Keronen *et al. EURASIP Journal on Advances in Signal Processing* (2015) 2015:76
DOI 10.1186/s13634-015-0259-1

EURASIP Journal on Advances in Signal Processing
a SpringerOpen Journal

RESEARCH Open Access

Feature enhancement of reverberant speech by distribution matching and non-negative matrix factorization

Sami Keronen[1*], Heikki Kallasjoki[1], Kalle J. Palomäki[1], Guy J. Brown[2] and Jort F. Gemmeke[3]

Abstract

This paper describes a novel two-stage dereverberation feature enhancement method for noise-robust automatic speech recognition. In the first stage, an estimate of the dereverberated speech is generated by matching the distribution of the observed reverberant speech to that of clean speech, in a decorrelated transformation domain that has a long temporal context in order to address the effects of reverberation. The second stage uses this dereverberated signal as an initial estimate within a non-negative matrix factorization framework, which jointly estimates a sparse representation of the clean speech signal and an estimate of the convolutional distortion. The proposed feature enhancement method, when used in conjunction with automatic speech recognizer back-end processing, is shown to improve the recognition performance compared to three other state-of-the-art techniques.

Keywords: Speech dereverberation; Feature enhancement; Non-negative matrix factorization; Distribution matching

1 Introduction

Automatic speech recognition (ASR) is becoming an effective and versatile way to interact with modern machine interfaces. However, in order to successfully adopt ASR in any practical application, high robustness to non-stationary speaker and environmental factors is required. While many noise-robust ASR techniques have been shown to meet the demands of specific applications (e.g., mobile communication), they often fail in more complex scenarios such as in the presence of room reverberation.

Recently, conventional Gaussian mixture model (GMM) and hidden Markov model (HMM)-based ASR systems have been superseded by hybrid multilayer-perceptron (MLP)-HMM systems [1], often referred to as deep neural network (DNN) systems. Despite all the successes obtained with DNNs, attributed to their ability to learn from large amounts of potentially noisy data, investigations have shown DNN systems can be quite sensitive to mismatched environments. For instance in [2], it was shown that even with state-of-the-art DNN systems, front-end processing is helpful in increasing ASR performance in mismatched conditions.

Previous studies have attempted to counteract the convolutional distortion caused by reverberation using a number of denoising methods, such as frequency domain linear prediction [3], modulation filtered spectrograms [4], or missing-data mask estimation designed for dereverberation [5]. All of these approaches make weak assumptions about the reverberant data (e.g., they do not require that the room impulse response is known) but they achieve only a moderate increase in ASR performance. More recent techniques include MLP-based feature enhancement systems; for example, a deep recurrent neural network (RNN) approach for log-spectral domain feature enhancement was recently proposed in [6] and applied to dereverberation. Similarly, an RNN exploiting long-range temporal context by using memory cells in the hidden units was applied to dereverberation in [7]. A further example is the reverberation modeling for speech recognition (REMOS) framework [8], which combines a clean speech model with a reverberation model to determine clean speech and reverberation estimates during recognition via a modified Viterbi algorithm. In conditions with relatively long reverberation times, REMOS provides higher recognition accuracy than a matched model.

*Correspondence: sami.keronen@aalto.fi
[1]Department of Signal Processing and Acoustics, Aalto university, P.O. Box 13000, 00076 Aalto, Finland
Full list of author information is available at the end of the article

This article focuses on one of the most powerful approaches for denoising of recent years—non-negative matrix factorization (NMF)—which models the speech spectrogram as a sparse non-negative linear combination of dictionary elements ("speech atoms"). NMF was formulated in [9] to decompose multivariate data and has been the basis of several sound source separation [10, 11] and denoising [12] systems. Noise robust ASR systems based on NMF were introduced in [13], using either feature enhancement or hybrid HMM decoding with so-called sparse classification. An alternative formulation of NMF, non-negative matrix factor deconvolution (NMFD), was introduced in [14] to take better advantage of temporal information. NMFD lends itself naturally to dereverberation; [15, 16] describe methods for blind dereverberation by decomposing the reverberated spectrum into a clean spectrum convolved with a filter, while constraining the properties of the speech spectrum.

Our previous work, published in two papers in the REVERB'14 workshop [17, 18], described two dereverberation techniques that are combined and extended in the current study. In the first paper [17], a technique was described for speech dereverberation that draws on the fundamental idea of NMF, in that it models speech as a linear combination of dictionary elements. However, the NMF-based approach was extended to incorporate a filter in the Mel-spectral domain that could be optimized for arbitrary convolutions. Furthermore, [17] used missing-data mask imputed (MDI) [19, 20] spectrograms to produce the initial estimate of the sparse representation of the clean speech signal, giving more effective dereverberation. Our second REVERB'14 workshop paper proposed a distribution matching (DM) scheme for unsupervised dereverberation of speech features [18]. This utilizes stacked and decorrelated spectro-temporal vectors containing a long time context. In the decorrelated transformation domain, the distributions of reverberant supervectors are equalized to match the a priori clean speech distribution by applying a non-parametric histogram equalization-based approach [21].

Bringing the ideas in our two workshop papers together, the current paper proposes a novel dereverberation feature enhancement method in a noise-robust ASR framework by combining the NMF and DM methods — a combination that was not tested in either of the workshop papers. More specifically, we present a single-channel source separation technique which extracts the speech signal from the observed mixture of speech and noise signals and train the ASR back-end with the enhanced (dereverberated) features to increase the recognizer tolerance for artifacts generated in denoising. Our previous work [17, 18] shows that DM outperforms MDI as a feature enhancement strategy. This brings us to the goal of the present study: to investigate whether the performance advantage of DM translates into better initial estimates of the sparse representation of the dereverberated speech features, compared to that obtained with MDI. The proposed method is evaluated on the reverberant 2014 REVERB Challenge data set [22] and shown to provide equal or higher ASR performance than three existing state-of-the-art feature enhancement methods, using similar back-end processing provided by the Kaldi toolkit [23]. Among the methods compared against our new approach, we include the RNN-based feature enhancement, a feature enhancement based on blind reverberation time estimation, and our previous system which used MDI to produce the initial clean speech estimate.

The remainder of the paper is structured as follows. Section 2 gives an overview of the proposed two-stage feature enhancement process. Sections 3 and 4 define the DM- and MDI-based initializations, used to estimate the initial sparse representation of the clean speech signal for NMF. Section 5 describes the procedure for non-negative matrix factorization of reverberant speech. Section 6 gives an overview of the experimental setup including the data set, the ASR system, parameter optimization, and brief descriptions of an additional multichannel feature enhancement and the computational requirements of the two-stage feature enhancement. Our results are presented and discussed in Sections 7 and 8, and conclusions from the study are presented in Section 9.

2 Overview of the dereverberation process

The flowchart of the dereverberation process and the overall ASR system is shown in Fig. 1. First, the speech signal is pre-processed (denoted by *1. Pre-processing*) into frames of Mel-scale filterbank energies, which are used as an input to the NMF part of the feature enhancement (denoted by *2. Feature enhancement*) for dereverberation. Conventional NMF feature enhancement would then be initialized with the previously described reverberant speech, but our implementation divides the feature enhancement into two stages: in the first stage, we construct an initial estimate of the non-reverberant speech that is used to initialize the NMF algorithm in the second stage. The factorization algorithm is initialized either with DM (described in Section 3) or MDI (briefly described in Section 4) dereverberated speech. The ASR back-end (denoted by *3. Back-end* in the figure) consists of either GMM- or DNN-based acoustic modeling of enhanced and transformed features and an HMM-based decoder.

3 Distribution matching initialization

The goal of the distribution matching (DM) method [18] is to recover the clean speech spectra $\mathbf{x}$ from the observed reverberant speech spectra $\mathbf{y}$ when the clean speech prior distribution $p(\mathbf{x})$ is assumed to be known and the distribution of the observed reverberant speech $p(\mathbf{y})$ can be

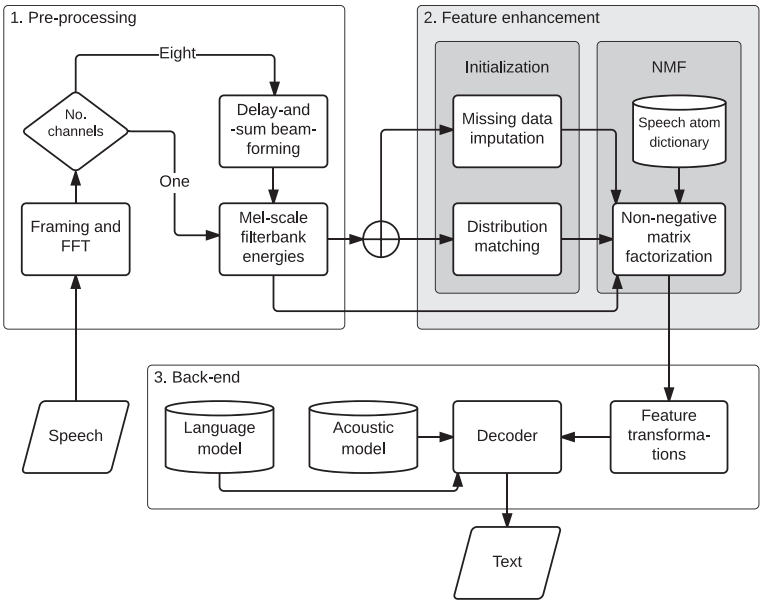

Fig. 1 General flowchart of the feature enhancement method and back-end ASR system. Overview of the speech signal processing pathway from time domain audio to decoding and recognition hypothesis

estimated during recognition. As our goal is to counteract the effects of reverberation, it is important to take into account the long time span in the effects of reverberation. In the following, we develop a method to map the distribution of reverberant speech observations to the clean speech prior. The DM method is also illustrated in Fig. 2. The method uses long time contexts decorrelated by linear transformations, after which a histogram equalization (HEQ) mapping can be utilized using one-dimensional distribution samples. HEQ was originally proposed for image processing [24] but subsequently has also been utilized in ASR to counteract noise and speaker changes over short temporal windows [21]. With a longer temporal context, as in the present study, HEQ has been used for feature space Gaussianization [25] to obtain a feature space that is easier to model with GMMs.

