RECOGNITION OF REVERBERANT SPEECH BY MISSING DATA IMPUTATION AND NMF FEATURE ENHANCEMENT

Heikki Kallasjoki[1], *Jort F. Gemmeke*[2], *Kalle J. Palomäki*[1], *Amy V. Beeston*[3], *Guy J. Brown*[3]

[1] Department of Acoustics and Signal Processing, Aalto University, Finland
[2] Department ESAT-PSI, KU Leuven, Belgium
[3] Department of Computer Science, University of Sheffield, UK

ABSTRACT

The problem of reverberation in speech recognition is addressed in this study by extending a noise-robust feature enhancement method based on non-negative matrix factorization. The signal model of the observation as a linear combination of sample spectrograms is augmented by a mel-spectral feature domain convolution to account for the effects of room reverberation. The proposed method is contrasted with missing data techniques for reverberant speech, and evaluated for speech recognition performance using the REVERB challenge corpus. Our results indicate consistent gains in recognition performance compared to the baseline system, with a relative improvement in word error rate of 42.6% for the optimal case.

Index Terms— Speech dereverberation, non-negative matrix factorization, missing data

1. INTRODUCTION

Practical automatic speech recognition (ASR) applications often require methods for coping with highly variable environmental distortions. In particular, when restricted to signals recorded with a single distant microphone, room reverberation typically causes severe degradation in the performance of conventional ASR systems. While the topic of speech recognition in noisy environments has been widely studied, many proposed systems are limited by an underlying assumption that the observed signal is an additive mixture of speech and noise, often with the latter having spectral characteristics unlike those of speech. The distortions introduced by the multiple reflected signals inherent in reverberation do not fit this model well.

In the missing data model [1], individual spectro-temporal components of the observed signal are classified as either reliable or unreliable. For the purposes of recognition, the regions considered unreliable can be entirely ignored, or used to provide an upper bound for the clean speech energy under the assumption of additive noise. In approaches based on data imputation, this model is realized by replacing the components that are considered unreliable with estimates derived from their surrounding context.

Modeling speech in terms of individual training samples has recently gained prominence in the field of speech recognition [2]. Non-negative matrix factorization (NMF) algorithms [3] are commonly utilized in these *exemplar-based* methods. A source separation feature enhancement method for noise-robust automatic speech recognition, based on representing observations as sparse linear combinations of exemplars, obtained with NMF, is presented in [4].

The NMF framework can also be extended to non-negative matrix factor deconvolution (NMFD) [5]. In previous research, the NMFD model has been applied to speech dereverberation in a spectral feature domain, by considering reverberated speech as a convolution of clean speech and room response features under different constraints [6, 7].

In the present study, we evaluate previous work on producing missing data masks specifically for reverberant speech, based on modulation filtering of spectral features [8, 9], on the REVERB challenge corpus [10]. Additionally, we adapt a method for room reverberation estimation, inspired by human auditory models [11], for use in mask estimation. A recently proposed bounded conditional mean imputation [12] algorithm is used for missing feature imputation.

We further propose an extension of the NMF-based noise-robust feature enhancement algorithm of [4] to the context of reverberant speech, and contrast its performance with the missing data methods. Compared to speech dereverberation based on the NMFD framework, our work is distinguished by retaining the model of speech as a linear combination of exemplars, but extending it to allow the inclusion and optimization of an arbitrary convolutional filter in the mel-spectrogram domain. The proposed NMF reverberant speech enhancement method also includes a missing data imputation step to produce an initial estimate of the sparse representation of the clean speech signal. Filtering in the domain of the sparse representation is used to bias the reconstruction, leading to stronger attenuation of reverberation.

The rest of this paper is organized as follows. In Sec-

This research was funded by the Hecse graduate school (H. Kallasjoki), Academy of Finland grants 251170 (H. Kallasjoki, K.J. Palomäki) and 136209 (K.J. Palomäki), the TEKES FuNeSoMo project (K.J. Palomäki), IWT-SBO project ALADIN, grant 100049 (J.F. Gemmeke), UK EPSRC, grant EP/G009805/1 (A.V. Beeston) and FP7-ICT project TWO!EARS, grant 618075 (G.J. Brown).

tion 2, we describe the evaluated missing data methods. The proposed NMF feature enhancement scheme is introduced in Section 3. The experimental setup is detailed in Section 4, and the obtained results presented in Section 5. The results are discussed in Section 6, and concluding remarks presented in Section 7.

