Abstract—In this paper, we propose a robust distant-talking speech recognition system with asynchronous speech recording. This is implemented by combining denoising autoencoder-based cepstral-domain dereverberation, automatic asynchronous speech (microphone or mobile terminal) selection and environment adaptation. Although applications using mobile terminals have attracted increasing attention, there are few studies that focus on distant-talking speech recognition with asynchronous mobile terminals. For the system proposed in this paper, after applying a denoising autoencoder in the cepstral domain of speech to suppress reverberation and performing Large Vocabulary Continuous Speech Recognition (LVCSR), we adopted automatic asynchronous mobile terminal selection and environment adaptation using speech segments from optimal mobile terminals. The proposed method was evaluated using a reverberant WSJCAM0 corpus, which was emitted by a loudspeaker and recorded in a meeting room with multiple speakers by far-field multiple mobile terminals. By integrating a cepstral-domain denoising autoencoder and automatic mobile terminal selection with environment adaptation, the average Word Error Rate (WER) was reduced from 51.8% of the baseline system to 28.8%, i.e., the relative error reduction rate was 44.4% when using multi-condition acoustic models.

I. INTRODUCTION

Many techniques have been proposed for robust automatic speech recognition in noise and reverberation, using multiple microphones such as a microphone array [1], [2], [3], [4], [5], [6], [7], [8], [9]. These techniques require the synchronous signals of multiple microphones and the cost and preparation of the microphone array is considerable. The synchronous microphone array device is not available in many meeting rooms. In this paper, we present a robust hands-free speech recognition system using a ubiquitous asynchronous smart terminal such as a smartphone. For this system, costs and preparation are minimal. A diagram of the proposed system is shown in Fig 1. The proposed system consists of three components: (1) Denoising Autoencoder (DAE)-based dereverberation, (2) Large Vocabulary Continuous Speech Recognition (LVCSR), and (3) Automatic asynchronous speech selection and environment adaptation. Here, we focus on (1) dereverberation-based pre-processing, and (3) post-processing based on automatic asynchronous mobile terminal selection and environment adaptation.

Many single-channel dereverberation methods have been proposed for robust distant-talking speech recognition [10], [11], [12], [13]. Cepstral mean normalization (CMN) [14], [15], [16], [17] may be considered the most general approach. It has been extensively examined and shown as a simple and effective way of reducing reverberation by normalizing cepstral features. However, the dereverberation of CMN is not completely effective in environments with late reverberation. Several studies have focused on mitigating the above problem [5], [11], [13], [18], [19], [20]. A reverberation compensation method for speaker recognition using spectral subtraction [21], in which late reverberation is treated as additive noise, was proposed in [11]. A method based on multi-step linear prediction (MSLP) was proposed by [5], [13] for both single and multiple microphones. The method first estimates late reverberations using long-term multi-step linear prediction, and then suppresses these with subsequent spectral subtraction. The drawback of this approach is that the optimal parameters for spectral subtraction are empirically estimated from a developing dataset, meaning that the late reverberation cannot be subtracted correctly as it is not precisely modeled. Recently, DAEs have been shown to be effective in many noise reduction applications because higher level representations and increased flexibility of the feature mapping function can be learned [22], [23]. Ishii et al. applied a DAE for spectral-domain dereverberation [24] and found the word accuracy of LVCSR was improved from 61.4 to 65.2% for the JNAS database [25]. However, the suppressed spectral-domain feature needs to be converted to a cepstral-domain feature, and this improvement is not sufficient. Previously, we found that Deep Neural Network (DNN) [26] -based cepstral-domain feature mapping is efficient for distant-talking speech processing [27]. In this paper, we apply a denoising autoencoder for cepstral-domain dereverberation because there are many LVCSR systems that adopt a cepstral-domain feature as the direct input.

After LVCSR, post-processing based on automatic mobile terminal selection and unsupervised environment adaptation were carried out. For automatic mobile terminal selection, the optimal mobile terminal of a speech segment based on Voice Activity Detection (VAD) was automatically selected.
according to the average power of the speech segment from each terminal. In this paper, an ideal VAD is applied. For environment adaptation, we applied the Constrained Maximum Likelihood Linear Regression (CMLLR) [28] based unsupervised model adaptation method using DAE-based suppressed speech of the selected mobile terminal.

