# LINEAR PREDICTION-BASED DEREVERBERATION WITH ADVANCED SPEECH ENHANCEMENT AND RECOGNITION TECHNOLOGIES FOR THE REVERB CHALLENGE

*Marc Delcroix, Takuya Yoshioka, Atsunori Ogawa, Yotaro Kubo, Masakiyo Fujimoto, Nobutaka Ito, Keisuke Kinoshita, Miquel Espi, Takaaki Hori, Tomohiro Nakatani, Atsushi Nakamura*

NTT Communication Science Laboratories, NTT Corporation, Japan, marc.delcroix@lab.ntt.co.jp

## ABSTRACT

This paper describes systems for the enhancement and recognition of distant speech recorded in reverberant rooms. Our speech enhancement (SE) system handles reverberation with blind deconvolution using linear filtering estimated by exploiting the temporal correlation of observed reverberant speech signals. Additional noise reduction is then performed using an MVDR beamformer and advanced model-based SE. We employ this SE system as a front-end for our advanced automatic speech recognition (ASR) back-end, which uses deep neural network (DNN) based acoustic models and recurrent neural network based language models. Moreover, we ensure good interconnection between the SE front-end and ASR back-end using unsupervised model adaptation to reduce the mismatch caused by, for example, front-end processing artifacts. Our SE front-end greatly improves speech quality and achieves up to a 60 % relative word error rate reduction for the real recordings of the REVERB challenge data, compared with a strong DNN-based ASR baseline.

***Index Terms***— Linear prediction-based dereverberation, model-based speech enhancement, DNN-based recognition.

## 1. INTRODUCTION

The use of distant microphones to capture speech remains challenging because noise and reverberation degrade the audible quality of speech and severely affect the performance of automatic speech recognition (ASR). Much research has been undertaken to tackle the effect of noise. However, dealing with reverberation has remained challenging because it has a long-term effect that covers several analysis time frames, and it induces highly non-stationary distortions. Consequently, mitigating reverberation requires dedicated approaches that exploit the long-term acoustic context and use efficient models of reverberation [1]. Such approaches differ fundamentally from conventional noise reduction techniques.

This paper presents our contribution to the REVERB challenge for the enhancement and recognition of distant speech recorded in reverberant rooms [2]. The REVERB challenge data cover various reverberation conditions (reverberation times between 0.25 and 0.7 s) and also include a significant amount of noise. Dealing with such severe conditions requires powerful dereverberation and noise reduction techniques. Our system combines speech enhancement (SE) techniques as a front-end to reduce reverberation and noise, and a state-of-the-art ASR back-end for optimal recognition performance. The front-end of our system exploits the *time* and *spatial* correlation of reverberant speech as well as clean speech *spectrum* characteristics, using a combination of SE processes. Moreover, we ensure good interconnection of the SE front-end and ASR back-end using unsupervised model adaptation to compensate for the mismatch caused by, for example, front-end processing artifacts.

A central part of our SE front-end consists of robust blind deconvolution based on long-term linear prediction, which aims at late reverberation reduction. The long-term effect of reverberation causes the *long-term time correlation* of the reverberant speech that can be exploited to estimate the late reverberation components using the weighted prediction error (WPE) algorithm [3, 4, 5]. This approach can be applied to single or multi-microphone cases and is very effective for reverberation suppression and robust to ambient noise. To reduce ambient noise and potential remaining reverberation components, we use a beamformer that employs the *spatial correlation* of the microphone array signals [6]. Finally, we further reduce noise using SE methods that rely on pre-trained *clean speech spectral models* [7, 8, 9]. Both our dereverberation and beamforming techniques employ *linear filtering* that guarantees low speech distortion. Moreover, model-based SE guarantees that the noise reduction is realized while keeping the enhanced speech spectrum characteristics close to that of clean speech. Consequently, our SE front-end greatly reduces reverberation and noise and improves both speech perceptual quality and ASR performance. Our SE front-end principally targets multi-channel tasks (2 channels (2ch) and 8 channels (8ch)) but we also provide some ASR results for the single channel (1ch) task.

To achieve good recognition performance, we use a state-of-the-art ASR back-end that consists of deep neural network (DNN) acoustic models (AMs) [10, 11] and recurrent neural network (RNN) based language models (LMs) [12, 13]. One issue with the REVERB challenge is that the provided multi-condition training data (trainData) is fairly similar to the simulated test data (SimData), but quite different from the real recordings (RealData) that are more severe in terms of noise and reverberation. DNNs are known to perform poorly when test conditions differ significantly from the training conditions [14]. Consequently, increasing the performance on the RealData set is particularly challenging. We tackle this issue by increasing the robustness of the DNNs to unseen conditions. Several approaches have been proposed for increasing the robustness of DNNs [15]. Here we simply augment the acoustic variations of the trainData set to expose the DNN-based AM to a larger variation of training samples. Moreover, we used unsupervised AM adaptation [16] to further compensate for the mismatch between test and training conditions as well as the effect of the processing artifacts introduced by the SE front-end.

We demonstrate the efficiency of the proposed front-end and back-end techniques with the REVERB challenge data [2], both for SE and ASR tasks. In particular, with our best set-up we achieve average word error rates (WER) of 4.2 % and 9.0 % for the SimData and RealData of the evaluation set of the challenge, respectively. Although we use a strong baseline that has already achieved high recognition performance with unprocessed distant speech, we obtain a large additional improvement using the proposed front-end (up to 60% relative WER reduction). This demonstrates that well

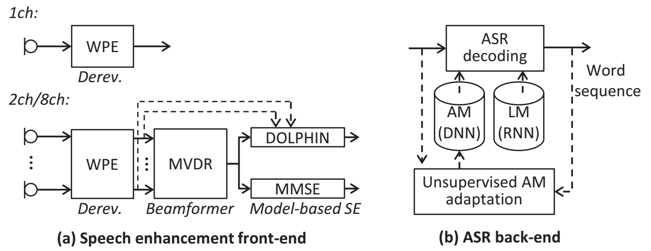

**Fig. 1**. Schematic diagram of the proposed system for enhancement and recognition of reverberant speech.

engineered SE front-ends still have a large impact when using DNN AMs, which may contrast with the results of some previous studies [15].

In the remainder of this paper, we provide a brief overview of our proposed system in Section 2, and discuss the components of the SE front-end and ASR back-end in Sections 3 and 4, respectively. Detailed experimental results are provided in Section 5. Finally, we conclude the paper in Section 6.

## 2. SYSTEM OVERVIEW

Figure 1 shows a schematic diagram of the proposed (a) SE front-end and (b) ASR back-end for 1ch, 2ch and 8ch set-ups. Note that we focus on the multi-microphone cases and the 1ch front-end is used only for recognition. The system consists of the following elements.