DM utilizes *three* steps that are applied in two iterations. *The first* step of the method is to find a signal representation that has a sufficiently long time context to counteract the effects of reverberation. Assuming that the effects of reverberation are linear and convolutive with the speech signal, we can represent them in the feature domain as linear transformations. Our basis features are C-dimensional Mel-spectral feature vectors of observed speech $\mathbf{y}$ that have been normalized to compensate for spectral distortion. The normalization is performed by estimating the reverberation-free spectral peaks to compute the normalization coefficients [5]. In the first iteration round, the

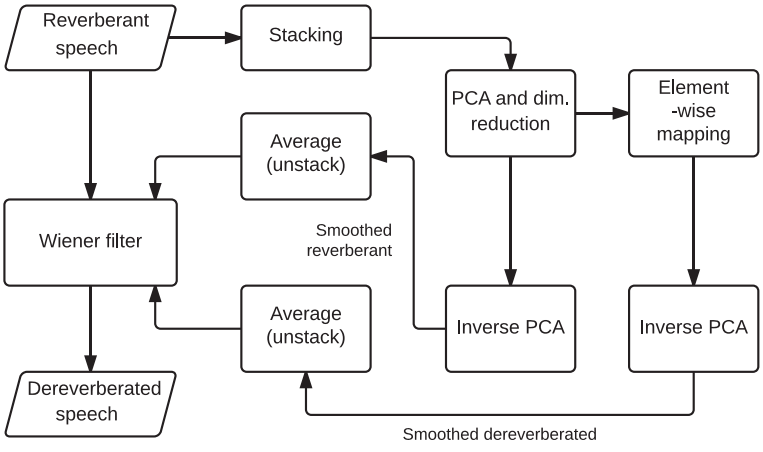

Fig. 2 Distribution matching (DM) flowchart. An overview of the DM dereverberation process. Here, the reverberant speech input and dereverberated speech output are represented as Mel-spectral feature vectors

observed speech $\mathbf{y}$ corresponds to original reverberated speech, whereas in the second iteration round, we use the dereverberated estimate as the observation $\mathbf{y} = \hat{\mathbf{x}}$. To take into account the duration of reverberation, the Mel-spectral observations are stacked over T consecutive frames to form CT-dimensional supervectors

$$\boldsymbol{y}_t = \left[\mathbf{y}_t^\top \ \cdots \ \mathbf{y}_{t+T-1}^\top\right]^\top, \tag{1}$$

where T is chosen large enough compared to the duration of the room impulse responses (RIRs) and $\top$ indicates transpose. Consequently, the speech features $\boldsymbol{y}$ affected by convolution can be formulated as $\boldsymbol{y} \approx \boldsymbol{H}\boldsymbol{x}$, where $\boldsymbol{H}$ is a filter matrix that performs convolution on the supervector $\boldsymbol{x}$ constructed from clean speech features.

The *second* step is to find a transformed feature domain that allows the use of one-dimensional mapping functions from the observed feature distribution to the non-reverberant target distribution. The supervector-based feature vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are highly correlated along the feature dimension because each vector includes spectral and temporal context, which introduces problems. In order to map such highly correlated features from the observed to the non-reverberant distribution, a complex multivariate mapping would be needed. However, the problem can be simplified by applying a decorrelating linear transformation to the spectral-temporal supervectors, after which it is possible to perform one-dimensional mappings. In this study, the applied transformation $\boldsymbol{D}$ is based on principal component analysis (PCA) to decorrelate the elements of the speech feature supervectors on a log-scale,

$$\boldsymbol{g}'_{\boldsymbol{y}} = \boldsymbol{D}\log\boldsymbol{y} \approx \boldsymbol{D}\log\boldsymbol{H}\boldsymbol{x}, \tag{2}$$

where $\boldsymbol{y}$ corresponds to reverberant speech in the first iteration and to the dereverberated speech estimate $\boldsymbol{y} = \hat{\boldsymbol{x}}$ in the second iteration. The quantity $\boldsymbol{g}'_{\boldsymbol{y}}$ denotes the observed speech supervector features in the decorrelated feature space, and the log operation is computed elementwise. The number of retained low-order principal components M of $\boldsymbol{D}$ can be treated as a tunable free parameter to obtain a more or less smoothed representation.

The *third* step is to develop the one-dimensional mapping functions that can be applied elementwise in the decorrelated feature domain. First, we make an assumption that the transformation $\boldsymbol{D}$ that decorrelates the non-reverberant speech supervectors $\boldsymbol{x}$ in the estimation of clean speech prior distribution also decorrelates all the observed speech supervectors $\boldsymbol{y}$ regardless of the extent of reverberation. Then, we can formulate one-dimensional elementwise bijective (one-to-one) mappings $F_{yx}^{(m)}$ from PCA-transformed reverberant supervector elements $g'_y(m)$ to dereverberated ones $\tilde{g}'_x(m)$ as follows

$$\tilde{g}'_x(m) = F_{yx}^{(m)}\left(g'_y(m)\right), \tag{3}$$

where m indexes the mapping for each feature element. As the PCA-transformed supervectors $\boldsymbol{g}'_{\boldsymbol{y}}$ represent sufficient temporal context relative to reverberation effects, it is possible to find effective mappings from reverberant speech to clean speech (see [18]).

In this work, functions for $F_{xy}^{(m)}$ are obtained by mapping the distribution of observed speech to match the distribution of the clean speech prior. In the first iteration step, we use the original reverberant speech as observations, and in the second step, we use the dereverberated estimate from the first iteration round. The mapping is easy to find if the distributions of clean and observed speech are represented by inverse cumulative distribution functions (ICDF) [21, 25]. In general, the empirical ICDF Φ_y^{-1} can be obtained simply by scaling and sorting the data samples. In our case, however, we omit the scaling as the data has already been equalized for spectral deviation. From now on, we simplify the notation and operate on individual components of the decorrelated supervectors by dropping all indices m. The mapping function F_{yx} from reverberant speech ICDF Φ_y^{-1} to clean speech ICDF Φ_x^{-1} is implemented by constructing a lookup table $\Phi_y^{-1} \underset{F}{\rightarrow} \Phi_x^{-1}$ with piecewise cubic Hermite interpolation (Section 3.3. in [26]). When applied in practice, the lookup table needs to be updated to reflect the current reverberation condition encountered during recognition. Assuming that reverberation conditions change slowly, a sample of reverberant data is collected during recognition to model reverberant distribution Φ_y^{-1}, which is the mapping input data distribution. While the input data distribution needs updating, the mapping target distribution Φ_x^{-1} is always represented using the same static clean speech sample. In the present study, the mapping input distribution is updated during recognition passes by using batches of development or test-set data. Each batch corresponds to a static reverberation condition in the REVERB Challenge data, described in Section 6.1.

We can now produce the estimate of the dereverberated log-spectral supervector $\tilde{\boldsymbol{x}}'$ as

$$\tilde{\boldsymbol{x}}' = \boldsymbol{D}^{-1}F_{yx}\left(\boldsymbol{g}'_{\boldsymbol{y}}\right), \tag{4}$$

where the mapping F_{yx} is realized using separate lookup tables $F_{yx}^{(m)}$ for each element m of $\boldsymbol{g}'_{\boldsymbol{y}}$ and $\boldsymbol{D}^{-1}$ is the inverse PCA transformation. Then, supervectors $\tilde{\boldsymbol{x}}'$ are unstacked to the linear Mel-spectral domain $\tilde{\mathbf{x}}$ with one frame time context using overlap adding, so that regions in adjacent supervectors containing Mel-spectra of the same time frame are averaged. Thus, linear Mel-spectral vectors $\tilde{\mathbf{x}}$ are obtained as

$$\tilde{\mathbf{x}} = \exp\left(1/T\sum_{t=1}^{T}\tilde{\boldsymbol{x}}'_{T-t}(\psi)\right), \tag{5}$$

where t indexes both the frames of adjacent supervectors and also the component Mel-spectral vectors within the range $\psi = [(t-1)C+1, \ldots, tC]$ in each supervector.

