

A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics

Gautham Jean Mysore, Paris Smaragdis

▶ **To cite this version:**

Gautham Jean Mysore, Paris Smaragdis. A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics. IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP 2011, May 2011, Prague, Czech Republic. pp.17-20, 10.1109/ICASSP.2011.5946317 . hal-01084331

HAL Id: hal-01084331
https://hal.science/hal-01084331

Submitted on 19 Nov 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A NON-NEGATIVE APPROACH TO SEMI-SUPERVISED SEPARATION OF SPEECH FROM NOISE WITH THE USE OF TEMPORAL DYNAMICS

Gautham J. Mysore

Advanced Technology Labs
Adobe Systems Inc.

Paris Smaragdis

University of Illinois at Urbana-Champaign,
Adobe Systems Inc.

ABSTRACT

We present a semi-supervised source separation methodology to denoise speech by modeling speech as one source and noise as the other source. We model speech using the recently proposed non-negative hidden Markov model, which uses multiple non-negative dictionaries and a Markov chain to jointly model spectral structure and temporal dynamics of speech. We perform separation of the speech and noise using the recently proposed non-negative factorial hidden Markov model. Although the speech model is learned from training data, the noise model is learned during the separation process and requires no training data. We show that the proposed method achieves superior results to using non-negative spectrogram factorization, which ignores the non-stationarity and temporal dynamics of speech.

Index Terms— Semi-supervised source separation, Denoising

1. INTRODUCTION

Denoising of speech has been a problem of interest for several decades [1, 2]. It has various applications such as improved speech recognition performance and enhanced intelligibility in telephony. We approach this problem using a semi-supervised source separation methodology. This is well suited to the problem of interest because we can model one source as speech and the other source as noise.

Non-negative spectrogram factorization techniques have had a great deal of success in the source separation literature. This class of techniques refers to the use of non-negative matrix factorization (NMF) [3] as well as its probabilistic counterparts such as probabilistic latent component analysis (PLCA) [4] for source separation [5]. These techniques use a single non-negative dictionary to model each source. They can be quite powerful in modeling the spectral structure of the sources. However, they ignore the non-stationarity and temporal dynamics of the sources, which is an important aspect of speech.

Wilson et. al [6] proposed a non-negative spectrogram factorization approach to denoising by including temporal dependency constraints across time frames. Although this method does capture some amount of temporal structure and shows an improved performance over naive non-negative spectrogram factorization, it still uses a single dictionary for each source (ignoring non-stationarity) and does not have a model of the temporal dynamics of speech.

We use a recently proposed source separation technique [7] that jointly models the spectral structure and temporal dynamics of speech using a non-negative hidden Markov model (N-HMM). Using this model, speech is modeled using multiple non-negative dictionaries and a Markov chain. The mixture of speech and noise is modeled using a non-negative factorial hidden Markov model (N-FHMM). Unlike the recently proposed methodology, we perform separation in a semi-supervised manner. Specifically, we learn a model of speech from training data but we learn the model of noise while actually performing separation. Therefore, no training data is required for the noise. We show superior results to using non-negative spectrogram factorization in the presence of heavy real world noise.

2. PROPOSED METHOD

In this section, we describe the proposed denoising method. We start with an overview of the modeling strategy. We then briefly describe the probabilistic models that we employ. Finally, we describe the actual denoising methodology.

2.1. Overview

Speech and noise are both modeled in the spectrogram domain. Particularly, each time frame of the spectrogram is modeled as a linear combination of the spectral components from a dictionary. Previous non-negative spectrogram factorization techniques use a single dictionary to model an entire source. Therefore, when the mixture is composed of two sources, every single column of the mixture spectrogram is modeled as a linear combination of spectral components from the concatenation of the dictionaries of the two individual sources.

We also model noise using a single dictionary. However, we use a N-HMM [7] to model speech. This model uses multiple dictionaries in order to model the non-stationarity of speech. The speech in a given time frame of the spectrogram is modeled by a linear combination of the spectral components from one (of the many) dictionaries of the speech model. However, the noise (in any time frame of the spectrogram) is modeled by a linear combination of the spectral components from a single dictionary. A given time frame of the noisy speech (mixture) is therefore modeled as a linear combination of spectral components from the concatenation of one of the dictionaries of speech and the dictionary of noise. This is illustrated in Fig.1. An example of dictionaries

