# A combined evaluation of established and new approaches for speech recognition in varied reverberation conditions

Sunit Sivasankaran, Emmanuel Vincent, Irina Illina

# A combined evaluation of established and new approaches for speech recognition in varied reverberation conditions

Sunit Sivasankaran[1,2,3,*], Emmanuel Vincent[1,2,3], Irina Illina[1,2,3]

[1] *Inria, Villers-lès-Nancy, F-54600, France*
[2] *Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54506, France*
[3] *CNRS, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54506, France*

**Abstract**

Robustness to reverberation is a key concern for distant-microphone ASR. Various approaches have been proposed, including single-channel or multichannel dereverberation, robust feature extraction, alternative acoustic models, and acoustic model adaptation. However, to the best of our knowledge, a detailed study of these techniques in varied reverberation conditions is still missing in the literature. In this paper, we conduct a series of experiments to assess the impact of various dereverberation and acoustic model adaptation approaches on the ASR performance in the range of reverberation conditions found in real domestic environments. We consider both established approaches such as WPE and newer approaches such as learning hidden unit contribution (LHUC) adaptations, whose performance has not been reported before in this context, and we employ them in combination. Our results indicate that performing weighted prediction error (WPE) dereverberation on a reverberated test speech utterance and decoding using an deep neural network (DNN) acoustic model trained with multi-condition reverberated speech with feature-space maximum likelihood linear regression (fMLLR) transformed features, outperforms more recent approaches and helps significantly reduce the word error rate (WER).

*Keywords:* Robust ASR, dereverberation, acoustic model adaptation, evaluation.

## 1. Introduction

Many automatic speech recognition systems (ASR) such as the Amazon Echo do not require the user to stay close to the microphone while interacting with the device. This raises two main problems, namely noise and reverberation. The performance of ASR systems is known to degrade under these external influences (Brutti and Matassoni, 2016; Petrick et al., 2014; Wölfel and McDonough, 2009; Yoshioka et al., 2012).

Various methods have been proposed to tackle ASR under noisy conditions (Li et al., 2013). Evaluation challenges have also been conducted on this specific issue (Barker et al., 2015). The problem can be addressed at different stages of ASR. At the signal level, single-channel and multichannel speech enhancement techniques have been applied (Virtanen et al., 2012; Sivasankaran et al., 2015). At the feature level, cepstral mean and variance normalization (CMVN) (Viikki

*Email address:* sunit.sivasankaran@inria.fr (Sunit Sivasankaran[1,2,3,*] )

et al., 1998), histogram equalization (Molau et al., 2003) and subband histogram normalization (Joshi et al., 2015) techniques have been employed. At the model level, multi-condition training and model adaptation techniques have been used. Another approach is to model the uncertainties created by the noise and compensate them across the ASR system in order to achieve noise robustness (Abdelaziz et al., 2015). An overview of noise robust techniques for ASR is described in (Li et al., 2013).

Similarly, ASR for reverberated speech has previously been studied (Kinoshita et al., 2013; Brutti and Matassoni, 2016; Yoshioka et al., 2012; Brutti and Matassoni, 2014). Room acoustics induces three components, namely: direct sound, early reflections and late reverberation. The boundary between early reflections and late reverberation is generally considered to be around 50 ms after the direct signal. The energy ratio of direct sound and early reflections with respect to the late reverberation is measured by the clarity index, $C_{50}$ (Kuttruff, 2009; Nishiura et al., 2007). This quantity is also sometimes referred to as the early-to-late ratio (ELR) and it is known to influence the ASR performance (Wölfel and McDonough, 2009). Another parameter which is commonly associated with reverberant signals is the reverberation time (RT) which is defined as the time taken for the audio signal to decay by 60 dB. Early reflections can be considered as a convolution of the speech signal with a stationary channel response, which can be handled by CMVN, and they have been reported to improve ASR performance under certain conditions (Petrick et al., 2014; Brutti and Matassoni, 2016). Late reverberation, by contrast, is considered to be uncorrelated to the speech signal and it contributes most to the degradation of the ASR performance.

Reverberation has been tackled across multiple fronts in ASR systems. Single-channel and multichannel dereverberation techniques have been investigated (Naylor and Gaubitch, 2010). Features which are more robust to reverberation have been proposed (Hermansky and Morgan, 1994; Kingsbury et al., 1998; Falk and Chan, 2010; Ganapathy et al., 2011). Time delay neural networks (Peddinti et al., 2015b,a) and polyphone (instead of triphone) acoustic models (Yoshioka et al., 2012), which take into account the longer term characteristics of speech, have shown promising results at the expense of a larger computational cost. I-vector based model adaptation (Peddinti et al., 2015b) and ROVER based system combination (Fiscus, 1997) have also been proposed as solutions. Evaluation campaigns such as REVERB (Kinoshita et al., 2013) and ASpIRE (Harper, 2015) have recently contributed to popularizing this research topic.

To the best of our knowledge there are only three studies in the literature which attempt to evaluate in detail the performance of various techniques for reverberant ASR. Brutti and Matassoni (2016) have analyzed the effect of reverberation on ASR and correlated its performance with the characteristics of room impulse responses (RIR). The RIRs used were synthetically generated and the results were validated using real RIRs. They showed the dependence of ASR accuracy on the features derived from the structure of early reflections and late reverberation. However, the experiments were conducted using multi-condition training only: no dereverberation or acoustic model adaptation was applied. Another work of interest is the summary of the REVERB challenge by Kinoshita et al. (2016). Many techniques were tried as part of this challenge. However, the test material covered a small range of ELR and RT conditions, which questions the validity of the results in the larger range of conditions found in real environments. Delcroix et al. (2015) detailed different techniques to counter reverberation using the REVERB dataset such as dereverberation using weighted prediction error (WPE) and minimum variance distortionless response (MVDR) beamforming. The effect on the performance using unsupervised adaptation and increased dataset size was also studied. This detailed work however lacks a comparison of the older WPE and MVDR methods to newer DNN based dereverberation and adaptation meth-

ods. Performance comparison using different features such as fMLLR are also missing. FMLLR features have however been used in Peddinti et al. (2015b) to obtain state of the art results as part of the ASpIRE challenge. Nevertheless a comparison of techniques and their combinations are missing in the literature.

In this work, we build upon these studies. We conduct a series of experiments to assess the impact of various dereverberation and acoustic model adaptation approaches on the ASR performance. We evaluate the validity of these approaches in varied ELR and RT reverberation conditions on data simulated from real RIRs. The best performing methods on the simulated data are then applied on the real data. We consider both established approaches and newer approaches, whose performance across different reverberation conditions has not been reported before, and we employ them in combination. As a side contribution, we evaluate the performance of DNN based dereverberation using fMLLR features, which hasnt been reported earlier to the best of our knowledge.

The rest of the paper is organized as follows. Section 2 details the dataset creation and sets up the baseline for our experiments. Section 3 explains how clean alignments for training are obtained and shows the relevant results. Section 5 details the different dereverberation techniques tried and Section 6 explains the model adaptation techniques. We then combine the best performing techniques and apply them on a real dataset in Section 7. We summarize our findings in Section 8.

## 2. Baseline

### *2.1. Dataset*

All experiments up to Section 7 are conducted on simulated data created in a similar way to the CHiME-3 simulated corpus (Barker et al., 2015). Speech consists of utterances from the Wall Street Journal (WSJ0) corpus (Garofalo et al., 2007). We use 7138 utterances from the original WSJ0 corpus for training, 410 utterances from the CHiME-3 corpus for development and 330 utterances from the CHiME-3 corpus for testing, all of which were recorded in clean conditions. The 7138 training utterances correspond to approximately 15 h of speech. The utterances are spoken by 83 speakers in the training set, 4 other speakers in the development set, and 4 other speakers in the test set. All data are sampled at 16 kHz.

188 real multichannel room impulse responses (RIR) recorded as part of the VoiceHome project (Bertin et al., 2016) are used to simulate reverberation [1]. The RIRs are recorded using two 8-microphone devices at various positions and distances in 12 real rooms. Recordings are made in 3 different homes, labeled as HOME1, HOME2 and HOME3. The RIRs from HOME1, HOME2 and HOME3 are used for training, development and testing respectively. The ELR and the RT were computed for each RIR. A scatter plot of the estimated ELR and RT is shown in Fig. 1. The details of the RIR recordings are described in (Bertin et al., 2016).