However, the dereverberated feature estimates $\tilde{\mathbf{x}}$ in this form are smoothed by the PCA and averaging operators. Therefore, we apply a Wiener filter to reintroduce some short-term variation that was present in the original reverberant observations $\mathbf{y}$ but was removed by the smoothing. For the Wiener filter, we also need a version of the reverberated features $\tilde{\mathbf{y}}$ that were smoothed by the same PCA transformation $\boldsymbol{D}$. The Wiener-filtered feature estimate $\hat{\mathbf{x}}$ is given by

$$\hat{\mathbf{x}} = \tilde{\mathbf{x}} \,./ \tilde{\mathbf{y}} \,.* \mathbf{y} \tag{6}$$

where ./ denotes elementwise division and .∗ elementwise multiplication. The importance of Wiener filtering is demonstrated in our previous work [18].

After progressing through the above three steps (Eqs. (1)–(6)) in the first iteration, the reverberant observation $\mathbf{y}$ is substituted with the current estimate $\hat{\mathbf{x}}$. After the second iteration, we obtain the estimates $\hat{\mathbf{x}}$ that are used either directly as enhanced features or as initialization estimates for the NMF processing.

4 Missing data imputation initialization

The missing data imputation method used here utilizes the bounded conditional mean imputation (BCMI) as proposed in [19]. The method uses a GMM model to capture the clean speech statistics for reconstructing the unreliable noisy regions of the observed speech spectrum. Here, we denote the noise-free reliable part of the speech spectrum by $\mathbf{x}_\mathrm{r}$ and the noisy unreliable part by $\mathbf{x}_\mathrm{u}$. The BCMI produces the clean speech estimate $\hat{\mathbf{x}}_\mathrm{u}$ using the conditional distribution $p(\mathbf{x}_\mathrm{u} \mid \mathbf{x}_\mathrm{r})$ with an assumption that the observed noisy speech $\mathbf{x}_\mathrm{u}$ acts as the upper bound for the underlying clean speech.

For estimating the missing data mask that specifies the reliable and unreliable regions, we use the approach proposed in [27]. The method uses a modulation band-pass filter along the time trajectory, tuned to the speech syllable rate. The filter emphasizes reverberation-free speech onsets so that they can be distinguished from reverberant segments of speech. Regions which are emphasized by the filter are labeled reliable, while regions that are de-emphasized are labeled as unreliable.

5 Non-negative matrix factorization of reverberant speech

Methods based on the non-negative matrix factorization (NMF) framework have been widely used for various speech processing tasks. A typical application of NMF is noisy speech feature enhancement via supervised source separation [13]. Given a pre-set *dictionary* of fixed-size magnitude spectrogram atoms of both speech and noise, an observed spectrogram is modeled by their non-negative linear combination. The individual reconstructions of both clean speech and noise spectrograms are based on estimating the corresponding dictionary atoms and their coefficients in the NMF representation. To account for observations of arbitrary length, the processing can be performed in (overlapping) windows of a fixed length of T frames.

In this work, we consider reverberant but relatively noise-free speech. Hence, we do not make use of the noise dictionary but still build on the same underlying speech model. We denote by $\mathbf{Y}$ the observed speech, represented by a $TC \times N$ matrix. Each column of $\mathbf{Y}$ is a collection of T frames of a C-dimensional Mel-scale spectrogram, stacked into a single vector. Under the NMF model, we have the approximation

$$\mathbf{Y} \approx \mathbf{SA}, \tag{7}$$

where $\mathbf{S}$ is a $TC \times K$ dictionary matrix of K spectrograms of clean speech, while $\mathbf{A}$ is the $K \times N$ activation matrix holding the linear combination coefficients.

The effect of reverberation extends across frame boundaries in the Mel-spectrogram domain. This can be approximated by a convolution of the samples of each frequency channel with a channel-specific T_f-sample filter. Using the stacked vector representation of the T-frame windows, the model of Eq. (7) can be extended to perform this convolution within each window. The resulting approximation is

$$\mathbf{Y} \approx \mathbf{RSA}, \tag{8}$$

where, denoting $T_r = T + T_f - 1$, $\mathbf{R}$ is a $T_r C \times TC$ matrix of the form

$$\mathbf{R} = \begin{pmatrix} r_{1,1} & 0 & 0 & & & & & \\ 0 & r_{1,2} & 0 & \cdots & & & & \\ 0 & 0 & r_{1,3} & & & & & \\ & \vdots & & \ddots & & & & \\ r_{2,1} & 0 & 0 & & r_{1,1} & 0 & 0 & \\ 0 & r_{2,2} & 0 & \cdots & 0 & r_{1,2} & 0 & \cdots \\ 0 & 0 & r_{2,3} & & 0 & 0 & r_{1,3} & \\ & \vdots & & \ddots & & \vdots & & \ddots \\ \underbrace{\qquad\qquad}_{C} & & & & & & & \end{pmatrix}. \tag{9}$$

The diagonal structure of $\mathbf{R}$ is designed so that a left multiplication of a stacked window vector of T frames results in the discrete convolution of the filter $\begin{bmatrix} r_{1,c} & r_{2,c} & \cdots & r_{T_f,c} \end{bmatrix}$ and the samples of the frequency channel c in that window. It is worth noting that Eq. (8) can be interpreted as either reverberating the clean speech estimate, $\mathbf{R}(\mathbf{SA})$, or making a linear combination of reverberated speech atoms, $(\mathbf{RS})\mathbf{A}$.

5.1 Optimization of the filter and activation matrices

Following the supervised NMF model, the dictionary matrix **S** is held constant. In the sliding window model, the values of the filter and activation matrices **R** and **A** are obtained independently for each window t. Denoting by $\mathbf{Y}_t$ and $\mathbf{A}_t$ the corresponding columns of **Y** and **A**, the filter and activation matrices are set to minimize

$$\sum_t \left(d(\mathbf{Y}_t, \mathbf{RSA}_t) + \lambda \left\| \mathbf{A}_t \right\|_1 \right), \tag{10}$$

where the $d(\mathbf{Y}_t, \mathbf{RSA}_t)$ term is a distance measure between the observation and the NMF approximation. The second term, which consists of the L^1 norm $\|\cdot\|$ of the activation weights multiplied by the sparsity coefficient λ, is intended to induce sparsity in **A** and thereby yield a sparse representation of the observation. In this work, the generalized Kullback-Leibler divergence is used for d.

The form of Eq. (10) admits the use of conventional iterative NMF optimization algorithms [9, 13] to perform multiplicative updates to both the **R** and **A** matrices. However, the optimization problem is not convex, and a simple scheme of alternately updating **R** and **A** did not yield results useful for dereverberation in earlier experiments [17]. The reasons behind this are hypothesized in Section 8. Accordingly, we use the following series of steps to obtain the factorization **RSA**:

1. A simpler dereverberation method is used to obtain an initial estimate of the non-reverberant speech of the observation, denoted by $\bar{\mathbf{X}}$. In this work, the estimate is obtained either through DM or MDI initialization, described in Sections 3 and 4, respectively.
2. The activation matrix **A** is initialized to all ones and iteratively updated for I_1 rounds to perform the factorization $\bar{\mathbf{X}} \approx \mathbf{SA}$.
3. While the dictionary atoms of **S** are strictly clean speech, the initial estimate $\bar{\mathbf{X}}$ is never perfectly dereverberated. Consequently, the activations **A** resulting from the preceding step will reflect the effects of reverberation, typically characterized by sequences of consecutive non-zero activations of the same dictionary atom. We therefore filter the time sequences of activations for each atom using a filter $H_A(z)$ and clamp the result to be non-negative. This filtering step has the effect of biasing the following estimation of **R** to emphasize the reverberation.
4. The filter matrix **R** is initialized to hold the constant T_f-sample filter $\frac{1}{T_f}\left[1 \cdots 1\right]$ for each frequency band. While keeping the **A** matrix fixed, **R** is iteratively updated for I_2 rounds to minimize the cost in the approximation $\mathbf{Y} \approx \mathbf{RSA}$. However, the multiplicative updates are neither guaranteed to preserve the filter structure described in Eq. (9), except for the zero elements, nor to result in a realizable filter. To enforce these properties, **R** is processed to have the form of Eq. (9) after each iteration: The new values of the filter coefficients $r_{t,c}$ are obtained by averaging over all their occurrences in the updated **R**, and clamping large values to satisfy $\forall t : r_{t+1,c} \leq r_{t,c}$. The coefficients are also uniformly scaled to $\sum_{t,c} r_{t,c} = C$.
5. As a final step, the **R** matrix is kept fixed, and the **A** matrix is iteratively updated for I_3 rounds based on $\mathbf{Y} \approx \mathbf{RSA}$.

