Speech Enhancement by Online Non-negative Spectrogram Decomposition in Non-stationary Noise Environments

Zhiyao Duan[1]†*, Gautham J. Mysore*[2]*, and Paris Smaragdis*[2,3]

[1]Department of EECS, Northwestern University, Evanston, IL, USA
[2]Advanced Technology Labs, Adobe Systems Inc., San Francisco, CA, USA
[3]Department of Computer Science, UIUC, Urbana, IL, USA

Abstract

Classical single-channel speech enhancement algorithms have two convenient properties: they require pre-learning the noise model but not the speech model, and they work online. However, they often have difficulties in dealing with non-stationary noise sources. Source separation algorithms based on non-negative spectrogram decompositions are capable of dealing with non-stationary noise, but do not possess the aforementioned properties. In this paper we present a novel algorithm that combines the advantages of both classical algorithms and non-negative spectrogram decomposition algorithms. Experiments show that it significantly outperforms four categories of classical algorithms in non-stationary noise environments.

Index Terms: speech enhancement, source separation, non-negative matrix factorization, online algorithm

1. Introduction

Speech enhancement, which aims to improve the quality and intelligibility of noisy speech signals by reducing noise, has been an active research problem for decades. Numerous algorithms have been proposed and commercialized. According to [1], existing single-channel methods can be generally categorized into four classes: spectral subtraction [2], Wiener filtering [3], statistical-model-based [4] and subspace algorithms [5].

Despite the differences between these categories, they share two important properties. First, they only require an estimation of the noise model (e.g. spectrum, subspace, etc.) from noise-only excerpts, and do not require a speech model before the enhancement process. This is appealing since noise-only excerpts are relatively easy to obtain from noisy environments when speech is silent. On the other hand, speech-only excerpts are difficult to obtain as we rarely encounter real-world segments with no noise. Also, the structure of noise is usually simpler and has less variations over time than speech, which makes pre-learning of noise models more feasible. The other important property is that they are online algorithms, i.e. they enhance speech signals frame by frame in a sequence. This makes them applicable in real time scenarios like telephony.

These algorithms, however, are intrinsically not able to deal with non-stationary noise. Three out of the four categories of classical algorithms (spectral subtraction, Wiener filtering and statistical-model-based) rely on the assumption that the noise is stationary. They model noise as a single spectral profile, or use a single Speech-to-Noise Ratio (SNR) in enhancing different noisy frames. However, common noises like typing on a computer keyboard during video chatting, or background speech in teleconferencing are non-stationary.

†This work was performed while interning at Adobe Systems Inc.

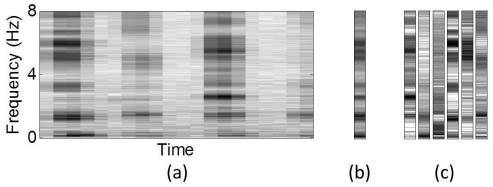

Figure 1: Comparison of noise models. (a) A spectrogram of 20 frames of computer keyboard noise. (b) Classical speech enhancement algorithms like spectral subtraction model noise with a single spectrum, e.g. the average spectrum. (c) Non-negative spectrogram decomposition uses a dictionary (multiple basis spectra). The latter is suitable for non-stationary noise.

These kinds of noise cannot be well represented by a single spectral profile and the SNR varies significantly over time. The category of subspace algorithms are theoretically able to model non-stationary noise using a high rank noise subspace. However, the assumption of orthogonality between the signal subspace and the noise subspace hinders noise suppression [1].

In [6], we proposed a source separation algorithm based on non-negative spectrogram decomposition, and applied it to speech enhancement in non-stationary environments. This algorithm pre-learns a noise model from noise-only excerpts beforehand, and separates noisy speech in an online fashion.

In this paper, we further this idea by explicitly modeling the tradeoff between noise reduction and speech distortion. We control this tradeoff through a single parameter. We also compare the proposed algorithm with four categories of classical speech enhancement algorithms. We show that the proposed algorithm achieves significantly better results in terms of both objective speech quality metrics as well as source separation metrics, in non-stationary noise environments with various SNRs.