II. DENOISING AUTOENCODER FOR CEPSTRAL-DOMAIN DEREVERBERATION

An autoencoder is a type of artificial neural network (NN) whose output is reconstruction of input, and is often used for dimensionality reduction. DAEs share the same structure as autoencoders, but input data is a noisy version of the output data. Autoencoders use feature mapping to convert noisy input data into clean output, and have been used for noise removal in the field of image processing [22]. Ishii et al. applied a DAE for spectral-domain dereverberation [24]. However, the suppressed spectral-domain feature needs to be converted to a cepstral-domain feature, and this improvement is not sufficient. In this paper, we apply a denoising autoencoder for cepstral-domain dereverberation because there are many LVCSR systems that adopt a cepstral-domain feature as the direct input.

Given a pair of speech samples: clean speech and corresponding reverberant speech, DAE learns the non-linear conversion function that converts reverberant speech features into clean speech. In general, reverberation is dependent on both current and several previous observation frames. In addition to the vector of the current frame, vectors of past frames are concatenated to form input.

For cepstral feature $X_i$ of observed reverberant speech of $i - th$ frame, cepstral features of $N - 1$ frames before the current frame are concatenated with the current frame to form a cepstral vector of $N$ frames. Output $O_l$ of the non-linear transformer based on the DAE is given by:

$$O_l = f_{l1}(...f_{l2}(f_{l1}(X_{i}, X_{i-1}, ..., X_{i-N})))$$  \hspace{1cm} (1)

where $f_l$ is the non-linear transformation function in layer $l$, $N$ is the number of frames to be used as the input features.

Topology of the cepstral-domain DAE for dereverberation is shown in Fig. 2. In this paper, the number of hidden layers is set to five. In Figure 2, $W_i (i = 1, 2, 3)$ shows the weighting of the different layers, and $W_l^T$ shows the transposition of $W_l$. That is to say, $W_1$, $W_2$ and $W_3$ are the encoder matrix and $W_1^T$, $W_2^T$ and $W_3^T$ are the decoder matrix, respectively. To train a deep neural network, Deep Belief Networks (DBNs) [26] are used for pre-training because they can obtain accurate initial values of the deep-layer neural networks. To obtain a pre-trained RBM, we trained the second hidden layer using the Bernoulli-Bernoulli RBM, and the third hidden layer using a Gaussian-Bernoulli RBM. DBNs are hierarchically configured by connecting these pre-trained RBMs. Here, $W_1$, $W_2$ and $W_3$ are learned automatically and $W_1^T$, $W_2^T$ and $W_3^T$ are generated from $W_1$, $W_2$ and $W_3$.

After pre-training, a backpropagation algorithm was applied to adjust the parameters. Backpropagation modifies the weights of the network to reduce the error of the teacher signal and the output value when a pair of signals (input signal and the ideal teacher signal, the cepstral feature of clean speech) are given. In this paper, the error in the input data is defined

$W_i$ and $W_i^T$ correspond to $f_{l2}$ in Eq. (1)
by the cross entropy of the teacher signal and the output of the output layer unit [22]. The steepest descent method was used to adjust the relative weighting of the units so as to minimize the error [22].

III. ASYNCHRONOUS SPEECH SELECTION AND ENVIRONMENT ADAPTATION

In distant-talking speech recognition, recognition accuracy is significantly degraded by noise and reverberation. It depends on the distance and direction of the microphone and the speaker. In this situation where asynchronous mobile terminals are used, it is important to select an optimal speech segment from an optimal mobile terminal. For the sake of simplicity, the ideal VAD was used in this paper. Thus, the challenge became how to automatically select an optimal mobile terminal.

We selected the speech from the mobile terminal with the maximum power, on the assumption that this terminal is nearest to the speaker and is most appropriate for recognition. Given the speech segment of \( n \)-th mobile terminal \( X_n \), the optimal mobile terminal is selected as:

\[
\hat{n} = \arg \max_n E_n, \quad n = 1, \ldots, N, \tag{2}
\]

where \( E_n \) is the average power of \( X_n \) and \( N \) is number of mobile terminals. As a result, \( X_{\hat{n}} \) is used as the input speech to LVCSR.