2. MISSING DATA TECHNIQUES

2.1. Missing Data Imputation

Several methods that model the clean speech with Gaussian mixtures have been introduced for the reconstruction task of missing data imputation. The bounded conditional mean imputation variant used in this work for reconstructing the unreliable regions of the observation was recently proposed in [12]. In conventional conditional mean imputation, the distribution of clean speech features $p(\mathbf{x})$ is modeled using a Gaussian mixture model. Denoting by $\mathbf{x}_\mathrm{r}$ and $\mathbf{x}_\mathrm{u}$, respectively, the reliable and unreliable components of a single observation frame or a window of consecutive frames, an estimate $\hat{\mathbf{x}}_\mathrm{u}$ is produced based on the conditional distribution $p(\mathbf{x}_\mathrm{u} \mid \mathbf{x}_\mathrm{r})$. Bounded conditional mean imputation (BCMI) extends the model by further assuming that the observed $\mathbf{x}_\mathrm{u}$ is an upper bound for the signal of interest. In [12], this approach is implemented by deriving an approximate parametric model for the posterior distribution of the bounded features.

2.2. Mask Estimation

A missing data mask that labels the reliable and unreliable regions of the observation is required in order to apply missing data methods. If the corresponding clean speech is known a priori, an *oracle mask* can be constructed by comparison with the observation. In general, however, this information is not available and the reliable regions must be identified based on the distorted observation alone. The accuracy of this mask estimation process has a major influence on the speech recognition performance, as shown by comparisons between the recognition performance obtained with estimated and oracle masks [1]. Three mask estimation schemes are evaluated for the missing data experiments presented in this work.

A mask estimation method designed for reverberant speech in particular, based on modulation filtering, has been presented in [8] and extended in [9]. In the original formulation, it is based on spectral features derived from a gammatone filterbank. For this work, the gammatone filterbank is replaced by the mel-spectral filterbank used in the feature extraction of the REVERB challenge baseline recognizer.

We denote by $y(t,b)$ the b'th mel channel component of frame t of the reverberant observation, compressed by raising it to the power 0.3. These features are further processed with a band-pass modulation filter with 3 dB cutoff frequencies of 1.5 Hz and 8.2 Hz and an automatic gain control step, and normalized by subtracting a channel-specific constant selected so that the minimum value for each channel over a single utterance is zero. The resulting signal is denoted by $y_\mathrm{bp}^\mathrm{agc}(t,b)$.

The missing data mask $m_\mathrm{R}(t,b)$ is derived by thresholding the AGC features,

$$m_\mathrm{R}(t,b) = \begin{cases} 1 & \text{if } y_\mathrm{bp}^\mathrm{agc}(t,b) > \theta(b), \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The threshold $\theta(b)$ for mel channel b is selected for each utterance based on the 'blurredness' metric B of [8], as

$$\theta(b) = \gamma \frac{\frac{1}{N}\sum_{t=1}^{N} y_\mathrm{bp}^\mathrm{agc}(t,b)}{1+\exp\left(-\alpha(B-\beta)\right)}, \quad (2)$$

where α, β and γ are set based on small-scale experiments.

An alternative mask estimation approach, denoted m_LP, is based on a room-reverberation estimation method for a computational auditory modeling task [11]. Reverberation tails are located in the signal $y(t,b)$ by first estimating the smoothed temporal envelope in each channel, $y_\mathrm{lp}(t-\tau_d,b)$, using a second-order low-pass Butterworth filter with cutoff frequency at 10 Hz, and identifying regions for which the derivative $y'_\mathrm{lp}(t-\tau_d,b) < 0$. The parameter τ_d corrects for the filter delay. The amount of energy in each decaying region of one frequency channel is quantified by

$$L(t,b) = \begin{cases} \frac{1}{|n(t,b)|}\sum_{k\in n(t,b)} y(k,b) & \text{if } y'_\mathrm{lp}(t-\tau_d,b) < 0, \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

where $n(t,b)$ is the set of contiguous time indices around t where the derivative for channel b remains negative. Under the assumption that reverberant signals result in greater $L(t,b)$ values than dry speech, the m_LP mask is defined as:

$$m_\mathrm{LP}(t,b) = \begin{cases} 1 & \text{if } L(t,b) < \theta_\mathrm{LP}, \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

A data-driven approach to the mask estimation problem is to label the unreliable regions with the aid of a binary classifier, trained on oracle masks based on a data set where the underlying clean speech signal, or an approximation of it, is known. In the context of noisy speech, such methods have been successfully applied using different classifiers, including Bayesian classifiers based on Gaussian mixture models (GMM) [13] and support vector machines (SVM) [14]. A large variety of acoustic features have been used for such classifier-based masks [13, 14, 15]. In this work, the data-driven mask estimation approach is evaluated with the following set of six features, denoted F1 through F6, related to the m_R and m_LP mask estimation methods.

