A Generalized Nonnegative Tensor Factorization Approach for Distant Speech Recognition with Distributed Microphones

Seyedmahdad Mirsamadi, Member, IEEE, John H.L. Hansen, Fellow, IEEE

Abstract—Automatic Speech Recognition using distant (far-field) microphones is a challenging task, in which room reverberation is one of the primary causes of performance degradation. This study proposes a multichannel spectral enhancement method for reverberation-robust ASR using distributed microphones. The proposed method uses the techniques of Nonnegative Tensor Factorization (NTF) in order to identify the clean speech component from a set of observed reverberant spectrograms from the different channels. The general family of alpha-beta divergences are used for the tensor decomposition task which provides increased flexibility for the algorithm and is shown to provide improvements in highly reverberant scenarios. Unlike many conventional array processing solutions, the proposed method does not require closely-spaced microphones and is independent of source and microphone locations. The algorithm can automatically adapt to unbalanced direct-to-reverberation ratios among the different channels, which is useful in blind scenarios in which no information is available about source-to-microphone distances. For a medium vocabulary distant ASR task based on TIMIT utterances, and using clean-trained deep neural network acoustic models, absolute WER improvements of +17.2%, +20.7% and +23.2% are achieved in single-channel, two-channel, and four-channel scenarios.

Index Terms—Distant speech recognition, Distributed far-field microphones, Reverberation, Nonnegative tensor factorization.

I. INTRODUCTION

RECENT advancements in Automatic Speech Recognition (ASR), both in areas of feature processing and acoustic modelling, have brought about major performance improvements, making voice-enabled technologies popular in many applications. However, to achieve satisfactory performance, it is still necessary to use close-talking microphones (e.g., headset or lapel microphones). In situations where a distant-talking (far-field) microphone is to be used, distortions in the received signal lead to considerable performance degradation. This is a major limitation because in many applications, wearing a headset microphone or speaking directly into a microphone that is placed at a fixed position is impossible or too limiting. It is desirable to let users talk to machines while freely moving in the environment, without having to wear cumbersome recording or transmitting devices.

There are various factors leading to degraded performance in distant speech recognition. These include room reverberation, additive environmental noise, speaker movements and head orientation, and possibly changes in speech production due to perceived distance by the human speaker. Reverberation, in particular, which is caused by the reflections of the sound wave off walls or other surfaces in the room, is a major challenge in distant speech recognition. The multiple signal replicas coming from different directions with different delays act as non-stationary and colored sources of interference which are correlated with the desired speech signal [1]. This renders most conventional noise-robust approaches inadequate for reverberant speech recognition, since uncorrelated noise is often an assumption in those approaches [2].

The effect of room reverberation on the speech signal is modelled by the Room Impulse Response (RIR) from the speaker’s location to the microphone. The effects from the speaker’s head orientation and movements can also be incorporated into this impulse response by allowing it to be time-variant. The goal in distant speech recognition is thus to suppress the effects of this (possibly time-varying) linear filtering from the speech signal or its spectrum.

The various existing approaches in the literature for achieving reverberation robustness in ASR can be broadly categorized into three distinct groups: Front-end enhancement approaches aim at recovering the clean speech features from the reverberant observations, which are then submitted to a clean-trained ASR back-end. This enhancement can be carried out either in the time or short-time Fourier transform (STFT) domain (known as linear filtering or blind deconvolution approaches [3], [4], [5]), in the power spectrum domain (spectral enhancement approaches [6], [7]), or in the Cepstrum domain (Cepstral enhancement approaches [8], [9]). The second category of approaches try to design feature extraction strategies that are invariant to the effects of reverberation, such as RASTA-PLP [10] or MHEC features [11]. Finally, model adaptation approaches try to adapt clean-trained models to better match a set of reverberant observations. These include conventional adaptation strategies like Maximum Likelihood Linear Regression (MLLR) [12] and Maximum a Posteriori (MAP) estimation [13] (which were originally developed for speaker adaptation), as well as adaptation strategies that are specifically designed for environmental mismatch (such as Vector Taylor Series (VTS) adaptation [14], REMOS adaptation [15], etc.). A more comprehensive overview of these approaches can be found in [1].

In addition to the above mentioned approaches, the use
of Deep Neural Network (DNN) acoustic models [16] has also been shown to provide improved robustness to noise and reverberation compared to conventional GMM-HMM acoustic models [17], [18]. This inherent robustness is due to the fact that the higher layers in a DNN produce features that are increasingly invariant to small changes in the input features. However, in spite of this enhanced robustness, typical word error rates on distant-talking microphones can still be as high as twice that of a close-talking microphone.

For distant speech recognition, microphone array processing techniques (i.e., multichannel approaches) are popular solutions to achieve increased robustness against reverberation and noise [19], [20]. Most conventional array processing solutions belong to the linear filtering category (i.e., they try to obtain an estimate of the clean speech waveform). By using a uniform array of closely-spaced microphones in a fixed and known configuration, these approaches are able to perform spatial filtering (in the form of fixed or adaptive beamforming) to enhance the desired signal originating from a target location, while suppressing the reflections and interfering noise signals. Although such beamforming techniques are one of the most common solutions for distant speech recognition, there are a number of factors which limit their applicability. First, they necessarily require closely-spaced microphones with fixed and known positions, and also with central processing that maintains synchrony between their signals. Second, these approaches require the source location to be known (or estimated from the array signals, which is itself prone to errors in reverberant environments). Finally, beamforming solutions are sub-optimal for reverberation robustness, since there are always reflections coming from the array’s look direction which will not be suppressed. More recently, an alternative solution has been proposed in which a neural network is trained to perform speech recognition directly from multichannel audio [21], [22]. This approach avoids many of the problems mentioned above since it does not depend on array geometry or speaker location, but it requires a large amount of multichannel training data. However, it is fast when decoding since it does not require front-end processing.

In this study, as an alternative to classical approaches, we consider situations where we have multiple microphones available, but they do not form a compact synchronous array. Instead, they are independent recording devices distributed in random unknown locations, and there is no assumed synchrony between their signals (they do not share the same clock). This is a more flexible situation that covers a wider range of applications (it eliminates the need to have pre-designed and calibrated arrays), but it gives rise to a number of new challenges which makes conventional techniques inapplicable [23], [24]. First, the lack of synchrony among the different channels means there is no meaningful time-delay information between them, hence making beamforming impractical. The delays between the signals of different recording devices in this case are directly dependent on segmentation decisions which mark the beginning of an utterance. Thus, they do not carry any meaningful spatial information. Fig. 1 illustrates this problem for two distributed microphones. While the sound wave reaches microphone 2 with a delay compared to microphone 1, the recorded waveforms falsely show the opposite, because there is no shared time reference. Another challenge with distributed arrays is the possibility of very different signal-to-noise ratios (SNR) or direct-to-reverberation ratios (DRR) among the various channels.

