ESTIMATING ROOM ACOUSTIC PARAMETERS FOR SPEECH RECOGNIZER ADAPTATION AND COMBINATION IN REVERBERANT ENVIRONMENTS

Feifei Xiong[⋆]*, Stefan Goetze*[⋆] *and Bernd T. Meyer*[†]

[⋆]Fraunhofer Institute for Digital Media Technology IDMT,
Project Group Hearing-, Speech- and Audio-Technology (HSA), 26129 Oldenburg, Germany
Email: {feifei.xiong,s.goetze}@idmt.fraunhofer.de
[†]Dept. of Medical Physics and Acoustics, University of Oldenburg, 26111 Oldenburg, Germany
Email: bernd.meyer@uni-oldenburg.de

ABSTRACT

This work analyzes the influence of reverberation on automatic speech recognition (ASR) systems and how to compensate its influence, with special focus on the important acoustical parameters i.e. room reverberation time T_{60} and clarity index C_{50}. A multilayer perceptron (MLP) using features of a spectro-temporal filter bank as input is employed to identify the acoustic conditions spanning various reverberant scenarios. The posterior probabilities of the MLP are used to design a novel selection scheme for adaptation in a cluster-based manner and for system combination achieved by recognizer output voting error reduction (ROVER). A comparison of word error rates is performed considering different training modes, and an average relative improvement of 7.1% is obtained by the proposed system compared to conventional multistyle training.

Index Terms— Automatic speech recognition (ASR), adaptation, reverberation, room reverberation time, clarity index

1. INTRODUCTION

Automatic speech recognition (ASR) systems have been substantially improved in the last few decades, which resulted in a large number of applications for mildly reverberant conditions, cf. e.g. [1, 2]. However, in scenarios with time-variant or a high amount of reverberation, ASR error rates are dramatically increased despite of the advances. This is especially true for mismatched training and test conditions that arise from ambient noises, variations of speaker characteristics, and channel distortions, as well as reverberation that is caused by multiple reflections inside an enclosed space and is usually modeled by a room impulse response (RIR). Since reverberation causes spectral changes as well as temporal smearing of consecutive frames, the above-mentioned mismatches are especially critical in the context of speech processing [3, 4].

One straightforward approach to use an ASR system in a new environment is to retrain the recognizer with new training data that has been collected in that new room, enabling an optimal match between the model and the data to be recognized. However, recording additional data is often not feasible and time-consuming at least. An alternative is to employ an adaptation scheme [5, 6, 7] to alleviate the mismatch of acoustic conditions between training and test data with a limited amount of data, as originally proposed for speaker adaptation. Several adaptation strategies have been adopted and evaluated to increase ASR robustness in reverberant environments by adapting the mean vectors of the clean hidden Markov models (HMMs) [8, 9, 10], where the state-level reverberation representation in HMMs is determined either by a maximum likelihood (ML) estimator with a few known calibration utterances [8], or by a strictly exponential energy decay model based on the room reverberation time T_{60} [9], while [10] turned to a statistical reverberation model in a feature-domain representation. Although they surpass the clean HMMs with a reduced quantity of adaptation data compared to conventional maximum likelihood linear regression (MLLR) [7] adaptation, still they can not be competitive to the ideal matched reverberant HMMs, even not to the multistyle trained models [11] in severely reverberant environments. Instead of assuming a reverberation model to tailor the HMMs to specific reverberant condition, this paper pursues an alternative approach that combines (i) the estimation of room acoustic parameters, (ii) the selection of appropriate ASR systems from a multitude of pre-trained adapted models, and (iii) a subsequent combination of the system outputs.