CMLLR [28] is the method for converting the mean and variance of the Gaussian distribution for each state of the hidden Markov models (HMMs) by using the regression matrix to reduce the mismatch between the adaptation data and model. This method is intended to obtain a transformation matrix for modifying the model parameters that maximize the likelihood of the adaptation data. In this paper, we applied CMLLR for unsupervised model adaptation, i.e., environment adaptation.

IV. EXPERIMENTS

A. Experimental setup

1) Training dataset: The training dataset provided by “REVERB challenge” (Reverberant Voice Enhancement and Recognition Benchmark) [29] was used. This dataset consists of the clean WSJCAM0 [30] training set and a multi-condition (MC) training set. Reverberant speech is generated from the clean WSJCAM0 training data by convolving the clean utterances with measured room impulse responses and adding recorded background noise. The reverberation times of the measured impulse responses range from approximately 0.1 to 0.8 sec. The number of speakers was 92 and the total number of utterances was 7861. This training dataset was used for both training of acoustic models and parameters of the DAE.

It should be noted that the recording rooms used for the multi-condition training data and test data were different.

2) Evaluation test set: To evaluate the proposed method, 100 utterances randomly selected from the evaluation test set of the WSJCAM0 corpus were emitted from a loudspeaker and recorded by three asynchronous mobile terminals (iPhone 4S) set in a seminar room. Fig 3 shows the structure of the seminar (recording) room and Fig 4 shows the positions of the loudspeakers and the placement of the mobile terminals. The speakers were fixed at eight positions from A to H shown in Fig 4. We recorded 100 utterances in total, at positions A-H, using an iPhone 4S application called “PCM recording” for speech recording.

3) Parameters for LVCSR and Dereverberation: In this study, Mel Frequency Cepstral Coefficients (MFCCs) were used for LVCSR and DAE-based dereverberation. The dimension of the MFCCs was 39 including 12 MFCCs plus power and their Delta and Delta-Delta coefficients.

For DAE-based dereverberation, feature vectors of the current frame and previous eight frames of reverberant speech were used as input. Thirty-nine MFCCs of the current frame of clean speech were used as teacher signals for output, i.e., the dimension of input was \( 39 \times 9 = 351 \). There were five hidden-layers of DAE, and 1024 hidden units in each hidden layer. MFCC features were normalized using the mean of the entire multi-condition training set. The DAE training was carried out using stochastic mini-batch gradient descent with a minibatch size of 256 samples. Fifty epochs with a learning rate of 0.002
were used for all layers during pre-training, and 100 epochs with a learning rate of 0.1 were used for all layers during fine-tuning.

In this study, we used a speech recognition system provided by the “REVERB challenge” task [29], which is based on the hidden Markov model tool kit (HTK) [31]. As an acoustic model, it employs tied-state HMMs with ten Gaussian components per state, trained according to the maximum-likelihood criterion. Both the clean acoustic model and multi-condition acoustic model were used for evaluating our proposed method.

B. Experimental Results

Tables I and II show the speech recognition results with the clean acoustic model and multi-condition acoustic model, respectively. “Single” shows the average recognition results of multiple mobile terminals when recognizing singly recorded speech for each mobile terminal. “Selection” shows the speech recognition results of applying post-processing based on automatic mobile terminal selection. When DAE-based cepstral-domain dereverberation was compared with CMN-based dereverberation and single channel MSLP-based dereverberation [13], a remarkable improvement was achieved in all cases. For MSLP-based dereverberation, the step size and the order of linear prediction were set to 500 and 750, respectively. For the multi-condition acoustic model, the average error rates of each mobile terminal (results corresponding to “Single” column) were improved from 49.0% of CMN and 47.3% of MSLP to 41.1% of cepstral-domain DAE. The improvement is larger than that of spectral-domain DAE. We consider that one of the reasons may be that cepstral-domain DAE was more effective than spectral-domain DAE. By integrating automatic mobile terminal selection and cepstral-domain DAE with environment adaptation using selected speech segments from optimal mobile terminals, the average WER was reduced from 51.8% of baseline system to 28.8%, i.e., the reduction rate was 44.4% for the multi-condition reverberant acoustic model.

In future, we will apply other parameters such as likelihood and not only power to the microphone input selection. We will also try to implement VAD into the system, and change the acoustic model to DNN-HMM [32].

VI. ACKNOWLEDGEMENTS

This work was partially supported by a research grant from the Telecommunications Advancement Foundation (TAF), Japan.

REFERENCES