In order to analyze the effect of different reverberation conditions, we cluster the RIRs into 6 classes based on ELR and RT. Specifically, we consider three ELR conditions:

1. low ELR $\leq$ 10 dB
2. medium ELR between 10 and 15 dB
3. high ELR $>$ 15 dB

[1] The recorded RIRs are freely available at `http://voice-home.gforge.inria.fr/voiceHome_corpus.html`

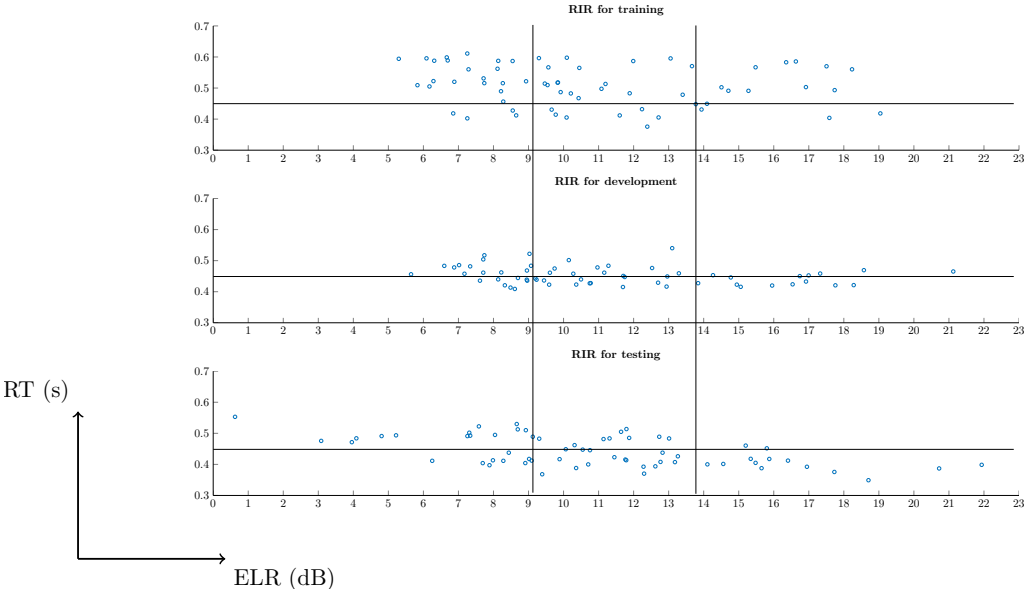



Figure 1: Distribution of ELR and RT of all the room impulse responses. The lines in the plot define the boundaries between RIR conditions used in this experiment.

and two RT conditions:

1. low RT $\leq$ 0.45 s
2. high RT $>$ 0.45 s.

The ranges of ELR and RT values for different classes were chosen in order to ensure that enough RIRs are available for training, development and testing in each of the classes.

Reverberated speech signals are then generated as follows. For every utterance, we pick a random multichannel RIR from the corresponding set (training, development, or testing) and from the desired reverberation class. The clean speech signal $s[n]$ is then normalized to unit power and convolved with the RIR $h_m[n]$ for each microphone $m$. Real noise $\eta[n]$, recorded in quiet conditions in the same room as the RIR, is added to every channel to make the simulation more realistic. Therefore the signal-to-noise ratio (SNR) matches the one observed in real condition in the room (and it depends on distance). We denote the reverberated speech signal by $y_m[n]$. As in CHiME-3, the normalization constant is computed after applying a high pass filter $f_{80}$ to remove frequency components below 80 Hz, which are generally associated with noise. The equations below summarize the reverberation simulation process:

$$s_{\text{high}}[n] = f_{80} * s[n] \tag{1}$$

$$s_{\text{pow}} = \frac{1}{N} \sum_{n=1}^{N} s_{\text{high}}[n]^2 \tag{2}$$

$$y_m[n] = \frac{1}{\sqrt{s_{\text{pow}}}} h_m * s[n] + \eta[n] \tag{3}$$

Here $*$ denotes convolution and $N$ the signal duration.

We generate 9 training, development and testing sets in total:

1. 1 condition-specific training, development and testing set for each ELR and RT class. We label the dataset as $CS^{i}_{(tr,dt,et)}$, where $i = 1 \ldots 6$, represents the six different reverberation classes mentioned in Section 2.1. $(tr, dt, et)$ represents training, development and test part of the dataset respectively.
2. 1 multi-condition training, development and testing set for ASR, generated by randomly picking RIRs across all reverberation classes. We label this dataset as $MC^{1}_{(tr,dt,et)}$.
3. 1 multi-condition training, development and testing set for deep neural network (DNN) based dereverberation. We label this dataset as $MC^{2}_{(tr,dt,et)}$. Note that $MC^{1}$ cannot be used for to train the dereverberation DNN since it would overfit this data and result in a mismatch between the training and test features for ASR.
4. 1 multi-condition training, development and testing set simulated by convolving the clean speech with the early part of the RIRs. We label this dataset as $MC^{E}_{(tr,dt,et)}$. This is done to evaluate the influence of early reflections on ASR.

The first 7 sets are used in all experiments, and the last two are used in Section 5 only. All training, development and test datasets contain 7138, 410 and 330 utterances, respectively. This is to ensure that the amount of data used while experimenting always remains the same. Apart from the above mentioned datasets we utilize the clean WSJ0 training and development dataset to obtain clean alignments. This dataset is labeled as $C^{WSJ}_{(tr,dt,et)}$.

### *2.2. ASR*

Speech recognition is done using a context-dependent DNN-HMM acoustic model. The default number of layers used for all models is 7 unless specified otherwise. Each layer is pretrained using restricted Boltzman machine (RBM). The RBMs are stacked together to form a deep belief network (DBN). The weights of the DBN are updated using the backpropagation algorithm with cross entropy as the loss function (Hinton et al., 2012). The Kaldi toolkit (Povey et al., 2011) was used to train the ASR system. The standard Kaldi CHiME-3 DNN recipe using nnet1 framework was modified for this work. The number of senones (DNN outputs) is set to 1983 throughout the paper. The development part of the dataset is used to ensure that the training does not overfit the training data. A trigram statistical language model was used as a language model (LM) while decoding. Results are reported only for the test dataset.

### *2.3. Baseline results*

Datasets used: $C^{WSJ}_{(tr,dt)}$, $MC^{1}_{(tr,dt,et)}$, $CS^{i}_{(et)}$, $i = 1 \ldots 6$

A first set of experiments were conducted to estimate the performance limits of the system. Logmel features of dimension 40 along with 5 frame left and right context were used as input features (until Section 4). This gave an input feature dimension of $40 \times 11 = 440$. The features were computed using a frame duration of 25 ms and frame shift of 10 ms. The input speech signal was filtered with an 80 Hz highpass filter before feature extraction in order to discard very low frequencies where the speech energy is minimal. Except in the cases when multichannel enhancement was applied (thereby reducing the multichannel input signal to a single-channel enhanced signal), only the first channel was used. No feature normalization was performed while

computing the logmel features. Using clean data for both training and testing, a word error rate (WER) of 2.78% was obtained. This gives the performance limit which should be aimed for while constructing an ASR system for reverberated speech. Using multicondition data, a WER of 12.70% was obtained and is considered as baseline in the rest of the paper. This sets the lower limit of acceptable performance for our system. For completeness we also evaluate the performance when recognizing clean data with the multicondition model. We obtain a WER of 71.6% [2]. The ASR performance of the multicondition model on different reverberation conditions $CS^{i}_{et}, i = 1 \ldots 6$ is shown in Fig. 5(a). The figure shows WER markings on a single dimensional axis. The bigger marker in black is used to mark the average performance of the system while the smaller markers in different colors are used to mark the performance in different reverberation conditions. As expected, the performance of the model in highly reverberant conditions (low ELR) is worse than in other reverberation conditions. Also the RT appears to have little influence on the ASR performance in high and medium ELR conditions, but it does have one in low ELR conditions.

### *2.4. Matched condition training*

Datasets used: $CS^{i}_{(tr,dt,et)}, i = 1 \ldots 6$

The idea behind matched condition training is to evaluate the model performance in the ideal case when the testing and training data are similar. In this section, training and testing are done on one of the 6 condition-specific datasets described above. For example, models trained on data with $RT \leq 0.45$ and $10 < ELR \leq 15$ are used only to decode data with the same RT and ELR range. The reverberation class is assumed to be known a priori in order to study the upper bound of the ASR performance with matched condition training (automatic estimation of the reverberation class is addressed in Section 6.2).

Results obtained using matched condition training are shown in Fig. 5(b). A small WER improvement is observed compared to multicondition training, but not as large as expected. This can be attributed to the fact that DNN acoustic models learn better when the training data exhibits more variability as compared to the smaller variability seen in matched condition training. Further experiments are to be conducted in the future to verify this claim as a function of the amount and the variability of the data. An improvement in performance was observed for high ELR conditions with high RT but a WER deterioration was observed in the case of high ELR and low RT. The RT values did not influence the performance in other reverberation conditions.

## 3. Improvements using clean alignment

Datasets used: $C^{WSJ}_{(tr,dt)}$, $MC^{1}_{(tr,dt,et)}$, $CS^{i}_{(et)}, i = 1 \ldots 6$

[2] Similar deterioration in ASR performance were observed while decoding clean speech with noisy multicondition acoustic model in CHiME and REVERB challenge baselines.