The estimation of room parameters (i) has been investigated in an earlier study [12], in which auditory features were employed to estimate T_{60} from acoustic speech signals. Complex 2D-Gabor features inspired by findings in the primary auditory cortex of mammal species were used to extract spectro-temporal patches of the signal of interest. T_{60} estimation was carried out with a multi-layer perceptron (MLP). This work exploits this estimate to improve ASR systems, and at the same time extends our previous approach to the estimation of the distance information between the speaker and the microphone, which is quantified by means of the clarity index C_{50} [13]. The posterior probabilities of classes obtained from the MLP (corresponding to different spatial configurations) are employed to select appropriate models for speech recognition (ii), where cluster-based adaptation achieved by MLLR [7] is employed to generate a series of cluster-dependent models. Furthermore, model-based feature normalization by constrained MLLR (CMLLR) [14] is sequently applied to these cluster-dependent models. Finally, a system combination (iii) is performed that is based on recognizer output voting error reduction (ROVER) [15]. The outputs of several models (that are considered to contribute to speech recognition through step (ii) and potentially carry complementary information) are integrated by ROVER to further enhance ASR performance.

The remainder of this paper is organized as follows: Section 2 introduces the T_{60} and C_{50} estimation based on MLPs and spectro-temporal modulation filtered features by 2D-Gabor filters. Adapta-

This work was partially supported by the project Dereverberation and Reverberation of Audio, Music, and Speech (DREAMS, project no. 316969) funded by the European Commission (EC), as well as by the DFG-Cluster of Excellence EXC 1077/1 "Hearing4all".

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

tion schemes and ROVER are briefly described in Section 3. The experimental procedure, the proposed selection scheme and results of estimation and ASR systems are addressed in Section 4. Concluding remarks are given in Section 5.

2. ROOM ACOUSTICAL PARAMETER ESTIMATION

The room reverberation time T_{60} [16] is one important characteristic in room acoustics, which is defined as the time interval for a 60 dB decay of the sound energy. It is often used as a criterion for speech recognition in reverberant environments. However, due to its independence on the position of the source and microphone within the room, T_{60} alone is not sufficient to fully describe the influence of reverberation on the recognition performance of state-of-the-art ASR systems [2]. It was shown that the energy ratio of the early part of an RIR, i.e. the direct sound and the early reflections, to its late reverberant tail is highly correlated to the ASR performance [17, 18]. This ratio is usually denoted as *clarity index* with a cut-off boundary of 50 ms [13],

$$C_{50} = 10 \log_{10} \left(\frac{\sum_{k=1}^{k_{50}} |h[k]|^2}{\sum_{k=k_{50}+1}^{\infty} |h[k]|^2} \right) , \qquad (1)$$

which mainly reflects the distance information between the sound source and the microphone for a given room. In (1), $h[k]$ denotes the RIR for the discrete time index k, and $k_{50} = \lceil 0.05 \cdot f_s \rceil$ is the time index after 50 ms at sampling rate f_s.

2.1. Estimation of T_{60} and C_{50}

Since it is not sufficient to separately consider either T_{60} or C_{50} for an ASR system design, a straightforward method is to jointly identify both parameters representing the room properties denoted as (T_{60}, C_{50}). As shown in [12], spectro-temporal modulations extracted from signals using a 2D-Gabor filter bank [19, 20] are well-suited for estimating T_{60}. This was achieved by using a multi-layer perceptron (MLP) as a classifier, where the output neurons correspond to specific T_{60} value ranges. *Diagonal* Gabor filters were especially sensitive to reverberation effects. The joint identification of (T_{60}, C_{50}) is therefore based on the setup outlined in [12]. The temporal context considered by the MLP is limited to 1 frame, 600 input neurons are used (corresponding to the Gabor feature dimension) as well as 400 hidden units. For the present study, the number of output neurons is 6, which is given by the different (T_{60}, C_{50}) pairs in the training data.

