interspeech 2008 incorporating SST 08

Regularized Non-negative Matrix Factorization with Temporal Dependencies for Speech Denoising

Kevin W. Wilson[1], *Bhiksha Raj*[1], *Paris Smaragdis*[2]

[1]Mitsubishi Electric Research Lab, Cambridge, MA
[2]Adobe Systems, Newton, MA

Abstract

We present a tecchnique for denoising speech using temporally regularized nonnegative matrix factorization (NMF). In previous work [1], we used a regularized NMF update to impose structure within each audio frame. In this paper, we add frame-to-frame regularization across time and show that this additional regularization can also improve our speech denoising results. We evaluate our algorithm on a range of nonstationary noise types and outperform a state-of-the-art Wiener filter implementation.

Index Terms: speech enhancement, source separation, speech modeling, speech processing

1. Introduction

This paper describes the use of a temporally regularized NMF update for denoising speech in nonstationary noise. Speech denoising in nonstationary noise is an important problem with increasingly broad applications as cellular phones and other telecommunications devices make electronic voice communication more common in a wide range of challenging environments, from urban sidewalk to construction site to factory floor. Standard approaches such as spectral subtraction and Wiener filtering require signal and/or noise estimates and therefore are typically restricted to stationary or quasi-stationary noise in practice.

Nonnegative matrix factorization, popularized by Lee and Seung [2], finds a locally optimal choice of W and H to solve the matrix equation $V \approx WH$ for nonnegative V, W, and H. This provides a way of decomposing a signal into a convex combination of nonnegative building blocks. When the signal, V, is a spectrogram and the building blocks, W, are a set of specific spectral shapes, Smaragdis [3] showed how NMF can be used to separate single-channel mixtures of sounds by associating different sets of building blocks with different sound sources. In Smaragdis's formulation, H becomes the time-varying activation levels of the building blocks. The building blocks in W constitute a model of each source, and because H allows activations to vary over time, this decomposition can easily model nonstationary noises. ([3] refers to its algorithm as probabilistic latent semantic analysis (PLSA). Under proper normalization and for the KL objective function used in this paper, NMF and PLSA are numerically equivalent [4], so the results in [3] are equally relevant to NMF or PLSA.)

NMF works well for separating sounds when the building blocks for different sources are sufficiently distinct. For example, if one source, such as a flute, generates only harmonic sounds and another source, such as a snare drum, generates only nonharmonic sounds, the building blocks for one source will be of little use in describing the other. In many cases of practical interest, however, there is much less separation between sets of building blocks. In particular, human speech consists of harmonic sounds (possibly at different fundamental frequencies at different times) and nonharmonic sounds, and it can have energy across a wide range of frequencies. For these reasons, many interfering noises can be represented, at least partially, by the speech building blocks. In a speech denoising application, where one "source" is the desired speech and the other "source" is interfering noise, this overlap between speech and noise models will degrade performance.

There is additional structure in speech and many other sounds, however. For example, a human speaker will never generate a simultaneous combination of two harmonic sounds with harmonically unrelated pitches. Our previous work, [1], exploited this type of signal structure by imposing a signal-specific covariance structure on the activation coefficients within each frame.

Another type of structure present in audio signals is frame-to-frame temporal structure. For example, two men with similar vocal tracts may, in any given audio frame, be producing sounds with very similar spectra, but if one is a fast-talking car salesman and the other is a laid-back surfer, we can use the differences in their speech rates to distinguish them. Many types of noise, such as a jackhammer at a construction site or loud music at a bar, have distinctive temporal structure that can be exploited to distinguish them from speech. In this paper, we exploit this temporal structure to improve the speech denoising performance of our algorithm.

This paper makes two contributions. First, we present an NMF update that is regularized both across audio frames and across activation coefficients within a frame. These regularization terms encourage the denoised output signal to have statistics similar to the known statistics of our source model across both time and frequency. Second, we evaluate the speech denoising performance of NMF with and without different forms of regularization and compare it to a state-of-the-art Wiener filter implementation.

2. Algorithm

Our technique for speech denoising consists of a training stage and an application (denoising) stage. During training, we assume availability of a clean speech spectrogram, V_s, of size $n_f \times n_{st}$, and a clean (speech-free) noise spectrogram, V_n, of size $n_f \times n_{nt}$, where n_f is the number of frequency bins, n_{st} is the number of speech frames, and n_{nt} is the number of noise frames. Different objective functions lead to different variants of NMF, a number of which are described in [5]. Kullback-Leibler (KL) divergence between V and WH, denoted $D(V||WH)$, was found to work well for audio source

separation in [3], so we will restrict ourselves to KL divergence in this paper. Generalization to other objective functions using the techniques described in [5] is straightforward.