To demonstrate the behavior of the algorithm described above, Fig. 3 illustrates the cost function of Eq. (10) as a function of the update iterations. All three iterative stages of the algorithm are shown: $I_1 = 50$ iterations of updating activations **A** based on the initial estimate $\bar{\mathbf{X}}$ in step 2, $I_2 = 50$ iterations of updating the filter matrix **R** in step 4, and finally $I_3 = 75$ further iterations to obtain the final values of **A** in step 5. The activation filtering in step 3 is reflected by a discontinuity in the cost function between steps 2 and 4. Note that the plotted cost function is based on the reverberant observation **Y**, which is not directly used as the optimization target in step 2. The cost function also measures only the accuracy of the reconstruction **RSA** and the sparsity of **A** and therefore does not indicate the dereverberation strength, which depends primarily on the filter represented by **R**.

A major drawback of this simple sliding window scheme in reverberant conditions occurs when the start of a window coincides with a silent interval in the underlying speech signal. In this case, the early frames of the window are dominated by observed reflections. When such a window is represented using a dictionary of individually reverberated atoms, the energy in the early frames is interpreted as direct sound and not properly attenuated.

To alleviate this issue, we use the NMFD [14] model, so that an individually reverberated dictionary atom activated in one window can "explain away" the energy of its reflections in succeeding overlapping windows. For the stacked vector representation, a computationally efficient implementation of the NMFD optimization scheme can be formulated by modifying the multiplicative update rule for the activation matrix **A** used in the iterative steps 2, 4, and 5 of the above algorithm.

For conventional NMF processing, the multiplicative update of matrix **A** that corresponds to the cost function given in Eq. (10) is defined as [13]

$$\mathbf{A} \leftarrow \mathbf{A} .* \frac{(\mathbf{RS})^\top \frac{\mathbf{Y}}{\mathbf{RSA}}}{(\mathbf{RS})^\top \mathbf{1} + \lambda}, \tag{11}$$

where $.*$ denotes elementwise multiplication, the division of two matrix operands is likewise performed elementwise, and $\mathbf{1}$ is a $T_r C \times N$ all-one matrix. We introduce the

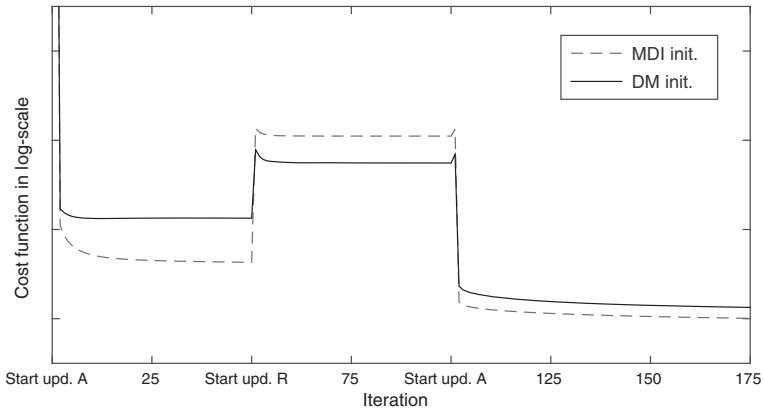

Fig. 3 The cost function of NMF parameter optimization in log-scale. The solid and dashed lines illustrate the NMF optimization with DM and MDI initializations, respectively. The cost function is computed as the average of ten randomly selected far recorded SimData utterances. Here, the activation matrix **A** is updated for the first 50 iterations, the reverberation matrix **R** is updated for the next 50 iterations, and finally the activation matrix **A** is updated again

dependencies between consecutive windows by adjusting the $\frac{\mathbf{Y}}{\mathbf{RSA}}$ term, so that the new update rule is

$$\mathbf{A} \leftarrow \mathbf{A} .* \frac{(\mathbf{RS})^\top s\left(\frac{\mathbf{y}}{o(\mathbf{RSA})}\right)}{(\mathbf{RS})^\top \mathbf{1} + \lambda}, \tag{12}$$

where $\mathbf{y}$ is the original, non-stacked observation spectrogram. In the update rule, the $o(\mathbf{Z})$ function denotes the result of overlap-adding the stacked vectors of matrix $\mathbf{Z}$ to a single spectrogram (in the same way as in Eq. (5)), while the $s(\mathbf{z})$ function denotes the conversion of spectrogram $\mathbf{z}$ to the stacked form. The corresponding change is also made to the update rule of the $\mathbf{R}$ matrix,

$$\mathbf{R} \leftarrow \mathbf{R} .* \frac{s\left(\frac{\mathbf{y}}{o(\mathbf{RSA})}\right)(\mathbf{SA})^\top}{\mathbf{1}(\mathbf{SA})^\top}. \tag{13}$$

5.2 NMF-based feature enhancement of reverberant speech

Based on the factorization, we can directly reconstruct the reverberant observation as $\tilde{\mathbf{Y}} = \mathbf{RSA}$ and the underlying clean speech as $\tilde{\mathbf{X}} = \mathbf{SA}$. By overlap-adding the stacked vectors, we obtain the corresponding Mel-scale spectrogram estimates $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{x}}$. While $\tilde{\mathbf{x}}$ could be used directly as input for a speech recognition system, in existing work on NMF-based source separation for speech in additive noise [13], better performance was obtained by using the same Wiener-filtering approach we have described for the DM-based initialization. Therefore, we compute the final enhanced features, as in the DM method, by filtering the original observation with the time-varying Mel-spectral filter defined as $\tilde{\mathbf{x}} ./ \tilde{\mathbf{y}}$, where $./$ denotes elementwise division.

The full NMF-based feature enhancement algorithm is provided in pseudo-code form in Algorithm 1.

Algorithm 1: The NMF-based feature dereverberation algorithm of Section 5.

Input : a reverberant Mel-scale spectrogram $\mathbf{y}$ a clean speech dictionary matrix $\mathbf{S}$ iteration limit parameters I_1, I_2, I_3 for the NMF factorization
Output: A dereverberated Mel-scale spectrogram estimate $\hat{\mathbf{x}}$

$\bar{\mathbf{x}} \leftarrow$ initial dereverberation estimate from DM or MDI processing of $\mathbf{y}$;
$\mathbf{Y} \leftarrow$ windowed stacked-vector representation of $\mathbf{y}$;
$\bar{\mathbf{X}} \leftarrow$ windowed stacked-vector representation of $\bar{\mathbf{x}}$;
perform the NMF factorization
    see Section 5.1 for a detailed description of the factorization algorithm;
    $\mathbf{A} \leftarrow$ an all-ones matrix;
    for $i \leftarrow 1$ **to** I_1 **do**
        update $\mathbf{A}$ to approximate $\bar{\mathbf{X}} \approx \mathbf{SA}$;
    end for
    filter each row of $\mathbf{A}$ with the activation filter $H_A(z)$;
    for $i \leftarrow 1$ **to** I_2 **do**
        update $\mathbf{R}$ to approximate $\mathbf{Y} \approx \mathbf{RSA}$;
        compute filter coefficients $r_{t,c}$ by averaging their occurrences in $\mathbf{R}$;
        clamp $r_{t+1,c} \leq r_{t,c}$ and scale $\sum_{t,c} r_{t,c} = C$;
        reconstruct the filter matrix $\mathbf{R}$;
    end for
    for $i \leftarrow 1$ **to** I_3 **do**
        update $\mathbf{A}$ to approximate $\mathbf{Y} \approx \mathbf{RSA}$;
    end for
end
$\tilde{\mathbf{y}} \leftarrow$ overlap-add the stacked vectors of $\mathbf{RSA}$;
$\tilde{\mathbf{x}} \leftarrow$ overlap-add the stacked vectors of $\mathbf{SA}$;
$\hat{\mathbf{x}} \leftarrow (\tilde{\mathbf{x}} ./ \tilde{\mathbf{y}}) .* \mathbf{y}$;

6 Experimental setup

6.1 Data set

The proposed feature enhancement method presented in the paper is evaluated on the 2014 REVERB Challenge data set [22]. The data set is only briefly described here. The first part of the data set, denoted by SimData, consists of an artificially reverberated British English version of the 5000-word Wall Street Journal corpus [28] mixed with recordings of background noise at a fixed signal-to-noise ratio (SNR) of 20 dB. SimData contains far and near microphone positions in three rooms of different size for a total of six recording scenarios. The second part of the REVERB Challenge data set contains real recordings, denoted by RealData, extracted from the multichannel Wall Street Journal audio visual corpus. The utterances of RealData have been recorded in a reverberant office space with background noise originating mostly from the air conditioning [29]. A summary of the SimData and RealData recording conditions is presented in the upper part of Table 1.