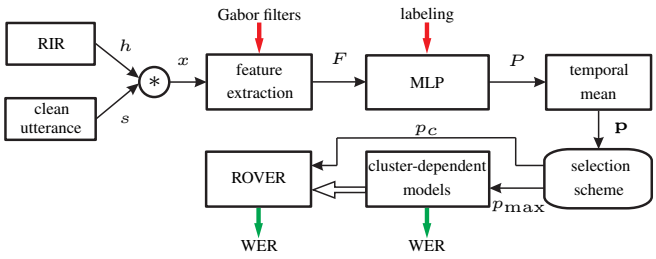

Fig. 1. Overview of the proposed system structure. The upper branch relates to the experimental setup for (T_{60}, C_{50}) estimation in the MLP network, while the lower branch illustrates the decisions of the proposed selection scheme transferred to cluster-dependent models either to select the matched model by $p_{\max}$ or to select beneficial models for ROVER by p_c to further reduce the WER.

2.2. System Structure

The MLP is trained using 2-D Gabor features F obtained from a reverberant version of the clean speech ($x = s * h$) with frame-wise labels for one class corresponding to that utterance. After the training procedure, frame-wise posterior probabilities P for all classes are obtained by the MLP forward run, which merge to the class-wise probability vector $\mathbf{p}$ averaged over time, as depicted in Fig. 1 where the upper branch gives an overview of the estimation procedure in the MLP network. A selection scheme (cf. Section 4.2) is proposed by leveraging the class-wise output probabilities $\mathbf{p}$, to select models from the adapted cluster-dependent models and for system combination achieved by ROVER, as briefly described in the following.

3. ADAPTATION AND SYSTEM COMBINATION

The aim of adaptation is to improve ASR performance either by feature normalization or by adapting the HMMs toward a particular test acoustic condition. Since there are various modes and different schemes for adaptation/adaptive training [21], cluster-based adaptation and model-based feature normalization are adopted in this task.

3.1. MLLR for Cluster-based Adaptation

Maximum likelihood linear regression (MLLR) was initially developed for speaker adaptation [7], which uses the maximum likelihood (ML) criterion to estimate a linear transform to adapt Gaussian mean and variance parameters of HMMs, i.e.,

$$\hat{\mu} = \mathbf{A}\mu + \mathbf{b}\,; \qquad \hat{\mathbf{\Sigma}} = \mathbf{H}\mathbf{\Sigma}\mathbf{H}^T\,, \qquad (2)$$

where $\hat{\mu}$ and $\hat{\mathbf{\Sigma}}$ are the adapted mean vector and covariance matrix from the pre-trained mean components μ and covariance matrix $\mathbf{\Sigma}$, respectively. The transform parameters $\mathbf{A}$, $\mathbf{b}$, and $\mathbf{H}$ are typically calculated by expectation maximization (EM) in an iterative process [22]. Note that only *mean* adaptation is applied in the following since pilot experiments have shown that *variance* adaptation does not help to increase the performance. In contrast to model-based adaptation which uses only one standard set of HMMs, cluster-based adaptation is based on a series of sets of HMMs as applied in this work. In general, several cluster-dependent models are built according to variant test acoustic conditions, e.g. speaker-dependent (SD) [5], which herein are determined by different combinations of (T_{60}, C_{50}).

3.2. CMLLR for Model-based Feature Normalization

As an alternative scheme to adapt both, mean vector and covariance matrix, constrained MLLR (CMLLR) [14] forces the linear transform of covariance matrix to be the same as the mean vector, denoted as follows compared to (2),

$$\hat{\mu} = \mathbf{A}_c\mu + \mathbf{b}_c\,; \qquad \hat{\mathbf{\Sigma}} = \mathbf{A}_c\mathbf{\Sigma}\mathbf{A}_c^T\,. \qquad (3)$$

It is straightforward to apply the transform $\mathbf{A}_c$ to the feature level [23] as model-dependent feature normalization, which is also closely related to MLLR model-based transform as described before.

