{(i)}$ and $\mathbf{n}^{(i)}$ follows.

We model the speech by a non-negative latent variable model,

$$\mathbf{s}^{(i)} = a^{(i)} \sum_{k=1}^{K_S} \boldsymbol{\sigma}_k b_k^{(i)}, \quad (2)$$

where $a^{(i)}$ is a binary variable that indicates if speech is present in frame i , $\{\boldsymbol{\sigma}_k\}_{k=1}^{K_S}$ is a set of basis vectors, and $\{\{b_k^{(i)}\}_{k=1}^{K_S}\}_{i=1}^I$ is a set of weights.

We make the assumption that both $\boldsymbol{\sigma}_k$ and $b_k^{(i)}$ are non-negative, similar to the approach in non-negative matrix factorization [8, 9]. We further assume that b is sparse, i.e., most of its elements are zero. It has been shown [9] that enforcing non-negativity in the decomposition leads naturally to a ‘‘parts based’’ representation. Since only additive combinations are allowed, cancellation of features cannot occur, and this often leads to meaningful decompositions. Enforcing sparsity of the weights leads to solutions where only a few basis vectors are active simultaneously yielding basis vectors that are more specific, since they are forced to resemble the data. It also allows for an over-complete factorization, i.e., computing more basis vectors than the dimensionality of the data matrix.

These assumptions about $\boldsymbol{\sigma}_k^{(i)}$ and $b_k^{(i)}$ are expressed in terms of prior distributions for the two sets of variables. For the basis vectors, $\boldsymbol{\sigma}_k$, we assume a flat (improper) prior over the non-negative reals. For the weights, $b_k^{(i)}$, we assume an i.i.d. one-sided exponential distribution with rate λ_B

$$p(b_k^{(i)}) = \lambda_B \exp(-\lambda_B b_k^{(i)}), \quad (3)$$

which leads to a sparse solution. We assume a flat prior over $a^{(i)}$.

We model the noise using a similar non-negative latent variable decomposition model,

$$\mathbf{n}^{(i)} = \sum_{k=1}^{K_N} \boldsymbol{\nu}_k c_k^{(i)}, \quad (4)$$

where the prior distribution for $c_k^{(i)}$ is an i.i.d. one-sided exponential with rate λ_C .

We assume that the residual, $\mathbf{r}^{(i)}$, is i.i.d. normal with zero mean and variance σ_r^2 ,

$$p(\mathbf{r}_n^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma_r} \exp\left(-\frac{(\mathbf{r}_n^{(i)})^2}{2\sigma_r^2}\right) \quad (5)$$

We can write the model more compactly in a matrix notation as

$$\mathbf{X} = \mathbf{SBA} + \mathbf{NC} + \mathbf{R}, \quad (6)$$

where $\mathbf{X} = [\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(I)}]$ is the data matrix, $\mathbf{A}_{ii} = a^{(i)}$ is a diagonal binary speech activity matrix, $\mathbf{S} = [\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_{K_S}]$ is a speech basis matrix, and $\mathbf{B}_{ki} = b_k^{(i)}$ is a weight matrix for the speech basis, $\mathbf{N} = [\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_{K_N}]$ is the noise basis matrix, $\mathbf{C}_{ki} = c_k^{(i)}$ is the noise weight matrix, and $\mathbf{R} = [\mathbf{r}^{(1)}, \dots, \mathbf{r}^{(I)}]$ is a matrix of residuals.

When we let $\boldsymbol{\theta} = \{\{a^{(i)}\}, \{b_k^{(i)}\}, \{c_k^{(i)}\}, \{\boldsymbol{\sigma}_k\}, \{\boldsymbol{\nu}_k\}\}$ denote all parameters in the model, we may then write the posterior distribution of the parameters as

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{X}) &= (2\pi\sigma_r^2)^{-IN/2} \exp\left(-\frac{\|\hat{\mathbf{X}}\|_F^2}{2\sigma_r^2}\right) + \\ &(\lambda_B)^{IK_S} \exp(-\lambda_B \|\mathbf{B}\|_1) + \\ &(\lambda_C)^{IK_N} \exp(-\lambda_C \|\mathbf{C}\|_1) \\ &\text{s.t. } \mathbf{S}, \mathbf{N} \geq \mathbf{0}, \end{aligned} \quad (7)$$

where $\hat{\mathbf{X}} = \mathbf{SBA} + \mathbf{NC}$.