The performance of a DNN acoustic model depends on the senone alignments which were used while training the model. The alignments for the training and the development data are obtained using a Gaussian mixture model-hidden Markov model (GMM-HMM) acoustic model. Since the performance on clean speech was better than the performance on reverberated speech, we believe that the alignments obtained using the clean GMM-HMM model (trained on clean data) are more reliable than those obtained from the GMM-HMM model trained on reverberated speech. Therefore, we use the alignment information obtained from clean data in order to train a reverberated DNN-HMM acoustic model from reverberated speech features. Similar ideas were tried earlier in (Sehr et al., 2014) in the context of GMM-HMM systems. A reduction of the training time and the WER was reported.

Results obtained using clean alignments to train a multicondition model are summarized in Fig. 5(c). An average relative WER improvement of 16.77% is observed when compared to the multicondition model trained using non-clean alignments. We observe RT values having an influence in WER values in low and medium ELR reverberation condition.

Matched condition training was also performed using clean alignments. The results obtained are shown in Fig. 5(d). Though an improvement in the WER performance compared to matched condition training without clean alignment is observed, multicondition training in general still performs better than matched condition training. Matched condition training for datasets with high RT seems to have particularly benefited with the usage of clean alignments. In the rest of the paper we use clean alignments for ASR acoustic model training.

## 4. Improved Features

---

Dataset used: $MC^{1}_{(tr,dt,et)}$, $CS^{i}_{(et)}, i = 1 \ldots 6$

---

Following improved ASR performance in reverberant conditions using feature-space maximum likelihood linear regression (fMLLR) in (Peddinti et al., 2015b), we apply the technique to our problem. To transform features using speaker-dependent fMLLR transforms, the 13-dimensional MFCC features are first normalized to zero mean per speaker using cepstral mean normalization (CMN). The current frame is spliced with three context frames on either side to form a $7 \times 13 = 91$ dimensional feature vector. Linear discriminant analysis (LDA) is used to transform the 91 dimensional vector to dimension 40. The fMLLR transform for every speaker is estimated from the LDA transformed features (Gales, 1998).

FMLLR features along with the clean alignments are then used to train the DNN acoustic model. The obtained results are summarized in Fig. 5(e). A WER of 7.04% which is an improvement of 44.56% over the baseline and 33.33% over the filterbank features with similar training and testing conditions, is observed. Improvement is seen across all the reverberation conditions. RT values influence the performance only in the medium ELR reverberation conditions. In the rest of this paper, all experiments are conducted using fMLLR transformed features

## 5. Dereverberation

One way to fight reverberation is to reduce it in the signal domain or the spectral domain. We therefore evaluate dereverberation techniques in this section. According to the notation in Section 2.1, the general formulation of the dereverberation problem is to estimate the acoustic signal $s[n]$

emitted by the source from the signals $y_m[n]$, $1 \leq m \leq M$, received by the $M$ microphones. Approaches to dereverberation can broadly be divided into two categories. One approach is to nonlinearly filter out reverberation in the power or magnitude time-frequency domain (Lebart et al., 2001; Habets et al., 2009) and the other (linear) approach is to invert the RIR in the time domain or the complex-valued time-frequency domain (Gannot and Moonen, 2003; Gillespie et al., 2001; Nakatani et al., 2008).

### *5.1. DNN based dereverberation*

Dataset used: $C^{WSJ}_{(tr,dt)}$, $MC^{2}_{(tr,dt,et)}$, $MC^{1}_{(tr,dt,et)}$, $MC^{E}_{(tr,dt,et)}$, $CS^{i}_{(et)}, i = 1 \ldots 6$

DNN based dereverberation is a recent single-channel technique which falls into the first category. Han et al. (2015) and Hsiao et al. (2015) proposed to train a DNN to estimate the clean speech spectrum from the reverberated speech spectrum. An improved WER along with improvements in speech quality measures such as the perceptual evaluation of speech quality (PESQ) were reported. We performed a series of experiments to understand the impact of this dereverberation technique. We remind the readers that all further ASR experiments are conducted using clean alignments and fMLLR features.

Our first experiment concerns the choice of the target. As mentioned above, Han et al. (2015) and Hsiao et al. (2015) considered clean speech $y^{\mathrm{D}}_m[n]$ (i.e., the direct part of the reverberated speech signal) as the target. However, other dereverberation techniques (not based on DNN) previously proposed in the literature considered the non-reverberant part of the signal (i.e., the sum of direct sound and early reflections) as the target instead (Habets et al., 2009). The non-reverberant part can be computed by convolving the speech signal in a similar manner as in (1)–(3), except that the RIR is reduced to the first 50 ms after direct sound: $y^{\mathrm{DE}}_m[n] = (1/\sqrt{s_{\mathrm{pow}}})h^{\mathrm{DE}}_m \star s[n]$. To decide which of these is to be used, we evaluate the WER achieved by performing ASR on the non-reverberant part of the speech signal (using dataset $MC^E$ without any dereverberation). The obtained WERs of 7.06% is significantly worse than the WER obtained for clean speech. This contrasts with the results observed by Petrick et al. (2014), which found early reflections to aid the ASR performance in some reverberation conditions. Further experiments are to be conducted to understand this apparent contradiction.

Based on these results, we argue that the non-reverberant part of the signal is a suboptimal target and we decide to learn a DNN mapping between the reverberated speech features and clean speech features instead. Before extracting the features from clean speech we normalize the energy of the target clean speech signal to that of the input reverberated speech signal.

This mapping can be learned in multiple ways, namely:

1. Learning a mapping between the fMLLR features of reverberated speech and clean speech. Here the output is the fMLLR of direct speech and the DNN output dimension is 40. DNN based dereverberation using fMLLR features has not been reported earlier in the literature.
2. Learning a mapping between the MFCCs of reverberated speech and clean speech (Weninger et al., 2014). FMLLR coefficients are then computed from the estimated clean MFCC features. The DNN output dimension here is 13 which corresponds to the MFCC dimension.

A simple multilayer perceptron (MLP) with three hidden layers was used to learn the mappings. Each hidden layer had 2048 nonlinear units. Mean squared error was employed as the objective function for optimization. A separate development set was used to ensure that the

model does not overfit the training data. The input features were normalized to zero mean and unit variance. Sigmoid activation functions were used for all hidden layers. The activation function for the output layer was set to a linear function. The weights of the network were initialized using RBM pretraining. Backpropagation with a minibatch size of 256 was used to train the network. Standard stochastic gradient descent was used as the optimizer with initial learning rate set to $\lambda = 0.08$.

The $MC^2$ dataset was used to train the dereverberation DNN while the $MC^1$ dataset, after dereverberation, was used to train the ASR models. The ASR results obtained after DNN-based dereverberation of the training, development and testing sets are summarized in Table 1. Direct estimation of fMLLR transformed features performed better than estimating the MFCC features and the WER result is 6.87%.

### *5.2. Beamforming*

Dataset used: $MC^1_{(tr,dt,et)}$

For comparison, we also performed multichannel dereverberation with weighted delay and sum beamforming using the Beamformit toolkit (Anguera et al., 2007). The toolkit uses Steered Response Power with the PHAse Transform (SRP-PHAT) algorithm for localization (Loesch and Yang, 2010). All eight channels of reverberated speech are used for beamforming. Speech dereverberation and ASR are performed on the same $MC^1$ dataset. We obtained a WER of 6.61%.

### *5.3. Single-channel spectral processing*

Dataset used: $MC^1_{(tr,dt,et)}$

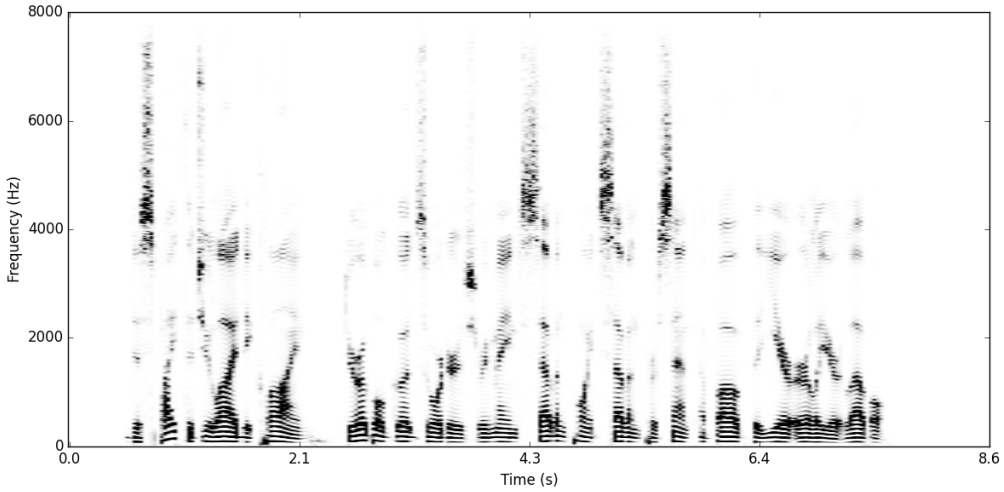
Spectral enhancement schemes based on room acoustics, minimum statistics, and temporal cepstrum smoothing can be employed to suppress noise and reverberation. We evaluate the single channel spectral enhancement technique proposed in (Cauchi et al., 2015) on one channel (0th channel) of the reverberated speech. The technique works by first computing a real valued spectral gain $g_m(\omega, t) \in [0, 1]$ and multiplying it with the reverberated speech spectrum $X(\omega, t)$ as:

$$\hat{S}_m(\omega, t) = g_m(\omega, t) \times X(\omega, t). \tag{4}$$

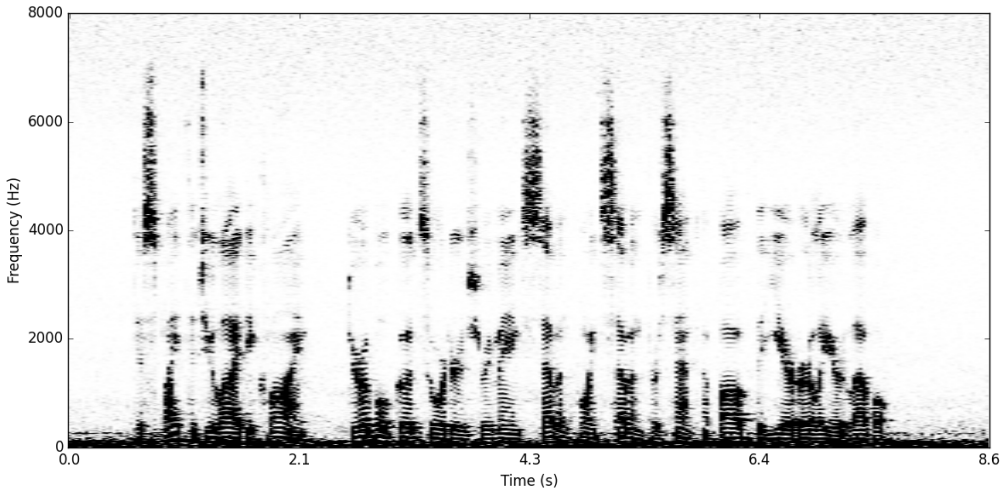
Here $\omega$ and $t$ are the frequency bin and frame index. The gain is computed using minimum mean square error (MMSE) estimators of clean speech spectral magnitudes. Speech dereverberation and ASR are performed on the same $MC^1$ dataset. We obtain a WER of 6.97%.

### *5.4. Weighted prediction error*

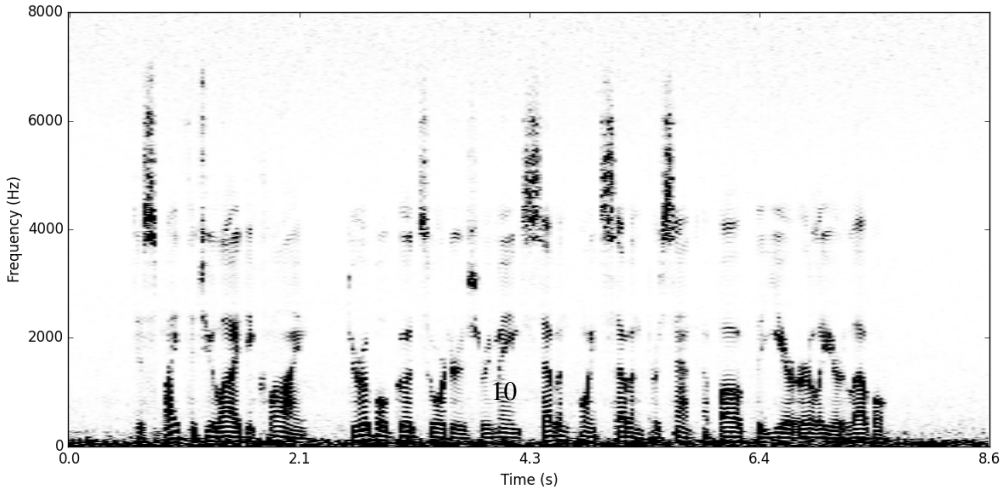
Dataset used: $MC^1_{(tr,dt,et)}$



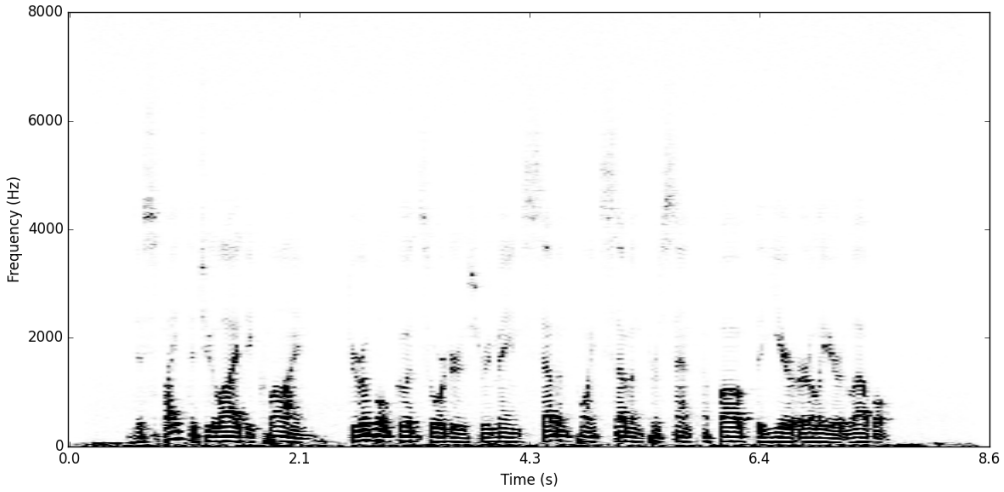

(a) Clean speech



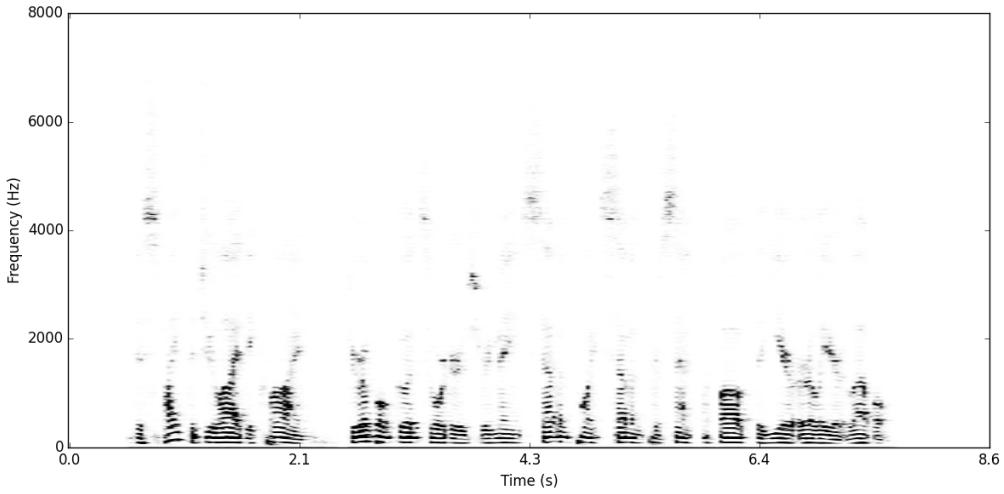

(b) Reverberated speech




(c) WPE




(d) Beamformit




(e) Spectral enhancement

Figure 2: Spectrogram of the clean, reverberated, dereverberated using WPE technique, dereverberated using Beamformit and dereverberated using spectral enhancement for the speech utterance: "DESPITE THE JULY DECLINE DURABLE GOODS ORDERS REMAINED SEVEN POINT SEVEN PERCENT ABOVE THE YEAR EARLIER LEVEL"

Finally, we evaluate the weighted prediction error (WPE) method in (Nakatani et al., 2008; Yoshioka et al., 2009, 2011), which falls into the second category of dereverberation approaches which attempt to invert the RIR. The WPE method operates in the complex-valued time-frequency domain by calculating multichannel linear prediction errors in each frequency bin. The derever-

berated spectral component $\hat{S}_m(\omega, t)$ can be computed as

$$\hat{S}_m(\omega, t) = Y_m(\omega, t) - Y'_m(\omega, t) \tag{5}$$

$$Y'_m(\omega, t) = \sum_{k=D_l}^{D_l+K_l-1} \sum_{m'=1}^{M} g_{mm'}(\omega, k) Y_{m'}(\omega, t-k), \tag{6}$$

where $\{g_{mm'}(\omega, k)\}_{D_l \leq k \leq D_l+K_l-1, 1 \leq m' \leq M}$ is the complex-valued $M$-input 1-output prediction filter at the $\omega$-th frequency bin, $K_l$ is the filter order and $D_l$ is the prediction step size. The prediction filter is estimated by minimizing the mean square error in a maximum likelihood framework. Speech dereverberation and ASR are performed on the same $MC^1$ dataset. We obtain a WER of 6.69%.