The data set is divided into speaker-exclusive training (clean speech), development, and evaluation sets. The RIRs are also different in the development and evaluation sets. The durations and the numbers of speakers and utterances of the sets are shown in the lower part of Table 1. In addition to the clean speech training set, an equal-sized multicondition (MC) training set is provided. The MC training data is artificially corrupted in the same manner as SimData but with unique impulse responses.

All the reverberant utterances in the REVERB Challenge data set are provided as single-channel, 2-channel, and 8-channel recordings. However, experiments in this study use either the single-channel setup, which is the main part of the study, or the 8-channel system in an additional experiment. The 8-channel system is constructed by applying a frequency domain delay-and-sum (DS) beamformer prior to the feature enhancement to investigate whether multichannel setups gain from the proposed method. The DS beamforming is briefly described in Section 6.4.

Table 1 Summary of recording conditions and data set parameters. SimData denotes artificially reverberated speech data with real RIRs and RealData denotes true recordings made in a reverberant room

Recording type	Room size	T_{60} (s)	Near mic. distance (m)	Far mic. distance (m)
SimData	Small	0.25	0.5	2.0
SimData	Medium	0.5	0.5	2.0
SimData	Large	0.7	0.5	2.0
RealData	Large	0.7	1.0	2.5
Data set	Recording type	Number of speakers	Utterances	Duration (h)
Training	Clean speech	92	7961	17.5
MC training	Similar to SimData	92	7961	17.5
Development	SimData	20	1484	3.3
	RealData	5	179	0.3
Evaluation	SimData	20	2176	4.8
	RealData	10	372	0.6

6.2 ASR system

In total, six feature enhancement, or front-end processing, combinations are applied in the evaluation; DM alone, NMF alone, DM-initialized NMF (denoted by *DM+NMF*), and MDI-initialized NMF (denoted by *MDI+NMF*). Moreover, the DM+NMF and MDI+NMF enhancements are combined with the additional DS beamformer in order to recognize the 8-channel audio. All systems with feature enhancements are trained on the MC training set.

The ASR back-end processing is performed using the publicly available Kaldi recognition toolkit [23] and the system utilized here is based on REVERB scripts provided in the toolkit. The use of Kaldi allows us to obtain results that are competitive with the state-of-the-art and also allows direct comparison with other studies that are based on the Kaldi back-end such as [6, 7, 30].

Two hybrid DNN-HMM and four GMM-HMM back-end systems of increasing acoustic model complexity are trained. The first back-end system, denoted by *LDA+MLLT*, is a triphone-based recognizer which uses feature vectors constructed from the first 13 of 23 Mel-frequency cepstral coefficients (MFCCs) drawn from nine consecutive frames. The feature vector dimensionality is reduced to 40 by linear discriminant analysis (LDA). Furthermore, a maximum likelihood linear transform (MLLT) is applied to improve the separability of acoustic classes in the feature space. The LDA+MLLT system is trained with the MC training set, but a similar system is also trained with the clean speech data for reference.

The second back-end system, denoted by *LDA+MLLT+SAT*, supplements the LDA+MLLT system with utterance-based speaker adaptive training (SAT). This is based on a variant of feature domain-constrained maximum likelihood linear regression (fMLLR) [31] designed for rapid adaptation on very small amounts of adaptation data.

The third back-end system, denoted by *LDA+MLLT+SAT+f-bMMI*, uses the acoustic model of the LDA+MLLT+SAT back-end to execute feature-space boosted maximum mutual information (f-bMMI) -based discriminative training [32]. The LDA+MLLT+SAT+f-bMMI is trained to obtain fully comparable single and 8-channel results with the feature enhancement proposed

in [30] and comparable 8-channel results with [6]. In the experiments, we set the boost factor to 0.1.

The fourth back-end system, denoted by LDA+MLLT+SA+bMMI+MBR, is based on the LDA+MLLT system and supplements it with utterance-based fMLLR speaker adaptation, boosted MMI (bMMI), and minimum Bayes risk (MBR) decoding [33]. The LDA+MLLT+SA+bMMI+MBR system is trained to obtain fully comparable results with the feature enhancement proposed in [6].

The fifth back-end is a hybrid DNN-HMM system, denoted by *LDA+MLLT+SAT+DNN*, trained with the adapted features of the LDA+MLLT+SAT back-end using a frame-based cross-entropy criterion and p-norm nonlinearities [34]. The DNNs consisted of 4 hidden layers and approximately 6.3 million parameters. The sixth back-end, denoted by *LDA+MLLT+SAT+DNN+SMBR*, supplements the LDA+MLLT+SAT+DNN back-end with state-level minimum Bayes risk (SMBR) criterion-based discriminative training [35] to obtain comparable results with the feature enhancement proposed in [30]. The SMBR training is applied only to the best performing LDA+MLLT+SAT+DNN back-end in the development set.

For the language model (LM), we use the 5000-word trigram model provided in the WSJ corpus. The LM weights are optimized separately for each back-end and for each feature enhancement combination, based on the averaged recognition word error rate (WER) over all eight test conditions in the development set. The optimized LM weights are also used in the estimation of fMLLR transformations for the first-pass recognition hypotheses.

6.3 Parameter setup

The parameter setups of the DM, MDI, and NMF methods use the same values as the best performing systems in the experiments of our previous studies [17, 18]. The settings are briefly summarized here. Mel-spectral features of $T = 20$ subsequent frames were collected for each DM supervector. The PCA transformation in Eq. (2) was estimated from 1000 randomly selected clean-speech training set utterances and applied to reduce the supervector dimensionality to $M = 40$ principal components. We have also conducted unpublished experiments utilizing both clean and reverberant data in the PCA training, which yielded slightly inferior ASR results compared to using only clean training data. The reasons behind this may be that is difficult to learn a transform that simultaneously decorrelates both clean speech and speech reverberated with a range of reverberation times, and it may be more important to decorrelate the target rather than the source domain prior to the mapping.

In the ASR experiments, the distribution mapping is applied in two iterations (see Section 3). The mapping function was updated every time that reverberation conditions changed and the ICDF Φ_y^{-1} of observations were collected from the full batch of utterances in each test condition. For the clean speech prior, we used a collection of random samples from the clean speech training set whose length was equal to that of the observation sample. Collectively there are three tunable paramaters in the DM initialization method (PCA-dimension M, stack dimension T and number of iterations).

Regarding the MDI system, the mask estimation stage requires three free parameters that were chosen to be the same ($\alpha = 19$, $\beta = 0.43$, and $\gamma = 1.4$) as in our earlier studies [17, 27]. In the imputation stage, we also utilized the same GMM-model as in our previous study; a 5-component GMM trained on a random 1000 utterance subset of the clean speech training set with a time context of three consecutive Mel-spectral feature frames. Taking together the parameters in the mask estimation as well as imputation stage totals to five tunable parameters.

For the NMF window length, we chose $T = 10$ frames, which offered a good balance between dictionary complexity and ASR performance. The length of the NMF $\mathbf{R}$ matrix initialization filter that functions as an upper bound on the reverberation time the update algorithm can handle was set to $T_f = 20$ samples to accommodate normal-sized rooms. The sparsity coefficient and iteration counts were set as follows: $\lambda = 1$, $I_1 = I_2 = 50$, and $I_3 = 100$. The clean speech dictionary consisting of $K = 7\,681$ atoms was constructed by selecting one random T-frame segment from each clean speech training set utterance. The filter in step 3 of the update algorithm was optimized to give the NMF feature enhancement low average WER on all reverberation conditions and therefore it is not optimal for all the separate conditions. Based on multiple small-scale experiments, the filter was selected as $H_A(z) = 1 - 0.9z^{-1} - 0.8z^{-2} - 0.7z^{-3}$. From dozens of candidates, the selected filter was the only one to work well on all reverberation conditions.

6.4 Delay-and-sum beamforming

For the delay-and-sum (DS) beamforming feature enhancement, we use the implementation of [36]. To describe DS beamforming in brief, it selects one of the channels as the reference signal and the differences between the arrival times of the reference and the other channel signals are estimated by generalized cross-correlation with a phase transformation [37]. By delaying the other channels by their estimated arrival times and summing all the signals, the coherent sound sources are amplified and the SNR of the speech signal is increased. In this work, DS is applied to the 8-channel data on the LDA+MLLT+SAT+f-bMMI back-end.

6.5 Computational requirements

The overall real-time factor for both DM+NMF and MDI+NMF feature enhancements is approximately 6.9 on one thread of an Intel Xeon E3-1230V2 processor. There is no significant difference between the computational costs of the DM and MDI initialization methods, and the real-time factors for both methods are less than one. In fact, the NMF enhancement is the most computationally demanding processing stage of the whole ASR system. Since both initialization methods also utilize the same amount of training data, the benefit of the DM method over MDI is that there are only three free parameters to tune instead of five. During recognition, the DM method operates in full batch mode, whereas MDI works on an utterance-by-utterance basis.