2.2. Reduction of non-stationary noise

In this section, we discuss how the proposed model can be used for reduction of non-stationary noise in a recording of a speech signal. The starting point is the non-negative time-frequency matrix, \mathbf{X} , of the noisy signal, and the task is to estimate the clean speech. In principle, to make inference about the speech using the model, we could marginalize the posterior distribution (7) and compute, for example, a posterior mean estimate of the speech. This leads to a difficult integral that can be computed for example using Monte Carlo methods; however, here we take a much simpler ad hoc approach.

We propose a three step procedure: First, we compute speech activity, $a^{(i)}$, using a state-of-the-art voice activity detector. Second, we compute the noise basis and weights using the frames of the signal identified as non-speech, which gives an initial estimate of the noise. By computing the gradient of the log-posterior distribution with respect to \mathbf{N} and \mathbf{C} we construct a multiplicative gradient descent algorithm, similar to the multiplicative algorithms for non-

negative matrix factorization [10]. This gives us the following updates, that are iterated until convergence,

$$\mathbf{N} \leftarrow \mathbf{N} \otimes (\mathbf{X} \tilde{\mathbf{A}} \mathbf{C}^\top) \oslash (\mathbf{N} \mathbf{C} \tilde{\mathbf{A}} \mathbf{C}^\top) \quad (8)$$

$$\mathbf{C} \leftarrow \mathbf{C} \otimes (\mathbf{N}^\top \mathbf{X}) \oslash (\mathbf{N}^\top \mathbf{N} \mathbf{C} + \lambda_C), \quad (9)$$

where \otimes and \oslash denotes element-wise multiplication and division, and $\tilde{\mathbf{A}}_{ii} = 1 - a^{(i)}$ is a diagonal binary matrix that identifies frames with no speech activity. Third, we jointly compute the final estimate of the noise weights and the basis and weights for the speech, using a similar iterative algorithm,

$$\mathbf{S} \leftarrow \mathbf{S} \otimes (\mathbf{X} \mathbf{A} \mathbf{B}^\top) \oslash (\hat{\mathbf{X}} \mathbf{A} \mathbf{B}^\top) \quad (10)$$

$$\mathbf{B} \leftarrow \mathbf{B} \otimes (\mathbf{S}^\top \mathbf{X}) \oslash (\mathbf{S}^\top \hat{\mathbf{X}} + \lambda_B) \quad (11)$$

$$\mathbf{C} \leftarrow \mathbf{C} \otimes (\mathbf{N}^\top \mathbf{X}) \oslash (\mathbf{N}^\top \hat{\mathbf{X}} + \lambda_C). \quad (12)$$

Finally, the speech and noise is estimated as

$$\hat{\mathbf{X}}_N = \mathbf{S} \mathbf{B} \quad \hat{\mathbf{X}}_S = \mathbf{N} \mathbf{C}. \quad (13)$$

These estimates are then combined with the original noisy signal using an appropriate suppression rule to compute the final refiltered speech waveform. In our experiments, we simply multiply the complex spectrum of the original noisy time frequency matrix by $\hat{\mathbf{X}}_S \oslash (\hat{\mathbf{X}}_S + \hat{\mathbf{X}}_N)$, and generate the time signal estimate by computing the inverse short-time Fourier transform.

3. EXPERIMENTAL RESULTS

Our experiments were aimed at evaluating the proposed method in comparison with classical spectral subtraction where the noise is estimated by averaging over non-speech frames.

We used four different types of noise: a machine gun firing shots in bursts, a string quartet playing Händel, background noise from a restaurant, and the noise of cars passing by in traffic. These four types of noise represent different degrees of non-stationarity. We mixed 100 sentences from the TIMIT speech database with the four types of noise at different SNRs ranging from -9 dB to 6 dB resulting in more than 4.5 hours of test audio. We inserted a one second pause before and after the speech signal to make sure every mixture contained a non-speech part for estimating the noise characteristics. The signals were sampled at 8 kHz, and we computed the short-time Fourier transform (STFT) using a 64 ms Hann window with 50 percent overlap. To reduce the dimensionality we mapped the magnitude STFT onto 32 MEL frequency bins between 20 Hz and 4 kHz. We used a $K_S = K_N = 256$ dimensional basis for the speech and noise, and optimal sparsity parameters were found as $\lambda_B = 0.1$ and $\lambda_C = 0$.