2. Non-negative spectrogram decomposition

Non-negative spectrogram decomposition has been quite successful in modeling non-stationary sound sources [7]. The basic idea is to represent the spectra in a sound spectrogram as convex combinations of several basis spectra forming a dictionary. In the language of Probabilistic Latent Component Analysis (PLCA) [7], it can be written as:

$$P_t(f) \approx \sum_z P(f|z)P_t(z) \tag{1}$$

where $P_t(f)$ represents the normalized magnitude spectrum at frame t, which is viewed as a probability distribution over

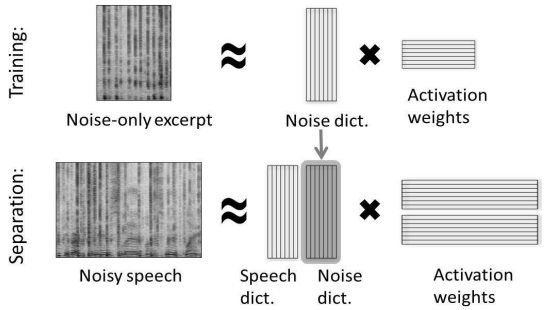

Figure 2: Illustration of semi-supervised source separation, based on non-negative spectrogram decomposition. During separation, the noise dictionary is fixed as what was trained, while the speech dictionary and activation weights are estimated.

frequency. $P(f|z)$ is a basis spectrum (component z) of the dictionary, and $P_t(z)$ is its activation weight at time t. This decomposition is achieved by minimizing the KL divergence $d(P_t(f)||Q_t(f))$ between the input spectra $P_t(f)$ and the reconstructed spectra $Q_t(f) = \sum_z P(f|z)P_t(z)$. All these distributions are discrete.

While the activation weights obtained from the decomposition might be different for different instances of the same source, the dictionary models invariant characteristics. Different spectral shapes can be approximated by different convex combinations of the dictionary basis spectra, making the dictionary suitable for modeling non-stationary sound sources. Figure 1 compares spectral subtraction and a non-negative spectrogram decomposition when modeling non-stationary noise.

Non-negative spectrogram decomposition has been successfully applied to source separation [8]. Applying it to speech enhancement where speech and noise are the sources, the idea is to first learn source dictionaries by decomposing clean training source excerpts using Eq. (1). The first row in Figure 2 shows the training process of the noise dictionary. The speech dictionary can be learned in the same way if speech-only excerpts are available. Now for each time frame of the noisy speech, we apply and fix these pre-learned dictionaries, and decompose its magnitude spectrum $P_t(f)$ as:

$$P_t(f) \approx \sum_{z \in \mathcal{S} \cup \mathcal{N}} P(f|z)P_t(z) \tag{2}$$

where $P(f|z)$ for $z \in \mathcal{S}$ represents the speech dictionary, and for $z \in \mathcal{N}$ represents the noise dictionary. The enhanced speech spectrum can be obtained by $\sum_{z \in \mathcal{S}} P(f|z)P_t(z)$.

If both the speech and noise dictionaries are pre-trained using Eq. (1) with clean source excerpts, then only the activation weights $P_t(z)$ need to be estimated in Eq. (2). This is called *supervised* separation [8]. In this case, Eq. (2), can be performed in each frame t independently, hence the algorithm is online.

However, speech-only training excerpts are often not available, as assumed in classical speech enhancement algorithms. In this case, only the noise dictionary can be trained beforehand, while the speech dictionary as well as the activation weights need to be estimated during the decomposition in Eq. (2). This is called *semi-supervised* separation [8], as shown in Figure 2. Existing algorithms of this type [8, 9] cannot be performed online, since the estimation of the speech dictionary needs to access the whole spectrogram of the noisy speech.