7 Results

The ASR results for the REVERB Challenge development set are collected in Table 2 and for the evaluation set in Table 3. This section primarily reviews the evaluation set results of our systems. Comparable ASR results from external studies [6, 30] are also gathered in Table 3 and analyzed in Section 8. The feature enhancement combinations are grouped by their respective back-end systems. In Table 2, the results are shown as average WERs separately for the SimData and RealData recordings. In Table 3, the results are also shown for each recording condition, but the comparisons between the feature enhancement methods are based on their respective average WERs. For reference, the REVERB Challenge baseline results, with and without MC training and batch-based MLLR speaker adaptation, are shown in the first two rows of the result tables. The Challenge baselines make use of MFCC features concatenated with their first- and second-order derivatives and bigram LMs.

For each back-end system, omitting the feature enhancement produces the highest error rates with the exception of DM enhancement on the LDA+MLLT+SAT+f-bMMI back-end, which gives the highest average error rate on RealData. For each back-end, the lowest error rates are obtained by taking advantage of either DM or MDI initialization in NMF feature enhancement, except for the LDA+MLLT back-end where NMF alone is the best performing feature enhancement. For each enhancement method, the corresponding average WERs are shown to decrease consistently on SimData while increasing the complexity of back-end processing. On RealData, however, none of the feature enhancements on the LDA+MLLT+SAT+DNN back-end is able to exceed their respective average results with the LDA+MLLT+SAT+f-bMMI back-end.

For both single-channel SimData and RealData, the proposed DM+NMF feature enhancement outperforms MDI+NMF for the majority of back-end systems.

Table 2 Average SimData and RealData word error rates for the REVERB Challenge development set

1-channel				
Feature enhancement	MC training	Back-end system	Avg. SimData	Avg. RealData
-	No	Challenge BL	51.86	88.38
-	Yes	Challenge BL + SA	25.16	47.23
-	No	LDA+MLLT	44.16	89.84
-	Yes		17.44	40.23
DM	Yes		15.23	36.16
NMF	Yes		13.09	34.31
MDI+ NMF	Yes		13.12	35.16
DM + NMF	Yes		12.80	36.23
-	Yes	LDA+MLLT+SAT	13.94	33.52
DM	Yes		11.87	31.96
NMF	Yes		9.99	29.75
MDI+ NMF	Yes		9.85	29.97
DM + NMF	Yes		9.97	28.05
-	Yes	LDA+MLLT+SAT+f-bMMI	11.15	29.65
DM	Yes		10.15	29.25
NMF	Yes		9.09	27.20
MDI+ NMF	Yes		8.83	28.11
DM + NMF	Yes		9.18	26.48
MDI+ NMF	Yes	LDA+MLLT+SA+bMMI+MBR	8.98	25.71
DM + NMF	Yes		9.05	26.18
-	Yes	LDA+MLLT+SAT+DNN	10.99	30.25
NMF	Yes		8.52	27.18
MDI+ NMF	Yes		8.01	25.99
DM + NMF	Yes		8.04	25.56
DM + NMF	Yes	LDA+MLLT+SAT+DNN+SMBR	7.97	25.40
8-channel				
DS + MDI + NMF	Yes	LDA+MLLT+SAT+f-bMMI	6.47	20.97
DS + DM + NMF	Yes		6.68	20.99

The following abbreviations are used in the table: multicondition (MC), baseline (BL), speaker adaptation (SA), linear discriminant analysis (LDA), maximum likelihood linear transform (MLLT), distribution matching (DM), missing data imputation (MDI), non-negative matrix factorization (NMF), speaker adaptive training (SAT), delay-and-sum (DS), feature domain boosted maximum mutual information criterion (f-bMMI), deep neural network (DNN), minimum Bayes risk (MBR) decoding and discriminative training with state-level minimum Bayes risk (SMBR) criterion. The dashed line separates the performance-wise comparable back-ends

The WER improvements for the proposed DM+NMF method over the MDI+NMF are 0.45 % and 0.9 % on LDA+MLLT+SAT+DNN and LDA+MLLT+SAT+f-bMMI back-ends, respectively. On 8-channel recordings, DS+DM+NMF produces the lowest average WER on SimData, whereas DS+MDI+NMF gives the best performance on RealData.

8 Discussion

We have shown that the proposed DM+NMF feature enhancement achieves the highest average performances on both single-channel SimData and RealData recordings. However, these highest performance figures are achieved by a small margin relative to MDI+NMF and NMF and with different back-ends. DM+NMF is also conceptually simpler than our previous MDI+NMF approach, with fewer parameters to optimize. It also gives a performance advantage compared to the systems of Weninger et al. [6] and Tachioka et al. [30]. In the following subsections, we

Table 3 Average SimData and RealData word error rates for the REVERB Challenge evaluation set

1-channel												
Feature	MC	Back-end system	SimData							RealData		
enhancement	training		R1F	R2F	R3F	R1N	R2N	R3N	Avg.	R1F	R1N	Avg.
-	No	Challenge BL	25.40	82.20	88.00	18.10	43.00	53.50	51.70	87.30	89.70	88.50
-	Yes	Challenge BL + SA	18.70	32.50	38.90	16.20	20.50	24.80	25.27	47.60	50.10	48.85
-	No	LDA+MLLT	18.79	75.21	87.70	10.55	26.57	40.45	43.21	90.04	90.29	90.17
-	Yes		12.32	24.19	29.96	13.11	12.57	15.88	18.01	38.69	40.18	39.44
DM	Yes		10.20	19.43	24.99	9.69	11.94	14.76	15.17	38.08	39.48	38.78
NMF	Yes		10.11	16.62	20.40	9.42	11.43	12.81	13.46	31.50	34.43	32.97
MDI+ NMF	Yes		10.54	17.34	21.14	10.35	11.38	12.91	13.94	33.32	35.64	34.48
DM + NMF	Yes		9.94	17.07	21.12	9.57	11.54	12.85	13.68	35.08	37.05	36.06
-	Yes	LDA+MLLT+SAT	9.96	18.94	23.81	8.62	10.66	12.38	14.06	33.49	33.82	33.66
DM	Yes		8.55	16.94	20.66	8.17	9.77	11.63	12.62	31.84	32.55	32.20
NMF	Yes		8.47	13.41	16.63	7.61	9.24	10.20	10.93	27.82	28.33	28.07
MDI+ NMF	Yes		7.96	13.73	16.84	7.44	9.32	10.37	10.94	28.26	29.38	28.82
DM + NMF	Yes		7.81	14.00	17.23	7.47	9.37	9.83	10.95	27.41	27.53	27.47
-	Yes	LDA+MLLT+SAT+f-bMMI	8.37	15.40	19.09	7.61	9.21	10.47	11.69	27.58	30.41	28.99
DM	Yes		8.00	14.66	17.13	7.20	8.62	10.12	10.96	29.51	29.67	29.59
NMF	Yes		8.00	12.62	15.18	7.17	8.75	9.31	10.17	**23.94**	27.05	25.50
MDI+ NMF	Yes		7.45	12.89	15.23	6.96	8.54	9.18	10.04	24.51	26.29	25.40
DM + NMF	Yes		7.13	12.62	15.61	6.86	8.18	9.21	9.94	24.92	25.65	25.29
Tachioka et al.	Yes		7.22	13.97	18.44	6.44	7.57	9.52	10.53	29.78	28.87	29.33
Weninger et al.	Yes	LDA+MLLT+SA+bMMI+MBR	7.52	14.15	15.30	6.39	8.41	9.47	10.21	28.06	**25.39**	26.73
MDI+ NMF	Yes		6.89	13.31	14.96	6.34	8.15	9.25	9.82	24.92	25.55	25.23
DM + NMF	Yes		7.30	12.63	14.98	6.40	8.71	9.38	9.90	24.17	26.09	**25.13**
-	Yes	LDA+MLLT+SAT+DNN	7.25	15.53	20.37	6.81	8.55	10.53	11.51	32.88	32.90	32.89
NMF	Yes		7.01	12.83	15.47	**5.66**	7.96	9.18	9.68	26.37	27.85	27.11
MDI+ NMF	Yes		**6.32**	12.25	**14.35**	6.30	7.65	8.53	9.23	26.54	27.18	26.86
DM + NMF	Yes		6.66	12.20	15.01	5.75	7.54	7.97	9.19	26.87	26.64	26.76
DM + NMF	Yes	LDA+MLLT+SAT+DNN+SMBR	6.57	**12.17**	15.01	5.85	7.56	**7.88**	**9.17**	26.87	26.25	26.56
Tachioka et al.	Yes	LDA+MLLT+SAT+DNN+bMMI	6.84	12.57	16.55	5.90	**7.35**	9.40	9.77	25.69	25.97	25.83
8-channel												
DS + MDI + NMF	Yes	LDA+MLLT+SAT+f-bMMI	6.22	8.91	10.70	5.79	7.10	7.25	7.66	**17.45**	17.85	**17.65**
DS + DM + NMF	Yes		6.56	**8.78**	**10.59**	5.86	6.77	**7.10**	**7.61**	17.99	18.08	18.03
BF + Tachioka et al.	Yes		6.64	10.13	13.15	6.17	**6.51**	7.40	8.33	23.67	20.63	22.15
BF + Weninger et al.	Yes	LDA+MLLT+SA+bMMI+MBR	**6.12**	9.69	11.28	**5.49**	6.80	7.13	7.75	22.52	**17.66**	20.09

Comparable single and eight channel feature enhancement results from studies by Weninger et al. [6] and Tachioka et al. [30] are also presented for the most complex 1- and 8-channel back-ends. The best results for each recording condition and for both 1- and 8-channel systems are printed in boldface. Here, in addition to the abbreviations of Table 2, BF denotes beamforming in general and as an example, R1F is decoded as Room 1, Far microphone. The dashed line separates the performance-wise comparable back-ends

discuss the principles underlying our approach and how these give rise to the performance gains observed and then compare our results with those from other studies.