The solution in such scenarios is to use multichannel extensions of spectrum enhancement approaches (for comparison, beamforming can be viewed as a multichannel extension of the linear filtering approach). Such solutions combine the information from the different channels in the power spectrum (or magnitude STFT) domain which makes them independent of the signal phases. Another advantage is that algorithms which operate on the magnitude spectra are less sensitive to speaker movements compared to linear filtering approaches, because the spectral envelope of the reverberation tail is fairly insensitive to the precise locations of the source and microphone.

In this study, the focus is on spectral enhancement approaches based on Nonnegative Matrix or Tensor Factorization (NMF or NTF) [25], [26]. Convolutive extensions of NMF [27] have previously been used for single-channel speech dereverberation [6] and for single-channel distant ASR [28]. Our early experiments in [29] showed the possibility of generalizing this approach to include additional channels by using tensor factorization. Here, we propose a more general framework for multichannel spectral enhancement with distributed microphones based on Convolutive Nonnegative Tensor Factorization (CNTF). The proposed generalization uses the family of alpha-beta divergences [30] for the factorization task, instead of the Euclidean distance measure which has typically been used for NMF-based dereverberation [28]. The two additional parameters in the alpha-beta divergence provide more flexibility to choose a divergence measure which better matches the distribution of the data. We will show that by selecting appropriate values for $\alpha$ and $\beta$, we can achieve increased robustness to reverberation and better ASR accuracy. Another goal in this study is to assess the performance improvements of the proposed CNTF algorithm with DNN-
based acoustic models which are known to be inherently more robust to noise and reverberation.

The remainder of this paper is organized as follows. In Section II, we describe the time-frequency model of reverberant speech which is used in the subsequent formulations. In Section III, using a tensor factorization interpretation for the described model, we present a multichannel spectrum enhancement algorithm for reverberation-robust distant ASR. We also present a method of divergence selection for the tensor factorization task in Section IV. The results of ASR experiments used to assess the performance of the proposed algorithm are reported in Section V, and Section VI provides a summary and concluding remarks.

II. TIME-FREQUENCY MODEL OF REVERBERANT SPEECH

The effect of the sound reflections received by a distant microphone in a reverberant environment is often described as a convolution between the clean speech signal and the room impulse response (RIR). The length of the RIR is a function of the room reverberation time ($T_{60}$) and is generally much longer than the duration of the short time segments used in speech recognition. The effect of the RIR on the spectrum can thus be viewed as a leakage of spectral content from one frame to subsequent frames. This long-term effect on subsequent frames is the reason why feature normalization techniques such as Cepstral Mean Normalization (CMN) or Relative Spectral filtering (RASTA) are not sufficient for addressing reverberation.

The relationship between reverberant and clean STFT coefficients can be formally expressed as [31],

$$
\hat{X}(k, m) = \sum_{k'=0}^{K-1} \sum_{p=0}^{L-1} H_{kk'}(p) \tilde{S}(k', m-p) + \tilde{E}(k, m),
$$

(1)

where $k$ is frequency index and $m$ is frame index, $\tilde{S}(k, m)$ and $\hat{X}(k, m)$ are the STFTs of the clean speech and received microphone signal, $\tilde{E}(k, m)$ is the STFT of the noise, and $\tilde{H}_{kk'}(m)$ is a time-frequency representation of the RIR effects using an analysis window of the form,

$$
w_{kk'}(n) = w_a(n)e^{j\frac{2\pi}{M}kn} * w_s(n)e^{j\frac{2\pi}{M}k'n},
$$

(2)

where $w_a(n)$ and $w_s(n)$ are the analysis and synthesis windows used in the STFT of the input speech, $N$ is the DFT length, and $*$ denotes convolution. Throughout all derivations in this paper, italic letters indicate scalars, lower-case bold letters represent vectors, and upper-case bold letters represent matrices.

Using the window given in (2), the time-frequency representation $H_{kk'}(m)$ can be expressed as

$$
H_{kk'}(m) = \sum_n h(n)w_{kk'}(mB - n),
$$

(3)

where $h(n)$ represents the time-domain RIR and $B$ is the skip period in STFT analysis. The summation is over length of the window. The cross-band filters $H_{kk'}(m)$ ($k \neq k'$) represent the aliasing effects due to the downsampling operation inherent in a STFT analysis with skip period of $B > 1$. The energy of the cross-band filters is small compared to the band-to-band filters $H_{kk}(m)$ due to the increasingly smaller overlap regions between the two modulated windows in (2) as $|k - k'|$ increases. As a result, it is common practice to consider the cross-band terms as additional noise terms, and write (1) in the magnitude STFT domain as

$$
X^{(i)}(k, m) = \sum_{p=0}^{L-1} H^{(i)}(k, p)S(k, m-p) + E^{(i)}(k, m),
$$

(4)

where $S(k, m)$ and $X^{(i)}(k, m)$ are the magnitude STFTs of the clean speech and reverberant speech of the $i$’th channel, $H^{(i)}(k, m) = |\tilde{H}^{(i)}_{kk}(m)|$ represents the spectral envelope of the RIR from the speaker position to the $i$’th microphone, and $E^{(i)}(k, m)$ represents the combined effects of additive noise, cross-band terms, and the error introduced by replacing the magnitude of the sum as required by (1) by a sum of magnitudes. The superscript $(i)$ is used to represent the channel index throughout this study. The length of the filters $H^{(i)}(k, m)$ is a function of room reverberation time. The model in (4) is a very useful representation which explicitly models the reverberation masking effect on the subband envelopes (using the filter $H^{(i)}(k, m)$), and serves as a basis for many studies on reverberant speech [6, 7, 15]. For the purposes of this study, we focus on reverberation robustness by assuming high-SNR recordings that allow us to neglect the additive noise term in (4). Dealing with additive interferences in the context of CNTF algorithm has been studied in [32].