Recently, researchers have proposed several online algorithms [10, 11, 12] for dictionary learning, from a large amount of data, which would be restrictive to do offline due to the memory and computational requirements. However, they are not suitable for learning the speech dictionary in real-time semi-supervised source separation. This is because they are designed to learn a good dictionary only after going through the entire data set so that they can reflect the entire data. Applying them to our situation, the intermediate speech dictionary estimated at frame t is not necessarily good enough to explain and separate the present frame. In fact, with these approaches, even the final dictionary learned after processing all the inputs once is often not good enough. It has been shown that cycling over the inputs several times and randomly permuting them at each cycle significantly improves the results [10, 11, 12].

3. The proposed algorithm

In [6], we proposed an online semi-supervised source separation algorithm based on the idea of Eq. (2). Applying it to speech enhancement, we train the noise dictionary using noise-only excerpts beforehand. For each noisy speech frame, we first perform Voice Activity Detection (VAD). If the frame contains prominent speech, we estimate and keep updating the speech dictionary while enhancing this frame. Otherwise, we only enhance the frame but keep the speech dictionary unchanged.

To do VAD on each frame, we decompose its spectrum using Eq. (1) with the pre-trained noise dictionary fixed and only estimate the activation weights. If the reconstruction error is smaller than a threshold, then it means the noise dictionary itself can explain the frame well enough, indicating that there is no prominent speech signals. In this case, we perform supervised separation on this frame, i.e. we decompose this frame using Eq. (2) with the noise dictionary fixed as what was pre-trained, and the speech dictionary fixed as what has been updated so far.

If the decomposition error is larger than the threshold, then the noise dictionary itself is not enough to explain the frame, hence speech signals are probably present. In this case, we perform semi-supervised separation in this frame, i.e. besides enhancing the speech signals, the speech dictionary is also estimated. We achieve this by decomposing the current frame together with a weighted buffer of some previous frames that are classified as having speech signals. Decomposing the current frame gives us the separation of speech and noise in the frame, while decomposing buffer frames provides us a constraint on the speech dictionary to prevent overfitting. Mathematically, we are solving the following optimization problem:

$$\underset{\substack{P(f|z) \text{ for } z \in \mathcal{S} \\ P_t(z) \text{ for } z \in \mathcal{S} \cup \mathcal{N}}}{\arg\min} \; d(P_t(f)||Q_t(f)) + \frac{\alpha}{L} \sum_{s \in \mathcal{B}} d(P_s(f)||Q_s(f)) \tag{3}$$

where $\mathcal{B}$ represents the set of the L buffer frames; α is the trade-off between the decomposition of the current frame t and the decomposition of buffer frames. This optimization problem can be solved by the Expectation-Maximization (EM) algorithm, which results in the following update equations:

$$\text{E}: \quad P_s(z|f) = \tfrac{1}{C_1} P_s(z)P(f|z), \text{ for } s \in \mathcal{B} \cup \{t\} \tag{4}$$

$$\text{M}: \quad P(f|z) = \tfrac{1}{C_2} \textstyle\sum_{s \in \mathcal{B} \cup \{t\}} V_{fs} P_s(z|f), \text{ for } z \in \mathcal{S} \tag{5}$$

$$P_t(z) = \tfrac{1}{C_3} \textstyle\sum_f V_{ft} P_t(z|f), \text{ for } z \in \mathcal{S} \cup \mathcal{N} \tag{6}$$

where V_{ft} is the input spectrum of the current frame t; V_{fs} for $s \in \mathcal{B}$ are input spectra of buffer frames weighted by α/L; C_1, C_2 and C_3 are normalization factors for these distributions.

This algorithm can be thought as a block-wise processing version (with overlaps) of the offline PLCA in Eq. (2), but

with different emphasis on the current frame and the buffer frames. Furthermore, the activation weights of the buffer frames are fixed as what has been estimated when decomposing these buffer frames before, instead of being estimated in Eq. (2).

4. Tradeoff between noise reduction and speech distortion

Speech enhancement algorithms often face the tradeoff between noise reduction and speech distortion [1]. In the proposed algorithm, the tradeoff is determined by the speech dictionary, because the enhanced speech is obtained as a convex combination of its basis spectra. Ideally, the speech dictionary should contain basis spectra whose convex hull can only cover speech spectra but not noise spectra. In practice, however, speech and noise often share some similar spectra (e.g. speech fricatives and some keyboard noise). Therefore, if the speech dictionary is very restrictive to avoid covering noise-like spectra, the enhanced speech might be noise-free, but severely distorted. On the other hand, if the speech dictionary is very broad to cover all possible speech spectra, the enhanced speech might not be distorted, but noise reduction might not be adequate as well.