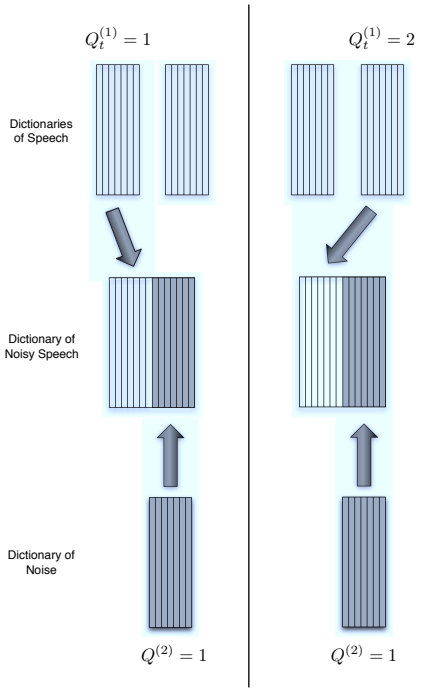


Fig. 1: Illustration of the possible combinations of dictionaries using the proposed method. In this simple example, speech is modeled using only two dictionaries.

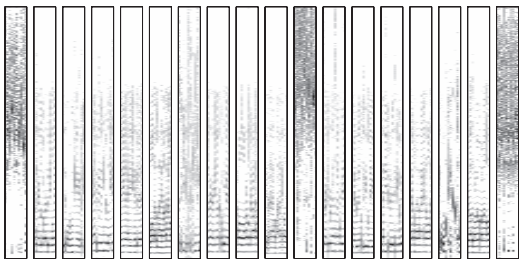


Fig. 2: Dictionaries of spectral components of speech. Eighteen (of the forty) dictionaries that were learned from a specific sample of speech are shown. Each dictionary contains ten spectral components and roughly corresponds to a subunit of speech.

that are used to model a specific sample of speech is shown in Fig.2. These dictionaries were learned from training data.

2.2. Probabilistic Model

The graphical model of the N-HMM is shown in Fig.3. Each dictionary corresponds to a state q. At time t, the N-HMM is in state q_t. Each spectral component of a given dictionary q is represented by z. A given spectral component is a multinomial distribution. Therefore, spectral component z of dictionary q is represented by $P(f|z,q)$. Since each column of the spectrogram of speech is modeled as a linear combination of spectral components, time frame t (modeled by state q) is

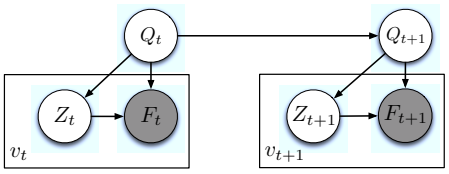


Fig. 3: Graphical model of the N-HMM.

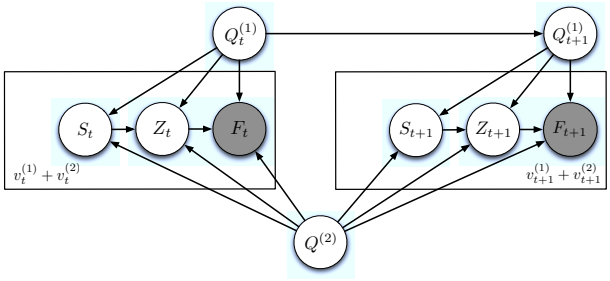


Fig. 4: Graphical model of the N-FHMM for denoising.

given by the following observation model:

$$P(f_t|q_t) = \sum_{z_t} P(f_t|z_t, q_t)P(z_t|q_t),$$

where $P(z_t|q_t)$ is a multinomial distribution of mixture weights for time t. The transitions between states are modeled with a Markov chain, given by $P(q_{t+1}|q_t)$.

We model the mixture of speech and noise with a N-FHMM [7], whose graphical model is shown in Fig.4. A N-HMM of speech can be seen in the upper half of the graphical model. A degenerate N-HMM (single state for all time frames) of noise can be seen in the lower half of the graphical model. A given time frame is modeled by a pair of dictionaries, $\{q_t^{(1)}, q^{(2)}\}$, one for each source. Of course, speech can be modeled by any one of many dictionaries whereas noise is modeled by a single dictionary. The interaction model (of the two sources) introduces a new variable s_t that indicates the ratio of the sources at a given time frame. $P(s_t|q_t^{(1)}, q^{(2)})$ is a Bernoulli distribution that depends on the states of the sources at the given time frame. The interaction model is given by:

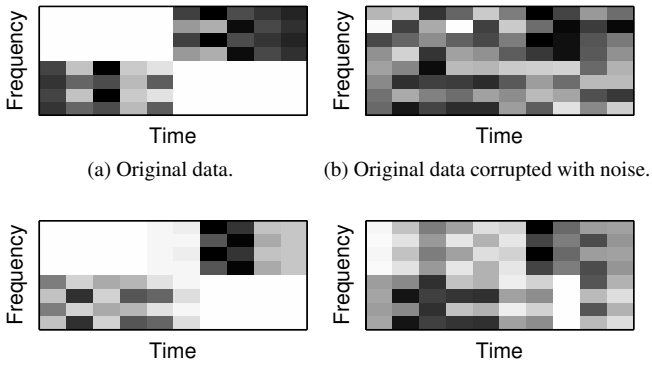
$$P(f_t|q_t^{(1)}, q^{(2)}) = \sum_{s_t}\sum_{z_t} P(f_t|z_t, s_t, q_t^{(s_t)})P(z_t, s_t|q_t^{(1)}, q^{(2)}),$$

where $P(f_t|z_t, s_t, q_t^{(s_t)})$ is spectral component z_t of state $q_t^{(s_t)}$ of source s_t. Of course, there is only one valid state for noise. $P(z_t, s_t|q_t^{(1)}, q^{(2)})$ gives the mixture weights for the spectral components of state $q_t^{(1)}$ of source $q^{(2)}$.

2.3. Denoising Methodology

The denoising procedure is as follows:

1. Learn the parameters of a N-HMM model of speech from clean speech training data (spectrogram), using the expectation-maximization (EM) algorithm.



(a) Original data.

(b) Original data corrupted with noise.

(c) Denoised data using the proposed method.

(d) Denoised data using non-negative spectrogram factorization.

Fig. 5: Illustration of denoising on a toy example using the proposed method and non-negative spectrogram factorization. The noise source is uniformly distributed random noise.

2. Model the spectrogram of noisy speech using a N-FHMM. Specify a subset of the parameters of the N-FHMM using the above learned parameters. Specifically, specify the parameters of the spectral components and transition matrix of speech.

3. Learn the remaining parameters of the N-FHMM using the EM algorithm.

4. Using these parameters, construct a soft mask by which to modulate the mixture spectrogram to obtain the denoised speech spectrogram.

5. Obtain the denoised speech waveform using the above denoised speech spectrogram.

The specific details of the EM equations and the construction of the soft mask can be found in [7]. The results of using this methodology are illustrated on toy data (representing toy spectrograms) in Fig.5. Two dictionaries of two spectral components each were used to model the clean toy data. As a comparison, the same experiment was repeated using non-negative spectrogram factorization [5]. In this case, four spectral components were used to model the clean toy data. Therefore, both methods use the same number of spectral components to model the clean training data. In both cases, an oracle example is used (noisy test data is obtained by adding random noise to the clean training data). As shown in the Fig.5, the proposed method does a better job of denoising than non-negative spectrogram factorization.

3. RESULTS AND DISCUSSION

We performed speech denoising experiments using the proposed method. As a comparison, we repeated the same experiments using non-negative spectrogram factorization. Specifically, we performed experiments on sixteen speakers (eight male and eight female) from the TIMIT database. For each speaker, we performed experiments using three different real world noises (ambient noise at an airport, traffic junction, and cafeteria). Therefore, we performed a total of forty-eight experiments for a given SNR.

For a given experiment, we obtained the training data by concatenating nine sentences of the given speaker. We obtained the spectrogram of this data using a window size of 1024 and a hop size of 256 (at Fs=16,000). We learned the N-HMM parameters from this data. Specifically, we learned 40 dictionaries of 10 spectral components each as well as the transition matrix. When using non-negative spectrogram factorization, we learned 1 dictionary of 30 spectral components each. These values were used because they were found to be optimal for source separation and a decrease in separation performance was observed when more than 30 spectral components per source were used for non-negative spectrogram factorization [7]. 1 dictionary of 10 spectral components was used to model noise in both cases.

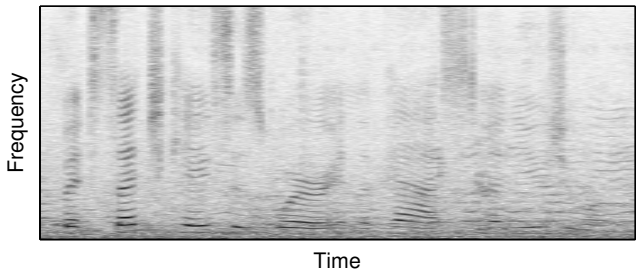
We then obtained the noisy speech by adding noise to an unseen sentence of the same speaker. Finally, we learned the unspecified parameters of the N-FHMM and reconstructed the clean speech. We repeated the experiment at three different SNRs.

Since this is essentially a source separation problem, we used the BSS-EVAL metrics [8] for evaluation. For each SNR, we report the average metrics (Table 1) on all sixteen speakers using all three types of noise. As a comparison, we report the same metrics using non-negative spectrogram factorization. One specific example is shown in Fig.6.