III. CNTF ALGORITHM FOR ROBUST ASR WITH DISTRIBUTED MICROPHONES

A. Nonnegative tensor model for reverberant spectrograms

The magnitude spectrogram matrices from the individual channels of a set of distributed microphones can be viewed as frontal slices of a third-order nonnegative tensor $X$ with dimensions $K \times M \times C$, where $K$ is the number of frequency bins in the STFT analysis, $M$ is the total number of frames in the utterance, and $C$ is the number of microphones. Based on the

![Fig. 2: Convolutive tensor model for multichannel reverberant speech recognition. $X$ and $H(p)$ are used to denote three-dimensional tensors as a whole, and the matrices $X^{(i)}$ and $H^{(i)}(p)$ represent the frontal slices of these tensors, i.e. the magnitude spectrogram and the RIR component for the $i$’th channel.](image)
signal model in (4), this tensor can be obtained by multiplying delayed versions of the source signal’s spectrogram matrix by different diagonal matrices representing the different taps of the RIR spectral envelopes:

\[ X^{(i)} \approx \sum_{p=0}^{L-1} H^{(i)}(p)S^{p\rightarrow}. \tag{5} \]

Here, \( H^{(i)}(p) \) is a \( K \times K \) nonnegative diagonal matrix whose diagonal elements are \( H_k^{(i)}(k) \) \( (k = 0, \ldots, K-1) \). \( S \) is the spectrogram matrix of the source signal, and the operator \( p \rightarrow \) shifts the rows of its argument matrix by \( p \) positions to the right, filling in zeros from the left. Multiplying these shifted spectrograms with diagonal matrices \( H^{(i)}(p) \) and adding the results is equivalent to the individual convolutions in each frequency band. Note that \( H^{(i)}(p) \) is assumed to be diagonal due to the model in (4) which does not use the cross-band filters. Fig. 2 illustrates the tensor model described above.

The described tensor model enables us to remove reverberation by using a convolutive extension of NTF in order to decompose the tensor \( X \) into a sum of tensor-matrix products, with the frontal slices of the factor tensors constrained to be diagonal. The resulting convolutive NTF (CNTF) algorithm will thus attempt to discover a common component between the frontal slices of \( X \), which is the clean speech spectrogram, and discard the differences as the contribution from the convolutive effects of \( H^{(i)}(p) \) (i.e. the RIR effects).

To achieve this decomposition, a divergence measure is minimized between the observed tensor \( X \) and its estimate \( Z \) given by the nonnegative factors. The objective function for the CNTF algorithm can thus be expressed in terms of the tensor elements as,

\[ J = \sum_{i,k,m} D[X^{(i)}(k,m)||Z^{(i)}(k,m)], \tag{6} \]

where

\[ Z^{(i)}(k,m) = \sum_{p=0}^{L-1} \hat{H}^{(i)}(k,p)\hat{S}(k,m-p). \tag{7} \]

Here, \( \hat{H}^{(i)}(k,p) \) and \( \hat{S}(k,m-p) \) represent the current estimates for the nonnegative factors. Note that since the base matrices \( H^{(i)}(p) \) in (5) are diagonal, the optimization can be carried out independently in each frequency bin \( k \) by operating on the scalars \( \hat{H}^{(i)}(k,p) \) and \( \hat{S}(k,p) \) instead of the entire matrices in (5). The remainder of the derivations here will therefore follow this scalar form.

### B. Alpha-beta divergence

The choice of the divergence measure for tensor factorization influences the performance of the algorithm because it is closely related to the distribution of the data. Most common choices in NMF are the Euclidean Distance (ED) and the (generalized) Kullback-Leibler (KL) divergence [26], although other divergence measures have also been studied [33], [34]. Each of these divergences correspond to a certain underlying generative model assumed for the input data through the following relationship (here we drop channel, frame, and frequency indices for simplicity):

\[ p(X|Z;\theta) = \frac{1}{f(\theta)} \exp \left( -D(X||Z;\theta) \right), \tag{8} \]

where \( D(X||Z;\theta) \) indicates a divergence measure with parameters \( \theta \), and \( f(\theta) \) is a normalizing factor (called the partition function) which makes \( p(X|Z;\theta) \) a valid probability density function:

\[ f(\theta) = \int_X \exp(-D(X||Z;\theta))dX. \tag{9} \]

Based on this interpretation, minimizing a divergence measure is actually a maximum-likelihood (ML) estimation of the nonnegative factors assuming the corresponding distribution given by (8) [35]. For example, NMF with the Euclidean distance measure gives ML estimates for the factors under a Gaussian assumption for the input matrix elements (Several other correspondences of this kind are listed in Table I).

In this study, we propose to use the general family of \( \alpha\beta \)-divergences [30] for the described tensor factorization task in multichannel dereverberation:

\[ D_{\alpha\beta}(X||Z) = -\frac{1}{\alpha\beta} (X^{\alpha}Z^{\beta} - \frac{\alpha}{\alpha + \beta} X^{\alpha + \beta} - \frac{\beta}{\alpha + \beta} Z^{\alpha + \beta}). \tag{10} \]

It has been shown in [30] that (10) is a valid divergence measure which is nonnegative for all values of the arguments and has a global minimum of zero only when \( X = Z \). This general definition can be extended by continuity to include the singularity points \( \alpha = 0, \beta = 0 \) and \( \alpha + \beta = 0 \), so that the divergence will be defined for all \( \alpha, \beta \in \mathbb{R} \). These continuity extensions (as well as some other particular values for \( \alpha \) and \( \beta \), as summarized in Table I), coincide with a number of popular divergence measures used in the literature. The alpha-beta divergence is thus a unifying generalization which interpolates between these measures, providing increased flexibility to match the actual data distribution according to (8) by appropriately selecting the parameters \( \alpha \) and \( \beta \). We will show in Sec. V that the use of \( \alpha\beta \)-divergence with appropriate values of the parameters improves performance of the CNTF algorithm, resulting in better ASR accuracy. This is because the Gaussianity assumption underlying the commonly used Euclidean distance is inaccurate for the observed power spectrum data, which results in sub-optimal ML estimates for the factors.

Using the \( \alpha\beta \)-divergence, our goal in CNTF enhancement is

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>( \beta )</th>
<th>divergence</th>
<th>underlying distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>Euclidean Distance</td>
<td>Gaussian</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>Generalized Kullback-Leibler (KL) Divergence</td>
<td>Poisson</td>
</tr>
<tr>
<td>1</td>
<td>-1</td>
<td>Itakura-Saito (IS) divergence</td>
<td>Gamma</td>
</tr>
</tbody>
</table>
to find the factors \( H^{(i)}(k, p) \) and \( S^{(i)}(k, m) \) which minimize the cost function in (6) subject to the nonnegativity constraints:

\[
H^{(i)}(k, p) \geq 0 \quad \text{and} \quad S^{(i)}(k, m) \geq 0, \quad \text{for all } i, k, p, m. \quad (11)
\]

At the point of minimum divergence, \( S^{(i)}(k, m) \) is expected to be an estimate the clean speech magnitude spectrum, while \( H^{(i)}(k, m) \) will represent the RIR spectral envelope for channel \( i \).

**C. Multiplicative update rules**

In this section, we derive update rules for \( S^{(i)}(k, m) \) and \( H^{(i)}(k, p) \) which yield the minimization of (6). As explained in Sec. III-A, the diagonality of the frontal slices of tensors \( \mathcal{H}(p) \) allows us to carry out the optimization independently in each frequency bin, thus breaking the matrix formulations into equivalent scalar forms with simpler gradient expressions.