The compressed mel-spectral features $y(t,b)$ are used directly as the first feature F1. Features F2 and F3 correspond to the GRAD and MOD features previously used in [15], respectively. GRAD is an estimate of the local slope of $y(t,b)$, while MOD is the bandpass-filtered spectrum $y_\mathrm{bp}^\mathrm{agc}(t,b)$ used in (1). Feature F4 is set to the ratio of $y_\mathrm{bp}^\mathrm{agc}(t,b)$ and the threshold $\theta(b)$, and F5 to the 'blurredness' metric B. Finally, $L(t,b)$ defined in (3) is used as feature F6.

Individual binary classifiers are trained for each mel channel based on oracle data. Masks generated with both GMM-based and SVM classifiers, denoted by m_{GMM} and m_{SVM}, respectively, are evaluated.

3. NMF FEATURE ENHANCEMENT

3.1. Speech model

Recently, approaches based on non-negative matrix factorization (NMF) have shown to be an effective approach to model additive noise. In NMF, observed spectrograms are modeled as a sparse, non-negative linear combination of spectrographic basis atoms, collected in a *dictionary*. With both speech and noise modeled as linear combinations of speech and noise atoms, respectively, noisy speech is modeled as the linear combination of noise and speech [4].

In this work, we do not rely on a noise model, but do exploit the same compositional speech model. In short, we write:

$$\mathbf{Y} \approx \mathbf{SA}, \tag{5}$$

with $\mathbf{Y}$ the observed speech, a dictionary $\mathbf{S}$ and activation weights $\mathbf{A}$. Here, $\mathbf{Y}$ is a $TC \times N$ matrix composed of a collection of C-dimensional mel-scale magnitude spectrograms, reshaped into column vectors by stacking T consecutive frames, with N the number of *windows* extracted from the observed utterance. $\mathbf{S}$ is a $TC \times K$ dictionary matrix with K atoms, similarly organized in windowed column vectors, and $\mathbf{A}$ a non-negative $K \times N$ activation matrix.

We obtain the dictionary $\mathbf{S}$ in advance by randomly extracting spectrograms from clean speech training data. This *exemplar-based* approach has been shown to be an effective method to obtain large dictionaries that can accurately model long temporal contexts [4].

3.2. Reverberant speech

To account for the effects of reverberation, we modify the model (5) to have the form

$$\mathbf{Y} \approx \mathbf{RSA}, \tag{6}$$

where $\mathbf{R}$ is a $T_r C \times TC$ matrix of the form

$$\mathbf{R} = \begin{pmatrix} \underbrace{\begin{matrix} r_{1,1} & 0 & 0 & \\ 0 & r_{1,2} & 0 & \cdots \\ 0 & 0 & r_{1,3} & \\ & \vdots & & \ddots \\ r_{2,1} & 0 & 0 & \\ 0 & r_{2,2} & 0 & \cdots \\ 0 & 0 & r_{2,3} & \\ & \vdots & & \ddots \end{matrix}}_{C} & & & & & \\ & & & & & \end{pmatrix} \begin{matrix} r_{1,1} & 0 & 0 & \\ 0 & r_{1,2} & 0 & \cdots \\ 0 & 0 & r_{1,3} & \\ & \vdots & & \ddots \end{matrix} . \tag{7}$$

Left multiplication of a T-frame spectrogram in the stacked column vector form by $\mathbf{R}$ is equivalent to the discrete convolution of the contents of each mel band b with the sequence $\begin{bmatrix} r_{1,b} & r_{2,b} & \dots & r_{T_f,b} \end{bmatrix}$ of length T_f, resulting in a stacked vector of $T_r = T + T_f - 1$ frames.

In this work, the dictionary $\mathbf{S}$ contains only non-reverberated speech exemplars, while the filter matrix $\mathbf{R}$ encodes the effect of the room reverberation on the signal. The approximation $\mathbf{RSA}$ can equivalently be interpreted as either the clean speech reconstruction distorted by reverberation, $\mathbf{R}(\mathbf{SA})$, or as the sparse representation of the observation using an artificially reverberated dictionary, $(\mathbf{RS})\mathbf{A}$.

3.3. Optimization

Given an observed utterance, $\mathbf{S}$ is kept fixed. Estimates of $\mathbf{R}$ and $\mathbf{A}$ are obtained by minimizing, for each window t, the cost function

$$d(\mathbf{Y}_t, \mathbf{RSA}_t) + \lambda \left\| \mathbf{A}_t \right\|_1 , \tag{8}$$

where the $d(\cdot, \cdot)$ term measures the distance between the observation and the approximation, with $\mathbf{Y}_t$ and $\mathbf{A}_t$ denoting the corresponding columns of $\mathbf{Y}$ and $\mathbf{A}$, respectively. The second term is a sparsity-inducing L^1-norm weighted by a sparsity coefficient λ. In this work we use the generalized Kullback-Leibler divergence for d.