**SE front-end**

- *Dereverberation based on long-term linear prediction:* We use the WPE algorithm to reduce late reverberation. WPE inputs 1ch, 2ch and 8ch observed signals and outputs the same number of dereverberated signals. WPE operates in the utterance batch mode. The real time factor (RTF) [1] of WPE is about 0.2, 0.5 and 2.8 for 1ch, 2ch and 8ch set-ups, respectively.

- *Beamformer:* Ambient noise and potential remaining reverberation components are reduced using a minimum variance distortionless response (MVDR) beamformer. The MVDR beamformer inputs the multi-channel signals dereverberated with WPE (2ch/8ch) and outputs a single-channel enhanced speech signal. MVDR operates in the utterance batch mode. The MVDR run time is negligible with RTFs of 0.01 and 0.03 for 2ch and 8ch set-ups, respectively.

- *Model-based SE:* We investigated two model-based SE approaches to further reduce noise, namely *dominance based locational and power-spectral characteristics integration* (DOLPHIN) and *model-based SE with minimum mean squared error (MMSE) estimates*. Both approaches use pre-trained spectral models of clean speech, trained using the clean speech training data set.

  DOLPHIN uses both the multi-channel output of WPE and the single channel output of MVDR to perform enhancement. It operates in the full batch mode (using all the data from a given test condition). The RTFs of DOLPHIN are about 6.1 and 10.5 for 2ch and 8ch set-ups, respectively.

  MMSE model-based SE uses only the output of the MVDR beamformer and operates in the utterance batch mode. Its RTF is about 0.5.

**ASR back-end**: We recognize the output of WPE for 1ch, and DOLPHIN for 2ch/8ch[2]. The RTF of the ASR decoding is about 6.

- *Acoustic model:* we used state-of-the-art DNN-based AMs. To create robust AMs, we augmented the amount of multi-condition training data to increase the acoustic conditions seen during training.

- *Unsupervised model adaptation:* The DNN AMs are adapted to the test environment to reduce the mismatch between the test and training conditions. Unsupervised adaptation is performed in the full batch mode to obtain a sufficient quantity of data to reliably estimate the AM parameters.

- *Language model:* we employed a state-of-the-art RNN-based LM with an on-the-fly rescoring technique to allow fast one pass decoding.

The following sections describe each component of our proposed system in more detail.

## 3. SPEECH ENHANCEMENT FRONT-END

### 3.1. Dereverberation based on long-term linear prediction

Dereverberation is the key component of our proposed SE front-end. We performed dereverberation based on long-term linear prediction in the short-time Fourier transform (STFT) domain by using the WPE algorithm, which was first described in [3] for a two-microphone one-output case and generalized later in [4, 5]. The nature of this algorithm and the relationship with other approaches are discussed in [1]. It is also noteworthy that this algorithm has been shown to improve the meeting transcription performance of DNN AMs trained on nearly matched training data [17]. In the following, we first describe the single-channel version of the algorithm and then extend it to the $M$-input $M$-output case to highlight the commonalities and differences between the single and multi-channel versions.

Since dereverberation processing acts on STFT coefficients, a single-channel observed signal $y(t)$, which is distorted by reverberation and background noise, is transformed into a set of (complex-valued) STFT coefficients $(y_n)_{n=1,\cdots,N}$ with $N$ being the number of time frames in an utterance. We have omitted a frequency bin index since different frequency bins are processed independently. The goal of dereverberation is to obtain a set of STFT coefficients $(x_n)_{n=1,\cdots,N}$, which is less reverberant than the input.

With the long-term linear prediction approach, dereverberation is achieved using a frequency-dependent linear prediction filter as follows:

$$x_n = y_n - \sum_{\tau=T_\perp}^{T_\top} g_\tau^* y_{n-\tau}, \qquad (1)$$

where $*$ stands for complex conjugation. With this formulation, the reverberant noise contained in $y_n$ is predicted from the past frames of observed speech, $y_{n-T_\perp}, \cdots, y_{n-T_\top}$, and then subtracted from $y_n$ to obtain an estimate of the dereverberated STFT coefficient. $T_\perp$ is normally set at 3 while $T_\top$ has a large value to deal with long-term reverberation (between 7 and 40). $G = (g_{T_\perp}, \cdots, g_{T_\top})$ is the set of prediction coefficients to be optimized, which is defined independently for each frequency bin. It is known that using a $T_\perp$ value

[1] All RTFs are calculated on modern CPUs (e.g. Intel Xeon, 2.6 GHz, using a Linux operating system)

[2] MMSE was not used for recognition as it performed slightly worse than DOLPHIN with our DNN-based ASR back-end.

greater than 1 prevents the processed speech from being excessively whitened while it leaves the early reflection distortion in the processed speech [18].

Using the concept of WPE minimization [5], the linear predictor $G$ can be optimized to minimize the following objective function, which can be derived assuming that the prediction error is Gaussian with a time-varying variance ($\theta_n$) corresponding to the short-time power spectrum of the source at a given frequency:

$$F_1 = \sum_{n=1}^{N} \left( \frac{\left| y_n - \sum_{\tau=T_\perp}^{T_\top} g_\tau^* y_{n-\tau} \right|^2}{\theta_n} + \log \theta_n \right), \tag{2}$$

where $\Theta = (\theta_1, \cdots, \theta_N)$ is a set of auxiliary variables that needs to be optimized jointly with $G$, which leads to interleaved updates of $G$ and $\Theta$. Each $\theta_n$ is updated simply by calculating $\theta_n = \left| y_n - \sum_{\tau=T_\perp}^{T_\top} g_\tau^* y_{n-\tau} \right|^2$ for a fixed $G$. Using notation $\boldsymbol{g} = [g_{T_\perp}, \cdots, g_{T_\top}]^T$, where the superscript $T$ indicates a non-conjugate transpose operation, $G$ can be updated as $\boldsymbol{g} = \boldsymbol{R}^{-1}\boldsymbol{r}$, where $\boldsymbol{R}$ and $\boldsymbol{r}$ are given by the following equations:

$$\boldsymbol{R} = \sum_{n=1}^{N} \frac{\vec{y}_{n-T_\perp} \vec{y}_{n-T_\perp}^H}{\theta_n}, \quad \boldsymbol{r} = \sum_{n=1}^{N} \frac{\vec{y}_{n-T_\perp} y_n^*}{\theta_n} \tag{3}$$

with the superscript $H$ representing a conjugate transposition and $\vec{y}_n$ being defined as $\vec{y}_n = [y_n, \cdots, y_{n-T_\top+T_\perp}]^T$. Two or three iterations provide good estimates and can be executed at a small computational cost.

The above-described single-channel dereverberation algorithm can be easily extended to $M$-microphones, with $M \geq 2$, by rewriting (1) in the form of a multi-channel linear prediction as follows:

$$\boldsymbol{x}_n = \boldsymbol{y}_n - \sum_{\tau=T_\perp}^{T_\top} \boldsymbol{G}_\tau^H \boldsymbol{y}_{n-\tau}, \tag{4}$$

where $\boldsymbol{y}_n$ denotes an $M$-dimensional vector of the STFT coefficients obtained from the $M$ microphones and $\boldsymbol{x}_n$ denotes a dereverberated STFT coefficient vector. Each prediction coefficient $g_n$ has been changed to an $M$-by-$M$ prediction matrix $\boldsymbol{G}_n$ to accept the multiple inputs and produce the same number of outputs.

