Multichannel Speech Dereverberation Based on Convolutional Nonnegative Tensor Factorization for ASR Applications

Seyedmahdad Mirsamadi and John H. L. Hansen*

Center for Robust Speech Systems (CRSS)
The University of Texas at Dallas, Richardson, TX 75080-3021, U.S.A.
{mirsamadi, john.hansen}@utdallas.edu

Abstract

Room reverberation is a primary cause of failure in distant speech recognition (DSR) systems. In this study, we present a multichannel spectrum enhancement method for reverberant speech recognition, which is an extension of a single-channel dereverberation algorithm based on convolutional nonnegative matrix factorization (NMF). The generalization to a multichannel scenario is shown to be a special case of convolutional nonnegative tensor factorization (NTF). The presented algorithm integrates information from across different channels in the magnitude short-time Fourier transform (STFT) domain. By doing so, it eliminates any limitations on the array geometry or a need for information concerning the source location, making the algorithm particularly suitable for distributed microphone arrays. Experiments are performed on speech data using actual room impulse responses from AIR database. Relative WER improvements using a clean-trained ASR system vary from +7.1% to +30.1% based on the number of channels and the source to microphone distances (1 to 3 meters).

Index Terms: reverberation, automatic speech recognition, nonnegative matrix and tensor factorization

1. Introduction

Automatic Speech Recognition (ASR) systems have now reached a performance level which enables them to be used in many real world applications. However, this is generally limited to the case where the speech signal is captured by a close-talking microphone. In the case of distant speech recognition (DSR), where the microphone is far from the speaker’s mouth, the recognition accuracy degrades drastically as a result of room reverberation and environmental noise. Reverberation, in particular, is a major challenge that must be addressed in DSR, occurring as a result of multiple sound reflections being captured by the distant-talking microphone. Although these reflections can be viewed as corruptive noise terms added to the actual desired speech signal, the reverberation problem is fundamentally different from conventional noise-robustness techniques, because the reflections serve as both nonstationary and colored noise which is also correlated with the desired speech signal.

The reverberation problem can potentially be addressed in three different stages of the ASR system front-end, namely (i) the waveform domain, (ii) magnitude Short Time Fourier Transform (STFT) domain, and (iii) feature (cepstrum) domain [1]. A multichannel solution (i.e., the use of a microphone array) can theoretically be exploited in any of the these stages by integrating information from different microphones into a single set of features. However, microphone array solutions have traditionally been used mostly in the waveform domain, in the form of either fixed or adaptive beamformers. The directional sound capture provided by a beamformer can provide a certain degree of signal enhancement which in turn improves the extracted features. However, this approach is not optimal for the task of reverberant speech recognition, mainly because in ASR we are interested in obtaining a set of features that closely resemble the clean training set, and not in recovering the actual speech waveform [2]. Other factors such as the requirement for the speaker’s location (which is difficult to estimate in a reverberant environment), and also certain limitations on the array geometry (for example to avoid spatial aliasing in high frequencies), further limit the applicability of beamforming approaches in realistic DSR scenarios.

In this paper, we propose a multichannel speech dereverberation method based on Nonnegative Tensor Factorization (NTF) which operates in the magnitude STFT domain. A multichannel approach which fuses information available from different sensors in the magnitude time-frequency domain is particularly useful in the context of distributed microphone array systems, because it does not depend on phase information which is difficult to preserve across independent channels of a distributed system [3, 4]. Furthermore, no assumptions are necessary about the specific location of the source or the individual sensors.

Nonnegative tensor factorization (NTF) [5] is a generalization of Nonnegative Matrix Factorization (NMF) to tensors (multi-way arrays). NMF is a multivariate data analysis technique popularized by the simple algorithms of [6] (referred to as Lee-Seung algorithms) which have been extended to the convolutive case in [7]. A special formulation of convolutional NMF was used for single-channel speech dereverberation in [8], shown to provide significant improvements in ASR accuracy in [9], and further extended to the Gammatone subband domain in [10]. In our study, we extend the method of [8] to a multichannel framework. We will show that while single-channel dereverberation is a special case of convolutive nonnegative matrix factorization, multichannel dereverberation can be considered as a special case of convolutive nonnegative tensor factorization with a third order tensor. The resulting algorithm is similar in nature to the multichannel version of the latent-variable decomposition based approach developed in [11], but without any statistical priors assumed for the input. The proposed multichannel extension is shown to provide recognition improvements in highly reverberant conditions compared with the original single-channel approach.

The remainder of this paper is organized as follows. In
Sec. 2 we describe a model for room reverberation in the short

time Fourier transform domain as the basis for spectrum en-
hancement approaches to dereverberation. In Sec. 3 we present

the NTF-based multichannel dereverberation algorithm, and ex-
plain the connections with a standard NTF problem. We pro-
vide experimental results in Sec. 4 which illustrate consistent

improvements in recognition accuracy. We discuss some inter-
esting properties of the proposed algorithm in Sec. 5, and finally

conclude the paper in Sec. 6.