During training, we separately perform standard NMF on the speech and the noise, minimizing $D(V_s||W_sH_s)$ and $D(V_n||W_nH_n)$, respectively. W_s and W_n are each of size $n_f \times n_b$, where n_b is the number of basis vectors chosen to represent each source. Each column of W is therefore one of the spectral "building blocks" we referred to earlier. H_s and H_n are of size $n_b \times n_{st}$ and $n_b \times n_{nt}$, respectively, and represent the time-varying activation levels of the basis vectors.

Also as part of the training phase, we estimate the statistics of H_s and H_n. In [1], we computed the empirical means and covariances of their log values assuming independence between frames, yielding μ_s, μ_n, Λ_{B_s}, and Λ_{B_n} where each μ is a length n_b vector and each Λ_B is an $n_b \times n_b$ covariance matrix. (The subscripted B indicates that this covariance is across *basis* functions within a frame.) We choose this implicitly Gaussian representation for computational convenience, and we choose to operate in the logarithmic domain because preliminary experiments showed better results for the log domain than the linear domain.

In this paper, we additionally compute empirical covariances across frames, assuming independence across basis functions. This corresponds to assuming that the log activation coefficient for each basis function evolves over time as a stationary Gaussian process. (Because a stationary Gaussian process can be equally well characterized by either its autocovariance or its power spectrum, our subsequent regularization can be thought of as regularizing the modulation spectrum of each log activation coefficient.) This results in separate covariance matrices for each basis function, ${}_k\Lambda_{T_s}$ and ${}_k\Lambda_{T_n}$, $k \in [1..n_b]$. (The subscripted T indicates that these inter-frame covariances are across *time*.) The size of these covariance matrices is $n_t \times n_t$, where n_t is the number of frames in the spectogram being analyzed. The details of how we construct and employ these covariance matrices are described below.

In the denoising stage, we fix W_s and W_n and assume that they will continue to be good basis functions for describing speech and noise. We concatenate the two sets of basis vectors to form W of size $n_f \times 2n_b$. This combined set of basis functions can then be used to represent a signal containing a mixture of speech and noise. Assuming the speech and noise are independent, we also concatenate to form $\mu = [\mu_s; \mu_n]$ and $\Lambda_B = [\Lambda_{B_s}\ 0; 0\ \Lambda_{B_n}]$. We now denote the combined set of interframe covariance matrices as ${}_k\Lambda_T$, $k \in [1..2n_b]$, where $k \in [1..n_b]$ correspond to speech basis functions and $k \in [n_b + 1..2n_b]$ correspond to noise basis functions. We then find an H to minimize the regularized objective function

$$D_{reg}(V||WH) = \sum_{ik}(V_{ik}\log\frac{V_{ik}}{(WH)_{ik}} + V_{ik} - (WH)_{ik}) - \alpha L_B(H) - \beta L_T(H) \quad (1)$$

$$L_B(H) = -\frac{1}{2}\sum_k\{(\log H_{:,k} - \mu)^\mathsf{T}\Lambda_B^{-1}(\log H_{:,k} - \mu) - \log[(2\pi)^{2n_b}|\Lambda|]\} \quad (2)$$

$$L_T(H) = -\frac{1}{2}\sum_k\{(\log H_{k,:} - \mu_k\ 1^\mathsf{T}) {}_k\Lambda_T^{-1}(\log H_{k,:} - \mu_k\ 1^\mathsf{T})^\mathsf{T} - \log[(2\pi)^{2n_b}|\Lambda|]\} \quad (3)$$

where 1^T is a row vector of ones of length n_t and where a colon subscript indicates the use of an entire row or column of H.