The objective function for minimization must also be modified accordingly. [5] derives the following objective function, which reduces to the single-channel objective function (2) when $M = 1$:

$$F_M = \sum_{n=1}^{N} \left( \left\| \boldsymbol{y}_n - \sum_{\tau=T_\perp}^{T_\top} \boldsymbol{G}_\tau^* \boldsymbol{y}_{n-\tau} \right\|_{\boldsymbol{\Lambda}_n}^2 + \log \det \boldsymbol{\Lambda}_n \right), \tag{5}$$

where, for vector $\boldsymbol{x}$ and matrix $\boldsymbol{\Lambda}$, $\|\boldsymbol{x}\|_{\boldsymbol{\Lambda}}^2 = \boldsymbol{x}^H \boldsymbol{\Lambda}^{-1} \boldsymbol{x}$. [5] describes how to efficiently optimize the set of prediction matrices so that (5) is (locally) minimized.

### 3.2. MVDR Beamforming

To suppress background noise (and possibly residual reverberation), MVDR beamforming was applied to the dereverberated signals for the multi-microphone set-up. As a result we obtain a single-channel speech signal, which is less distorted by background noise and reverberation than the input dereverberated signals.

In this work, the MVDR beamformer was implemented as described in [6]. This implementation is suitable for the REVERB Challenge task since it does not require explicit transfer functions between a target speaker and microphones, which change from utterance to utterance in the task being considered.

Instead of relying explicitly on the transfer functions, our beamformer needs a noise covariance matrix for each frequency bin. These statistics can be computed from the initial and final 10 frames of each utterance, from which speech sounds are assumed to be absent. See [6] (in particular, Eq. (24)) for details of the beamforming algorithm.

### 3.3. DOLPHIN

DOLPHIN is a model-based multi-channel SE technique that we use here to reduce residual ambient noise. DOLPHIN efficiently combines conventional direction-of-arrival (DOA) [19, 20] feature based enhancement [21] and spectral feature based approaches through a dominant source index (DSI) that indicates whether noise or speech is dominant at each time/frequency bin. The algorithm is detailed in [7]. Here we briefly explain its use with the REVERB challenge data.

DOLPHIN uses DOA feature models and spectral models of speech and noise to determine the DSI by using the expectation-maximization (EM) algorithm. The DOA feature models consist of a mixture of Watson distributions, whose parameters are learned on a per utterance basis. The speech DOA feature model is learned from the dereverberated speech DOA obtained from the WPE output. To obtain the noise DOA feature model, we assume that ambient noise is diffusive and therefore the distribution of the DOA features of the ambient noise is approximately the same as that of late reverberation. Given this assumption, we can approximate the noise DOA feature model parameters using the estimated late reverberation components that are obtained as a side output of WPE.

The speech and noise spectral models consist of Gaussian mixture models. The speech spectral model is trained using the clean speech training data provided by the challenge. Then to reduce the mismatch between training and test conditions, unsupervised channel adaptation is performed using all the utterances of a given test condition (full batch mode) following the procedure described in [7]. We use the MVDR output to calculate the adaptation parameters. The noise spectral model parameters are estimated on a per utterance basis. Finally, we perform noise reduction on the MVDR output.

DOLPHIN is used for 2ch and 8ch SE and in our SE front-end for recognition.

### 3.4. Model-based SE with MMSE estimates

In this section, we briefly describe the principle of our proposed single channel model-based SE with the joint processing of noise model estimation and speaker adaptation [8, 9], which we use here to reduce residual ambient noise remaining in the MVDR output signal. The method provides an alternative to DOLPHIN and has the merit of operating in the utterance-batch mode.

Most techniques for model-based SE, e.g. vector Taylor series (VTS) [22], create a noisy speech observation model by combining clean speech and noise models through an approximated observation model. With such an approximated observation model, accurate noise model parameter estimation is a challenge. In addition, variation of the speaker characteristics requires speaker adaptation of the clean speech model to ensure good noise suppression performance. However, the joint estimation of noise and speaker adaptation parameters is computationally intractable due to the direct unobservability of clean speech and noise signals with conventional techniques.

To overcome these issues, we propose a way of achieving joint unsupervised processing by using MMSE estimates of clean speech and noise [8]. First a rough observation model is created using VTS approximation to combine speech and noise models. This observation model is used to obtain MMSE estimates of the clean speech and noise signals. These signals are then used to calculate precisely speech and noise statistics that can be employed to refine the observation model. This recursive procedure is formulated with the EM algorithm. Ten iterations of the EM algorithm are generally sufficient to obtain good performance.

MMSE estimates of clean speech and noise include some estimation errors that often degrade the parameter estimation accuracy. Namely, in a period with a high segmental signal to noise ratio (SNR), the MMSE estimates of the clean speech signal become highly reliable, whereas the MMSE estimates of the noise signals become unreliable, and vice versa. Thus, it is desirable to eliminate unreliable estimates if we are to obtain accurate parameters for the noise model and speaker adaptation. To deal with this problem, we employ a reliable data selection method based on voice activity detection that consists of segmental SNR-based feature and k-means clustering [9]. This process implies that the model-based MMSE SE approach operates in the utterance batch mode.

## 4. ASR BACK-END

### 4.1. DNN-based acoustic model

We used a conventional context-dependent (CD) DNN-HMM based AM, obtained with layer-wise RBM pre-training followed by fine-tuning using backpropagation [10, 11]. We used log mel filter-bank coefficients as DNN input features. The multi-condition training data provided by the REVERB challenge are similar to the SimData set, but present a large mismatch with the RealData set. In particular the SNR of the training data is fixed at 20 dB. Therefore, obtaining good performance on the RealData set is a challenge. We address this issue by extending the training data set to cover more environmental variations, i.e. by using multi-condition training data obtained with various SNR levels without using any SE front-end. Here, to obtain a robust AM, we do not use the SE front-end to preserve the acoustic variations during training as it has been shown that enhancing the training data may degrade the DNN performance [15].

### 4.2. Unsupervised AM adaptation

There is a large mismatch between the training and testing conditions because of the difference in the acoustic conditions and also because we do not use an SE front-end during training. Unsupervised AM adaptation can be used to mitigate such a mismatch. There have been only few investigations of the adaptation of DNN-based AMs [23, 16, 24]. A simple but efficient approach consists of performing a few additional fine tuning steps (with backpropagation) on the adaptation data, using labels estimated from a first recognition pass [16]. This technique has been investigated for speaker adaptation where it was demonstrated that retraining the whole network would provide better performance improvement compared with retraining only the input or the output layers. We propose using a similar approach for environmental adaptation. Here we perform unsupervised full batch adaptation (using all the data from a given test condition), which implies environmental adaptation with only limited speaker adaptation as the adaptation data cover several speakers. In contrast to [16], for environmental adaptation, we confirmed experimentally the superiority of adapting only the input layer.