To optimize (8) a simple alternative multiplicative update algorithm can be used [4, 3]. The optimization problem is not convex, however, and preliminary experiments revealed it is difficult to obtain stable estimates for both $\mathbf{R}$ and $\mathbf{A}$. Instead, we use the following algorithm to derive the factorization $\mathbf{RSA}$:

1. An initial estimate $\tilde{\mathbf{X}}$ of the non-reverberant speech is constructed using the m_{R} missing data mask estimation and BCMI imputation methods described in Section 2.2 and Section 2.1, respectively.
2. Using the initial estimate, the activation matrix $\mathbf{A}$ elements are obtained from the non-negative factorization $\tilde{\mathbf{X}} \approx \mathbf{RSA}$, with I_1 rounds of multiplicative updates and $\mathbf{R}$ fixed to the identity matrix.
3. As an ad-hoc method to bias the reverberation filter estimation, the resulting sequences of activation values for each individual exemplar are filtered with $H_A(z) = 1 - 0.9z^{-1} - 0.8z^{-2} - 0.7z^{-3}$ and clamped to be non-negative. This filtering step has the effect of suppressing consecutive activations of a single exemplar, which is typical of reverberant signals.
4. The $\mathbf{R}$ matrix is initialized to contain the constant T_f-length filter $\frac{1}{T_f}[1 \ \dots \ 1]$ on all mel channels.
5. Keeping $\mathbf{A}$ fixed, the filter matrix $\mathbf{R}$ is updated for I_2 iterations to approximate the reverberant observation, based on the factorization $\mathbf{Y} \approx \mathbf{RSA}$. While the multiplicative update of matrix $\mathbf{R}$ does preserve elements set to zero, it does not necessarily result in a matrix of the form shown in (7), or one corresponding to a physically plausible filter. To prevent this, we extract

the filter coefficients $r_{t,b}$ from $\mathbf{R}$ after every iteration by averaging across all occurrences of each coefficient. These are restricted to satisfy $\forall t : r_{t+1,b} < r_{t,b}$ by clamping overly large values and normalized by scaling to $\sum_{t,b} r_{t,b} = C$. The $\mathbf{R}$ matrix is then reinitialized to have the form of (7).

6. Finally, the $\mathbf{R}$ matrix is kept fixed, and the $\mathbf{A}$ parameters are updated for another I_3 iterations.

In the simple sliding window model presented here, individual windows are processed independently of each other, and averaging over the overlapping frames is used to form the final clean and reverberant speech reconstructions. In reverberant conditions, if the start of a window coincides with the start of a pause in the speech, the early frames of the window are filled with sound energy originating only from reflections, with no direct speech component. When such a window is represented using a reverberated dictionary, this energy is interpreted as direct sound, and is therefore not properly attenuated in the enhanced features.

To avoid this issue, we replace the use of the reverberant speech reconstruction $\mathbf{RSA}$ in the iterative updates of [3], in steps 2, 5 and 6 of the above algorithm, by a version obtained by summing over the overlapping frames in $\mathbf{RSA}$. This makes it possible for a reverberated exemplar activated in an earlier window to "explain away" the reverberant energy in later windows it overlaps with. The resulting model is similar to that of non-negative matrix factor deconvolution (NMFD) [5].

3.4. Feature enhancement

After obtaining $\mathbf{R}$ and $\mathbf{A}$, we make a clean speech reconstruction, $\hat{\mathbf{X}} = \mathbf{SA}$, and a reconstruction of the reverberated observation, $\hat{\mathbf{Y}} = \mathbf{RSA}$. Using the same Wiener filter approach as in [4], we use these reconstructions to construct a time-varying filter as the ratio of $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$, after overlap-add averaging across the overlapping windows. Feature enhancement is then carried out by filtering the observed reverberant speech with the obtained mel-spectral filter.

4. EXPERIMENTAL SETUP

4.1. Data

The data sets for all experiments described in this work are provided by the REVERB challenge, and described in detail in [10]. All test utterances are taken from the WSJCAM0 British English continuous speech recognition corpus [16]. In order to test speech recognition performance in reverberant environments, both clean speech artificially distorted using measured room impulse responses ("SimData") as well as speech recorded in a reverberant room ("RealData") are used.

The SimData data set contains utterances from 6 simulated reverberation conditions, resulting from the combination of three different rooms (small, medium, large) and two speaker-to-microphone distances (near, far). The T_{60} reverberation times of the small, medium and large rooms are approximately 0.25 s, 0.5 s and 0.7 s, while the near and far microphone distances are 0.5 m and 2.0 m, respectively. In all six conditions, measured noise signals are also added to the distorted utterances at a fixed signal-to-noise ratio (SNR) of 20 dB.

The RealData data set consists of real recordings of speakers in a reverberant (T_{60} of 0.7 s) meeting room, recorded at two microphone distances: approximately 1.0 m and 2.5 m for the near and far conditions, respectively. Prompts from the WSJCAM0 corpus are used for the content of the utterances.