When α and β are zero, this is equal to the standard KL divergence objective function [5]. For nonzero α and β, there is an added penalty proportional to the negative log likelihood under our jointly Gaussian models for $\log H$. This term encourages the resulting H to be consistent with the statistics of H_s and H_n as empirically determined during training. Varying α and β allows us to control the trade-off between fitting the observed spectrogram of mixed speech and noise, V, and achieving high likelihood under our prior model, with α controlling within-frame regularization and β controlling inter-frame regulariztion. Following [5], the multiplicative update rule for H is

$$H_{ab} \leftarrow H_{ab}\frac{\sum_i W_{ia}V_{ib}/(WH)_{ib}}{[\sum_k W_{ka} + \alpha\varphi_B(H) + \beta\varphi_T(H)]_\varepsilon} \quad (4)$$

$$\begin{aligned}\varphi_B(H_{ab}) &= -\frac{\partial L_B(H)}{\partial H_{ab}} \\ &= -\frac{(\Lambda_B^{-1}(\log H_{:,b} - \mu))_a}{H_{ab}} \\ \varphi_T(H_{ab}) &= -\frac{\partial L_T(H)}{\partial H_{ab}} \\ &= -\frac{({}_a\Lambda_T^{-1}(\log H_{a,:} - \mu_a\ 1^\mathsf{T})^\mathsf{T})_b}{H_{ab}}\end{aligned}$$

where $[\,]_\varepsilon$ indicates that any values within the brackets less than the small positive constant ε should be replaced with ε to prevent violations of the nonnegativity constraint and avoid divisions by zero.

Finally, to reconstruct the denoised spectrogram, we compute $\hat{V}_{speech} = W_{speech}H_{1:n_b}$, using the speech basis functions and the top n_b rows of H to approximate the target speech.

In [1], we demonstrated the usefulness of within-frame regularization. In this paper, we focus on the usefulness of inter-frame regularization. Figure 1 gives a simple toy example of separating with and without inter-frame regularization. Here we set $n_f = n_b = 2$, and we assume that for both speech and noise, one basis function represents the high frequency and the other represents the low frequency. The original signals are in the left column, the unregularized NMF reconstructions are in the center column, and the regularized NMF reconstructions are in the right column. Source 1 is like a laid-back "surfer dude" with slowly varying spectra, and source 2 is like a fast-talking car salesman. Because the basis functions for the speech and noise are the same, unregularized NMF is completely unable to reconstruct the individual sources. Note however that its chosen reconstructions do sum to accurately model the mixture signal, indicating that it successfully minimized $D(V||WH)$. (Although not shown, within-frame regularization also fails completely because the within-frame statistics of the two sources are identical.) The temporally regularized NMF is able to exploit the differences in modulation rates to accurately reconstruct the two signals given only the mixture signal and their statistical models. This example is extreme in that the two sources' bases

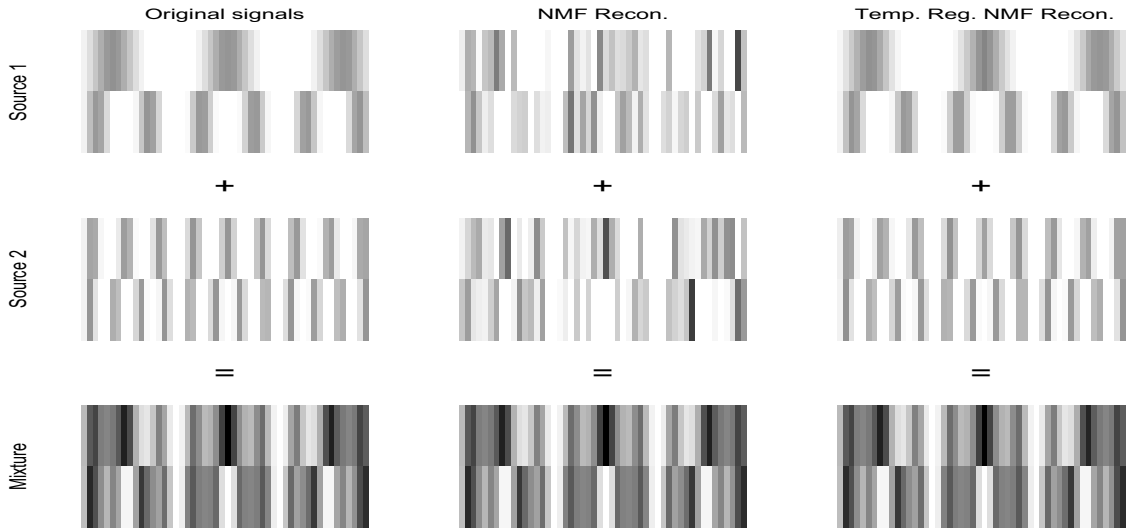


Figure 1: A toy example showing the advantage of regularizing across frames. Each panel is a spectrogram, where the horizontal axis represents time and the vertical axis represents frequency. Darker colors represent higher intensity. The leftmost column shows the original signals. For source 1, high- and low-frequency energies are modulated slowly and independently. For source 2, high- and low- frequency energies are modulated rapidly and independently. In the middle column, unregularized NMF finds a reconstruction that perfectly models the mixture signal, but each individual source is poorly reconstructed. In the rightmost column, near-perfect reconstruction of individual sources is achieved by regularization.