The voice-activity detector is a crucial part of the noise reduction system, and by comparing a state of the art system with the ideal we can assess the importance of an accurate VAD. The VAD used in the Qualcomm-ICSI-OGI [11]

noise reduction frontend represents the state of the art, and was used to estimate the regions of each mixture containing speech. The ideal VAD was simply computed from the knowledge of the clean speech signal in the synthesis process.

In the classical spectral subtraction results, we simply computed the noise spectrum as the average over frames classified as non-speech by the VAD. Spectral bins where the result was negative were set to zero, [2]. The estimate was then combined with the phase of the noisy signal, and then the final signal estimate was computed by inverse STFT.

An example of the proposed method is given in Figure 1. The improvement in average SNR for the four different types of noise and two different VADs is shown in Figure 2.

4. DISCUSSION

The results in Figure 2 show a clear trend. The proposed method outperforms the classical approach in almost all conditions, especially for the most non-stationary types of noise and when the signal-to-noise ratio is low. Both approaches perform better with the ideal voice activity detector over the Qualcomm-ICSI-OGI VAD, and the proposed method appears somewhat more sensitive to a non-ideal voice activity detection.

The machine gun noise posed a serious problem for the classical spectral subtraction approach, whereas the proposed method handled this type of noise very well. Qualitatively, the classical approach “muffles” the sound of the machine gun without providing a significant attenuation. In addition it gives a distortion of the speech. The proposed method attenuates the machine gun sound significantly and introduces only a very small speech distortion.

In the case of restaurant noise, the string quartet (noise), and the traffic noise, the results are very similar. With the ideal voice activity detector, the proposed method significantly outperforms the classical approach in terms of signal-to-noise ratio in almost all conditions, especially at low signal-to-noise ratios.

Qualitatively there is a big difference between the two methods: the classical approach reduces the noise but introduces musical noise — a well known problem with spectral subtraction, which can be reduced using a more advanced suppression rule as well as post processing. The output of the proposed method has no musical noise; rather, the residual noise sounds like attenuated white noise that is slightly distorted and fluctuates a bit with the speech.

Audio samples related to this article are available online¹.