### 4.3. RNN-based language modeling

Diverse acoustic environments induce an acoustic mismatch between a training set and evaluation sets. In such a case, an improved language modeling technique is expected to be helpful since the linguistic characteristics can be considered invariant with respect to acoustic environment variations if the use case of the system does not change.

RNN-LMs [12] enhanced with a one-pass decoding technique based on an on-the-fly rescoring strategy [13] is a good choice for improving LM accuracy since it can accurately capture the long-term dependency between words without greatly increasing computational costs. To estimate the RNN-LMs, we prepared text data sets by extracting sentences from the WSJ text corpora [25] distributed by the Linguistic Data Consortium (LDC), while ensuring that the sentences in the evaluation and development sets were not employed. The training data set for RNN-LM consists of 716,951 sentences.

Following [12], we interpolated the optimized RNN-LMs with conventional trigram LMs to enhance the word prediction performance. We confirmed that the RNN-LMs were capable of greatly reducing perplexities, i.e. the development set perplexities of the trigram LM, the RNN-LM, and the interpolated LM were 56.24, 60.83, and 41.73, respectively. Thus, the use of the improved LMs based on RNN-LMs is expected to be advantageous for reverberant speech recognition.

## 5. EXPERIMENTS

### 5.1. Experimental settings

**SE Front-end**: For WPE, we set $T_\perp = 3$ and $T_\top = 40, 30, 7$ for 1ch, 2ch and 8ch, respectively. We used a window length of 32 ms and a frame shift of 8 ms for both WPE and MVDR. The settings for DOLPHIN are described in [7] Section V. B. 2). The settings of MMSE are similar to those in [9] Section 5.2. The results were evaluated in terms of cepstral distance (CD), speech to reverberation modulation energy ratio (SRMR), log likelihood ratio (LLR), frequency-weighted segmental signal to noise ratio (FWSegSNR) and PESQ.

**ASR back-end:** We used two different CD-DNN-HMM based AMs, one trained with the multi-condition training data provided by the challenge (AM 1), and one using extended training data (AM 2). The extended training data consisted of the WSJCAM0 [26] clean training data, WSJCAM0 training data recorded with the second microphone (table microphone) and noisy and reverberant training data obtained with the script provided by the REVERB challenge to generate multi-condition training data but by setting the SNR at 10 and 15 dB in addition to the original 20 dB. This extended data set is about 5 times the size of the REVERB challenge training data set (about 85 hours). It consisted of the same utterances and any variation originated solely from the acoustic environment. Note that all the elements to create the extended training data set were released with the challenge data.

The features used for ASR consist of 40 log mel filter-bank coefficients, with their delta and acceleration (120 coefficients in total). We used 5 left and 5 right context windows as the DNN input, corresponding to a total of 1320 visible input units. There were 7 hidden layers each with 2048 units. There were 3129 output HMM states. For the training of the DNN we used HMM state alignment obtained using the clean training data with an HMM-GMM based ASR system trained with the ML criterion. The validation set used for the training of the data was created by randomly selecting 5 % of the training data.

We investigated two types of LMs. The first one consisted of the WSJ tri-gram LM that is distributed with the American version of WSJ. The second LM was an RNN-LM. The interpolation coefficient for the RNN-LM was set at 0.5. For decoding, we used an LM weight of 11 and a relatively large search beam of 400 for optimal

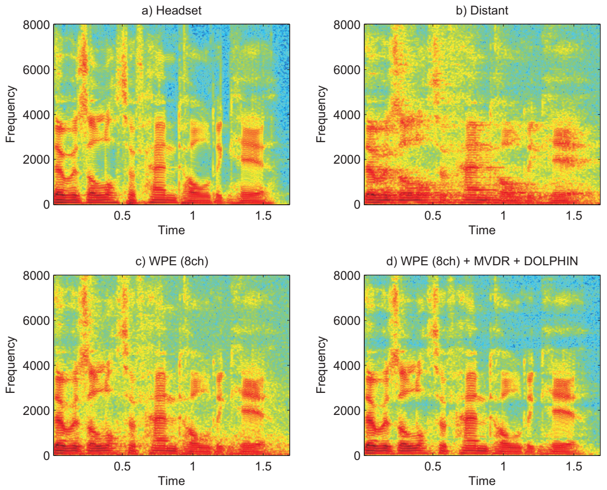

**Fig. 2**. Spectrograms for an utterance from the RealData evaluation set.

ASR performance.

For unsupervised batch adaptation of the first layer of the DNNs, we performed a few backpropagation iterations assuming that the labels obtained from a first recognition pass were true references. The learning rate was set at 0.0005 and the number of iterations at about 15 epochs.

All the parameters of the SE front-end and ASR back-end were tuned using the development set.

## 5.2. Results

A summary of the results obtained on the development set can be found in Appendix 7. In the following, we discuss the results for the evaluation set.

### *5.2.1. SE results*

We used the 2ch and 8ch set-ups for the SE task. Table 1 shows the results obtained with the SE objective measures for the evaluation set. Note that the front-end was essentially tuned for optimal ASR performance using the DNN-based ASR back-end, and thus the results may not be the best for the SE task.

All the systems operate in the utterance batch mode, except for DOLPHIN, which operates in the full batch mode. The results we submitted for the REVERB challenge consist of those for system III (2ch) and VII (8ch) for the utterance batch mode, and IV (2ch) and VIII (8ch) for the full batch mode. The results in Table 1 confirm that each component of the SE front-end consistently improves performance. The results for DOPLHIN and MMSE are somewhat similar in terms of objective measures. However, an informal listening test revealed that MMSE tends to reduce more noise than DOLPHIN at the expense of slightly more perceived artifacts.

For the most severe acoustic conditions (i.e., room3 far and RealData), we noticed that the presence of noise affected the dereverberation performance and that in some cases it resulted in the reverberation tail remaining perceivable after processing.

Figure 2 shows the spectrograms of part of an utterance of the RealData evaluation set, processed with our 8ch set-up. Due to space constraints we have only provided the most relevant spectrograms. Figure 2-c) clearly reveals the strong dereverberation effect of WPE. The remaining ambient noise is reduced significantly using MVDR and DOLPHIN as shown in 2-d).