In the proposed algorithm, the speech dictionary is estimated using the EM algorithm when separating the noisy speech instead of being pre-learned from speech-only excerpts. This makes it easier to bias to one of the two cases. In this section, we explore controlling the speech dictionary through a prior during the EM iterations. The strength of the prior controls the tradeoff between noise reduction and speech distortion.

Each basis spectrum $P^{(t)}(f|z)$ of the speech dictionary at frame t is a discrete distribution. Since the Dirichlet distribution is the conjugate prior distribution of the discrete distribution, it can be used to impose priors on the dictionary. The Dirichlet distribution is defined by a set of positive and real hyperparameters, each of which corresponds to an element of the discrete distribution. We set these parameters as $\gamma P^{(t-1)}(f|z)$, i.e. the basis spectrum of the speech dictionary estimated at frame $t-1$, scaled by a positive number γ. Then the priors for all the basis spectra $\Lambda_f^{(t)}$ at time t are:

$$P(\Lambda_f^{(t)}) \propto \prod_z \prod_f P(f|z)^{\gamma P^{(t-1)}(f|z)} \tag{7}$$

The idea of this prior is intuitive. Since the speech signals do not change much over two adjacent time frames $t-1$ and t, the two underlying speech dictionaries are likely to be very similar. Through the hyperparameter, the estimated speech dictionary $P^{(t-1)}(f|z)$ at frame $t-1$ serves as an exemplar for the to-be-estimated speech dictionary $P^{(t)}(f|z)$ at frame t. In this way, the estimation of $P^{(t)}(f|z)$ is guided, which avoids it being trapped in local minima in early iterations. In addition, the information obtained at frame $t-1$ is passed to frame t, which speeds up the convergence of the iterations at frame t.

With this prior, the EM algorithm remains the same except that Eq. (5) needs to be modified as:

$$P(f|z) = (1-\beta)\frac{\sum_{s\in\mathcal{B}\cup\{t\}} V_{fs}P_s(z|f)}{C_2} + \beta P^{(t-1)}(f|z) \tag{8}$$

which is a convex combination of two terms through $\beta \in [0,1]$. The first term is a scaled version of the right hand side of Eq. (5), which is resulted from processing the t-th frame, and can be viewed as the likelihood part. The second term is a scaled version of the basis spectrum estimated in frame $t-1$, which is the prior part. The scaling parameter γ is absorbed into β, where $\gamma = 0$ implies $\beta = 0$ and $\gamma \to \infty$ implies $\beta \to 1$.

We linearly decrease β from 1 to 0 throughout the EM iterations. This is intuitive, as in the early iterations the estimates resulted from the likelihood part are noisy and a stronger prior to guide the optimization process is favored. This avoids the estimates being trapped to far-away local optima. Gradually, as the likelihood estimates are localized to some close local optima, the prior needs to be switched off to let the estimates be freely tuned to fit the current frame t. We denote the number of iterations that $\beta > 0$ by τ, which controls the speed at which the prior is decreased, hence controls the overall strength of the prior. When $\tau = 0$, no prior is imposed in any iteration and the speech dictionary is always randomly initialized. When $\tau = 1$, the prior only affects the first iteration, i.e. the speech dictionary in the current frame is initialized as the one estimated in the previous frame. When $\tau > 1$, the initialization is the same as when $\tau = 1$, and the prior will also affect following EM iterations in Eq. (8) until β vanishes. With the increase of the prior strength, the estimated speech dictionary will be more restrictive, leading to better noise reduction but more speech distortion.