¹<http://www.mikkelschmidt.dk/mlsp2008>

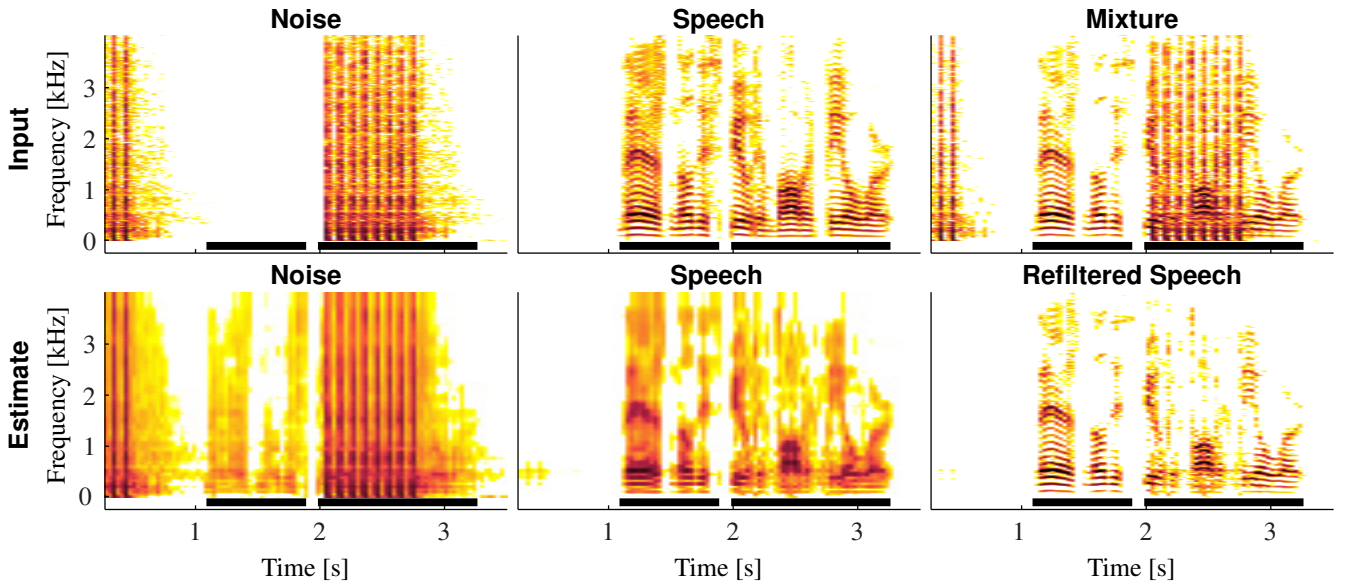


Fig. 1. Example of the separation of speech from machine gun noise. The solid black lines indicate frames classified as speech. The top row shows the noise, the clean speech, and the mixed input signal. The bottom row shows the estimated noise and speech, and refiltered final result. The estimates appear slightly blurred because of the dimensionality reduction.

5. CONCLUSION

We have presented a non-negative latent variable decomposition model for reduction of non-stationary noise as well as a simple procedure for estimating the model parameters and performing noise reduction. The model makes only weak assumptions about the noise and the speech, and all parameters are estimated directly from a noisy mixed signal. In future work we will focus on improving the model in two areas: We expect that the noise reduction can be improved by i) specifying more detailed prior distributions, for example a harmonic prior for the speech basis vectors and a hidden Markov prior for the speech activity variable; and ii) by making direct inference in the model, for example using Markov chain Monte Carlo methods. Furthermore, it might be interesting to use a more accurate model for the residual, along the lines of Parry and Essa [12, 13].

The results presented in this paper suggest that the proposed noise reduction method based on non-negative latent variable decomposition has an advantage over the classical approach, especially in the case of low SNR and highly non-stationary noise. The method was successfully evaluated for a range of noise types mixed with speech from the TIMIT database. The results show that the voice activity detector is an essential component in the suggested method. The state-of-the-art voice activity detector, used in the Qualcomm-ICSI-OGI system, gives a signal-to-noise performance which is slightly worse than ideal voice activity detection in most conditions. However, it is well-known

that signal-to-noise ratio performance does not necessarily correlate well with subjective measures such as with perceived sound quality or speech intelligibility, which is also confirmed by our informal listening tests.

Many state of the art noise reduction systems are based on estimating the noise by averaging over noise frames combined with advanced post processing. We believe these systems can be improved by using a more advanced noise estimator such as the one presented here.