For alpha and beta values in the region \( 1 - \frac{\beta}{\alpha} \in [0, 1] \), the cost function (6) is ensured to be convex with respect to either \( S^{(i)}(k, m) \) or \( H^{(i)}(k, p) \), but not both [30]. Consequently, the optimization of the nonnegative components are carried out in an alternating fashion, i.e., by keeping \( S^{(i)}(k, m) \) constant and updating \( H^{(i)}(k, p) \), and vice versa.

The derivatives of the cost function (6) with respect to the factor variables are:

\[
\frac{\partial J}{\partial H^{(i)}(k, p)} = -\frac{1}{\alpha} \sum_{m} \left[ X^{(i)}(k, m)^{\alpha} Z^{(i)}(k, m)^{\beta - 1}
- Z^{(i)}(k, m)^{\alpha + \beta - 1} \right] S^{(i)}(k, m - p),
\]

\[
\frac{\partial J}{\partial S^{(i)}(k, l)} = -\frac{1}{\alpha} \sum_{i} \sum_{m} \left[ X^{(i)}(k, m)^{\alpha} Z^{(i)}(k, m)^{\beta - 1}
- Z^{(i)}(k, m)^{\alpha + \beta - 1} \right] H^{(i)}(k, m - l).
\]

Gradient descent updates using the above derivatives do not necessarily preserve nonnegativity of the results. However, similar to conventional NMF algorithms [26], we can derive multiplicative update rules (which guarantee the preservation of nonnegativity) by assuming the following adaptive learning rates for the gradient descent optimizations:

\[
\eta_H = \frac{H^{(i)}(k, p)}{\frac{1}{\alpha} \sum_{m} Z^{(i)}(k, m)^{\alpha + \beta - 1} S^{(i)}(k, m - p)},
\]

\[
\eta_S = \frac{S^{(i)}(k, l)}{\frac{1}{\alpha} \sum_{i} \sum_{m} Z^{(i)}(k, m)^{\alpha + \beta - 1} H^{(i)}(k, m - l)}.
\]

Using (14) and (15) with the gradients in (12) and (13) results in the following gradient descent update equations:

\[
H^{(i)}(k, p) \leftarrow H^{(i)}(k, p) \frac{\sum_{m} Y^{(i)}(k, m) S^{(i)}(k, m - p)}{\sum_{m} Z^{(i)}(k, m)^{\alpha + \beta - 1} S^{(i)}(k, m - p)},
\]

\[
S^{(i)}(k, l) \leftarrow S^{(i)}(k, l) \frac{\sum_{i} \sum_{m} Y^{(i)}(k, m) H^{(i)}(k, m - l)}{\sum_{i} \sum_{m} Z^{(i)}(k, m)^{\alpha + \beta - 1} H^{(i)}(k, m - l)},
\]

where,

\[
Y^{(i)}(k, m) = X^{(i)}(k, m)^{\alpha} Z^{(i)}(k, m)^{\beta - 1}.
\]

Note that both numerators and denominators in update equations (16) and (17) are in the form of correlations between the estimated factors \( H^{(i)}(k, p) \) and \( S^{(i)}(k, m) \) and the intermediate variables \( Z^{(i)}(k, m) \) and \( Y^{(i)}(k, m) \). In practice, similar to [36], these correlations are computed via FFT multiplication which considerably reduces the computational complexity of the CNTF algorithm. Also note that by setting \( \alpha = \beta = 1 \) in (16) and (17), we obtain the update equations of [29] for CNTF with Euclidean Distance objective function. Moreover, by setting the number of channels to \( C=1 \), the algorithm simplifies to the single-channel CNMF algorithms of [6] and [28].

To address the scale indeterminacy inherent in the decomposition of (7), we impose the additional constraint \( \sum_{i} \sum_{p} H^{(i)}(k, p) = 1 \), which is satisfied by performing the following normalization for the RIR spectral envelopes after each update:

\[
H^{(i)}(k, p) \leftarrow \frac{H^{(i)}(k, p)}{\sum_{i} \sum_{p} H^{(i)}(k, p)}. \quad (19)
\]

We prefer to use this normalization strategy which is shared among the RIR components of different channels instead of individually normalizing \( H^{(i)}(k, p) \) for each channel (i.e. normalizing only over different lags \( p = 1, \ldots, L - 1 \) for each channel \( i \)). This is because the normalization in (19) allows the algorithm to automatically adjust the gains of the filters \( H^{(i)}(k, p) \) according to the DRR of the corresponding channel. This will in turn adjust the contribution of each individual channel to the final estimate of \( S^{(i)}(k, m) \) according to the distance between the corresponding microphone and the speaker. This is required in a blind scenario where we have no information about speaker or microphone locations. We believe such a strategy is beneficial compared to channel selection algorithms which identify noisy channels and completely eliminate their contribution [37], [38] (see the results in Sec. V for more details about this automatic adjustment of channel contributions).
of $\alpha$ and $\beta$ in Table I. In [39], an alternative method called score-matching, is introduced for the estimation of such non-normalized models in which the pdf is only known up to a multiplicative constant ($f(\alpha, \beta)$ in our case). The score matching technique was extended to the case of nonnegative data in [40], and was successfully used for divergence selection in the task of music analysis in [41]. Here, we apply the score-matching technique for selecting the parameters of $\alpha\beta$-divergence in the CNTF algorithm.

A. The score-matching principle

The score function of a distribution is defined as the derivative of its log-density:

$$\psi(X; \theta) = \frac{\partial \log p(X; \theta)}{\partial X}, \quad (20)$$

The point in using the score function is that it does not depend on the normalization term $f(\theta)$ (the partition function). It has been shown in [39] that the parameters of non-normalized models can be estimated by minimizing the expected distance between the score function resulting from the observed data and the score function given by the model. An extension of this principle for nonnegative data [40] states that the score-matching (SM) estimator for the parameters $\theta$ is given by:

$$\hat{\theta} = \arg\min_{\theta} L_{SM}(\theta), \quad (21)$$

$$L_{SM}(\theta) = \int_{\mathbb{R}^+} p(X; \theta) [2X\psi(X; \theta) + X^2\xi(X; \theta)] + \frac{1}{2}\psi^2(X; \theta)X^2 dX, \quad (22)$$

where $\xi(X; \theta)$ indicates the derivative of the score function (i.e., the second derivative of the log-density):

$$\xi(X; \theta) = \frac{\partial^2 \log p(X; \theta)}{\partial X^2}. \quad (23)$$

In practice, the integral mean in (22) is replaced by a sample average over the data.