The spectrograms obtained after dereverberation for the various algorithm choices along with the spectrogram of clean speech are shown in Fig. 2[3]. Beamformit was observed to result in energy suppression in the high frequency range for this and many other utterances.

### *5.5. Combining dereverberation techniques*

Dataset used: $MC^1_{(tr,dt,et)}$, $MC^2_{(tr,dt)}$

The ASR results obtained when applying the above considered dereverberation techniques to the training, development and testing sets are shown in Table 1. Improvements in the WER performance are observed with all dereverberation techniques but difference in the improvements were small. The best result of 6.61% was observed using Beamformit that is a 6.1% relative improvement compared to training and testing on reverberated data.

We further evaluate the ASR performance using combinations of dereverberation techniques. We start with Beamformit followed by DNN based fMLLR estimation. This resulted in a WER of 6.59% while WPE followed by DNN based fMLLR estimation gave a WER of 6.65%. The best performing system was a combination of Beamformit and DNN based fMLLR estimation. The performance of the system with WPE and DNN based fMLLR estimation is shown for all reverberation conditions in Fig. 5(f). The results follow previous trends namely, RT was not a factor if the ELR is sufficiently high but the RT has an influence on the WER for low and medium ELR conditions.

## 6. Model adaptation

There are multiple techniques which are used to adapt acoustic models to different conditions. I-vectors (Dehak et al., 2009) are typically used to adapt to speakers and environments. Learning hidden unit contributions (LHUC) (Swietojanski and Renals, 2014) is another approach to adapt acoustic models to speakers. Linear input network (LIN) (Neto et al., 1995; Abrash et al., 1995), which defines an additional layer between the input features and the hidden layer of the network can be seen as a linear transform of the input feature, which has a similar effect to feature-domain maximum likelihood linear regression (fMLLR) feature transformation.

[3]Spectrograms from DNN could not be plotted since they estimate the MFCCs or fMLLR features directly

Table 1: WER (%) obtained with different speech dereverberation techniques along with number of channels used. With the exception of baseline results, all other ASR experiments are conducted using fMLLR features and clean alignments.

| **Dereverberation Technique** | **# Channels** | **WER** |
|---|---|---|
| No dereverb baseline (filterbank feats) | 1 | 12.70 |
| No dereverb baseline (fMLLR feats) | 1 | 7.04 |
| DNN (fMLLR feats) | 1 | 6.87 |
| DNN (MFCC feats) | 1 | 7.58 |
| Beamformit | 8 | 6.61 |
| WPE | 8 | 6.69 |
| Spectral Enhancement | 1 | 6.97 |
| Beamformit + DNN (fMLLR) | 8 | 6.59 |
| WPE + DNN (fMLLR) | 8 | 6.65 |

Another adaptation technique which is similar to fMLLR is described in Yao et al. (2012) except that the adaptation takes place in a higher dimension space as compared to fMLLR. This method requires estimation of parameters for a fully connected affine layer which is in the order of $2048 \times 2048$. This is in contrast to the LHUC method which requires estimation of 2048 parameters. Since the amount of available data for adaptation was minimal (around 200 utterances), we select i-vectors and LHUC as representative of the aforementioned model adaptation techniques and investigate their impact on the ASR performance in a reverberant environment.

### *6.1. Learning Hidden Unit Contributions*

Dataset used: $MC^{1}_{(tr,dt,et)}$, $CS^{i}_{(dt,et)}, i = 1 \ldots 6$

LHUC is a new model adaptation technique which was introduced for speaker adaptation. In this work we analyze the detailed performance of LHUC in reverberant conditions using fMLLR features (without applying any kind of dereverberation). The key aspect of LHUC is to learn a speaker-dependent scaling factor for each hidden unit in the DNN acoustic model. If $\phi$ is a nonlinear activation function, $W^l$ and $b^l$ are the weight matrix and bias components for the $l$-th layer of a classical DNN and $x^l$ is the input to the layer, the $j$-th hidden unit activation for the $l$-th layer is classically given by

$$h^l_j = \phi(W^l_j x^l + b^l_j). \tag{7}$$

LHUC learns an additional set of parameters for each hidden unit, such that

$$h^l_j = \zeta(\theta^l_j) \circ \phi(W^l_j x^l + b^l_j) \tag{8}$$

where $\theta$ are the speaker-dependent LHUC parameters and $\zeta$ is a scaling function which constrains the value range of the scaling factors and can either be a sigmoid, tanh or exponential function. The value $\zeta(\theta_j^l)$ is the scaling factor for the $j$-th activation unit of the $l$-th layer. In our experiment, the function $\zeta$ was set to exponential as advocated in (Swietojanski et al., 2016). An advantage of LHUC adaptation is that the weights and bias values which are learned using the full training set remain untouched. This preserves the feature representations learned by the network.

In this paper, experiments are performed using the supervised approach applied to the problem of reverberation. We assume that the reverberation condition (i.e., the ELR and RT class among the 6 possible classes) is known for each utterance in the training, development, and testing sets. Half of the development data (205 utterances) for each reverberation condition is used to learn the LHUC parameters and the rest is used for validation. This ensures that the model does not overfit the training data. The frame classification error rate on the development data was used to decide the improvement in the model. Based on the senone error rate of the validation data, we stop training after 15 iterations.

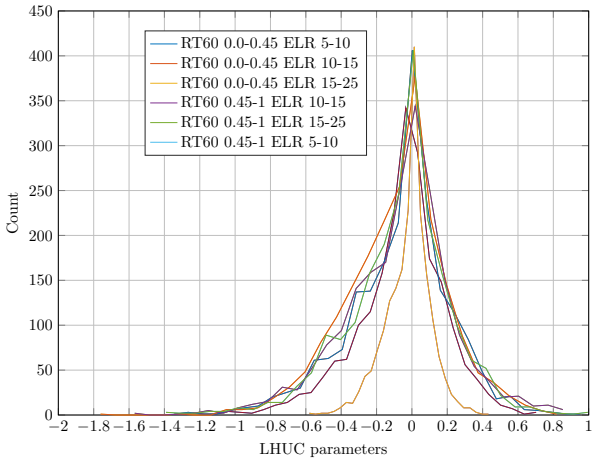



Figure 3: Histogram of LHUC parameters in the first layer.

The statistics of the learned LHUC parameters vary according to the reverberation condition as seen for the first layer LHUC parameters in Fig. 3. Most of the learning happens in the bottom layers. The statistics of LHUC parameters in the top layers are found to be similar for different reverberation conditions. An example histogram for the fourth layer can be as seen in Fig. 4. This observation is consistent with the ones reported in (Swietojanski et al., 2016).

The ASR results obtained using LHUC adaptation given knowledge of the true reverberation condition are summarized in Fig. 5(g). LHUC adaptation outperforms the baseline for all reverberation conditions.

### *6.2. Identifying reverberation conditions*

| Dataset used: $CS^{i}_{(tr,dt,et)}, i = 1 \ldots 6$ |
|---|

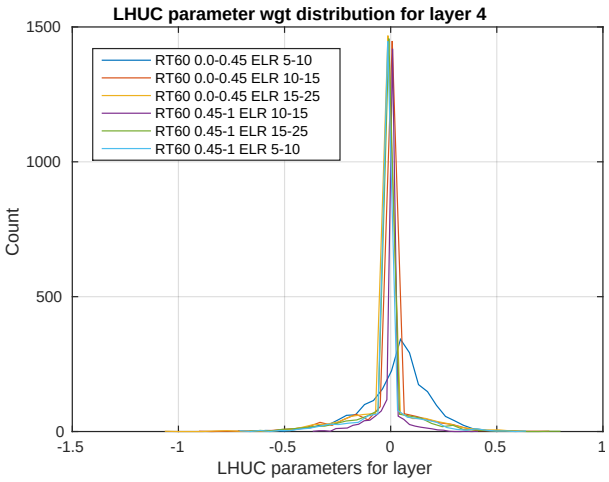



Figure 4: Histogram of LHUC parameters in the fourth layer.

In order to employ LHUC in a practical scenario when the true reverberation condition is unknown, the reverberation condition must be automatically identified. As mentioned above, i-vectors do encode some relevant information about the reverberant environment. Therefore, we use i-vectors computed on reverberated speech as features along with support vector machines (SVM)(Vapnik, 1998) as a classifier to identify the reverberation condition of each utterance.