5. Experiments

We evaluate the proposed algorithm using 300 noisy speech files, about 1.25 hours long in total. These files are obtained by adding clean speech files with noise-only files in five SNR conditions: -10dB, -5dB, 0dB, 5dB and 10dB. The clean speech files are the full speech corpus in the NOIZEUS dataset [1], which has thirty short English sentences spoken by three female and three male speakers. We concatenate all sentences from the same speaker into one long sentence, resulting in six long sentences, each of which is about fifteen seconds long. We collected ten types of noise, including *birds*, *casino*, *cicadas*, *computer keyboard*, *eating chips*, *frogs*, *jungle*, *machine guns*, *motorcycles* and *ocean*. Some are more stationary, e.g. cicadas and ocean noise, while others are very non-stationary, e.g. birds and computer keyboard noise. Each noise is at least one minute long. The first twenty seconds are used to train the noise dictionary beforehand. The rest is used to generate noisy speech files. The sampling rate of all the files is 16kHz.

In the proposed algorithm, we segment each noisy speech file into frames of 64ms with 48ms overlap. We set the speech dictionary size as 20, since we find it is enough to get a perceptually good reconstruction of the clean speech files. We choose the noise dictionary size from $\{1, 2, 5, 10, 20, 50, 100, 200\}$, according to the complexity of the noise. We set the buffer size L to 60 frames, and choose the buffer tradeoff factor α from $\{1, 2, \cdots, 20\}$ the one that achieves the best enhancement result in the condition of SNR of 0dB for each noise. The number of EM iterations in each frame is set to 20, which almost always assures convergence in our experiments. It takes about 25 seconds to denoise each file (about 15 seconds long) in a Matlab implementation in a computer with a 4-core 2.13GHz CPU.

We compare the proposed algorithm with four classical speech enhancement algorithms: spectral subtraction (*MB*) [2], Wiener filtering (*Wiener-as*) [3], statistical-model-based (*log-MMSE*) [4] and subspace algorithm (*KLT*) [5]. We use Loizou's implementations of these algorithms, as provided in [1]. To make comparisons fair, noise models of these algorithms are also learned from the twenty seconds noise training excerpts.

We use two kinds of evaluation metrics. The first kind is *PESQ* [13], which is a broadly used objective speech quality metric. It ranges from 0.5 to 4.5, with a larger value for

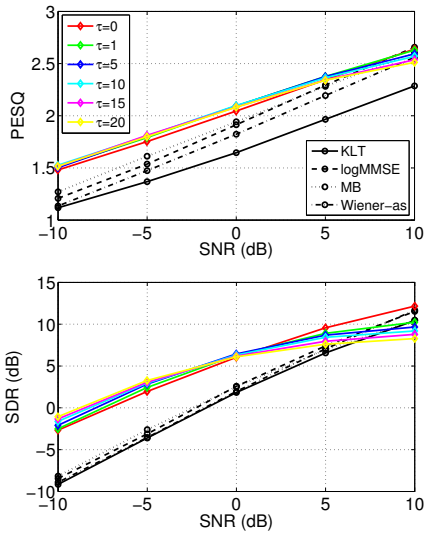

Figure 3: Comparison of the proposed algorithm as the prior ramp length τ varies, with four classical algorithms.

better quality. The second kind is *BSS-EVAL* [14], which is broadly used in source separation. It has three metrics. *Source-to-Interference Ratio (SIR)* reflects noise reduction, *Source-to-Artifacts Ratio (SAR)* reflects artifacts introduced by the separation process, and *Source-to-Distortion Ratio (SDR)* reflects the overall separation quality. For all three metrics, larger values mean better quality. Note that speech distortion in the speech enhancement literature is actually reflected by SAR, while SDR accounts for both noise reduction and speech distortion.

Figure 3 shows the average results of the proposed algorithm and classical algorithms for various SNR settings[1]. It can be seen that in terms of both PESQ and SDR, the proposed algorithm with different prior ramp length parameter τ achieves significantly better results than all the four classical algorithms in most SNR settings. The lower the SNR is, the larger the differences are. This suggests that the proposed algorithm is especially suitable to deal with non-stationary noise in low SNR settings. When SNR is 10dB, the performance of proposed algorithm varies in a large range with different τ parameters. With the increase of τ (hence increase of the prior strength), the performance gets worse, and finally worse than some of the classical algorithms. The reason is that when the mixture contains little noise, the speech dictionary estimated purely from the likelihood part in Eq. (8) is good enough, while a too strong prior prevents it achieving this good estimate.