**Table 2**. Mean WER for the evaluation test set using the HTK baseline system with an acoustic model trained on clean data. Adap. means unsupervised batch adaptation using constrained maximum likelihood linear regression (CMLLR). The results are shown only for systems submitted to the SE task, i.e. 2ch and 8ch set-ups.

| | Proc. | Adap. | SimData | RealData |
|---|---|---|---|---|
| 0 | Distant | - | 51.7 | 88.5 |
| | | X | 39.2 | 81.5 |
| I | WPE(2ch) | - | 28.2 | 65.8 |
| | | X | 19.3 | 52.7 |
| II | I+MVDR | - | 23.6 | 58.0 |
| | | X | 17.2 | 44.4 |
| III | II+MMSE | - | 20.3 | 46.2 |
| | | X | 16.9 | 39.7 |
| IV | II +DOLPHIN | - | 22.1 | 52.4 |
| | | X | 17.1 | 39.8 |
| V | WPE (8ch) | - | 25.6 | 61.2 |
| | | X | 18.0 | 48.2 |
| VI | V +MVDR | - | 17.6 | 41.9 |
| | | X | 14.5 | 32.6 |
| VII | VI + MMSE | - | 17.2 | 35.4 |
| | | X | 14.6 | 31.0 |
| VIII | VI + DOLPHIN | - | 17.1 | 37.4 |
| | | X | 14.3 | 29.5 |

In addition to the objective measures, in Table 2 we also provide the WER for the evaluation obtained with the HTK [27] baseline system using clean AMs. Table 2 reveals the potential improvement of the SE front-end with a clean AM, but this should be interpreted carefully since the SE front-ends were not tuned for optimal WER performance with this recognizer.

### *5.2.2. ASR results*

Table 3 shows the WER for the evaluation set for the 1ch, 2ch and 8ch ASR systems described in Figure 1. All systems operate in the utterance batch mode except for those using DOLPHIN and the adaptation, which operate in the full batch mode. The results we submitted for the REVERB challenge consist of those for system I-d (1ch) III-d (2ch) and VI-d (8ch) for the utterance batch mode, and I-e (1ch) IV-e (2ch) and VII-e (8ch) for the full batch mode. The other results are provided to attest to the contribution of each component of our proposed system. In particular, the results with AM 1 are given for comparison with other participants' results but should be interpreted carefully since the parameter tuning was not performed with this AM.

Table 3 also shows the WER of clean speech for SimData and that of speech recorded using headset and lapel mics for RealData. The headset recordings are almost clean and consequently the performance difference between clean (SimData) and headset (RealData) speech seems to indicate that the mismatch between the training data and the RealData set originates not only from noise and reverberation but also from other factors related to the spoken utterances such as speaking style.

The results in Table 3 demonstrate that dereverberation using WPE plays an essential role in our SE front-end. Indeed, WPE alone is responsible for relative WER improvements of up to 22%, 33% and 38% for 1ch, 2ch and 8ch, respectively. We observe a larger performance improvement when using multi-microphone processing. For dereverberation, most of the performance gain in terms of WER is already observed when using two microphones, but MVDR and DOLPHIN work particularly well when using eight microphones. Note that for moderate reverberation and noise conditions (i.e. Room

**Table 1**. SE scores for the evaluation set. Systems submitted to the SE task of the REVERB challenge are highlighted in bold fonts. Numbers with an asterisk are best scores.

| | | | SimData | | | | | | | RealData | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Room1 | | Room2 | | Room3 | | Ave. | Room1 | | Ave. |
| | | | Near | Far | Near | Far | Near | Far | - | Near | Far | - |
| 0 | Distant | CD | 1.99 | 2.67 | 4.63 | 5.21 | 4.38 | 4.96 | 3.97 | | | |
| | | SRMR | 4.50 | 4.58 | 3.74 | 2.97 | 3.57 | 2.73 | 3.68 | 3.17 | 3.19 | 3.18 |
| | | LLR | 0.35 | 0.38 | 0.49 | 0.75 | 0.65 | 0.84 | 0.58 | | | |
| | | FWSegSNR | 8.12 | 6.68 | 3.35 | 1.04 | 2.27 | 0.24 | 3.62 | | | |
| | | PESQ | 2.14 | 1.61 | 1.40 | 1.19 | 1.37 | 1.17 | 1.48 | | | |
| I | WPE(2ch) | CD | 2.04 | 3.66 | 4.38 | 4.86 | 3.93 | 4.32 | 3.66 | | | |
| | | SRMR | 4.68 | 5.12 | 4.38 | 4.48 | 4.44 | 3.93 | 4.50 | 4.29 | 4.68 | 4.48 |
| | | LLR | 0.34* | 0.33 | 0.46 | 0.60 | 0.51 | 0.58 | 0.47 | | | |
| | | FWSegSNR | 8.59 | 8.09 | 5.17 | 3.29 | 4.35 | 2.60 | 5.35 | | | |
| | | PESQ | 2.50 | 1.99 | 1.77 | 1.46 | 1.77 | 1.43 | 1.82 | | | |
| II | I+MVDR | CD | 1.84 | 2.21 | 3.90 | 4.46 | 3.46 | 3.92 | 3.30 | | | |
| | | SRMR | 4.85 | 5.49 | 4.47 | 4.95 | 4.61 | 4.35 | 4.79 | 5.01 | 5.40 | 5.21 |
| | | LLR | 0.34* | 0.35 | 0.42 | 0.52 | 0.50 | 0.55 | 0.45 | | | |
| | | FWSegSNR | 9.50 | 8.89 | 6.58 | 4.73 | 5.42 | 3.60 | 6.45 | | | |
| | | PESQ | 2.93 | 2.26 | 2.05 | 1.59 | 2.12 | 1.57 | 2.09 | | | |
| **III** | **II+MMSE** | CD | 1.87 | 2.05 | 2.26 | 2.93 | 2.18 | 2.73 | 2.34 | | | |
| | | SRMR | 5.00 | 5.66 | 4.80 | 5.32 | 4.91* | 4.64 | 5.06 | 6.55 | 6.71 | 6.63 |
| | | LLR | 0.38 | 0.39 | 0.32 | 0.40 | 0.48* | 0.50 | 0.41* | | | |
| | | FWSegSNR | 10.54 | 10.55* | 11.65 | 9.72 | 9.95* | 8.79 | 10.20 | | | |
| | | PESQ | 3.20 | 2.48 | 2.67 | 1.79 | 2.71 | 1.75 | 2.43 | | | |
| **IV** | **II +DOL** | CD | 1.55* | 1.94* | 3.01 | 3.68 | 2.60 | 3.13 | 2.65 | | | |
| | | SRMR | 4.88 | 5.53 | 4.64 | 5.21 | 4.75 | 4.58 | 4.93 | 6.04 | 6.26 | 6.15 |
| | | LLR | 0.35 | 0.35 | 0.38 | 0.48 | 0.49 | 0.52 | 0.43 | | | |
| | | FWSegSNR | 11.18* | 10.53 | 8.72 | 6.68 | 7.82 | 5.94 | 8.48 | | | |
| | | PESQ | 3.11 | 2.41 | 2.34 | 1.72 | 2.41 | 1.71 | 2.28 | | | |
| V | WPE (8ch) | CD | 1.97 | 2.35 | 4.37 | 4.86 | 3.90 | 4.31 | 3.63 | | | |
| | | SRMR | 4.67 | 5.05 | 4.38 | 4.93 | 4.46 | 4.35 | 4.64 | 4.32 | 4.79 | 4.55 |
| | | LLR | 0.34* | 0.32* | 0.48 | 0.57 | 0.50 | 0.57 | 0.46 | | | |
| | | FWSegSNR | 8.74 | 8.30 | 5.02 | 3.61 | 4.36 | 2.86 | 5.48 | | | |
| | | PESQ | 2.56 | 2.18 | 1.78 | 1.52 | 1.80 | 1.51 | 1.89 | | | |
| VI | V +MVDR | CD | 1.82 | 2.14 | 3.14 | 3.70 | 2.73 | 3.16 | 2.78 | | | |
| | | SRMR | 5.31 | 6.05 | 4.57 | 5.49 | 4.72 | 4.96 | 5.18 | 6.00 | 6.28 | 6.14 |
| | | LLR | 0.40 | 0.43 | 0.38 | 0.41 | 0.50 | 0.51 | 0.44 | | | |
| | | FWSegSNR | 10.09 | 9.39 | 8.39 | 6.96 | 7.07 | 5.81 | 7.95 | | | |
| | | PESQ | 3.21 | 2.81 | 2.50 | 1.98 | 2.64 | 2.13 | 2.54 | | | |
| **VII** | **VI + MMSE** | CD | 2.09 | 2.25 | 2.13* | 2.47* | 2.18 | 2.36* | 2.25* | | | |
| | | SRMR | 5.45* | 6.20* | 4.82* | 5.82* | 4.91* | 5.16* | 5.39* | 7.31* | 7.37* | 7.34* |
| | | LLR | 0.44 | 0.48 | 0.31* | 0.34* | 0.50 | 0.49* | 0.43 | | | |
| | | FWSegSNR | 9.97 | 9.56 | 11.99* | 10.82* | 9.90 | 9.60* | 10.31* | | | |
| | | PESQ | 3.33* | 2.94* | 2.97* | 2.23* | 3.05* | 2.39* | 2.82* | | | |
| **VIII** | **VI + DOL** | CD | 1.67 | 1.97 | 2.43 | 3.10 | 2.14* | 2.53 | 2.31 | | | |
| | | SRMR | 5.34 | 6.08 | 4.67 | 5.64 | 4.80 | 5.08 | 5.27 | 6.98 | 7.09 | 7.04 |
| | | LLR | 0.45 | 0.47 | 0.38 | 0.40 | 0.56 | 0.54 | 0.47 | | | |
| | | FWSegSNR | 10.88 | 10.30 | 10.36 | 8.61 | 9.23 | 8.13 | 9.58 | | | |
| | | PESQ | 3.23 | 2.91 | 2.72 | 2.12 | 2.85 | 2.33 | 2.69 | | | |