and within-frame statistics are identical while their inter-frame statistics are quite different, but it makes the potential of temporal regularization clear. We show in the following section that incorporating this regularizing prior term does improve speech denoising in practice.

3. Results

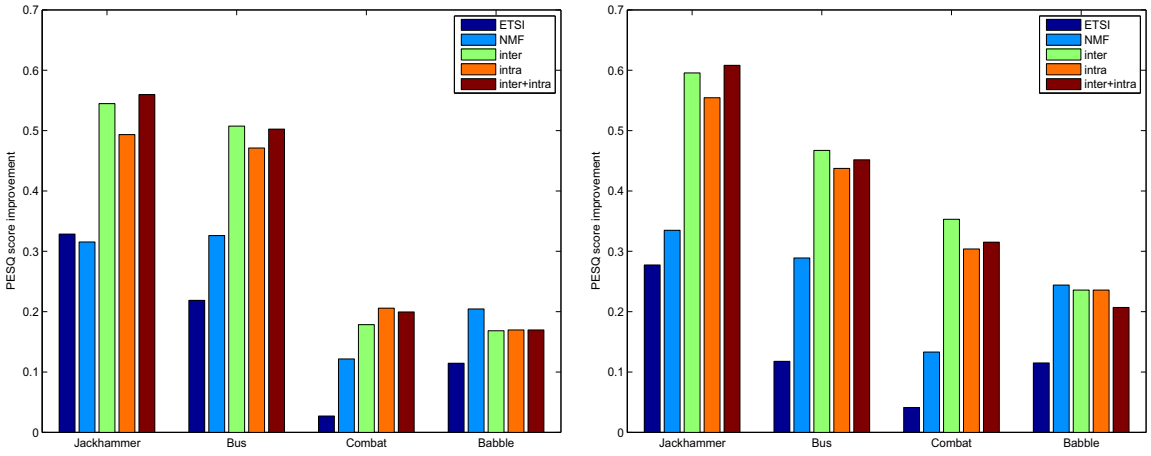
We tested NMF and regularized NMF on a variety of speakers and with four different types of nonstationary background noise (jackhammer noise, bus/street noise, combat noise, and speech babble noise). All parameters remained at fixed values across all experiments. We used 16 kilohertz audio with $n_f = 513$ and $n_b = 80$. When within-frame regularization was used $\alpha = 0.25$. When inter-frame regularization was used $\beta = 0.05$. (The numerical values of α and β are meaningless without knowing the magnitude of the spectrogram values, but we want to emphasize that they remained fixed throughout.) To estimate Λ_B, we simply estimate the empirical covariance of the activation coefficients in our clean training data, assuming that each frame is an independent observation. To estimate the ${}_k\Lambda_T$, we first compute the empirical autocovariance of our clean training data out to a fixed maximum lag d_{max} (12 frames in our experiments). We then form ${}_k\Lambda_T$, a symmetric Toeplitz band matrix with bandwidth $2d_{max} + 1$, by putting the autocovariance at lag zero on the main diagonal and the autocovariance for other lags on the corresponding minor diagonals. Because there are n_b separate ${}_k\Lambda_T$, and because of its special structure, it is much more efficient to exploit this structure in solving for the elements of $\varphi_T(H)$ than to invert ${}_k\Lambda_T$ explicitly. We simply use Matlab's sparse solver, but additional optimizations may be possible. Matlab implementations of NMF and regularized NMF run in near real-time on a 3GHz PC.

We used speech from the TIMIT database [7], testing two sentences from each of ten speakers in each of our four chosen types of background noise. We normalized speech and noise so that the average signal-to-noise ratio (SNR) for each mixture was 0 dB. We trained a separate noise model for each of the four noise types, and we trained a single speaker-independent speech model on a group of several speakers from outside our test set. This single model was then used to denoise noisy signals from all test speakers.