Figure 4 shows the influence of the prior ramp length to the separation results in the 0dB SNR condition. It can be seen that the overall separation result SDR does not change significantly with τ, although the highest value 6.3dB is achieved when $\tau = 5$. However, a clear tradeoff between SIR and SAR is shown. With the increase of τ, SIR significantly increases from 9.3dB to 17.4dB, while SAR significantly decreases from 9.6dB to 6.6dB. A higher SIR indicates better noise reduction and a higher SAR indicates less speech distortion. It is useful to see that the tradeoff can be controlled by the single parameter τ. In practice, τ could be implemented as a knob that users can tune according to their own preferences in real time.

[1]Audio examples can be found at `http://www.cs.northwestern.edu/~zdu459/is2011/examples.html`.

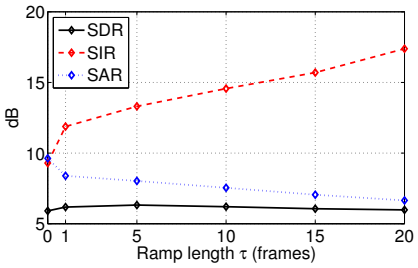

Figure 4: The tradeoff between noise reduction (SIR) and artifacts introduction (SAR) as the prior ramp length τ varies.

6. Conclusions

In this paper, we presented a novel speech enhancement algorithm based on non-negative spectrogram decomposition for non-stationary noise environments. Like typical denoising algorithms, the proposed approach also possesses the properties of "requiring a pre-training of only the noise model" and "working online". Experiments show the superiority of the proposed algorithm over four classical algorithms in non-stationary environments with different SNRs.

7. References

[1] Loizou, P.C. *Speech Enhancement: Theory and Practice*, Taylor and Francis, 2007.

[2] Kamath, S. and Loizou, P.C., "A multi-band spectral subtraction method for enhanceing speech corrupted by colored noise," in *Student Research Abstracts of Proc. ICASSP*, 2002.

[3] Scalart, P. and Filho, J., "Speech enhancement based on a priori signal to noise estimation," in *Proc. ICASSP*, pp. 629–632, 1996.

[4] Ephraim, Y. and Malah, D., "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, 33:443–445, 1985.

[5] Hu, Y. and Loizou, P.C., "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Process.*, pp. 334–341, 2003.

[6] Duan, Z., Mysore, G.J. and Smaragdis P., "Online PLCA for real-time semi-supervised source separation," in *Proc. LVA/ICA, LNCS*, 7191:34–41, 2012.

[7] Smaragdis, P., Raj, B. and Shashanka, M., "A probabilistic latent variable model for acoustic modeling," in *Workshop of Advances in Models for Acoustic Processing, NIPS*, 2006.

[8] Smaragdis, P., Raj, B. and Shashanka, M., "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. ICA 2007, LNCS*, 4666:414–421, 2007.

[9] Schmidt, M.N. and Larsen, J., "Reduction of non-stationary noise using a non-negative latent variable decomposition," in *Proc. MLSP*, 2008.

[10] Mairal, J., Bach, F., Ponce, J. and Sapiro, G., "Online learning for matrix factorization and sparse coding," *J. Machine Learning Research*, 11:19–60, 2010.

[11] Wang, F., Tan, C., König, A.C. and Li, P., "Efficient document clustering via online nonnegative matrix factorizations," in *Proc. SDM*, 2011.

[12] Lefèvre, A., Bach, F. and Févotte, C., "Online algorithms for non-negative matrix factorization with the Itakura-Saito divergence," in *Proc. WASPAA*, 2011.

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

[14] Vincent, E., Fevotte, C. and Gribonval, R., "Performance measurement in blind audio source separation," *IEEE Trans. on Audio Speech Lang. Process.*, 14(4):1462–1469, 2006.