As the room impulse responses of SimData and test utterances of RealData are measured with similar 8-channel circular microphone arrays, multi-microphone methods can be used with both data sets. The data sets also share the 5000-word WSJ vocabulary, so a single set of acoustic and language models can be used for all utterances.

The test sets are divided to separate development and evaluation subsets. In addition to the test sets, a multi-condition training set of simulated reverberant speech is also provided. T_{60} reverberation times of the utterances range approximately from 0.1 s to 0.8 s. The room impulse responses used in the construction of the training set are separate from the three rooms of the SimData data set. Duration of the training data set is approximately 17.5 hours (7861 utterances), while the SimData and RealData evaluation sets have durations of approximately 4.8 hours (2176 utterances) and 0.6 hours (372 utterances), respectively.

Subsets of the multi-condition training set, along with the corresponding clean speech signals, are used in all experiments involving oracle data. This use of aligned clean and reverberant speech is limited to the construction of oracle masks as training targets for missing data masks m_{GMM} and m_{SVM}, and the verification (but not selection) of parameters for m_{R}. No other experiments assume the availability of such data. The oracle mask training subset contains 240 utterances, selected to have an equal amount of samples distorted with each of the room responses used in the multi-condition training set construction.

4.2. Speech Recognition System

The speech recognition performance evaluation is performed with the REVERB challenge baseline recognizer, built on the HTK toolkit [17]. The features used for recognition are 13 mel-frequency cepstral coefficients (MFCCs), along with their first and second time derivatives. These features are modeled with tied-state hidden Markov models, with mixtures of 10 Gaussian components as the emission distributions.

Two sets of acoustic models are used for the experiments. The clean speech acoustic model is trained with the WSJCAM0 corpus training set, while the multi-condition model is based on retraining the clean model using the REVERB challenge multi-condition training set. When the presented

feature enhancement methods are used, the enhancement is also performed on the utterances of the multi-condition training set.

In addition, constrained maximum likelihood linear regression (CMLLR) adaptation is optionally applied in conjunction with the multi-condition model. The adaptation is performed over the whole test set of a single test condition in an unsupervised manner, by using initial recognition results as the transcriptions. A regression tree of 256 regression classes is optimized for each condition.

4.3. Feature Enhancement Processing

The proposed feature enhancement methods operate in the mel-spectral domain, and are performed during the corresponding stage of the feature extraction processing of the REVERB challenge baseline recognizer system. All the methods operate independently on each utterance, and therefore fall in the class of utterance-based batch processing. As a result, the full recognition system performs either utterance-based or full batch processing, depending on whether the CMLLR adaptation is disabled or enabled, respectively.

A robust channel normalization method presented in [8] is used to counteract simple convolutional distortions between clean speech models and test data caused by, e.g., differences in recording equipment. With the exception of the baseline results, this normalization step is consistently applied to both training and test data prior to feature enhancement processing.

In order to take advantage of the provided multi-channel microphone array recordings, we incorporate the PHAT-DS beamformer implementation of [18]. One channel is chosen as the reference signal, and the relative time delays of arrival for the direct sound component of the speech in the other recording channels are estimated based on the PHAT-weighted generalized cross-correlation. The signals are then aligned and summed, so that constructive interference amplifies the clean speech signal. Experiments involving the multi-channel recordings are performed by generating corresponding single-channel signals using the PHAT-DS beamformer for all reverberant data sets, including the multi-condition training, development and evaluation data sets. The recognition process is identical to the single-channel experiments in all other respects.

Regarding the computational cost of the presented methods, missing data imputation has relatively modest requirements, with a real-time factor of approximately 0.1 for the feature enhancement processing, as measured on a single thread on a conventional workstation (Intel Xeon E3-1230 V2). By contrast, the proposed NMF feature enhancement is more costly, having a real-time factor of 6.9 on the same system, for the implementation used in this study. However, in related work on the NMF feature enhancement for noisy speech, significant performance improvements have been obtained by modifying the implementation to, e.g., take advantage of GPU computation.

4.4. Missing Data Imputation

For the missing data mask m_{R}, parameters $\alpha = 19$, $\beta = 0.43$ and $\gamma = 1.4$ in (2) are chosen based on previous work [9]. A small-scale grid search of α, β and γ on a subset of 25 utterances of the multi-condition training set is used to confirm the suitability of the selected parameter values, by comparing the resulting missing data masks to the corresponding oracle masks. The m_{LP} mask threshold of (4) is set to $\theta_{\mathrm{LP}} = 0.5$, based on inspection of generated missing data masks. A grid search with a word error rate criterion is used to select the threshold for generating oracle masks for the training of the m_{GMM} and m_{SVM} methods. 32-component GMMs are used by the m_{GMM} mask classifiers. The imputation algorithm of Section 2.1 is performed using a 5-component GMM trained on a random 1000-utterance subset of the clean speech training set, with a time context of 3 consecutive frames for each window.