3.3. System Combination

National institute of standards and technology (NIST)'s ROVER [15] has been shown to be effective to further reduce the word error rate (WER) by combining the outputs of multiple recognizers. Note that it is important to select beneficial and complementary recognizers to obtain the improved results among the cluster-dependent

HMMs. Instead of generating a new adapted model using interpolation weights, e.g. in cluster-based adaptive training [24], alternatively, a selector based on the MLP posterior probabilities is proposed to assist the model selection for ROVER, as illustrated by the lower branch in Fig. 1.

4. EXPERIMENTS AND RESULTS

We use the WSJCAM0 British English corpus [25] as database of clean speech utterances. It contains 7861 utterances for training and another 742 for testing at a sampling rate of 16 kHz. Overlapping speech segments of 25 ms length with 10 ms shift are used for log-mel-spectrogram calculation and feature extraction. The RIRs are measured in real-world scenarios [26], including 3 different rooms and 2 different positions, representing 3 types of T_{60} and 2 types of C_{50}, respectively. Reverberation times of the small, medium and large volume rooms are 218, 507 and 710 ms, respectively, calculated using [16]. The distances between the source and the microphone are about 50 cm for the *near* position and 200 cm for the *far* position. Table 1 lists the room acoustic parameters.

4.1. Estimation Performance

48 diagonal 2D-Gabor filters [12] are applied to the log-mel-spectrogram to extract feature vectors for the MLP classifier. They cover temporal modulations from 2 to 25 Hz and spectral modulations from -0.25 to 0.25 cycle/channel, respectively. Adjusting the parameter setting from a sampling rate of 8 kHz in [12] to 16 kHz, 31 mel-frequency channels for mel-spectrogram calculation are used, so that 600-dimensional feature vectors are obtained. For details on Gabor feature extraction, the reader is referred to [20].

Room	T_{60}(ms)	C_{50}(dB)		Class	$\mathrm{E}_{\mathrm{est}}$(%)	Avg. $p_{\max}$
Small	218	near	30.5	C1	1.48	0.80
		far	20.1	C2	0.27	0.84
Medium	507	near	17.3	C3	0.13	0.76
		far	9.1	C4	0.00	0.80
Large	710	near	14.9	C5	4.99	0.61
		far	6.8	C6	0.00	0.88

Table 1. Room acoustic parameters of the measured RIRs for evaluation. MLP estimation performance is measured by the average estimation error rate ($\mathrm{E}_{\mathrm{est}}$ %) based on a winner-takes-all rule and the average maximal probabilities $p_{\max}$ for acoustic configurations.

As summarized in Table 1, according to a winner-takes-all rule, i.e. class-wise probability $p_{\max}$, the estimation error rates ($\mathrm{E}_{\mathrm{est}}$) of all 6 classes of ($T_{60}$, C_{50}) are smaller than 5% (average $\mathrm{E}_{\mathrm{est}}$ is 1.15%), which verifies the effectiveness of the MLP classifier to distinguish reverberation effects from different room configurations of (T_{60}, C_{50}). An analysis of error patterns suggests that of the two parameters under consideration, C_{50} emerges as the dominant one. Fig. 2 (b) shows a confusion matrix, with the off-diagonal elements corresponding to errors. The observed errors mainly arise from the contiguous C_{50} values, while even the same value of T_{60} shows less impact to the classifier decision.

4.2. Selection Scheme

The winner-takes-all decision rule based on $p_{\max}$ is used as the selector for the aforementioned $\mathrm{E}_{\mathrm{est}}$ evaluation, which can be also considered as the selector to the matched HMM from the adapted cluster-dependent models, while other HMMs will not be considered. However, for system combination, the discarded models may carry useful messages, either due to the estimation error, e.g. Fig. 2 (a) omitting the true matched model, or because of the potential beneficial information during recognition, e.g. Fig. 2 (c) with similar sensitivity to reverberation. Therefore, in order to select as many beneficial models as possible for ROVER combination (cf. the lower branch in Fig. 1), a dedicated selection scheme is proposed by means of a threshold value ∇_p for the relative difference to the maximal probability $p_{\max}$, represented as