B. Score-matching estimator for $\alpha$ and $\beta$ in CNTF algorithm

Using the alpha-beta divergence and based on the assumed distribution in (8), the score function for the magnitude STFT coefficients and its derivative will be given by:

$$\psi(X|Z; \alpha, \beta) = -\frac{\partial D_{\alpha\beta}(X||Z)}{\partial X} = \frac{1}{\beta}X^{\alpha-1}(X^\beta - Z^\beta), \quad (24)$$

$$\xi(X|Z; \alpha, \beta) = -\frac{\partial^2 D_{\alpha\beta}(X||Z)}{\partial X^2}$$

$$= \frac{1}{\beta}X^{\alpha-2}\{(\alpha - 1)Z^\beta - (\alpha + \beta - 1)X^\beta\}. \quad (25)$$

Inserting (24) and (25) into the SM objective function (22), and replacing the integral by a sample mean, we obtain:

$$L_{SM}(\alpha, \beta) = \sum_{i,k,m} \left[\frac{\alpha + 1}{\beta} X^\alpha Z^\beta - \frac{\alpha + \beta + 1}{\beta} X^{\alpha+\beta} \right.$$

$$\left. + \frac{1}{\beta^2} X^{2\alpha}(Z^\beta - X^\beta)^2\right]. \quad (\beta \neq 0) \quad (26)$$

Algorithm 1 CNTF dereverberation

**Input:** $X^{(i)}(k, m)$, $Y^{(i)}(k, m)$, $H^{(i)}(k, f)$

$m \in \{0, \ldots, M - 1\}$

$k \in \{0, \ldots, K - 1\}$

$i \in \{1, \ldots, C\}$

$p \in \{0, \ldots, L - 1\}$

Set $\alpha = 1, \beta = 1$

Set $F = M + L - 1$

**repeat**

  **Initialize** $H^{(i)}(k, p) = 1 - \frac{p}{\pi r}$

  **Initialize** $S(k, m) = X^{(i)}(k, m)$

  **for** $\text{iter} = 1$ to $N$ **do**

  **> Computing** $Z^{(i)}(k, m)$ via FFT multiplication:

  $$H^{(i)}(k, f) = \sum_{p=0}^{L-1} H^{(i)}(k, p) \exp(-j \frac{2\pi}{F} f)$$

  $$S(k, f) = \sum_{m=0}^{M-1} S(k, m) \exp(-j \frac{2\pi}{F} f)$$

  $$Z^{(i)}(k, m) = \sum_{f=0}^{F-1} H^{(i)}(k, f) \exp(j \frac{2\pi}{F} f)$$

  **> Intermediate variables**:

  $$Y^{(i)}(k, m) = X^{(i)}(k, m)^\alpha Z^{(i)}(k, m)^{\beta-1}$$

  $$V^{(i)}(k, m) = Z^{(i)}(k, m)^{\alpha+\beta-1}$$

  **> Compute correlations via FFT multiplication**:

  $$C_{YS}^{(i)}(k, p) = \sum_{f=0}^{F-1} Y^{(i)}(k, f) S^*(k, f) \exp(j \frac{2\pi}{F} f)$$

  $$C_{XYS}^{(i)}(k, p) = \sum_{f=0}^{F-1} Y^{(i)}(k, f) S^*(k, f) \exp(j \frac{2\pi}{F} f)$$

  **> Update nonnegative factors**:

  $$H^{(i)}(k, p) \leftarrow H^{(i)}(k, p) \frac{C_{YS}^{(i)}(k, p)}{C_{XYS}^{(i)}(k, p)}$$

  $$S(k, m) \leftarrow S(k, m) \frac{\sum_{i=1}^{C} C_{YS}^{(i)}(k, m)}{\sum_{i=1}^{C} C_{XYS}^{(i)}(k, m)}$$

  **> Normalization**:

  $$H^{(i)}(k, p) \leftarrow \frac{H^{(i)}(k, p)}{\sum_{p=0}^{L-1} H^{(i)}(k, p)}$$

  **end for**

  **> Update** $\alpha$ and $\beta$:

  $$L_{SM}(\alpha, \beta) = \sum_{i,k,m} \left[\frac{\alpha + 1}{\beta} X^\alpha Z^\beta - \frac{\alpha + \beta + 1}{\beta} X^{\alpha+\beta} \right.$$

  $$\left. + \frac{1}{\beta^2} X^{2\alpha}(Z^\beta - X^\beta)^2\right].$$

  $$(\alpha, \beta) = \arg\min_{(\alpha, \beta)} L_{SM}(\alpha, \beta) \quad 1 - \frac{\beta}{\alpha} \in [0, 1]$$

  until $\alpha$ and $\beta$ converge, or maximum iterations is reached

**Output:** $S(k, m)$

Superscript * represents complex conjugate. All variables with a bar represent FFT domain variables.

Note that in (26) we have dropped all indices from $X^{(i)}(k, m)$ and $Z^{(i)}(k, m)$ (i.e., replaced them with $X$ and $Z$) for simplicity. For the singularity point $\beta = 0$, we extend (26) by continuity which yields:

$$L_{SM}(\alpha, 0) = \sum_{i,k,m} \left[\left(-1 + (\alpha + 1) \log \frac{Z}{X}\right) X^\alpha \right.$$

$$\left. + \frac{1}{2} \left(X^\alpha \log \frac{Z}{X}\right)^2\right]. \quad (27)$$

The optimum values for $\alpha$ and $\beta$ are thus found by searching for a $(\alpha, \beta)$ pair in the convexity region $(\frac{1-\beta}{\alpha} \in [0, 1])$ which minimizes $L_{SM}(\alpha, \beta)$.

The problem with the above procedure for divergence
V. EXPERIMENTS

A. Speech Data

To provide a flexible framework in choosing different microphone configurations and speaker locations, we use room impulse responses from Aachen Impulse Response (AIR) database [42]. This is a collection of RIRs recorded in real rooms with different $T_{60}$ values at different source-to-microphone distances. These RIRs have been measured using maximum length sequences of degree 15 as the excitation signal. The measurements have been performed at a sampling rate of 48 kHz with an accuracy of 24 bits, using professional audio equipment providing high-quality and low-noise results. Detailed information about AIR database can be found in [42].

To generate reverberant speech data, we convolve TIMIT utterances with different RIRs from the AIR databases. We use the standard TIMIT data partitions, consisting of 462 speakers for train and 168 speakers for test. This produces a medium-vocabulary ASR task for distant read speech which will allow us to concentrate on the evaluation of our front-end enhancement framework.

The AIR dataset provided RIRs from different rooms with different reverberation times. These include an office room ($T_{60} \approx 430$ ms), a stairway area ($T_{60} \approx 800$ ms) and a lecture hall ($T_{60} \approx 800$ ms). For the majority of experiments reported here, we use the RIRs from the stairway area which have been recorded at different distances (1m, 2m, 3m) and different azimuth angles for each distance (ranging from $0^\circ$ to $180^\circ$). This allows us to test the algorithm for different source and microphone configurations. The microphone locations used in this study are illustrated in Fig. 3, where they have been numbered for easy referencing.