2. Reverberation Model

A reverberant speech signal is modelled in the time domain by a

convolution between the clean speech signal and the room

impulse response (RIR) from the source location to the micro-

phone. The length of the RIR is often much longer than the typi-

ical time segments used in speech recognition systems, and this

causes the spectral content of each frame to influence the sub-

sequent frames (referred to as the “spectral smearing” effect).

This effect can be modelled in the magnitude STFT domain by a

convolution of the form,

\[ X^{(i)}(m, k) = \sum_{p=0}^{L-1} H^{(i)}_{k}(p)S(m - p, k), \]

where \( m, k \), and \( i \in \{1, \cdots, N\} \) are frame, frequency and

channel indices, \( N \) is the number of microphones, \( S(m, k) \) and \( X^{(i)}(m, k) \) are the magnitude STFTs of the clean speech and

\( i \)’th microphone signal, and \( \hat{H}^{(i)}_{k}(m) \) is the subband envelope

of the RIR from the source location to the \( i \)’th microphone.

The reverberant spectrogram model of Eq. (1) has been the basis for

many recent studies on speech dereverberation [12, 13].

3. NTF-based dereverberation algorithm

3.1. Problem Formulation

Based on the model in Eq. (1), the multichannel dereverberation

problem can be stated as finding the nonnegative factors

\( \hat{H}^{(i)}(m) \) for all channels, together with the single common non-

negative factor \( S(m, k) \), which jointly minimize an error crite-
nion between the reverberant spectrograms \( X^{(i)}(m, k) \) and their

estimates given by \( \hat{X}^{(i)}(m, k) = \hat{H}^{(i)}(m) \ast S(m, k) \) (here, \( \ast \)

denotes convolution on frame index \( m \)). We therefore define the

following cost function using the Euclidian distance between the two

spectra as the error criterion:

\[ E = \sum_{m,k} \|X(m, k) - \hat{Z}(m, k)\|_{p}^{2}, \]

where

\[ \hat{Z}(m, k) = \sum_{p=0}^{L-1} H_{k}(p)\hat{S}(m - p, k), \]

\[ \hat{H}_{k}(m) = [\hat{H}^{(1)}_{k}(m), \cdots, \hat{H}^{(N)}_{k}(m)]^{T}, \]

\[ X(m, k) = [X^{(1)}(m, k), \cdots, X^{(N)}(m, k)]^{T}. \]

Note that the cost function of Eq. (2) is the sum of all the error

terms associated with each individual channel. The minimiza-
tion of Eq. (2) subject to the nonnegativity constraints,

\[ H^{(i)}_{k}(p) > 0 \text{ and } S(m, k) > 0, \text{ for all } i, p, m, k, \]

is expected to yield an estimate of the clean speech spectrogram

as well as the subband envelopes of the RIRs. To address the

scaling indeterminacy inherent in the problem, we impose the

following additional constraint:

\[ \sum_{i=1}^{N} \sum_{p=0}^{L-1} H^{(i)}_{k}(p) = 1, \quad k = 1, \cdots, K. \]

It is important to use the above normalization strategy instead of

individual normalization of each filter \( H^{(i)}_{k}(m) \) (as is done in

[8] and [10]), because this will allow the algorithm to adjust the

gain of each filter according to the SNR of the corresponding

microphone signal (see Sec. 5 for more details).

3.2. Relations with the standard NTF

As illustrated in Fig. 1, the set of magnitude spectrograms for

different channels can be considered as the frontal slices of a

third order tensor \( X \). In this view, the dereverberation problem

described above is equivalent to the following convolutive NTF

problem:

\[ X^{(i)} = \sum_{p=0}^{L-1} H^{(i)}(p) \rightarrow \hat{X}^{(i)}(m, k), \quad i = 1, \cdots, N, \]

where \( X^{(i)} \) is the magnitude spectrogram matrix of the \( i \)’th

channel which forms the \( i \)’th frontal slice of the tensor \( X \). The

base matrices \( H^{(i)}(p) \) are all diagonal matrices of the form,

\[ H^{(i)}(p) = \begin{bmatrix} H^{(1)}_{1}(p) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & H^{(N)}_{L}(p) \end{bmatrix}, \]

which constitute the \( i \)’th frontal slice of the tensor \( H(p) \), as

shown in Fig. 1. The matrix \( S \) in Eq. (8) is the magnitude spec-

trogram matrix of the clean speech signal, and the operator \( \rightarrow \)

shifts the rows of its argument matrix by \( p \) positions to the right,

filling in zeros from the left.