4.5. NMF Feature Enhancement

The window and filter length, sparsity coefficient and iteration count parameters of the NMF feature enhancement method of Section 3 are set to $T = 10$, $T_f = 20$, $\lambda = 1$, $I_1 = 50$, $I_2 = 50$ and $I_3 = 100$, all based on small-scale experiments on development set data. A clean speech dictionary $\mathbf{S}$ of $K = 7681$ exemplars is constructed by selecting a single random T-frame segment from each of the utterances in the WSJCAM0 clean speech training set.

5. RESULTS

5.1. Missing Data Imputation

Word error rates for the evaluation of the competing mask estimation methods on the SimData and RealData development data sets are presented in Table 1. The recognition was performed using clean speech acoustic models and without CMLLR adaptation. The best performing mask is highlighted in bold, and comparable results for the NMF feature enhancement method of Section 3, denoted as "NMF", are provided for reference.

As the m_{R} mask estimation method is the best overall method, producing competitive results for the SimData data set while slightly outperforming the other methods for RealData, it is used as the method of choice for the missing data initialization step in the NMF feature enhancement algorithm. The m_{LP} masks are relatively unsuitable for conditions involving low amounts of reverberation. The classifier-based masks (m_{GMM}, m_{SVM}) show some evidence of overfitting, as they both outperform the m_{R} mask estimation method on the oracle mask training set, but are on average approximately equally efficient on the SimData development set and fall behind in the RealData experiments. The effect is more notable for the m_{SVM} mask than the m_{GMM} mask.

5.2. NMF Feature Enhancement

Final results on the REVERB challenge evaluation data set are presented in Table 2. Four scenarios are considered: single-channel feature enhancement using acoustic models trained on clean speech with no adaptation (denoted "Clean"); acoustic models trained on multi-condition data both without ("MC") and with ("MC+ad.") CMLLR adaptation; and multi-channel feature enhancement with the DS-PHAT beamformer in the multi-condition training and CMLLR adaptation case ("8-ch."). Within each scenario, the best obtained result is indicated in bold.

In addition to the REVERB challenge baseline (denoted "Baseline"), results are provided for both the BCMI imputation with m_{R} missing data masks ("BCMI"), and with the proposed NMF feature enhancement method ("NMF"). Relative word error rate reductions over the baseline system in the average results for the SimData and RealData data sets for all four scenarios are summarized in Table 3.

Table 1. Recognition word error rates (%) for evaluated mask estimation methods on development set data using clean speech acoustic models. The best performing mask is highlighted in bold. Comparable results for the NMF feature enhancement, with the m_{R} mask used for initialization, are also shown for reference.

	SimData							RealData		
	Room 1		Room 2		Room 3		Ave.	Room 1		Ave.
	Near	Far	Near	Far	Near	Far	–	Near	Far	–
Baseline	15.29	25.29	43.90	85.80	51.95	88.90	51.81	88.71	88.31	88.51
BCMI, mask m_{R}	15.00	24.53	32.49	**64.56**	37.69	**66.35**	40.07	**68.37**	**67.40**	**67.88**
BCMI, mask m_{LP}	22.49	27.53	48.56	67.37	49.95	72.35	48.01	73.67	72.45	73.06
BCMI, mask m_{GMM}	**14.97**	23.06	**30.93**	64.75	37.78	68.37	**39.94**	71.62	70.13	70.87
BCMI, mask m_{SVM}	15.19	**21.78**	34.09	67.91	**35.04**	70.87	40.78	74.80	73.48	74.14
NMF	*13.13*	*18.88*	*22.95*	*41.66*	*26.11*	*46.93*	*28.26*	*53.21*	*54.48*	*53.84*

Table 2. Word error rates for evaluation data experiments. Results are presented for clean speech acoustic models ("Clean") and multi-condition models both without ("MC") and with ("MC+ad.") CMLLR adaptation, as well as multi-condition training, adaptation and testing on the microphone array signals preprocessed with the DS-PHAT beamformer ("8-ch."). The best performing method within each category is highlighted in bold.