1 near), optimal performance is already achieved with a single microphone. We also confirm that the use of the extended training data set, the RNN-LM and unsupervised AM adaptation consistently improves performance. Although the parameters for adaptation (learning rate, number of iterations, etc.) were tuned on the development set, we obtained a larger improvement with the evaluation set than with the development set, since the evaluation set contains a larger number of data.

It is noteworthy that the ASR performance of our best system is almost equivalent to that of speech recorded with a close talking mic (lapel-mic). Nevertheless, the performance gap between enhanced speech and clean/headset speech is much smaller for SimData than for RealData, suggesting that room remains for improvement with the SE front-end if we are to further reduce the WER for RealData.

Using DNN with RNN-LM and adaptation, we have already been able to obtain relatively high recognition performance for distant speech even without any SE front-end. Nevertheless, our best proposed SE front-end provided a large additional improvement in performance, namely relative WER reduction of about 30 % and 60 % for SimData and RealData, respectively. This demonstrates that well designed speech enhancement front-ends can have a great impact on recognition performance when using DNN-based ASR especially in multi-microphone processing scenarios.

## 6. CONCLUSION

In this paper we proposed an SE and ASR system for speech recorded in noisy and reverberant rooms. We showed that dereverberation plays a key role in improving the recognition of distant speech. Moreover, by combining a dereverberation algorithm, advanced noise reduction techniques and a state-of-the-art ASR system we obtained excellent performance for both SE and ASR tasks.

## 7. APPENDIX

Tables 4 and 5 show the results on the development set for the SE and ASR tasks, respectively.

**Table 3**. WER for the evaluation set. The systems submitted to the REVERB challenge are highlighted in bold fonts. Numbers with an asterisk are best scores.