Note that in the single-channel case (i.e., \( N = 1 \)) the NTF

problem in Eq. (8) simplifies to a NMF problem, similar to [8].

Because of the diagonal form of the base matrices in Eq. (8),

the optimization of the components can be carried out indepen-
dently in each subband, as detailed in the next section.
3.3. Multiplicative update rules

A standard gradient descent minimization of Eq. (2) does not necessarily preserve the nonnegativity of the components. However, it has been shown in [6] that by using a variable step size in each iteration, simple multiplicative update formulas can be derived which ensure the nonnegativity of the results. The gradient of Eq. (2) with respect to filter coefficients is,

$$\frac{\partial E}{\partial H_k(p)} = 2 \sum_m [Z(m, k) - X(m, k)] S(m - p, k).$$  

(10)

We choose the following vector of step sizes for the gradient descent optimization of filter coefficients:

$$\eta_H = H_k(p) \odot 2 \sum_m Z(m, k) S(m - p, k),$$  

(11)

where $\odot$ represents element-wise division. Each element of $\eta_H$ is the step size used for the corresponding channel. Using Eq. (10) and (11), the update rule for $H_k(p)$ is,

$$H_k(p) \leftarrow H_k(p) \odot \sum_m X(m, k) \hat{S}(m - p, k)$$  

\[\odot \sum_m Z(m, k) \hat{S}(m - p, k),\]

(12)

where $\odot$ represents element-wise multiplication. Similarly, using the derivative with respect to $S(l, k)$,

$$\frac{\partial E}{\partial S(l, k)} = 2 \sum_m [Z(m, k) - X(m, k)]^T H_k(m - l),$$  

(13)

and choosing the following step size parameter

$$\eta_S = \frac{\hat{S}(l, k)}{2 \sum_m Z^T(m, k) H_k(m - l)},$$  

(14)

we obtain the multiplicative update formula for clean STFT estimates:

$$\hat{S}(l, k) \leftarrow \frac{\sum_m X^T(m, k) H_k(m - l)}{\sum_m Z^T(m, k) H_k(m - l)}.$$  

(15)

At the end of each iteration, all filter coefficients are normalized according to Eq. (7). Note that similar to the single-channel case in [8], both final update rules of Eq. (12) and (15) contain cross-correlation terms in their numerators and denominators. These correlations can be computed via FFT multiplication in the modulation frequency domain in order to reduce the computational complexity of the algorithm.

4. Experiments

We conduct speech recognition experiments to assess the performance of the proposed multichannel dereverberation algorithm for ASR applications. The CMU Sphinx3 system was used for the recognition experiments. We used 13-dimensional Mel-frequency cepstral coefficients (MFCCs) along with their delta and double-delta extensions as speech features. Acoustic models (3-state HMMs with 8 Gaussians per state) were trained on clean utterances from the TIMIT database. Cepstral mean normalization (CMN) and a trigram language model have been used in all experiments.

5. Discussion

The NTF-based multichannel dereverberation algorithm presented in this study has properties that make it attractive for distributed array systems, which are emerging as effective solutions for distance-based speech recognition in applications such as smart home or office environments. In a distributed array,
microphones are generally in unknown random locations in the room, and different channels have different gains and signal-to-noise ratios (SNRs). Another challenge is a possible lack of synchrony among different channels, as a result of independent processing units of the recording devices [4].

As mentioned in Sec. 1, the NTF-based dereverberation method is independent of signal phases and therefore circumvents the synchronization problem. Another interesting property observed for the proposed algorithm is that the update rules automatically adjust the filter taps of each channel (i.e. $H_{(i)}^{(m)}(m)$) according to the corresponding SNR. To illustrate this, we performed an experiment for a 4-channel scenario in which three of the microphones are at a distance of 1 m to the source (microphones 1, 2 and 3 in Fig. 3), but one of the microphones is located 3 m away from the source (microphone 11 in Fig. 3), thus having a much lower signal to reverberation ratio (SRR). Fig. 6 shows the normalized subband filters after 10 iterations of the algorithm (for an example frequency bin $k = 10$). It is observed that the weights corresponding to the low-SRR channel have been set to small values in the adaptation process compared to the other channels, minimizing the effect of this low-SRR channel on the final clean STFT estimates.

6. Conclusions

In this study, we presented a multichannel dereverberation algorithm for ASR applications. The proposed algorithm generalizes a single-channel NMF-based method to a general and not necessarily uniformly spaced multichannel framework by using nonnegative tensor factorization. The proposed algorithm was experimentally shown to provide relative WER improvements of up to +30% in highly reverberant conditions. This gain was achieved by the generalized reformulation of the single-channel NMF approach to an improved multichannel solution. The algorithm was also shown to be robust against varying signal qualities among the different channels (occurring due to different spatial locations of the microphones).
7. References