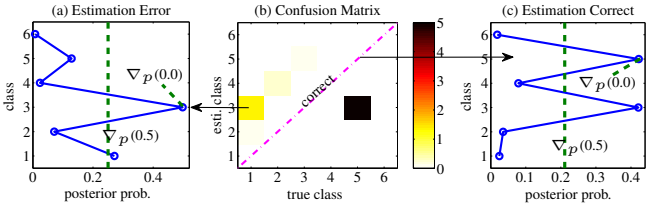

Fig. 2. (a) estimation error example from true class C1 to estimated class C3; (b) estimation confusion matrix according to Table 1 (diagonal/correct elements set to 0 for visibility of the error rates); (c) estimation correct example from true class C5 to estimated class C5.

$$\underset{c\in\mathrm{C}}{\arg}\left(\frac{p_{\max}-p_c}{p_{\max}}\leq\nabla_p\right), \tag{4}$$

where p_c denotes the class-wise probability of any class c from all defined classes C. A choice of $\nabla_p = 0.0$ corresponds to the winner-takes-all rule, while $\nabla_p = 1.0$ means to choose all the cluster-dependent models. Pilot experiments have shown that $\nabla_p = 0.5$ is a reasonable compromise that effectively selects a sufficient number of models (which usually includes the matched cluster-dependent model that corresponds to the true acoustic configuration).

4.3. ASR Performance

Mel-frequency cepstral coefficients (MFCCs) with delta and double-delta coefficients (dimension of 39) together with cepstral mean and variance normalization (CMVN) are employed in the ASR experiments. Multistyle training [11] is applied on the basis of all non-homogeneous data built by the clean training utterances convolved with the RIRs from Table 1, resulting in 6 cluster-dependent models after cluster-based adaptation. Context-dependent triphone HMMs with 3 states per model are applied together with 12 Gaussian mixture models (GMMs) per state and a language scaling factor of 14.0 for the 5k-word-bigram language model. MLLR and CMLLR schemes involve two passes [27]. The first pass is a global adaptation that builds a global transform used in the second pass, for which a regression class tree with up to 256 leaf nodes is generated. The ideal matched training models are generated by sharing exactly the same reverberant condition between training and test data. The amount of the training data is kept the same for different training modes and adaptation strategies for the sake of fair comparison.

As seen in Table 2, ASR systems do suffer from the reverberation effects. Even though the ideal matched training models are applied, WERs in severely reverberant environments are at a high level, e.g. 31.70% for class C6 with (T_{60}, C_{50}) of (710 ms, 6.8 dB). In general, the WER raises as T_{60} increases and C_{50} decreases. C_{50} has a higher impact than T_{60} as seen e.g. from WER comparison between C4 and C5, which is in line with the error patterns observed for the MLP estimates.

Cluster-dependent HMMs are generated based on MLLR mean adaptation in a *supervised* mode, i.e. the transcription of the adaptation data is known during multistyle training. The results with the

	Test/Class	C1	C2	C3	C4	C5	C6
	clean	15.44	26.07	28.49	61.76	36.46	78.58
	multistyle	15.00	16.91	18.31	27.17	21.02	35.78
oracle	ideal matched	**12.29**	**15.11**	**16.84**	**24.66**	**19.28**	**31.70**
oracle	mllr C1	**14.03**	17.21	18.39	29.26	21.51	38.03
oracle	mllr C2	15.04	**16.53**	18.99	27.26	21.98	36.49
oracle	mllr C3	14.76	17.82	**17.87**	28.24	**20.48**	37.12
oracle	mllr C4	16.21	17.75	19.60	**26.55**	22.53	35.41
oracle	mllr C5	15.14	18.15	18.29	28.47	**20.68**	37.15
oracle	mllr C6	17.04	18.71	19.94	27.20	23.26	**34.15**
	mlp $\nabla_p = 0.0$	14.08	**16.53**	**17.87**	**26.55**	20.70	**34.15**
rover	$\nabla_p = 0.5$	**14.00**	**16.52**	**17.84**	**26.52**	**20.53**	**34.15**
rover	$\nabla_p = 1.0$	14.82	17.04	18.18	27.18	20.78	35.94