In addition, the experiments at the end of this section (V-D) use RIRs from the office area and lecture hall as well. This will enable us to have diverse reverberation characteristics for experiments with multi-condition training.

### Table II: Word Error Rates (%) in ASR experiments with clean-trained models

<table>
<thead>
<tr>
<th>number of channels</th>
<th>Feature Processing</th>
<th>$d = 2m$</th>
<th>$d = 3m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C=1</td>
<td>No enhancement</td>
<td>21.0</td>
<td>29.9</td>
</tr>
<tr>
<td></td>
<td>CNTF ($\alpha = 1, \beta = 1$) (Euclidean Distance)</td>
<td>8.7</td>
<td>13.0</td>
</tr>
<tr>
<td></td>
<td>CNTF ($\alpha = 1, \beta = 0$) (KL Divergence)</td>
<td>8.5</td>
<td>12.8</td>
</tr>
<tr>
<td></td>
<td>CNTF ($\alpha = 1, \beta = -1$) (IS Divergence)</td>
<td>9.4</td>
<td>13.6</td>
</tr>
<tr>
<td></td>
<td>CNTF (SM estimates for $\alpha, \beta$)</td>
<td>8.5</td>
<td>12.7</td>
</tr>
<tr>
<td>C=2</td>
<td>CNTF ($\alpha = 1, \beta = 1$) (Euclidean Distance)</td>
<td>7.2</td>
<td>10.2</td>
</tr>
<tr>
<td></td>
<td>CNTF ($\alpha = 1, \beta = 0$) (KL Divergence)</td>
<td>6.6</td>
<td>9.8</td>
</tr>
<tr>
<td></td>
<td>CNTF ($\alpha = 1, \beta = -1$) (IS Divergence)</td>
<td>7.0</td>
<td>10.7</td>
</tr>
<tr>
<td></td>
<td>CNTF (SM estimates for $\alpha, \beta$)</td>
<td>6.2</td>
<td>9.2</td>
</tr>
<tr>
<td>C=4</td>
<td>CNTF ($\alpha = 1, \beta = 1$) (Euclidean Distance)</td>
<td>5.9</td>
<td>8.1</td>
</tr>
<tr>
<td></td>
<td>CNTF ($\alpha = 1, \beta = 0$) (KL Divergence)</td>
<td>5.5</td>
<td>7.6</td>
</tr>
<tr>
<td></td>
<td>CNTF ($\alpha = 1, \beta = -1$) (IS Divergence)</td>
<td>5.9</td>
<td>8.3</td>
</tr>
<tr>
<td></td>
<td>CNTF (SM estimates for $\alpha, \beta$)</td>
<td>4.8</td>
<td>6.7</td>
</tr>
</tbody>
</table>

* Single-channel experiments use microphones at $90^\circ$ (i.e. microphones 6,10) in Fig. 3. Dual-channel experiments use microphones at $30^\circ$ and $90^\circ$ (i.e. microphones 6,8,10,12).

B. ASR system setup

All ASR experiments reported here use hybrid DNN-HMM acoustic models and a trigram language model created using TIMIT transcriptions (resulting in a medium-vocabulary task with approximately 6000 words). Using a frame size of 25 ms and a skip rate of 10 ms, we extract Mel-filterbank features with 24 Mel filters along with their first and second order derivatives (i.e., delta and double-delta features). We choose Mel-filterbank coefficients because they have been shown to provide consistently better accuracies with DNN-based acoustic models compared to Mel-frequency cepstral coefficients (MFCCs) which are commonly used in GMM-HMM systems [43]. Utterance based mean and variance normalization has been used for these features in all experiments. We initially train tied-state trigram GMM-HMM acoustic models and a trigram language model created using TIMIT transcriptions (resulting in a medium-vocabulary task). The trained models are then used for forced-alignment of the training data to obtain frame-level senone labels for the training features. Using these labels, we train a deep neural network acoustic model. The inputs to the DNN are concatenated features from a context window of 11 frames. Considering the limited training data in the TIMIT corpus and in order to prevent overfitting, we use a DNN with 3 hidden sigmoid layers each containing 1024 nodes. A softmax output layer in the DNN converts the activations of the last hidden layer into senone posteriori. The DNN parameters are tuned by stochastic gradient descent (SGD) using error back-propagation to minimize the frame-level cross-entropy.
between the DNN outputs and the ground-truth senone labels from the forced alignment. The DNN training consists of 30 epochs over the training data with a fixed learning rate of 0.04, using a minibatch size of 256 features. The described clean-trained acoustic model results a word error rate of 3.1% on the clean test set of TIMIT, which is expected for a medium-vocabulary matched train and test experiment.

For recognizing reverberant test data, we first compute magnitude STFTs for all channels using a frame size of 64 ms and a skip rate of 16 ms. These are then jointly processed by 10 iterations of the CNTF dereverberation algorithm to produce an estimate of the clean speech spectrogram. The choice of 10 iterations was experimentally verified to be adequate to provide best ASR accuracy. The CNTF algorithm uses a filter length of \( L = 16 \) for all \( H^{(i)}(k,p) \), and the filters are initialized with \( H^{(i)}(k,p) = 1 - p / 2L, (p = 0, \ldots, L-1) \). We perform experiments both with the commonly used Euclidean distance measure (and some other well-known fixed divergences) as well as with \( \alpha \beta \)-divergence using optimum values for \( \alpha \) and \( \beta \) given by the score-matching estimator discussed in Sec. IV. The clean speech magnitude STFT is used together with the phases from one of the channels (first microphone) to reconstruct an estimate for the clean speech waveform, which is finally used for Mel-filterbank feature extraction.

C. ASR results and algorithm analysis

Table II shows the results of ASR experiments performed to assess the performance of the CNTF algorithm in different DRR scenarios with different microphone configurations. All experiments in this section use a subset of the microphones shown in Fig. 3. The word error rates shown in the table indicate the effectiveness of the proposed dereverberation strategy in both single-channel and multi-channel scenarios. Using a single-channel version of the algorithm in Table II (i.e., setting \( C = 1 \)) with the Euclidean distance measure, relative improvements of +58.5% and +56.5% are provided over the raw filterbank features for source-to-microphone distances of \( d = 2m \) and \( d = 3m \). The relative WER improvements provided by dual-microphone and four-microphone configurations over the single-channel scenario are +17.2% and +32.2% for \( d = 2m \), and +21.5% and +37.7% for \( d = 3m \).