8.1 The principles of the approach

The main features of the enhancement method proposed in the current study are that it is unsupervised and makes only weak assumptions about the reverberation in both the DM and NMF stages. In contrast to DM, the MDI front-end requires a measurement of the extent of reverberation which is mapped to masked thresholds utilizing a function with three experimentally adjusted free parameters [27]. In the DM initialization, the two main assumptions are that reverberation effects are convolutive and long term, and that the same transformations can be used to decorrelate each reverberation condition. In the NMF stage, reverberation is again assumed to be convolutive with a long-term effect. The activation filtering assumes certain characteristics of temporal modulation patterns of activations that are common to all rooms. Therefore, neither the DM initialization nor NMF make assumptions relating to any specific room.

That said, the unsupervised nature of the proposed method also raises some challenges. The cost function we use measures the success of reconstructing the original observed speech, but its relation to the dereverberation

or room characteristics is indirect (see Fig. 3). Therefore, it is possible for the cost function to converge even when the method does not apply dereverberation. This also explains why we needed to modify the iterative update rules to implement the NMFD model—our preliminary experiments conducted with and without initialization showed that the cost function converged, but the NMF dereverberation was not successful.

The filtering of the activation matrix by H_A, done in step 3 of the NMF update algorithm, is motivated by the need to remove traces of reverberation that remain in matrix **A**. These traces are caused by imperfections in the initial estimation stage and by the first stage of NMF reconstruction before the filter update is applied (step 2 and Fig. 3). More specifically, filtering the activation matrix by H_A serves to move the traces of reverberation that remain in **A** to matrix **R**, which is updated in the next stage of iterations (step 4). The filtering scheme is similar to other approaches that apply modulation filters to counteract reverberation (e.g. [5]). It emphasizes reverberation-free speech onsets through a smoothed derivator filter along the time trajectory; not in spectrograms as in earlier studies, but in the activations **A**. The filtering also increases the sparsity of **A**. After the matrix **R** update iterations (step 4), the following activation matrix **A** update (step 5) does not use activation filtering. The filtering scheme is motivated by the notion that it is more useful to model reverberation as much in matrix **R** as possible. The reason for this, as discussed above, is that the NMF cost function measures the precision of reconstruction of the original reverberant speech, rather than dereverberation that should be left for matrix **R**. Note that matrix **R** is updated only once, as our preliminary experiments revealed that by alternating the **R** and **A** updates, it is difficult to obtain stable estimates for both matrices. Our hypothesis is that either the cost function optimized by NMF is not optimal for reverberation or that the optimization algorithm gets easily stuck in local minima. Evidence supporting the former explanation is that increasing the iteration counts did reduce the cost function but impaired the recognition performance.

Considering the initialization step in the NMF algorithm on the most complex LDA+MLLT+SAT+f-bMMI and LDA+MLLT+SAT+DNN back-ends, the results indicate that it is beneficial to apply dereverberation during initialization. However, on the less complex LDA+MLLT+SAT back-end, the benefit is negligible and on the least complex LDA+MLLT back-end, the initialization step is detrimental as the NMF alone provides the lowest average WERs on both SimData and RealData.

Our previous studies [17, 18] have shown that DM outperforms MDI by a small margin in feature enhancement as it achieves 37.87 % and 72.25 % average WERs on the REVERB Challenge SimData and RealData recordings, respectively, while MDI yields 39.14 and 71.67 %. This observation may also explain why DM is better than MDI when applied as the initialization method. However, we cannot conclude that any better dereverberation method used to initialize NMF would also lead to better factorization. For instance in [17], experiments were conducted using NMF and MDI as separate feature enhancement methods for a system with acoustic models trained on unenhanced MFCCs. For non-reverberant speech signals, the MDI feature enhancement had no notable impact on performance compared to the clean-speech-trained baseline (the authors report WERs of 12.70 and 12.55 %, respectively). However, the MDI-initialized NMF feature enhancement severely degraded the clean speech recognition accuracy (17.37 % WER), because the NMF introduced prominent artifacts in the speech signals.

8.2 Comparison to similar studies

As discussed in Section 8.1, one key factor of our two step feature enhancement is the ability to generalize. Our approach is based on unsupervised learning, in which a filter with an arbitrary impulse response can be learned from data, and arbitrary speech utterances can be modeled through the combination of dictionary atoms using NMF. Accordingly, the dereverberation approach generalizes well to unseen data. In contrast, the RNN-based system in [6] requires supervised training and may become over-trained to particular reverberation conditions or speaker attributes. This may limit its ability to generalize to unseen data. Evidence that our system generalizes comparatively well to unseen room conditions can be found by comparing the SimData and RealData results for our system and the Weninger et al. system. Relative error reduction (calculated between average results of our DM+NMF method and LDA+MLLT+SA+bMMI+MBR back-end and the Weninger et al. system) for our system compared to the Weninger et al. system is twice as large for RealData (6.0 %) than for SimData (3.0 %), indicating better performance for our system in mismatched conditions.

A closer examination of the results obtained with LDA+MLLT+SAT+f-bMMI and LDA+MLLT+SAT+DNN back-ends reveals that although the MDI+NMF and DM+NMF feature enhancements benefit the DNN-based back-end system in terms of SimData performance, the improvements on RealData are not as large as with f-bMMI discriminative training. This may be due to non-optimal DNN training, as the risk of over-training is relatively prominent with DNNs.

The feature enhancement method of Tachioka et al. [30] is based on blind reverberation time estimation for a dereverberation process similar to spectral subtraction. Our method, on the other hand, does not make use of reverberation time but makes only weak assumptions about

the reverberation conditions, as discussed in Section 8.1. With the LDA+MLLT+SAT+f-bMMI back-end, the DM-only feature enhancement achieves nearly as good a performance as Tachioka et al., with a relative average error increase of 4.1 % on SimData and 0.9 % on RealData. In our previous study [17], the MDI system based on the same mask estimation method as in the current study was shown to outperform an MDI method with mask estimation based on assessment of room reverberation. These findings imply that the final recognition performance can be significantly degraded by inaccuracies in reverberation estimates. In multichannel recordings, the Tachioka et al. system invokes DS beamforming with a cross-spectrum phase analysis and a peak-hold process for the direction of arrival estimation. While the beamforming in [30] is essentially an improved version of our DS implementation, our results indicate that a conventional DS performs better for the REVERB Challenge data. This is apparent from the observation that the relative difference between the average error rates of Tachioka et al. and DM+NMF are larger on 8-channel than on single-channel setups, for both SimData and RealData.

Even though our average DNN+SMBR discriminative training-based results (9.17 %) are slightly better than the comparable DNN+bMMI results of Tachioka et al. (9.77 %) on SimData, the Tachioka et al. system provides higher average performance on RealData (26.56 % vs. 25.83 %, respectively). It is also noteworthy that in our experiments, discriminative training brought little benefit to the DNN system, whereas a more significant improvement was seen for Tachioka et al.'s DNN back-end. The best single-channel results in the study of Tachioka et al. are obtained by combining the results from 16 separate recognition systems by using recognizer output voting error reduction (ROVER). The average WERs for the ROVER system are 8.51 % for SimData and 23.70 % for RealData. To put things in perspective, the best performing single-channel recognizer in the REVERB Challenge, proposed by Delcroix et al. [38], achieved average WERs of 5.2 % on SimData and 17.4 % on RealData. The most significant benefit of the Delcroix et al. system compared to ours lies in the acoustic model, which has higher input dimensionality and was trained on an extended data set approximately five times the size of the REVERB Challenge training data set. The Delcroix et al. system also operated in full-batch mode.

9 Conclusions

This paper proposed a two-stage feature enhancement method for dereverberation of speech for noise robust ASR, based on a combination of distribution matching and non-negative matrix factorization. The proposed method was evaluated with modern ASR back-ends based on variants of the GMM-HMM and DNN-HMM frameworks and shown to outperform our previous combination of missing data imputation and NMF [17] by a small margin. In several instances, the proposed method also gave higher recognition accuracy than the state-of-the-art reference approaches by [6, 30] with similar back-end processing. The main benefit of the proposed method over the reference approaches is that it generalizes well to unseen reverberation conditions. This was reflected in the most difficult real-data scenarios in the REVERB Challenge, where our DM+NMF-based ASR systems achieve the largest performance gains over reference approaches. Moreover, the NMF alone and MDI+NMF-based systems were also shown to perform well with respect to the reference approaches.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The research was supported by the Academy of Finland projects 136209 (Sami Keronen, Kalle J. Palomäki) and 251170 (Heikki Kallasjoki and Kalle J. Palomäki). Guy J. Brown was supported by the EU project Two!Ears under grant agreement ICT-618075.