Table 2. Word error rate (WER %) of each test class with clean-condition, multistyle and ideal matched training modes, as well as the MLLR mean cluster-based adaptation with the *oracle* method and the proposed selection scheme $\nabla_p = 0.0$. ROVER is used for system combination with selection schemes $\nabla_p = 0.5$ and 1.0.

oracle method (the reverberant condition for test data is assumed to be known in advance) in Table 2 show that a match of the cluster-dependent model with its test condition results in the lowest WERs. On the other hand, the performance is prone to be even worse than the results of multistyle training when some mismatched adapted models are selected, especially when their (T_{60}, C_{50}) values differ significantly, which indicates that a smart model selector is important. By means of the proposed selector with $\nabla_p = 0.0$, each test utterance is classified to the specific cluster-dependent model for recognition. The comparable WERs to the *oracle* situation testify the effectiveness of this selector based on MLP estimator for optimal matched model selection among cluster-dependent models.

Interestingly, some mismatched models also fit other test sets, e.g. test C5 achieves even lower WER when using an adapted HMM from C3, most likely due to their similar C_{50} condition as shown in Table 2 and Fig. 2 (c). Motivated by this, ROVER is used to combine these beneficial models which still work fine compared to the matched one. In other words, these beneficial models can be derived from the similarities of the class-wise posterior probabilities to $p_{\max}$ as described in (4). Results show that the model selector with $\nabla_p = 0.5$ offers ROVER the potentially beneficial cluster-dependent models to further reduce WERs. When increasing the threshold further to $\nabla_p = 1.0$, detrimental models are selected as well, which results in average performance.

	+ cmllr	C1	C2	C3	C4	C5	C6
oracle	multistyle	13.95	16.25	17.59	25.94	20.82	33.08
oracle	mllr matched C	13.56	15.42	17.54	25.59	20.45	32.71
	mlp $\nabla_p = 0.0$	13.58	15.42	17.54	25.59	20.42	32.71
rover	$\nabla_p = 0.5$	**13.53**	**15.39**	17.45	25.56	20.36	32.68
rover	$\nabla_p = 1.0$	13.70	15.92	**17.40**	**25.33**	**20.31**	**32.66**

Table 3. Word error rate (WER %) of each test class with an unsupervised CMLLR adaptation on the basis of multistyle training and matched cluster-dependent adapted models with the *oracle* method. MLP-based scores for one selected model ($\nabla_p = 0.0$) are reported, as well as WERs for ROVER-based systems that consider a medium number of system outputs ($\nabla_p = 0.5$) or all models ($\nabla = 1.0$).

Finally, we explore the integration of an *unsupervised* adaptation scheme in the cluster-based adaptation, in which the recognition results of the test data in a *batch* mode (all the adaptation data is available before adaptation) will be used to further adapt the models that recognize the test data again. CMLLR is an established method for this adaptation or model-based feature normalization, and hence applied for the following experiments. Under this *batch* mode, compared to the *oracle* method to group all test data by a-priori known reverberant condition, the proposed system applies the selector $\nabla_p = 0.0$ to classify the test data into 6 classes, so as to determine the CMLLR adaptation transforms w.r.t. each cluster-dependent model. CMLLR adaptation reduces the WERs by additional 1 to 2% (Table 3) compared to the corresponding results in Table 2. Still, WERs with the matched model selector $\nabla_p = 0.0$ behave nearly the same as the *oracle* method, again indicating that the proposed selector is efficient for optimal matched cluster-dependent model selection. In contrast to results in Table 2, the best results are obtained when ROVER considers all models ($\nabla_p = 1.0$) for C3-C6. It seems, that although the unsupervised adaptation introduces a hypothesis bias [21], at the same time it results in increased complementary information when applied to cross-adapt different cluster-dependent models so that ROVER can benefit from those complementary recognition outputs after combining more recognizers.