The experiment was conducted with 7138 training utterances for each of the reverberation conditions. This gave us $7138 \times 6 = 42828$ utterances and implicitly, 42828 i-vectors for training. Similarly, 2460 i-vectors were used for validation and the 1980 i-vectors for testing. The i-vectors were computed as described in Section 6.3 below. A radial basis function was used as the kernel and the parameters $C = 512$ and $\gamma = 0.0078125$ were set using the validation set. LibSVM (Chang and Lin, 2011) was used to train the SVM. An accuracy of 53.3% was obtained. The confusion matrix is shown in Table 2. Good classification accuracy in terms of ELR and greater confusion in terms of RT can be observed. This is consistent with the range of ELR and RT values observed in our dataset and with the error rates (expressed in terms of the relative error in dB and %) reported by Eaton et al. (2015).

We then use these identified labels to pick up the respective LHUC adapted models for ASR decoding. The results obtained are shown in Fig. 5(h). Strangely, a small improvement in WER can be observed compared to Fig. 5(g), where real reverberation condition labels were used. Nevertheless the results are worse than the ASR system without LHUC adaptation for all reverb conditions.

### *6.3. I-vectors*

---

Dataset used: $MC^{1}_{(tr,dt,et)}$, $CS^{i}_{(et)}$, $i = 1 \ldots 6$

---

For comparison, we evaluate i-vector based adaptation, which is traditionally employed for speaker adaptation of DNN acoustic models in the context of ASR. The i-vector associated with each utterance represents a point in the subspace of speaker variability and it is appended to

Table 2: Confusion matrix (%) for reverberation condition classification. True conditions are in rows and estimated conditions in columns.

| | RT [<0.45] ELR [<10] | RT [<0.45] ELR [10-15] | RT [<0.45] ELR [>15] | RT [>0.45] ELR [<10] | RT [>0.45] ELR [10-15] | RT [>0.45] ELR [>15] |
|---|---|---|---|---|---|---|
| RT [<0.45] ELR [<10] | **8** | 0 | 0 | 58 | 34 | 0 |
| RT [<0.45] ELR [10-15] | 3 | **58** | 1 | 1 | 35 | 1 |
| RT [<0.45] ELR [>15] | 0 | 27 | **47** | 0 | 1 | 26 |
| RT [>0.45] ELR [<10] | 0 | 1 | 0 | **95** | 4 | 0 |
| RT [>0.45] ELR [10-15] | 0 | 16 | 0 | 2 | **81** | 2 |
| RT [>0.45] ELR [>15] | 0 | 37 | 1 | 0 | 32 | **30** |

Table 3: WER (%) for different model adaptation techniques. The results are obtained by averaging over $6 \times 330 = 1980$ utterances.

| I-vector (unknown reverb condition) | LHUC (known reverb condition) | LHUC (unknown reverb condition) |
|---|---|---|
| 7.36 | 7.25 | 7.00 |

the fMLLR features in all time frames while training the acoustic model. A 10% relative WER improvement was reported using i-vectors for speaker adaptation in (Saon et al., 2013). I-vectors are also known to enable adaptation to environmental acoustic conditions (Rouvier and Favre, 2014) and have been used previously to adapt to reverberant conditions (Alam et al., 2014) in the context of the REVERB challenge.

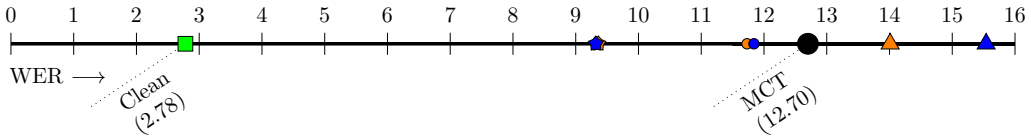
In our experiment, the universal background model (UBM) is built using a subset of multicondition training data. MFCCs are used as features to build the UBM. MFCC features normalized with CMVN along with 3 context features on either side are appended to form a single vector. The dimension of the features was reduced to 40 by linear discriminant analysis (LDA). The total variability matrix was learned using the full multicondition training set. I-vectors of dimension 100 are eventually appended to the acoustic features to train the DNN. The input dimension of the DNN is now 540 (100 i-vectors and $40 \times 11$ fMLLR). ASR results obtained using i-vector adaptation method are shown in Table 3. A WER of 7.36% was obtained which is worse than the ASR system without i-vectors. I-vectors seems to have a negative influence when used in combination with fMLLR features even though they were observed to improve performance when used in combination with filterbank features in similar settings (result not shown in the figure).

A summary comparison of the tested model adaptation techniques is given in Table 3. None of the adaptation schemes resulted in WER improvements with fMLLR features. For this reason, we do not combine LHUC or i-vectors adaptation methods with speech dereverberation. A snapshot of all the ASR performance obtained using all techniques on simulated data is shown in Fig. 5 for convenience. For easier comparison, the results on the simulated data are summarized in Table 4.

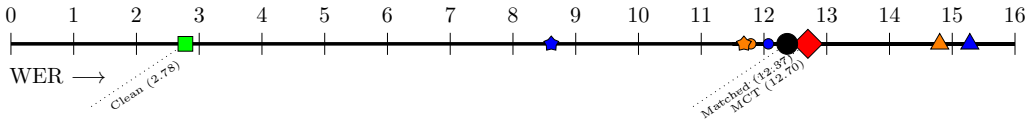
## 7. Experiments on real data

To confirm whether the techniques described above work in the case of real reverberated speech, we performed experiments on the real data of the REVERB challenge (Kinoshita et al., 2016). The challenge provides 1, 2 and 8 channel training, development and test data at a sampling rate of 16 kHz. The training set contains 7861 utterances of clean speech. Reverberation

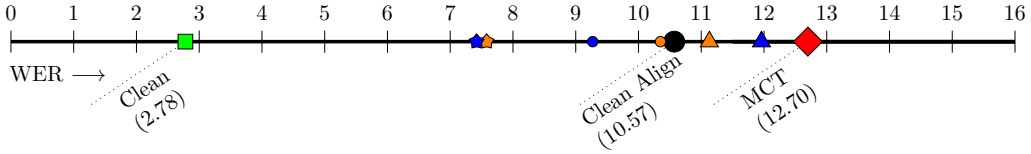
(a) Baseline: multicondition training with non-clean alignments and filterbank features.



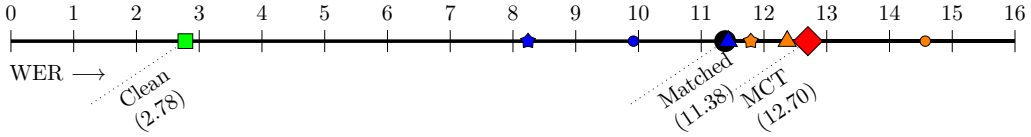

(b) Matched condition training with non-clean alignments and filterbank features.



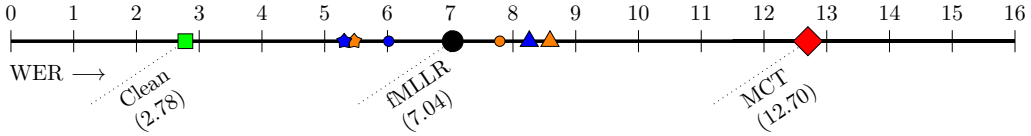

(c) Multicondition training with clean alignments and filterbank features.



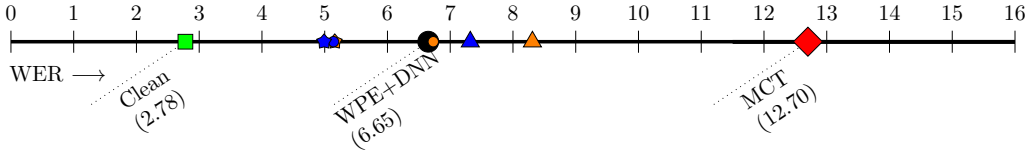

(d) Matched condition training with clean alignments and filterbank features.



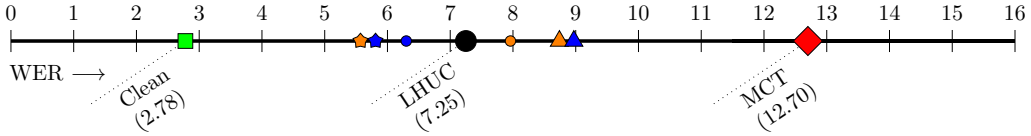

(e) Multicondition training with clean alignments and fMLLR features.




(f) Multicondition training with clean alignments, fMLLR features, and WPE+DNN dereverberation.




(g) Multicondition training with clean alignments, fMLLR features, and LHUC with known reverberation labels.




(h) Multicondition training with clean alignments, fMLLR features, and LHUC with unknown labels.