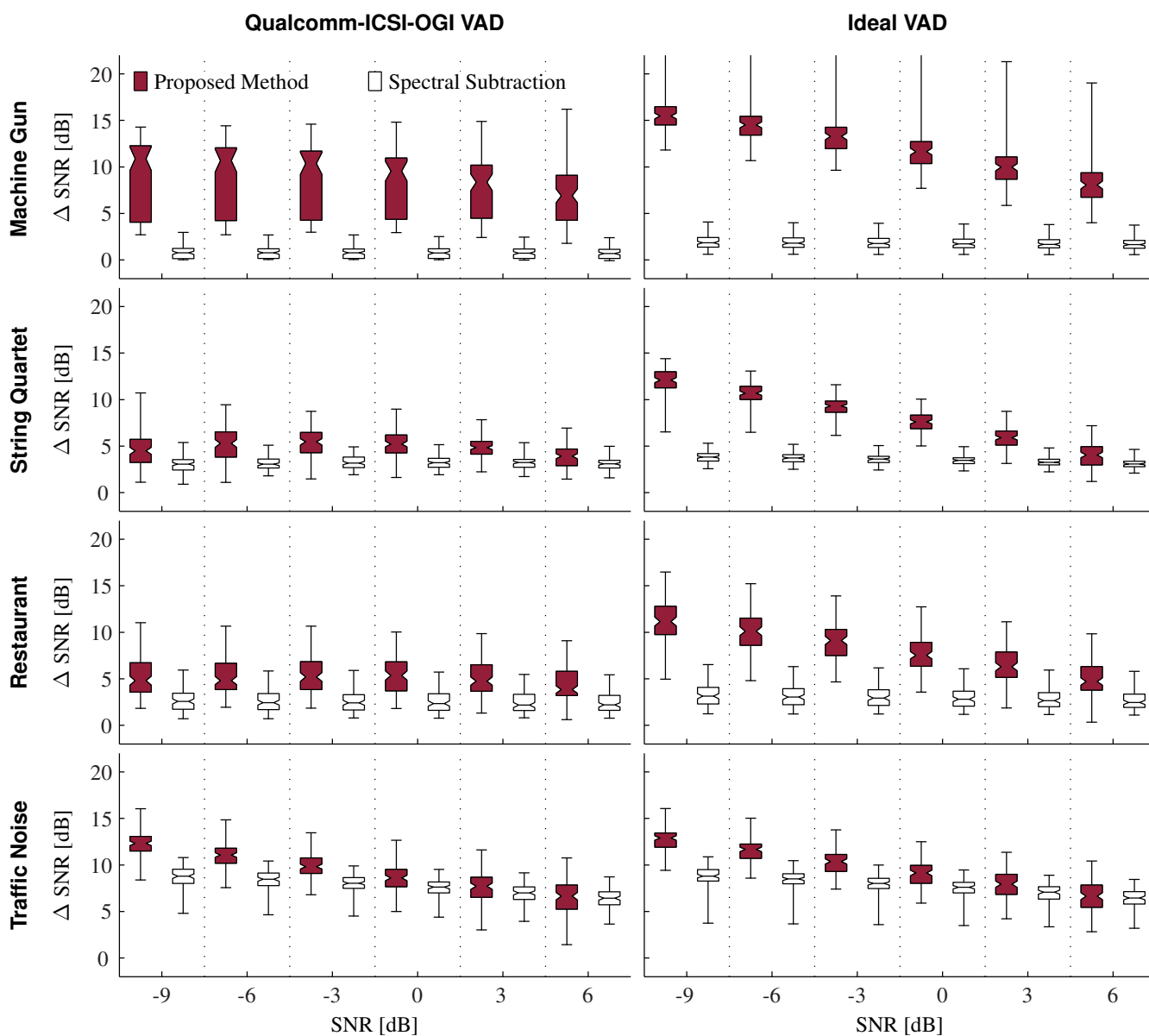


Fig. 2. Improvement in average signal-to-noise ratio (SNR) for the proposed method compared with classical spectral subtraction at different input SNRs. The left column is results using the Qualcomm-ICSI-OGI VAD and the right column is results using an ideal VAD. The rows indicate different types of noise. The proposed method outperforms spectral subtraction in almost all conditions, especially for the most non-stationary types of noise and for the worst SNRs.

6. REFERENCES

- [1] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on (ICASSP)*, vol. 4, 1979, pp. 208–211.
- [2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, 1979.
- [3] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 5, pp. 504–513, 2001.
- [4] —, "Spectral subtraction based on minimum statistics," in *European Signal Processing Conference, Proceedings of (EUSIPCO)*, 1994, pp. 1182–85.
- [5] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and wiener filtering," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on (ICASSP)*, vol. 6, 2000, pp. 1875–78.
- [6] K. Manohar and R. Preeti, "Speech enhancement in nonstationary noise environments using noise properties," *Speech Communication*, vol. 48, no. 1, pp. 96–109, 2006.
- [7] M. N. Schmidt, J. Larsen, and F.-T. Hsiao, "Wind noise reduction using non-negative sparse coding," in *Machine Learning for Signal Processing, IEEE International Workshop on*, Aug 2007, pp. 431–436.
- [8] P. Paatero and U. Tapper, "Positive matrix factorization: A nonnegative factor model with optimal utilization of error-estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, Jun 1994.
- [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] —, "Algorithms for non-negative matrix factorization," in *Neural Information Processing Systems, Advances in (NIPS)*, 2000, pp. 556–562.
- [11] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivasdas, "Qualcomm-icsi-ogi features for asr," in *Spoken Language Processing, ISCA International Conference on (INTERSPEECH)*, 2002, pp. 21–24.
- [12] R. M. Parry and I. Essa, "Incorporating phase information for source separation via spectrogram factorization," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on (ICASSP)*, vol. 2, 2007, pp. 664–664.
- [13] —, "Phase-aware non-negative spectrogram factorization," in *Independent Component Analysis and Blind Signal Separation, International Conference on (ICA)*, ser. Lecture Notes in Computer Science (LNCS), vol. 4666. Springer, Sep 2007, pp. 536–543.