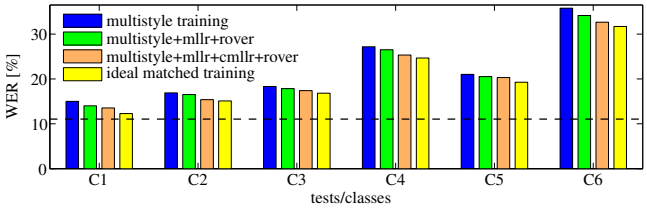

Fig. 3. WER comparison of each test class for multistyle training, MLLR adaptation with ROVER cooperated by the proposed selector, and its enhanced scenario by CMLLR, as well as the ideal matched training model. The performance without reverberation (clean HMMs and clean test data) is 11.06% (dashed line).

The reduced WERs as mentioned above by the synergy of matched cluster-dependent model selection and system combination based on the proposed selection scheme are summarized and illustrated in Fig. 3. Compared to multistyle training, the ideal matched training system reaches a relative improvement of 10.9%, which can be seen as an upper bound for the given task. For the proposed system that does not rely on a-priori information but is alternative to the *oracle* matched cluster-dependent model mode, the average relative improvements are 3.5% and 7.1% when applying MLLR adaptation together with ROVER and further enhancement by CMLLR, respectively. The ideal scores could be further approached by modeling the neighboring reverberant feature vectors to be conditional dependent in HMMs according to [28, 4].

5. CONCLUSION

This contribution combines the estimation of the room acoustic parameters T_{60} and C_{50} with the adaptation and system combination for ASR systems in reverberant environments. Results indicate the clarity index C_{50} to be of higher importance than T_{60} to measure the reverberation effects to ASR. By means of an MLP-based estimator, a novel selection scheme has been proposed to select the optimal cluster-dependent model to reach the performance of an *oracle* matched-model system based on MLLR cluster-based adaptation and CMLLR model-based feature normalization. As well, it assists ROVER to further reduce WERs by combining the recognizer outputs from optimal cluster-dependent models and models containing complementary information to approach the performance of the ideal matched training system. Compared to the multistyle training approach, a relative WER improvement of 7.1% is obtained by the proposed strategy for ASR systems in reverberant environments with T_{60} ranging from 218 to 710 ms and C_{50} from 6.8 to 30.5 dB.

6. REFERENCES

[1] D. Gelbart and N. Morgan, "Double the Trouble: Handling Noise and Reverberation in Far-Field Automatic Speech Recognition," in *Proc. Int. Conf. Spoken Language Process*, Colorado, USA, Sep. 2002, pp. 2185–2188.

[2] M. Wölfel and J. McDonough, *Distant Speech Recognition*, John Wiley & Sons Ltd, 2009.

[3] A. Sehr, *Reverberation Modeling for Robust Distant-Talking Speech Recognition*, Ph.D. thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany, Oct. 2009.

[4] 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 Processing Magazine*, vol. 29, no. 6, pp. 114–126, Nov. 2012.

[5] S. Furui, "Unsupervised Speaker Adaptation Method based on Hierarchical Spectral Clustering," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Glasgow, UK, May 1989, vol. 1, pp. 286–289.

[6] X. Huang and K.F. Lee, "On Speaker-Independent, Speaker-Dependent and Speaker-Adaptive Speech Recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 1, no. 2, pp. 150–157, 1993.

[7] C.J. Leggetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," *Computer Speech and Language*, vol. 9, pp. 171–185, Feb. 1995.