Model	Method	SimData							RealData		
		Room 1		Room 2		Room 3		Ave.	Room 1		Ave.
		Near	Far	Near	Far	Near	Far	–	Near	Far	–
Clean	Baseline	18.32	25.77	42.71	82.71	53.56	87.97	51.82	90.07	88.01	89.04
	BCMI	17.77	24.72	30.89	56.63	39.34	65.62	39.14	73.87	69.48	71.67
	NMF	**16.77**	**20.13**	**23.75**	**39.03**	**29.21**	**49.65**	**29.74**	**61.00**	**57.26**	**59.13**
MC	Baseline	20.79	21.38	23.36	38.69	28.25	45.21	29.60	57.97	55.20	56.58
	BCMI	20.01	20.86	23.06	34.35	27.44	37.83	27.25	52.12	50.51	51.31
	NMF	**18.30**	**18.74**	**20.90**	**28.48**	**23.56**	**34.76**	**24.11**	**47.46**	**46.66**	**47.06**
MC+ad.	Baseline	16.27	18.67	20.65	32.64	24.71	39.36	25.37	49.82	47.94	48.88
	BCMI	16.42	18.84	21.47	30.60	24.48	35.73	24.58	45.86	46.25	46.05
	NMF	**15.89**	**17.15**	**19.22**	**26.24**	**21.24**	**31.75**	**21.91**	**40.79**	**42.03**	**41.41**
8-ch.	Baseline	**14.42**	16.40	15.84	23.73	18.38	29.86	19.76	38.74	41.69	40.21
	BCMI	15.15	16.52	16.59	23.09	18.32	26.75	19.40	37.85	38.72	38.28
	NMF	14.47	**15.83**	**15.55**	**19.92**	**16.64**	**24.42**	**17.80**	**34.43**	**35.15**	**34.79**

Table 3. Relative reduction of average word error rates over the baseline recognizer in Table 2.

	Clean		MC.		MC+ad.		8-ch.	
	SimData	RealData	SimData	RealData	SimData	RealData	SimData	RealData
BCMI	−24.5%	−19.5%	−7.9%	−9.3%	−3.1%	−5.8%	−1.8%	−4.8%
NMF	−42.6%	−33.6%	−18.5%	−16.8%	−13.6%	−15.3%	−9.9%	−13.5%

6. DISCUSSION

With the exception of the near-distance Room 1 test condition of the multi-channel experiments, where the results of all systems are very close, the proposed NMF feature enhancement method consistently outperforms both the challenge baseline as well as the evaluated missing data imputation system. The missing data imputation also provides performance gains over the baseline system in most test conditions, especially in cases involving high levels of reverberation. While the relative improvements are lower compared to the clean speech acoustic model case, both methods remain beneficial for reverberant speech even when used in conjunction with multi-condition training, CMLLR adaptation and the DS-PHAT beamformer.

For speech signals with negligible amounts of reverberation, missing data imputation with the m_{R} mask estimation has a relatively minor effect, and the corresponding recognition performance is essentially equivalent to the baseline system. By contrast, the NMF feature enhancement causes a marked distortion for clean speech signals, reducing the recognition performance in non-reverberant environments. Word error rates for both missing data imputation and NMF feature enhancement for the three clean speech conditions of the REVERB evaluation data set are presented in Table 4. Results are provided for clean speech acoustic models with no adaptation, and multi-condition acoustic models both with and without CMLLR adaptation. When multi-condition acoustic models are used, both feature enhancement methods reduce the mismatch between the processed multi-condition training set and the clean speech test set, and consequently outperform the baseline system. Finally, if CMLLR adaptation is enabled, all three systems perform comparably, with an accuracy slightly less than that of the clean speech acoustic model baseline.

In order to establish an upper bound for the performance of the missing data imputation approach, the speech recognition performance was measured also for the oracle mask training subset. In these experiments, the baseline recognizer had a word error rate of 59.8%, while missing data imputation with oracle masks achieved a word error rate of 40.6%.

Table 4. Clean speech recognition performance of evaluated methods in all considered single-channel feature enhancement scenarios.

Model	Method	Room			Ave.
		1	2	3	
Clean	Baseline	**12.89**	12.64	**12.13**	**12.55**
Clean	BCMI	12.91	**12.59**	12.61	12.70
Clean	NMF	18.08	17.50	16.53	17.37
MC	Baseline	30.38	30.47	30.35	30.40
MC	BCMI	**22.53**	**21.96**	**22.73**	**22.40**
MC	NMF	24.50	23.74	23.62	23.95
MC+ad.	Baseline	16.23	15.58	15.87	15.89
MC+ad.	BCMI	**15.15**	**14.94**	15.48	**15.18**
MC+ad.	NMF	16.50	15.68	**15.45**	15.87

For comparison, corresponding results for estimated masks were 50.1–52.3%, and 45.4% for the proposed NMF feature enhancement method. While the recognition performance of imputation with oracle masks exceeded that of NMF feature enhancement, realistic mask estimation methods typically fall far short of oracle performance, as seen both in this study and previous work [1, 15].

While the REVERB corpus contains microphone array signals, our main focus for this study is in the single-channel feature enhancement scenario. However, binaural features have been used in several missing data mask estimation methods [15]. Mask generation based on features derived from binaural and 8-channel signals would be a natural extension for future work on similar data.