| | Proc. | AM | Adap. | RNN-LM | SimData | | | | | | | RealData | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | Room1 | | Room2 | | Room3 | | Ave. | Room1 | | **Ave.** |
| | | | | | Near | Far | Near | Far | Near | Far | - | Near | Far | - |
| | Clean / | 2 | - | X | 3.3 | | 3.5 | | 3.8 | | 3.5 | 7.5 | 6.2 | 6.9 |
| | Headset mic | 2 | X | X | 3.4 | | 3.5 | | 3.9 | | 3.6 | 6.5 | 5.3 | 5.9 |
| | Lapel mic | 2 | - | X | - | - | - | - | - | - | **-** | 9.7 | 10.0 | 9.8 |
| | | 2 | X | X | - | - | - | - | - | - | **-** | 8.2 | 8.3 | 8.3 |
| 0 - a | Distant | 1 | - | - | 5.9 | 6.6 | 7.9 | 12.2 | 8.7 | 13.2 | 9.1 | 32.6 | 32.3 | 32.5 |
| b | | 2 | - | - | 5.1 | 5.6 | 6.7 | 11.5 | 7.6 | 11.6 | 8.0 | 27.1 | 27.9 | 27.5 |
| c | | 2 | X | - | 4.7 | 5.4 | 6.4 | 10.3 | 7.5 | 10.9 | 7.5 | 22.1 | 24.7 | 23.4 |
| d | | 2 | - | X | 4.1 | 4.7 | 5.5 | 9.7 | 5.9 | 9.9 | 6.6 | 25.7 | 26.9 | 26.3 |
| e | | 2 | X | X | 3.8 | 4.4 | 5.3 | 8.5 | 5.8 | 9.5 | 6.2 | 21.1 | 23.3 | 22.2 |
| **I** - a | WPE (1ch) | 1 | - | - | 6.2 | 6.0 | 7.3 | 10.6 | 7.7 | 10.4 | 8.0 | 27.9 | 27.7 | 27.8 |
| b | | 2 | - | - | 4.9 | 5.2 | 6.4 | 9.3 | 6.8 | 8.9 | 6.9 | 21.4 | 22.1 | 21.7 |
| c | | 2 | X | - | 4.5 | 4.8 | 6.1 | 8.7 | 6.7 | 8.2 | 6.5 | 18.1 | 19.9 | 19.0 |
| **d** | | **2** | **-** | **X** | **3.9** | **4.2** | **5.0** | **7.7** | **5.7** | **7.3** | **5.6** | **20.0** | **20.6** | **20.3** |
| **e** | | **2** | **X** | **X** | **3.5*** | **4.0** | **4.6** | **6.8** | **5.1** | **7.2** | **5.2** | **16.4** | **18.4** | **17.4** |
| II - a | WPE (2ch) | 1 | - | - | 6.6 | 6.3 | 7.1 | 8.6 | 7.6 | 8.8 | 7.5 | 25.7 | 25.3 | 25.5 |
| b | | 2 | - | - | 4.9 | 5.1 | 6.1 | 7.5 | 6.5 | 7.4 | 6.3 | 18.5 | 19.1 | 18.8 |
| c | | 2 | X | - | 4.6 | 4.8 | 5.9 | 7.1 | 6.3 | 7.0 | 5.9 | 15.5 | 16.7 | 16.1 |
| d | | 2 | - | X | 3.9 | 4.0 | 4.9 | 5.8 | 5.4 | 6.1 | 5.0 | 17.7 | 17.4 | 17.5 |
| e | | 2 | X | X | 3.6 | 3.6* | 4.6 | 5.6 | 5.0 | 6.0 | 4.7 | 14.6 | 15.1 | 14.9 |
| **III** - a | II + MVDR | 1 | - | - | 6.5 | 6.5 | 6.8 | 7.8 | 6.7 | 7.6 | 7.0 | 22.3 | 22.1 | 22.2 |
| b | | 2 | - | - | 4.9 | 5.0 | 6.1 | 7.0 | 6.3 | 6.6 | 6.0 | 16.1 | 16.6 | 16.4 |
| c | | 2 | X | - | 4.5 | 4.7 | 5.6 | 6.4 | 6.0 | 6.4 | 5.6 | 13.6 | 15.2 | 14.4 |
| **d** | | **2** | **-** | **X** | **4.1** | **4.0** | **4.5** | **5.6** | **5.0** | **5.6** | **4.8** | **14.9** | **15.2** | **15.0** |
| e | | 2 | X | X | 3.6 | 3.7 | 4.2 | 5.1 | 4.8 | 5.2 | 4.4 | 12.4 | 13.4 | 12.9 |
| **IV** - a | III + DOL | 1 | - | - | 6.6 | 6.6 | 6.6 | 7.7 | 6.1 | 7.1 | 6.8 | 21.1 | 21.2 | 21.2 |
| b | | 2 | - | - | 4.7 | 5.1 | 5.9 | 6.4 | 6.0 | 6.7 | 5.8 | 16.2 | 16.6 | 16.4 |
| c | | 2 | X | - | 4.5 | 4.6 | 5.5 | 6.3 | 5.7 | 6.4 | 5.5 | 13.8 | 15.2 | 14.5 |
| d | | 2 | - | X | 4.1 | 4.0 | 4.5 | 5.4 | 5.0 | 5.5 | 4.8 | 14.8 | 15.9 | 15.4 |
| **e** | | **2** | **X** | **X** | **3.7** | **3.6*** | **4.2** | **5.1** | **4.8** | **5.2** | **4.4** | **12.0** | **13.4** | **12.7** |
| V - a | WPE (8ch) | 1 | - | - | 6.3 | 6.5 | 7.3 | 7.8 | 7.3 | 8.8 | 7.3 | 25.2 | 24.2 | 24.7 |
| b | | 2 | - | - | 4.6 | 5.1 | 6.3 | 6.7 | 6.1 | 6.8 | 6.0 | 17.2 | 18.5 | 17.8 |
| c | | 2 | X | - | 4.4 | 4.8 | 6.0 | 6.4 | 5.9 | 6.7 | 5.7 | 14.4 | 15.6 | 15.0 |
| d | | 2 | - | X | 4.0 | 4.2 | 4.8 | 5.2 | 5.2 | 5.8 | 4.9 | 16.6 | 17.2 | 16.9 |
| e | | 2 | X | X | 3.7 | 3.7 | 4.4 | 5.1 | 4.6 | 5.8 | 4.5 | 13.4 | 14.2 | 13.8 |
| **VI** - a | V + MVDR | 1 | - | - | 7.0 | 7.0 | 6.1 | 6.7 | 6.6 | 7.2 | 6.8 | 17.2 | 17.4 | 17.3 |
| b | | 2 | - | - | 4.6 | 5.0 | 5.2 | 5.9 | 5.7 | 6.2 | 5.4 | 12.2 | 14.0 | 13.1 |
| c | | 2 | X | - | 4.6 | 4.8 | 5.0 | 5.5 | 5.6 | 6.0 | 5.3 | 10.8 | 11.7 | 11.2 |
| **d** | | **2** | **-** | **X** | **4.0** | **4.0** | **3.8*** | **4.5*** | **4.5** | **4.9** | **4.3** | **11.4** | **12.4** | **11.9** |
| e | | 2 | X | X | 3.7 | 3.8 | 3.9 | 4.5* | 4.3* | 5.0 | 4.2* | 9.8 | 10.3 | 10.0 |
| **VII** - a | VI + DOL | 1 | - | - | 7.3 | 7.5 | 6.6 | 7.1 | 6.9 | 7.2 | 7.1 | 15.4 | 16.3 | 15.8 |
| b | | 2 | - | - | 4.6 | 4.9 | 5.3 | 5.8 | 6.0 | 6.0 | 5.4 | 12.0 | 13.7 | 12.8 |
| c | | 2 | X | - | 4.5 | 4.8 | 5.2 | 5.4 | 5.7 | 5.9 | 5.2 | 10.1 | 11.4 | 10.8 |
| d | | 2 | - | X | 4.0 | 4.1 | 4.0 | 4.5* | 4.3* | 4.8* | 4.3 | 10.0 | 12.0 | 11.0 |
| **e** | | **2** | **X** | **X** | **3.7** | **4.0** | **4.0** | **4.5*** | **4.4** | **4.8*** | **4.2*** | **8.8*** | **9.3*** | **9.0*** |

## 8. REFERENCES

[1] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: robustness against reverberation for automatic speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 114–126, 2012.

[2] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, A. Sehr, W. Kellermann, S. Gannot, R. Maas, R. Haeb-Umbach, V. Leutnant, and B. Raj, "The REVERB challenge: a common evaluation framework for dereverberation and recognition of reverberant speech," in *Proc. of WASPAA'13*, New Paltz, NY, USA, October 2013.

[3] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," in *Proc. of ICASSP'08*, 2008, pp. 85–88.

[4] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 1, pp. 69–84, 2011.

[5] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 10, pp. 2707–2720, 2012.

[6] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 2, pp. 260–276, 2010.

[7] T. Nakatani, T. Yoshioka, S. Araki, M. Delcroix, and M. Fujimoto, "Dominance based integration of spatial and spectral features for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 12, pp. 2516–2531, 2013.

[8] M. Fujimoto, S. Watanabe, and T. Nakatani, "Noise suppression with unsupervised joint speaker adaptation and noise mixture model estimation," in *Proc. of ICASSP '12*, 2012, pp. 4713–4716.