Our results are shown in Figure 2. All results are shown as improvement relative to the score of the unprocessed 0 dB SNR mixture, and each bar represents an average value over ten speakers. To quantify our results, we use the ITU Perceptual Evaluation of Speech Quality (PESQ) [8], a metric designed to match mean opinion scores of perceptual quality. PESQ scores range from 1 through 5, and PESQ improvements on the order of 0.5, which we achieve in many cases, are quite noticeable.

In addition to NMF and regularized NMF, we processed each example with the ETSI Aurora front end's Wiener filtering [6], a European telecommunications standard which has been carefully tuned for good performance in denoising speech. It is important to note that, in contrast to the ETSI Wiener filter, all of our NMF variants use both a training and a testing stage, so they benefit from environment-specific noise models. The ETSI Wiener filter has no training stage, so its noise model must be estimated online using a voice activity detector and assumptions about the stationarity of the noise. However, the ETSI Wiener filter has an advantage as long as its voice activity detector works properly because it can then completely silence intervals with no speech activity, yielding very good denoising in those intervals. Because of the major differences between the two types of denoising, detailed comparisons of the results are of limited use, but we feel that it is important to compare to an established baseline and that some general conclusions are possible. The PESQ scores for both regularized and unregularized NMF are almost always greater than for the ETSI Wiener filter, and in many cases are substantially greater.

Overall, there is little difference between the results for



(a) Male

(b) Female

Figure 2: Speech denoising performance for our chosen noise types. "ETSI" is the front end Wiener filtering described in [6]. "NMF" is applying the iterative update in Equation 4 with no regularization, $\alpha = \beta = 0$. "Inter" is applying Equation 4 with inter-frame regularization only, $\alpha = 0, \beta = 0.05$. "Intra" is applying Equation 4 with within-frame regularization only, $\alpha = 0.25, \beta = 0$. "Intra+inter" is applying Equation 4 with both types of regularization, $\alpha = 0.25, \beta = 0.05$.

male speakers and female speakers. Note that regularization in any form almost always substantially improves on the unregularized NMF results. This shows that the additional structure imposed by regularization (within-frame, across-frame, or both) consistently improves the denoising performance across a variety of background noises. Inter-frame and intra-frame regularization results are comparable, with inter-frame performing better on jackhammer and bus noise. Jackhammer noise has obvious temporal structure, and the bus noise includes a lot of slowly idling engine noise, which also has pronounced temporal structure. We had hoped to see an additional jump in performance by combining the two regularization terms, but with the exception of a small improvement for jackhammer noise, this did not happen. We speculate that the values of α and β that were chosen for each individual type of regularization may not be the best choices for the combined regularization. In the future, we plan to more carefully explore the joint parameter space.

The aforementioned trends are relatively consistent across three of the four noise types, but performance on "babble" noise departs from these trends. For "babble," it appears that regularization is not as helpful. We believe that regularization is not as useful for babble because the distribution of the speech-like babble noise is very similar to the distribution for speech itself, both within a frame and between frames.

4. Conclusion

We have shown that NMF can be used to denoise speech in the presence of nonstationary noise, and we have shown that by regularizing NMF based on a simple statistical model of speech and noise, we can exploit additional signal structure to improve performance. In particular, we showed that the use of inter-frame regularization can improve speech denoising, especially when the interfering noise has a pronounced temporal structure. Our results equal or surpass results from a state-of-the-art Wiener filter implementation on a range of noise types.

In the future, we would like to more carefully explore the combination of intra- and inter-frame regularization, and we plan to explore multi-scale temporal regularization to capture even longer-term patterns.

5. References

[1] K. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008.

[2] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *NIPS*, 2000, pp. 556–562. [Online]. Available: citeseer.ist.psu.edu/lee01algorithms.html

[3] P. Smaragdis, "From learning music to learning to separate," in *Forum Acusticum*, 2005.

[4] E. Gaussier and C. Goutte, "Relation between plsa and nmf and implications," in *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2005.

[5] A. Cichocki, R. Zdunek, and S. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, 2006, pp. 621–625.

[6] "Speech processing, transmission and quality aspects (stq); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms," Tech. Rep. ETSI ES 202 050 V1.1.3, 2003.

[7] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom documentation," National Technical Information Service, Tech. Rep. PB93-173938, 1993.

[8] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Tech. Rep. ITU-T P.862, 2001.