The improvements provided by adding more microphones is more significant in low-DRR scenarios. It can be observed that although the DNN-based acoustic model shows some inherent robustness to reverberation (indicated by the starting WER of 29.9% in the low-DRR condition of \( d = 3m \)), the proposed multichannel front-end enhancement strategy can provide significant improvements over the clean-trained DNN baseline. Also shown in Table II are the WERs obtained with different divergence measures in each scenario. Although all measures are able to provide considerable improvements, the best ASR accuracy in most cases is achieved by an \( (\alpha, \beta) \) pair given by the score-matching estimator introduced in Sec. IV.

To better understand the operation of the algorithm, we have plotted the RIR spectral envelopes \( (H^{(i)}(k,p)) \) discovered by the CNTF algorithm in Figures 4 and 5 for different configurations. In the first experiment, we use a 3-channel version of the algorithm for three microphones located at distances of 1m, 2m, and 3m from the source (microphones 3, 7, and 11 in Fig. 3). The resulting RIR envelopes estimated by the CNTF algorithm for each channel are then averaged across all frequencies and plotted in Fig. 4. It can be observed that the algorithm has correctly identified the RIR spectral envelopes consistently across the reverberant spectrograms. Microphone 3 at \( d = 1m \) has a fairly high DRR which results in a fast exponential decay of the corresponding RIR’s spectral envelope. In contrast, the RIR spectral envelope of microphone 11 at \( d = 3m \) has a much slower energy decay due to the lower DRR of this channel.

We also perform a second experiment which is designed to illustrate the algorithm’s performance when we have highly unbalanced DRRs among the channels. We test a 4-channel example of such situations, in which three channels (microphones 1, 2, and 3) are fairly close to the source \( (d = 1m) \), but the fourth channel (microphone 11) is farther away at \( d = 3m \). The resulting estimates of \( H^{(i)}(k,p) \) given by the CNTF algorithm are averaged across all frequencies and plotted in Fig. 5. It is observed that the algorithm has automatically identified the low-DRR channel and has assigned much smaller values to the corresponding \( H^{(i)}(k,p) \). Considering the update equation of (17), this will in turn reduce the contribution of the low-DRR channel to the final estimate of the clean speech spectrogram. Thus, the algorithm is capable of blindly adjusting the
channel contributions according to their signal quality. This is a very useful characteristic in a blind scenario where there is no information about the source and microphone locations.

D. Comparisons with training on multi-condition or enhanced data

A DNN acoustic model trained on multi-condition data from various environments is becoming increasingly popular to handle far-field ASR tasks [17]. If such multi-condition data is available, a DNN with a sufficient number of hidden layers usually provides very good results, outperforming most other approaches. In this section, we study the proposed algorithm’s performance in situations where a multi-condition training data is available. We compare three different cases. The first method uses clean-trained acoustic models and applies CNTF to reverberant test data in order to reduce the mismatch. The second approach uses reverberant data from multiple different rooms and distances to train multi-condition models, and uses reverberant features directly for decoding (no feature processing). The third approach applies CNTF to both training and test data to further reduce the mismatch.

The experiments in this section use RIRs from all three environments in the AIR dataset (office room, lecture hall, stairway). Using different source-to-microphone distances available for each room, we have a collection of 62 different RIRs to create a multi-condition training data which is diverse in terms of reverberation characteristics. The multi-condition data is created by convolving each utterance in the TIMIT training set with a randomly selected RIR from this pool. The resulting reverberant signal is time-synchronized with the original clean utterance so that the ground-truth senone alignments from the clean data can be used for training. We consider three different test environments. The first set of experiments are in an office room ($T_{60} \approx 0.43s$) with two microphones placed at $d = 2m$ and $d = 3m$ from the source. The second set of experiments are conducted in the stairway area which possesses a longer reverberation time ($T_{60} \approx 0.8s$), keeping the same distances of $d = 2m$ and $d = 3m$. For the last set of experiments, we use a lecture hall which has a similar reverberation time of $T_{60} \approx 0.8s$, but we increase source-to-microphone distances to $d = 4m$ and $d = 5.5m$. For all of the experiments in this section, fixed values of $\alpha = 1$ and $\beta = 1$ have been used for CNTF updates.

The results of ASR experiments in the above three cases are shown in Table III. All multichannel experiments here have unbalanced DRRs, i.e., one microphone is located farther from the source compared to the other (note the distances mentioned in Table III). An important observation from the table is that although the reverberant microphone has a more corrupted signal (lower DRR), including it in a joint CNTF processing with the closer channel is still superior to a channel selection strategy in which only the closer microphone is identified and used [37]. In this case, based on the discussion in Section V-C, although the reverberation filters $H^{(1)}(k, p)$ place a smaller weight on the low-DRR channel, they are still able to draw useful information from this channel which helps ASR accuracy. Another result to be noted in Table III is that although the multi-condition DNN baseline (third row) provides considerable robustness to reverberation and good accuracy, the error rates can be further reduced by applying CNTF processing to both train and test data, because this will further reduce the mismatch. Therefore, while CNTF front-end processing is necessary in mismatched conditions with clean-trained models, it can also be beneficial for multi-condition models. We believe that for very large datasets spanning very diverse environmental conditions, multi-condition trained DNN acoustic models are sufficient without front-end processing. However, for moderate data sizes which inevitably exclude many possible reverberation conditions, reducing the mismatch by applying CNTF processing will prove helpful, as verified by the results in Table III.

<table>
<thead>
<tr>
<th>Train data</th>
<th>Features</th>
<th>Office Room ($T_{60} \approx 0.43s$)</th>
<th>Stairway Area ($T_{60} \approx 0.8s$)</th>
<th>Lecture Hall ($T_{60} \approx 0.8s$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>clean</td>
<td>fbank + CNTF (1ch)</td>
<td>8.8</td>
<td>10.8</td>
<td>15.9</td>
</tr>
<tr>
<td>clean</td>
<td>fbank + CNTF (2ch)</td>
<td>8.4</td>
<td>9.2</td>
<td>12.9</td>
</tr>
<tr>
<td>multi-cond.</td>
<td>fbank (1ch)</td>
<td>6.6</td>
<td>7.3</td>
<td>10.6</td>
</tr>
<tr>
<td>multi-cond. (w/ CNTF)</td>
<td>fbank + CNTF (1ch)</td>
<td>6.3</td>
<td>6.6</td>
<td>9.1</td>
</tr>
<tr>
<td>multi-cond. (w/ CNTF)</td>
<td>fbank + CNTF (2ch)</td>
<td>5.9</td>
<td>6.3</td>
<td>8.1</td>
</tr>
</tbody>
</table>

* Single channel experiments use the microphone closer to the source.