[9] M. Fujimoto and T. Nakatani, "A reliable data selection for model-based noise suppression using unsupervised joint speaker adaptation and noise model estimation," in *Proc. of ICSPCC '12*, 2012, pp. 148–153.

[10] A. Mohamed, G.E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 14–22, 2012.

[11] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, "Deep

**Table 4**. Mean SE scores for the development set.

| | | | SimData | RealData |
|---|---|---|---|---|
| 0 | Distant | CD | 3.88 | |
| | | SRMR | 3.67 | 3.79 |
| | | LLR | 0.58 | |
| | | FWSegSNR | 3.48 | |
| | | PESQ | 1.42 | |
| I | WPE(2ch) | CD | 3.58 | |
| | | SRMR | 4.45 | 5.55 |
| | | LLR | 0.48 | |
| | | FWSegSNR | 5.18 | |
| | | PESQ | 1.72 | |
| II | I+MVDR | CD | 3.22 | |
| | | SRMR | 4.70 | 6.54 |
| | | LLR | 0.45 | |
| | | FWSegSNR | 6.36 | |
| | | PESQ | 1.96 | |
| III | II+MMSE | CD | 2.39 | |
| | | SRMR | 4.98 | 8.28 |
| | | LLR | 0.41 | |
| | | FWSegSNR | 9.86 | |
| | | PESQ | 2.26 | |
| IV | II +DOL | CD | 2.13 | |
| | | SRMR | 4.83 | 7.79 |
| | | LLR | 0.42 | |
| | | FWSegSNR | 8.74 | |
| | | PESQ | 2.13 | |
| V | WPE (8ch) | CD | 3.53 | |
| | | SRMR | 4.54 | 5.68 |
| | | LLR | 0.47 | |
| | | FWSegSNR | 5.35 | |
| | | PESQ | 1.79 | |
| VI | V +MVDR | CD | 2.62 | |
| | | SRMR | 5.07 | 7.86 |
| | | LLR | 0.43 | |
| | | FWSegSNR | 8.27 | |
| | | PESQ | 2.41 | |
| VII | VI + MMSE | CD | 2.34 | |
| | | SRMR | 5.29 | 9.54 |
| | | LLR | 0.43 | |
| | | FWSegSNR | 10.03 | |
| | | PESQ | 2.62 | |
| VIII | VI + DOL | CD | 2.15 | |
| | | SRMR | 5.15 | 9.16 |
| | | LLR | 0.45 | |
| | | FWSegSNR | 10.13 | |
| | | PESQ | 2.53 | |

**Table 5**. Mean WERs for the development set.

| | Proc. | AM | Adap. | RNN-LM | SimData | RealData |
|---|---|---|---|---|---|---|
| 0 - a | Distant | 1 | - | - | 8.7 | 26.3 |
| b | | 2 | - | - | 7.8 | 23.2 |
| c | | 2 | X | - | 7.4 | 22.7 |
| d | | 2 | - | X | 6.5 | 21.2 |
| e | | 2 | X | X | 6.0 | 20.0 |
| I - a | WPE (1ch) | 1 | - | - | 7.9 | 23.5 |
| b | | 2 | - | - | 7.0 | 19.3 |
| c | | 2 | X | - | 6.4 | 18.5 |
| d | | 2 | - | X | 5.7 | 17.8 |
| e | | 2 | X | X | 5.3 | 16.5 |
| II - a | WPE (2ch) | 1 | - | - | 7.2 | 23.7 |
| b | | 2 | - | - | 6.3 | 18.3 |
| c | | 2 | X | - | 5.8 | 17.4 |
| d | | 2 | - | X | 5.1 | 16.1 |
| e | | 2 | X | X | 4.7 | 14.8 |
| III - a | II + MVDR | 1 | - | - | 6.5 | 21.9 |
| b | | 2 | - | - | 5.7 | 17.0 |
| c | | 2 | X | - | 5.4 | 15.7 |
| d | | 2 | - | X | 4.7 | 14.6 |
| e | | 2 | X | X | 4.5 | 13.6 |
| IV - a | III + DOL | 1 | - | - | 6.4 | 20.8 |
| b | | 2 | - | - | 5.5 | 16.3 |
| c | | 2 | X | - | 5.3 | 15.1 |
| d | | 2 | - | X | 4.5 | 13.9 |
| e | | 2 | X | X | 4.3 | 12.7 |
| V - a | WPE (8ch) | 1 | - | - | 7.1 | 23.4 |
| b | | 2 | - | - | 6.0 | 16.9 |
| c | | 2 | X | - | 5.5 | 15.7 |
| d | | 2 | - | X | 4.8 | 14.7 |
| e | | 2 | X | X | 4.4 | 14.0 |
| VI - a | V + MVDR | 1 | - | - | 6.4 | 17.5 |
| b | | 2 | - | - | 5.0 | 13.7 |
| c | | 2 | X | - | 4.8 | 13.1 |
| d | | 2 | - | X | 4.1 | 11.7 |
| e | | 2 | X | X | 3.9 | 11.3 |
| VII - a | VI + DOL | 1 | - | - | 6.8 | 17.0 |
| b | | 2 | - | - | 5.0 | 13.4 |
| c | | 2 | X | - | 4.8 | 12.6 |
| d | | 2 | - | X | 4.1 | 11.4 |
| e | | 2 | X | X | 3.9 | 10.7 |

neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[12] M. Thomáš, *Statistical Language Models Based on Neural Networks*, Ph.D. thesis, Brno University of Technology, 2012.

[13] T. Hori, Y. Kubo, and A. Nakamura, "Real-time one-pass decoding with recurrent neural network language model for speech recognition," in *Proc. of ICASSP '14*, 2014.

[14] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks - studies on speech recognition tasks," in *Proc. of ICLR'13*, 2013.

[15] M.L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. of ICASSP'13*, 2013, pp. 7398–7402.

[16] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *Proc. of ICASSP'13*, 2013, pp. 7947–7951.

[17] T. Yoshioka, X. Chen, and M. J. F. Gales, "Impact of single-microphone dereverberation on DNN-based meeting transcription systems," in *Proc. ICASSP'14*, 2014.

[18] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 4, pp. 534–545, 2009.

[19] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[20] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 3, pp. 516–527, 2010.

[21] S. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proc. EUROSPEECH*, 2003, pp. 1009–1012.

[22] P.J. Moreno, B. Raj, and R.M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc of ICASSP'96*, 1996, vol. 2, pp. 733–736 vol. 2.

[23] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, "Adaptation of context-dependent deep neural networks for automatic speech recognition," in *Proc. of SLT'12*, 2012, pp. 366–369.

[24] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proc. of ICASSP'13*, 2013, pp. 7893–7897.

[25] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proc. SNL'92*, Morristown, NJ, USA, 1992, pp. 357–362, Association for Computational Linguistics.

[26] 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. of ICASSP'95*. 1995, pp. 81–84, IEEE.

[27] 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*, Cambridge University Engineering Department, Cambridge, UK, 2006.