[8] C.K. Raut, T. Nishimoto, and S. Sagayama, "Model Adaptation for Long Convolutional Distortion by Maximum Likelihood based State Filtering Approach," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France., May 2006, vol. 1, pp. 1133–1136.

[9] H.G. Hirsch and H. Finster, "A New Approach for the Adaptation of HMMs to Reverberation and Background Noise," *Speech Commun.*, vol. 50, no. 3, pp. 244–263, 2008.

[10] A. Sehr, M. Gardill, and W. Kellermann, "Adapting HMMs of Distant-Talking ASR Systems using Feature-Domain Reverberation Models," in *Proc. European Signal Processing Conf. (EUSIPCO)*, Glasgow, Scotland, Aug. 2009, pp. 540–543.

[11] M.J.F. Gales, "Adaptive Training for Robust ASR," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Madonna di Campiglio, Italy, Dec. 2001, pp. 15–20.

[12] F. Xiong, S. Goetze, and B.T. Meyer, "Blind Estimation of Reverberation Time based on Spectro-Temporal Modulation Filtering," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 443–447.

[13] H. Kuttruff, *Room Acoustics*, Spon Press, London, 4th edition, 2000.

[14] V.V. Digalakis, D. Rtischev, and L.G. Neumeyer, "Speaker Adaptation using Constrained Estimation of Gaussian Mixtures," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 3, no. 5, pp. 357–366, Sep. 1995.

[15] J.G. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, CA, Dec. 1997, pp. 347–354.

[16] M.R. Schroeder, "New Method of Measuring Reverberation Time," *J. Acoust. Soc. Amer.*, vol. 37, no. 3, pp. 409–412, 1965.

[17] T. Nishiura, Y. Hirano, Y. Denda, and M. Nakayama, "Investigations into Early and Late Reflections on Distant Talking Speech Recognition toward Suitable Reverberation Criteria," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007, pp. 1082–1085.

[18] T. Fukumori, M. Morise, and T. Nishiura, "Performance Estimation of Reverberant Speech Recognition based on Reverberant Criteria RSR-Dn with Acoustic Parameters," in *Proc. Interspeech*, Makuhari, Japan, Sep. 2010, pp. 562–565.

[19] B.T. Meyer, S.V. Ravuri, M.R. Schädler, and N. Morgan, "Comparing Different Flavors of Spectro-Temporal Features for ASR," in *Proc. Interspeech*, Florence, Italy, Aug. 2011, pp. 1269–1272.

[20] M.R. Schädler, B.T. Meyer, and B. Kollmeier, "Spectro-Temporal Modulation Subspace-Spanning Filter Bank Features for Robust Automatic Speech Recognition," *J. Acoust. Soc. Am.*, vol. 131, no. 5, pp. 4134–4151, 2012.

[21] K. Yu, *Adaptive Training for Large Vocabulary Continuous Speech Recognition*, Ph.D. thesis, Hughes Hall College and Cambridge University Engineering Department, University of Cambridge, UK, Jul. 2006.

[22] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[23] M.J.F. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, Apr. 1998.

[24] M.J.F. Gales, "Cluster Adaptive Training for Speech Recognition," in *Int. Conf. on Spoken Language Processing*, Sydney, Australia, Dec. 1998, pp. 1783–1786.

[25] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJ-CAM0: A British English Speech Corpus for Large Vocabulary Continuous Speech Recognition," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Detroit, Michigan, USA, May 1995, pp. 81–84.

[26] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The REVERB Challenge: A Common Evaluation Framework for Dereverberation and Recognition of Reverberant Speech," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, Oct. 2013.

[27] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X.A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*, Cambridge University Engineering Department, Cambridge, 2009, http://htk.eng.cam.ac.uk/.

[28] A. Sehr, R. Maas, and W. Kellermann, "Reverberation Model-based Decoding in the Logmelspec Domain for Robust Distant-Talking Speech Recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1676–1691, Sep. 2010.