Author details

[1]Department of Signal Processing and Acoustics, Aalto university, P.O. Box 13000, 00076 Aalto, Finland. [2]Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello, S1 4DP Sheffield, UK. [3]Audience, Inc., 331 Fairchild Drive, Mountain View 94043, CA, USA.

Received: 13 February 2015 Accepted: 31 July 2015
Published online: 20 August 2015

References

1. G Hinton, L Deng, D Yu, G Dahl, A Mohamed, N Jaitly, A Senior, V Vanhoucke, T Sainath, B Kingsbury, Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Proc. Mag. **29**(6), 82–97 (2012)
2. JT Geiger, JF Gemmeke, B Schuller, G Rigoll, in *Proc. INTERSPEECH*. Investigating NMF speech enhancement for neural network based acoustic models (IEEE, Singapore, Singapore, 2014)
3. S Thomas, S Ganapathy, H Hermansky, Recognition of reverberant speech using frequency domain linear prediction. IEEE Signal Proc. Let. **15**, 681–684 (2008)
4. B Kingsbury, N Morgan, S Greenberg, Robust speech recognition using the modulation spectrogram. Speech Commun. **25**, 117–132 (1998)
5. KJ Palomäki, GJ Brown, JP Barker, Techniques for handling convolutional distortion with 'missing data' automatic speech recognition. Speech Commun. **43**(1–2), 123–142 (2004)
6. F Weninger, S Watanabe, J Le Roux, JR Hershey, Y Tachioka, J Geiger, B Schuller, G Rigoll, in *Proc. REVERB Workshop (REVERB'14)*. The MERL/MELCO/TUM system for the REVERB Challenge using deep recurrent neural network feature enhancement, Florence, Italy, 2014
7. JT Geiger, E Marchi, B Schuller, G Rigoll, in *Proc. REVERB Workshop (REVERB'14)*. The TUM system for the REVERB Challenge: recognition of reverberated speech using multi-channel correlation shaping dereverberation and BLSTM recurrent neural networks, Florence, Italy, 2014
8. A Sehr, R Maas, W Kellermann, Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognition. IEEE Trans. Audio, Speech, Language Process. **18**(7), 1676–1691 (2010)
9. DD Lee, HS Seung, in *Adv. Neur. In. 13*, ed. by TK Leen, TG Dietterich, and V Tresp. Algorithms for non-negative matrix factorization (MIT Press, Cambridge, 2001), pp. 556–562
10. T Virtanen, Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. IEEE T. Audio Speech. **15**(3), 1066–1074 (2007)

11. P Smaragdis, JC Brown, in *IEEE Workshop Applicat. Signal Process. Audio and Acoust*. Non-negative matrix factorization for polyphonic music transcription (IEEE, New Paltz, NY, USA, 2003), pp. 177–180
12. KW Wilson, B Raj, P Smaragdis, A Divakaran, in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. Speech denoising using nonnegative matrix factorization with priors (IEEE, Las Vegas, NV, USA, 2008), pp. 4029–4032
13. JF Gemmeke, T Virtanen, A Hurmalainen, Exemplar-based sparse representations for noise robust automatic speech recognition. IEEE T. Audio Speech. **19**(7), 2067–2080 (2011)
14. P Smaragdis, in *Independent Component Analysis and Blind Signal Separation. Lecture Notes in Computer Science*, ed. by CG Puntonet, A Prieto. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs, vol. 3195 (Springer, Berlin Heidelberg, 2004), pp. 494–499
15. H Kameoka, T Nakatani, T Yoshioka, in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms (IEEE, Taipei, Taiwan, 2009), pp. 45–48
16. K Kumar, R Singh, B Raj, R Stern, in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. Gammatone sub-band magnitude-domain dereverberation for ASR (IEEE, Prague, Czech Republic, 2011), pp. 4604–4607
17. H Kallasjoki, JF Gemmeke, KJ Palomäki, AV Beeston, GJ Brown, in *Proc. REVERB Workshop (REVERB'14)*. Recognition of reverberant speech by missing data imputation and NMF feature enhancement, Florence, Italy, 2014
18. K Palomäki, H Kallasjoki, in *Proc. REVERB Workshop (REVERB'14)*. Reverberation robust speech recognition by matching distributions of spectrally and temporally decorrelated features, Florence, Italy, 2014
19. U Remes, in *Proc. INTERSPEECH*. Bounded conditional mean imputation with an approximate posterior (ISCA, Lyon, France, 2013), pp. 3007–3011
20. AV Beeston, GJ Brown, in *UK Speech Conf*. Modelling reverberation compensation effects in time-forward and time-reversed rooms, Cambridge, UK, 2013
21. S Dharanipragada, M Padmanabhan, in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*. A non-linear unsupervised adaptation technique for speech recognition (ISCA, Beijing, 2000)
22. K Kinoshita, M Delcroix, T Yoshioka, T Nakatani, E Habets, R Haeb-Umbach, V Leutnant, A Sehr, W Kellermann, R Maas, S Gannot, B Raj, in *Proc. IEEE Workshop Applicat. Signal Process. Audio and Acoust. (WASPAA)*. The REVERB challenge: a common evaluation framework for dereverberation and recognition of reverberant speech (IEEE, New Paltz, NY, USA, 2013)
23. D Povey, A Ghoshal, G Boulianne, L Burget, O Glembek, N Goel, M Hannemann, P Motlicek, Y Qian, P Schwarz, J Silovsky, G Stemmer, K Vesely, in *IEEE Automat. Speech Recognition and Understanding Workshop*. The Kaldi speech recognition toolkit (IEEE, Waikoloa, HI, USA, 2011)
24. SM Pizer, EP Amburn, JD Austin, R Cromartie, A Geselowitz, T Greer, JB Zimmerman, K Zuiderveld, Adaptive histogram equalization and its variations. Comput. Vision Graph. **39**(3), 355–368 (1987)
25. G Saon, S Dharanipragada, D Povey, in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. Feature space Gaussianization, vol. 1 (IEEE, Montreal, Canada, 2004), pp. 329–332
26. CB Moler, *Numerical Computing with MATLAB, Revised Reprint Paperback*. (Society of Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 2008)
27. KJ Palomäki, GJ Brown, JP Barker, in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. Recognition of reverberant speech using full cepstral features and spectral missing data (IEEE, Toulouse, France, 2006)
28. T Robinson, J Fransen, D Pye, J Foote, S Renals, in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. WSJCAM0: a British English speech corpus for large vocabulary continuous speech recognition (IEEE, Detroit, MI, USA, 1995)
29. M Lincoln, I McCowan, J Vepa, HK Maganti, in *IEEE Automat. Speech Recognition and Understanding Workshop*. The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments (IEEE, Cancún, Mexico, 2005)
30. Y Tachioka, T Narita, F Weninger, S Watanabe, in *Proc. REVERB Workshop (REVERB'14)*. Dual system combination approach for various reverberant environments with dereverberation techniques, Florence, Italy, 2014
31. D Povey, K Yao, in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. A basis method for robust estimation of constrained MLLR (IEEE, Prague, Czech Republic, 2011), pp. 4460–4463
32. D Povey, D Kanevsky, B Kingsbury, B Ramabhadran, G Saon, K Visweswariah, in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. Boosted MMI for model and feature-space discriminative training (IEEE, Las Vegas, NV, USA, 2008), pp. 4057–4060
33. H Xu, D Povey, L Mangu, J Zhu, Minimum Bayes risk decoding and system combination based on a recursion for edit distance. Comput. Speech Lang. **25**(4), 802–828 (2011)
34. X Zhang, J Trmal, D Povey, S Khudanpur, in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. Improving deep neural network acoustic models using generalized maxout networks (IEEE, Florence, Italy, 2014)
35. B Kingsbury, in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling (IEEE, Taipei, Taiwan, 2009), pp. 3761–3764
36. MF Font, Multi-microphone signal processing for automatic speech recognition in meeting rooms. Master's thesis, Universitat Politècnica de Catalunya, Spain, 2005
37. CH Knapp, GC Carter, The generalized correlation method for estimation of time delay. IEEE T. Acoust. Speech. **24**(4), 320–327 (1976)
38. M Delcoix, T Yoshioka, A Ogawa, Y Kubo, M Fujimoto, N Ito, K Kinoshita, M Espi, T Hori, T Nakatani, A Nakamura, in *Proc. REVERB Workshop (REVERB'14)*. Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB Challenge, Florence, Italy, 2014