VI. CONCLUSION

We have presented an algorithm for reverberation-robust distant speech recognition using distributed far-field microphones which is based on convolutive nonnegative tensor factorization (CNTF). The developed algorithm attempts to remove the convolutional effects of the room impulse response from the subband envelopes of the received signals. In the single channel case, the algorithm simplifies to nonnegative matrix factorization and attempts to decompose each subband envelope into a convolution of two components, one being the clean speech subband envelope and the other representing the convolutive effects of the RIR. In the multichannel case, the algorithm makes use of the additional observations to improve this decomposition. In each subband, the algorithm tries to identify a common component among the subband envelopes of the different channels (the clean speech component) and discards the convolutive differences between the channels as RIR distortions.

The proposed algorithm was shown to provide considerable improvements for a clean-trained ASR system in both single-channel and multichannel scenarios. It is important to note that although the DNN acoustic models used in this work show some inherent robustness to feature distortions, they are still unable to provide satisfactory results in clean-trained...
scenarios, illustrating the importance of front-end enhancement solutions for mismatched train and test conditions. This is especially important for reverberation robustness, since reverberation is a long-term effect that spans a large number of short frames used in speech recognition. Although a DNN acoustic model tries to partially compensate this by using concatenated features from neighboring frames, this is not adequate for sufficiently suppressing the reverberation component. The proposed CNTF algorithm explicitly addresses this long-term effect by attempting to remove the long-term correlations introduced in the subband envelopes by the RIR. Moreover, it was shown that the proposed algorithm can also be helpful with multi-condition training data, since applying CNTF to both train and test signals will help reduce the mismatch. Our future work would include assessing the performance of the CNTF algorithm in the presence of environmental noise and also in conversational ASR tasks. We believe that the improvements resulting from the use of optimized values of $\alpha$ and $\beta$ in this study will be more accentuated in the presence of noise, since it helps to adapt to the statistics of the observed features.

REFERENCES


John H.L. Hansen (IEEE S’81-M’82-SM’93-F’07) received the Ph.D. and M.S. degrees in Electrical Engineering from Georgia Institute of Technology, Atlanta, Georgia, in 1988 and 1983, and B.S.E.E. degree from Rutgers University, College of Engineering, New Brunswick, N.J. in 1982. He was awarded honorary degree Doctor Technics Honoris Causa from Aalborg University (Aalborg, DK) (April 2016) in recognition of his contributions to speech signal processing and speech/language/hearing science. He joined University of Texas at Dallas (UTDallas), Engineering and Computer Science in 2005, where he presently serves as Jonsson School Associate Dean for Research, as well as Professor of Electrical Engineering and also holds the Distinguished University Chair in Telecommunications Engineering. He previously served as Department Head of Electrical Engineering from Aug. 2005 Dec. 2012, overseeing a +4x increase in research expenditures ($4.5M to $22.3M) with a 20% increase in enrollment along with 18 additional T/TT faculty, growing UTDallas to be the 8th largest EE program from ASEE rankings in terms of degrees awarded. He also holds a joint appointment as Professor in the School of Behavioral and Brain Sciences (Speech & Hearing). At UTDallas, he established the Center for Robust Speech Systems (CRSS) which is part of the Human Language Technology Research Institute. Previously, he served as Dept. Chair and Professor of Dept. of Speech, Language and Hearing Sciences, and Professor of the Dept. of Electrical & Computer Engineering, at Univ. of Colorado Boulder (1998-2005), where he co-founded and served as Associate Director of the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTDallas. He has been named IEEE Fellow (2007) for contributions in “Robust Speech Recognition in Stress and Noise,” Inter. Speech Communication Association (ISCA) Fellow (2010) for contributions on research for speech processing of signals under adverse conditions, and received The Acoustical Society of Americas 25 Year Award (2010) in recognition of his service, contributions, and membership to the Acoustical Society of America. He is currently serving as the elected Vice-President of ISCA and member of ISCA Board. He was also selected and is serving as Vice-Chair on U.S. Office of Scientific Advisory Committees (OSAC) for OSAC-Speaker in the voice forensics domain (2015-2017). Previously he served as IEEE Technical Committee (TC) Chair and Member of the IEEE Signal Processing Society: Speech-Language Processing Technical Committee (SLTC) (2005-08; 2010-14); elected IEEE SLTC Chair for 2011-13, Past-Chair for 2014), and elected ISCA Distinguished Lecturer (2011/12). He has served as member of IEEE Signal Processing Society Educational Technical Committee (2005-08; 2008-10); Technical Advisor to the U.S. Delegate for NATO (IST/G1-01); IEEE Signal Processing Society Distinguished Lecturer (2005/06), Associate Editor for IEEE Trans. Speech & Audio Processing (1992-99), Associate Editor for IEEE Signal Processing Letters (1998-2000), Editorial Board Member for IEEE Signal Processing Magazine (2001-03); and guest editor (Oct. 1994) for special issue on Robust Speech Recognition for IEEE Trans. Speech & Audio Proc. He has served on Speech Communications Technical Committee for Acoustical Society of America (2000-03), and previously on ISCA Advisory Council. His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human-computer interaction. He has supervised 73 PhD/MS thesis candidates (36 PhD, 37 MS/MA), was recipient of The 2005 University of Colorado Colorado Teacher Recognition Award as voted on by the student body, author/co-author of 600 journal and conference papers including 11 textbooks in the field of speech processing and language technology, coauthor of textbook Discrete-Time Processing of Speech Signals, (IEEE Press, 2000), co-editor of DSP for In-Vehicle and Mobile Systems (Springer, 2004), Advances for In-Vehicle and Mobile Systems: Challenges for International Standards (Springer, 2006), In-Vehicle Corpus and Signal Processing for Driver Behavior (Springer, 2008), and lead author of the report The Impact of Speech Under Stress on Military Speech Technology, (NATO RTO-TR-10, 2000). He also organized and served as General Chair for ISCA Interspeech-2002, Sept. 16-20, 2002, Co-Organizer and Technical Program Chair for IEEE ICASSP-2010, Dallas, TX, March 15-19, 2010, and Co-Chair and Organizer for IEEE SLT-2014, Dec. 7-10, 2014 in Lake Tahoe, NV.

Seyedmahdad (Matt) Mirsamadi (IEEE S’08) received his B.S. and M.S. degrees in Electrical and Electronic Engineering from Amirkabir University of Technology, Tehran, Iran, in 2009 and 2011, respectively. During his Masters studies, he joined the Multimedia Signal Processing Research Lab at Amirkabir University, where he worked on acoustic signal processing with a focus on microphone arrays and multichannel audio systems. In 2013, he joined the Center for Robust Speech Systems (CRSS) at the University of Texas at Dallas (UTD), where he is currently pursuing his Ph.D. degree as a research assistant. He was a recipient of the best selected paper award in robust speech recognition at IEEE Spoken Language Technology (SLT) Workshop in 2014. His research interests include speech signal processing, speech recognition, speech enhancement, machine learning and artificial intelligence, data science, and applications of machine learning in speech and audio processing.