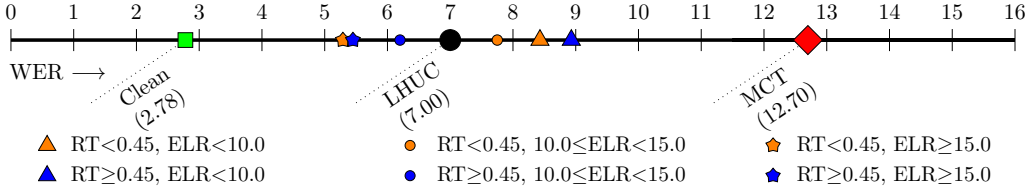



Figure 5: A snapshot of WER scores of all methods across all reverberation conditions. The figure shows WER markings on a single dimensional axis. The bigger marker in black is used to denote the average performance of the system while the smaller markers in different colors and shapes are used to mark the performance in different reverberation conditions. The green marker is used to denote the performance in clean conditions.

Table 4: Summary of results in simulated conditions.

| Index | Methods | WER |
|---|---|---|
| 1 | Multicondition | 12.70 |
| 2 | Matched condition | 12.37 |
| 3 | Multicondition (clean aligns) | 10.57 |
| 4 | 3 + fMLLR feats | 7.04 |
| 5 | 4 + WPE | 6.69 |
| 6 | 4 + Beamformit | 6.61 |
| 7 | 4 + DNN | 6.87 |
| 8 | 4 + Spectral Ench. | 6.97 |
| 9 | 6 +DNN | 6.59 |
| 10 | 5 + DNN | 6.65 |
| 11 | 3 + i-vectors | 7.36 |
| 12 | 4 + LHUC (known reverb) | 7.25 |
| 13 | 4+ LHUC (unknown reverb) | 7.00 |

and noise were simulated using these clean utterances as part of the REVERB dataset. The development and test sets contains both simulated and real reverberated noisy speech. The development set contains 1484 simulated utterances and 179 real utterances. In our experiments we use the full training set (7861 utterances) and the full development set (1663 utterances) to train the DNN acoustic model. Testing was carried out on the real part of the test set only, which contains 186 utterances with high reverberation conditions and 186 utterances with low reverberation conditions. No speaker information is provided during the testing phase.

A DNN based acoustic model was trained in a similar fashion as mentioned in Section 2.2. 40 dimensional fMLLR features were given as input. The DNN output dimension was 2079 which corresponds to the total number of senones. A trigram statistical LM was used as a language model.

The results obtained are shown in Table 5. A baseline WER of 78.54% was obtained using clean alignments and filterbanks as features. Advanced fMLLR features resulted in improved performance of 35.81% WER. It must be noted that no speaker information was provided as part of the REVERB challenge and hence the fMLLR features were computed per utterance instead of per speaker. Applying Beamformit improved the ASR performance to 28.20%, an improvement of 21.25% relative. A relative WER improvement of 18.3% was obtained using WPE based dereverberation which is slightly worse than using Beamformit. I-vectors in both the multicondition and WPE dereverberated setting, deteriorated the ASR performance as was observed in the simulated experiments. Finally, performance evaluation was conducted in mismatched conditions. Speech signals after WPE and Beamformit based dereverberations were decoded using an acoustic model trained on multicondition data. Relative WER improvements of 37.0% and 31.7% were observed, respectively. These are our best performing systems on the real test set of the REVERB challenge. The best performing systems of the REVERB challenge reported an average WER of 18.6% (across far and near field reverberations), under similar training condi-

Table 5: WER (%) on real data. All experiments are conducted using FMLLR as features.

| Train | Test | Far Condition | Near Condition | Average |
|---|---|---|---|---|
| Multi condition | Multi condition | 35.99 | 35.64 | 35.81 |
| Beamformit | Beamformit | 29.07 | 27.34 | 28.20 |
| WPE | WPE | 27.68 | 31.97 | 29.82 |
| Multi condition with i-vectors | Multi condition with i-vectors | 41.73 | 38.33 | 40.03 |
| WPE with i-vectors | WPE with i-vectors | 33.29 | 35.96 | 34.62 |
| **Multi condition** | **WPE** | **20.83** | **24.34** | **22.58** |
| Multi condition | Beamformit | 24.27 | 24.59 | 24.43 |

tions (8 channels, no boosting of dataset size). Given that we do not use sequence-discriminative training and RNN language modeling which were used in the top performing systems, the results obtained in this paper are competitive.

## 8. Conclusion

We presented a study of various techniques to tackle the ASR problem in a wide range of reverberation conditions, considering both established techniques and new techniques whose performance in such conditions had not been analyzed so far. We started off by creating a baseline for our dataset. The dataset was created by corrupting clean speech with real impulse responses. Six different reverberation classes were defined by drawing boundaries using the RT and ELR range. The performance of different techniques was computed in each of these reverberation conditions.

Matched condition training did not improve the ASR performance compared to multicondition training. Improvements in ASR performance were observed when clean alignments were used as ground truth to train the acoustic model. Switching features from filterbanks to fMLLR gave a huge gain in ASR performance. WPE, beamforming, spectral enhancement and DNN based dereverberation improved the ASR performance whereas adaptation methods such as LHUC and i-vectors did not provide any gains. Combining clean alignment of training data, fMLLR features and WPE/beamforming based dereverberation gave an overall relative improvement of 44% on the simulated test set. A significant improvement was similarly observed across all reverberation conditions. Experiments were also performed using the REVERB dataset where the test set contained speech recorded in real reverberant conditions with the acoustic model trained using simulated data. Combining clean alignment of training data, fMLLR features and WPE based dereverberation under mismatched training conditions gave a relative improvement of 37.0%. The obtained results were close to the best performing systems on REVERB challenge under similar settings.

Though the improvement is significant we consider the ASR for reverberated speech problem as unsolved. This is because of a relative WER difference of 59.0% between the performance with clean speech and our best performing combination of techniques on the simulated dataset. Further research on dereverberation and model adaptation is needed.

## 9. Acknowledgments

We acknowledge the support of Bpifrance (FUI voiceHome). Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`). The authors would also like to thank Pawel Swietojanski and co-authors for sharing the LHUC code, Takuya Yoshioka and co-authors for sharing the WPE code, and Aditya Arie Nugraha for useful discussions. We thank Ziteng Wang for running our experiments on the REVERB dataset. Finally, we thank the editor and anonymous reviewers for carefully reading our manuscript and offering valuable suggestions.

## 10. References

Abdelaziz, A. H., Watanabe, S., Hershey, J. R., Vincent, E., Kolossa, D., 2015. Uncertainty propagation through deep neural networks. In: Interspeech. pp. 3561–3565.

Abrash, V., Franco, H., Sankar, A., Cohen, M., 1995. Connectionist speaker normalization and adaptation. In: Eurospeech. pp. 2183–2186.

Alam, M. J., Gupta, V., Kenny, P., Dumouchel, P., 2014. Use of multiple front-ends and i-vector based speaker adaptation for robust speech recognition. In: Proceedings of REVERB Challenge.

Anguera, X., Wooters, C., Hernando, J., 2007. Acoustic beamforming for speaker diarization of meetings. IEEE Transactions on Audio, Speech, and Language Processing 15 (7), 2011–2022.

Barker, J., Marxer, R., Vincent, E., Watanabe, S., dec 2015. The third 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines. In: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015). Scottsdale, AZ, United States, pp. 504–511.

Bertin, N., Camberlein, E., Vincent, E., Lebarbenchon, R., Peillon, S., Lamandé, É., Sivasankaran, S., Bimbot, F., Illina, I., Tom, A., Fleury, S., Jamet, É., 2016. A French corpus for distant-microphone speech processing in real homes. In: Interspeech. San Francisco, United States, pp. 2781–2785.

Brutti, A., Matassoni, M., 2014. On the use of Early-to-Late Reverberation Ratio for ASR in reverberant environments. In: Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on. IEEE, pp. 4638–4642.

Brutti, A., Matassoni, M., 2016. On the relationship between Early-to-Late Ratio of Room Impulse Responses and ASR performance in reverberant environments. Speech Communication 76, 170–185.

Cauchi, B., Kodrasi, I., Rehr, R., Gerlach, S., Jukić, A., Gerkmann, T., Doclo, S., Goetze, S., 2015. Combination of MVDR beamforming and single-channel spectral processing for enhancing noisy and reverberant speech. EURASIP Journal on Advances in Signal Processing 2015 (1), 1–12.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2 (3), 1–27.

Dehak, N., Dehak, R., Kenny, P., Brümmer, N., Ouellet, P., Dumouchel, P., 2009. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In: Interspeech. Vol. 9. pp. 1559–1562.

Delcroix, M., Yoshioka, T., Ogawa, A., Kubo, Y., Fujimoto, M., Ito, N., Kinoshita, K., Espi, M., Araki, S., Hori, T., Nakatani, T., 2015. Strategies for distant speech recognitionin reverberant environments. EURASIP Journal on Advances in Signal Processing 2015 (1), 1–15.
URL `http://dx.doi.org/10.1186/s13634-015-0245-7`

Eaton, J., Gaubitch, N., Moore, A., Naylor, P., 2015. The {ACE} challenge — Corpus description and performance evaluation. In: Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015 IEEE Workshop on. IEEE, pp. 1–5.