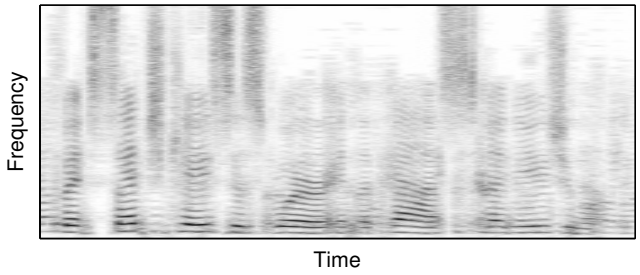
The actual suppression of the noise is reflected in the source to interference ratio (SIR). As shown in Table 1, the proposed method achieves superior results at all noise levels. The superior performance of the proposed method is more pronounced as the noise level increases (lower SNR). The artifacts that are introduced by the denoising process are reflected in the source to artifacts ratio (SAR). Non-negative spectrogram factorization introduces less artifacts than the proposed method. However, the difference is quite small for high noise levels (0dB and -3dB SNR). The overall performance is reflected in the source to distortion ratio (SDR). The proposed method performs better at high noise levels due to the higher noise suppression and only a small increase in artifacts. At 3dB SNR, non-negative spectrogram factorization does better due to a smaller difference in noise suppression and a larger difference in artifacts

The main reason for increased noise suppression capability (SIR) when using the proposed method is that it is a more structured and constrained model than non-negative spectrogram factorization. Particularly, every time frame of the speech part of the noisy speech spectrogram is explained by one (out of forty) dictionaries. Each dictionary contains spectral components that correspond roughly to one specific subunit of speech. Therefore, unless a given time frame corresponds to an unvoiced phoneme, the speech model will not be able to explain the noise very well, thus suppressing it. Moreover, the temporal dynamics help in deciding which dictionary is used to explain a given time frame of the data.

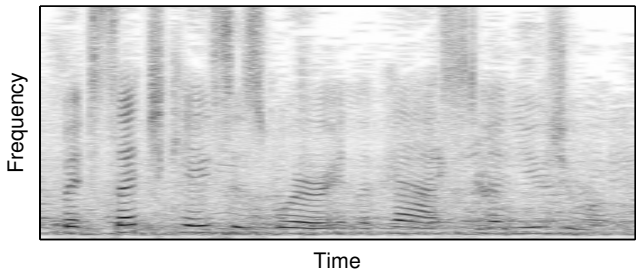
On the other hand non-negative spectrogram factorization uses a single dictionary to explain all of the voiced and unvoiced phonemes. Therefore, even if a time frame of the speech part of the noisy speech spectrogram corresponds to a voiced phoneme, certain spectral components from the dictionary will be able to explain noise in the same time frame.



(a) Noisy speech.



(b) Denoised speech using the proposed method.



(c) Denoised speech using non-negative spectrogram factorization.

Fig. 6: Illustration of speech denoising using the proposed method and non-negative spectrogram factorization. The noise source is ambient noise in an airport.

A similar argument can be used to explain the slightly higher artifacts that are introduced when using the proposed method. Since non-negative spectrogram factorization can use more spectral components to explain a given time frame of speech (thirty components rather than ten components), it can sometimes do a better job of explaining some of the subtleties of speech. It should be stressed that the difference in SAR is quite small at high noise levels.

These results suggest that the proposed method is preferable to non-negative spectrogram factorization for denoising at high noise levels.

4. CONCLUSIONS

We have presented a new method for denoising of speech by framing it as a semi-supervised source separation problem, using recently proposed probabilistic models. We have shown that it achieves superior results to using non-negative spectrogram factorization in heavy noise conditions. This is attributed mainly to the fact the N-HMM and N-FHMM are more structured models than non-negative spectrogram factorization, as they explicitly model non-stationarity and temporal dynamics of speech.

SIR (dB)	3dB	0dB	-3dB
Proposed Method	**17.24**	**12.95**	**6.66**
Factorization	12.26	6.49	0.61

SAR (dB)	3dB	0dB	-3dB
Proposed Method	8.17	7.82	4.70
Factorization	**12.27**	**8.90**	**5.38**

SDR (dB)	3dB	0dB	-3dB
Proposed Method	7.41	**6.22**	**1.49**
Factorization	**8.81**	3.83	-1.89

Table 1: Source separation metrics for various mixing levels (SNR) of speech and noise.

5. REFERENCES

[1] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

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

[4] P. Smaragdis, B. Raj, and M. Shashanka, "Probabilistic latent variable model for acoustic modeling," in *Advances in Models for Acoustic Processing Workshop, Neural Information Processing Systems*, Dec. 2006.

[5] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proceedings of the International Conference on Independent Component Analysis and Signal Separation*, Sept. 2007.

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

[7] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden markov modeling of audio with application to source separation," in *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation*, Sept. 2010.

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