For the optimization algorithm described in Section 3.3, the activation matrix filtering performed in step 3 seems crucial for obtaining a solution that is capable of significantly attenuating reverberation, though at the cost of the noted degradation of performance for clean speech signals. In comparable development set experiments, omitting step 3 resulted in speech recognition performance essentially equivalent to that of the missing feature imputation alone.

In the related noise-robust NMF feature enhancement system [4], a joint dictionary of speech and noise samples is used to provide a noise model. While a noise dictionary was not used in this study, as the SNR of 20 dB for the background noise in the REVERB corpus is relatively high, the possibility of combining both methods to account for a reverberant, noisy environment remains a potential topic for future work.

A widely used approach for further improving the efficiency of feature enhancement methods is to take the uncertainty of the enhanced features into account during the recognition process, via frameworks such as observation uncertainties or uncertainty decoding [19]. For the missing data imputation applied in this work, such information is readily available in the posterior distribution of the imputed features [12]. While the proposed NMF feature enhancement method does not inherently provide a way to estimate the variance of the enhanced features, heuristic uncertainty estimates have been successfully used in conjunction with the related noise-robust NMF feature enhancement system [20]. The applicability of similar extensions in the context of reverberant speech could also be investigated in future work.

7. CONCLUSIONS

In this study, we presented an extension of a NMF-based noise-robust feature enhancement scheme to account for speech distorted by reverberation. The speech recognition performance of the proposed method, along with previous missing data techniques for reverberant speech, were evaluated on the REVERB challenge corpus. Significant word error rate reduction was observed under both clean speech as well as multi-condition acoustic models, with the proposed NMF-based feature enhancement outperforming the evaluated missing data methods.

8. REFERENCES

[1] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, 2001.

[2] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, D. Van Compernolle, K. Demuynck, J. F. Gemmeke, J. R. Bellegarda, and S. Sundaram, "Exemplar-based processing for speech recognition: An overview," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 98–113, 2012.

[3] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *In NIPS*. 2000, pp. 556–562, MIT Press.

[4] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[5] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Independent Component Analysis and Blind Signal Separation*, C. G. Puntonet and A. Prieto, Eds., vol. 3195 of *Lecture Notes in Computer Science*, pp. 494–499. Springer Berlin Heidelberg, 2004.

[6] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2009)*, 2009, pp. 45–48.

[7] K. Kumar, R. Singh, B. Raj, and R. Stern, "Gammatone sub-band magnitude-domain dereverberation for ASR," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, 2011, pp. 4604–4607.

[8] K. J. Palomäki, G. J. Brown, and J. P. Barker, "Techniques for handling convolutional distortion with 'missing data' automatic speech recognition," *Speech Communication*, vol. 43, no. 1-2, pp. 123–142, 2004.

[9] K. J. Palomäki, G. J. Brown, and J. P. Barker, "Recognition of reverberant speech using full cepstral features and spectral missing data," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006)*, 2006.

[10] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-13)*, 2013.

[11] A. V. Beeston and G. J. Brown, "Modelling reverberation compensation effects in time-forward and time-reversed rooms," in *UK Speech Conference*, Cambridge, UK, Sept. 2013.

[12] U. Remes, "Bounded conditional mean imputation with an approximate posterior," in *Proc. INTERSPEECH*, 2013, pp. 3007–3011.

[13] M. L. Seltzer, B. Raj, and R. M. Stern, "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, no. 4, pp. 379–393, 2004.

[14] J. F. Gemmeke, Y. Wang, M. Van Segbroeck, B. Cranen, and H. Van Hamme, "Application of noise robust MDT speech recognition on the SPEECON and SpeechDat-Car databases," in *Proc. INTERSPEECH*, 2009, pp. 1227–1230.

[15] S. Keronen, H. Kallasjoki, U. Remes, G. J. Brown, J. F. Gemmeke, and K. J. Palomäki, "Mask estimation and imputation methods for missing data speech recognition in a multisource reverberant environment," *Computer Speech & Language*, vol. 27, no. 3, pp. 798 – 819, 2013.

[16] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAM0: a British English speech corpus for large vocabulary continuous speech recognition," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1995)*, 1995.

[17] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, "The HTK book, version 3.4," Tech. Rep., Cambridge University Engineering Department, 2006.

[18] M. F. Font, "Multi-microphone signal processing for automatic speech recognition in meeting rooms," M.S. thesis, Universitat Politècnica de Catalunya, Spain, 2005.

[19] H. Liao and M. J. F. Gales, "Issues with uncertainty decoding for noise robust automatic speech recognition," *Speech Communication*, vol. 50, no. 4, pp. 265–277, 2008.

[20] H. Kallasjoki, J.F. Gemmeke, and K.J. Palomaki, "Estimating uncertainty to improve exemplar-based feature enhancement for noise robust speech recognition," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 2, pp. 368–380, 2014.