Falk, T. H., Chan, W.-Y., 2010. Modulation spectral features for robust far-field speaker identification. Audio, Speech, and Language Processing, IEEE Transactions on 18 (1), 90–100.

Fiscus, J. G., 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In: Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on. IEEE, pp. 347–354.

Gales, M. J. F., 1998. Maximum likelihood linear transformations for HMM based speech recognition. Computer Speech & Language 12 (2), 75–98.

Ganapathy, S., Pelecanos, J., Omar, M. K., 2011. Feature normalization for speaker verification in room reverberation. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on. IEEE, pp. 4836–4839.

Gannot, S., Moonen, M., 2003. Subspace methods for multimicrophone speech dereverberation. EURASIP Journal on Applied Signal Processing 2003, 1074–1090.

Garofalo, J., Graff, D., Paul, D., Pallett, D., 2007. CSR-I (WSJ0) Complete, linguistic Data Consortium, Philadelphia.
Gillespie, B. W., Malvar, H. S., Florêncio, D. A. F., 2001. Speech dereverberation via maximum-kurtosis subband adaptive filtering. In: Acoustics, Speech, and Signal Processing, IEEE International Conference on. Vol. 6. IEEE, pp. 3701–3704.
Habets, E. A. P., Gannot, S., Cohen, I., 2009. Late reverberant spectral variance estimation based on a statistical model. Signal Processing Letters, IEEE 16 (9), 770–773.
Han, K., Wang, Y., Wang, D., Woods, W. S., Merks, I., Zhang, T., 2015. Learning spectral mapping for speech dereverberation and denoising. Audio, Speech, and Language Processing, IEEE/ACM Transactions on 23 (6), 982–992.
Harper, M., dec 2015. The Automatic Speech recogition In Reverberant Environments (ASpIRE) challenge. In: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). pp. 547–554.
Hermansky, H., Morgan, N., 1994. RASTA processing of speech. Speech and Audio Processing, IEEE Transactions on 2 (4), 578–589.
Hinton, G., Deng, L., Yu, D., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Dahl, G., Kingsbury, B., nov 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine 29 (6), 82–97.
Hsiao, R., Ma, J., Hartmann, W., Karafiát, M., Grézl, F., Burget, L., Szöke, I., Černocký, J. H., Watanabe, S., Chen, Z., Mallidi, S. H., Hermansky, H., Tsakalidis, S., Schwartz, R., dec 2015. Robust speech recognition in unknown reverberant and noisy conditions. In: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU). pp. 533–538.
Joshi, V., Bilgi, R., Umesh, S., Garcia, L., Benitez, C., 2015. Sub-band based histogram equalization in cepstral domain for speech recognition. Speech Communication 69, 46–65.
Kingsbury, B. E. D., Morgan, N., Greenberg, S., 1998. Robust speech recognition using the modulation spectrogram. Speech communication 25 (1), 117–132.
Kinoshita, K., Delcroix, M., Gannot, S., Habets, E. A. P., Haeb-Umbach, R., Kellermann, W., Leutnant, V., Maas, R., Nakatani, T., Raj, B., Others, 2016. A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research. EURASIP Journal on Advances in Signal Processing 2016 (1), 1–19.
Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Sehr, A., Kellermann, W., Maas, R., 2013. The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech. In: Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on. IEEE, pp. 1–4.
Kuttruff, H., 2009. Room Acoustics, 5th Edition. CRC Press.
Lebart, K., Boucher, J.-M., Denbigh, P. N., 2001. A new method based on spectral subtraction for speech dereverberation. Acta Acustica united with Acustica 87 (3), 359–366.
Li, J., Deng, L., Gong, Y., Haeb-Umbach, R., 2013. An overview of noise-robust automatic speech recognition. Audio, Speech, and Language Processing, IEEE/ACM Transactions on 22 (4), 745–777.
Loesch, B., Yang, B., 2010. Adaptive segmentation and separation of determined convolutive mixtures under dynamic conditions. In: Latent Variable Analysis and Signal Separation. Springer, pp. 41–48.
Molau, S., Hilger, F., Ney, H., 2003. Feature space normalization in adverse acoustic conditions. In: Acoustics, Speech, and Signal Processing, IEEE International Conference on. Vol. 1. IEEE, pp. 656–659.
Nakatani, T., Yoshioka, T., Kinoshita, K., Miyoshi, M., Juang, B.-H., 2008. Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation. In: Acoustics, Speech and Signal Processing, IEEE International Conference on. IEEE, pp. 85–88.
Naylor, P. A., Gaubitch, N. D. (Eds.), 2010. Speech Dereverberation. Springer.
Neto, J., Almeida, L., Hochberg, M., Martins, C., Nunes, L., Renals, S., Robinson, T., 1995. Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system. In: Eurospeech. pp. 2171–2174.
Nishiura, T., Hirano, Y., Denda, Y., Nakayama, M., 2007. Investigations into early and late reflections on distant-talking speech recognition toward suitable reverberation criteria. In: Interspeech. Antwerp, Belgium, pp. 1082–1085.
Peddinti, V., Chen, G., Manohar, V., Ko, T., Povey, D., Khudanpur, S., 2015a. JHU ASpIRE system: Robust LVCSR with TDNNs, iVector adaptation and RNN-LMs. In: Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on. pp. 539–546.
Peddinti, V., Chen, G., Povey, D., Khudanpur, S., 2015b. Reverberation robust acoustic modeling using i-vectors with time delay neural networks. In: Interspeech. pp. 2440–2444.
Petrick, R., Lohde, K., Wolff, M., Hoffmann, R., 2014. The harming part of room acoustics in automatic speech recognition. In: Interspeech. pp. 1094–1097.
Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Others, 2011. The Kaldi speech recognition toolkit. In: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding. IEEE Signal Processing Society.
Rouvier, M., Favre, B., 2014. Speaker adaptation of DNN-based ASR with i-vectors: Does it actually adapt models to speakers? In: Interspeech. pp. 3007–3011.
Saon, G., Soltau, H., Nahamoo, D., Picheny, M., 2013. Speaker adaptation of neural network acoustic models using i-

vectors. In: Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, pp. 55–59.
Sehr, A., Barfuss, H., Hofmann, C., Maas, R., Kellermann, W., may 2014. Efficient training of acoustic models for reverberation-robust medium-vocabulary automatic speech recognition. In: Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on. pp. 177–181.
Sivasankaran, S., Nugraha, A. A., Vincent, E., Morales Cordovilla, J. A., Dalmia, S., Illina, I., Liutkus, A., 2015. Robust ASR using neural network based speech enhancement and feature simulation. In: 2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015). Arizona, United States, pp. 482–489.
Swietojanski, P., Li, J., Renals, S., 2016. Learning hidden unit contributions for unsupervised acoustic model adaptation. IEEE/ACM Transactions on Audio, Speech and Language Processing (Submitted).
Swietojanski, P., Renals, S., 2014. Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models. In: Spoken Language Technology Workshop (SLT). IEEE, pp. 171–176.
Vapnik, V. N., 1998. Statistical Learning Theory. Wiley, New York.
Viikki, O., Bye, D., Laurila, K., 1998. A recursive feature vector normalization approach for robust speech recognition in noise. In: Acoustics, Speech and Signal Processing. Proceedings of the IEEE International Conference on. Vol. 2. IEEE, pp. 733–736.
Virtanen, T., Singh, R., Raj, B., 2012. Techniques for noise robustness in automatic speech recognition. John Wiley & Sons.
Weninger, F., Watanabe, S., Tachioka, Y., Schuller, B., 2014. Deep recurrent de-noising auto-encoder and blind de-reverberation for reverberated speech recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, pp. 4623–4627.
Wölfel, M., McDonough, J., 2009. Distant Speech Recognition. Wiley.
Yao, K., Yu, D., Seide, F., Su, H., Deng, L., Gong, Y., 2012. Adaptation of context-dependent deep neural networks for automatic speech recognition. In: Spoken Language Technology Workshop (SLT), 2012 IEEE. IEEE, pp. 366–369.
Yoshioka, T., Nakatani, T., Miyoshi, M., 2009. Integrated speech enhancement method using noise suppression and dereverberation. Audio, Speech, and Language Processing, IEEE Transactions on 17 (2), 231–246.
Yoshioka, T., Nakatani, T., Miyoshi, M., Okuno, H. G., 2011. Blind separation and dereverberation of speech mixtures by joint optimization. Audio, Speech, and Language Processing, IEEE Transactions on 19 (1), 69–84.
Yoshioka, T., Sehr, A., Delcroix, M., Kinoshita, K., Maas, R., Nakatani, T., Kellermann, W., nov 2012. Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition. Signal Processing Magazine, IEEE 29 (6), 114–126.