REVERBERATION ROBUST SPEECH RECOGNITION BY MATCHING DISTRIBUTIONS OF SPECTRALLY AND TEMPORALLY DECORRELATED FEATURES

Kalle J. Palomäki and Heikki Kallasjoki

Department of Signal Processing and Acoustics, Aalto University, Finland;
e-mail: {kalle.palomaki,heikki.kallasjoki}@aalto.fi.

ABSTRACT

This paper addresses dereverberation of speech using an unsupervised approach utilizing speech prior and taking only weak assumptions on reverberation. Our approach uses a long time context representation of reverberated speech in spectral-temporal supervectors which are decorrelated by PCA. In the decorrelated domain, supervectors are mapped from the reverberant speech distribution to the clean speech distribution and then to mel-spectral vectors. A mel-domain Wiener filter is applied as post processing. Our results demonstrate performance gains over the provided baseline recognizer, and show that the method can be coupled to CMLLR adaptation with cumulative benefits for clean trained models. Furthermore, we show that using dimensionality reduction coupled with the Wiener filter is better than using full-dimensional PCA in representing small variance components in speech.

Index Terms— dereverberation, speech recognition, supervector, decorrelation, unsupervised

1. INTRODUCTION

Considering automatic speech recognizers (ASR) used in practical applications, we cannot often control the recording conditions but are reliant on hardware which the end-users have. The quality of microphones, recording environments and distance between speaker and microphone can vary a great deal. The very same ASR system may need to cope with data from a single distant microphone, sophisticated arrays and close-talk microphones. This calls for methods that are unsupervised and not reliant on prior information about environments, arrays or specific microphones.

The conventional method to counteract effects of reverberation or transmission lines has been to produce robust features using cepstral mean normalization [1], modulation filtered spectrograms [2] or frequency domain linear prediction [3]. Their advantage is simplicity and wide applicability, since they make only weak prior assumptions on the data, but used alone they yield only modest performance gains. Other approaches that can be used to improve reverberation robustness that also make only weak assumptions are missing feature methods with masks designed for reverberation [4, 5] or simply unsupervised adaptation [6, 7]. The advantage of Bayesian dereverberation approaches over the above-mentioned is that they can flexibly utilize either coarse or more precise source (speech) and filter (reverberation) models jointly [8, 9]. However, precise modeling of the filter and the source require computationally expensive methods such as Monte Carlo Markov Chain sampling [9].

Almost regardless of the enhancement method, in practical ASR it is often common to utilize an adaptation method as the last step to counteract variations from speakers or environments. The adaptation can be based on, e.g., the acoustic model distributions in ASR [6], or data distributions using powerful non-linear distribution matching methods independently of the acoustic models [7, 10]. Non-linear adaptation can also be combined with acoustic model domain adaptation with cumulative benefits [7, 10].

The purpose of the present study is to develop a new adaptation method based on distribution matching that is suitable for dereverberation. Due to the long lasting effects of reverberation, we utilize decorrelated spectral-temporal supervectors that include time context information. Methods related to the present study are non-linear adaptation [7] and feature space gaussianization [10]. The present study extends the previous work specifically by addressing the problem of dereverberation, whereas the previous work dealt primarily with speaker adaptation [7] or general purpose feature space gaussianization [10] designed to produce features that are easy to model with Gaussian mixtures, with no specific intention to speech enhancement. Focusing on dereverberation leads us to use longer feature contexts. We also utilize post-filtering methods that were not addressed in the above-mentioned previous studies. Furthermore, we discuss the mathematical motivation why the proposed method is suitable for dereverberation.

This work was supported by TEKES FuNeSoMo (KJP), KAUTE and Emil Aaltonen foundations (KJP), and by the Academy of Finland grants 136209 (KJP), and 251170 (KJP, HK, Finnish Centre of Excellence program 2012-2017) the Hecse graduate school (HK). The recognition experiments presented in this work were performed in part using computational resources within the Aalto University "Science-IT" project.

2. DEREVERBERATION METHOD

Dereverberation can be considered as a Bayesian inverse problem, in which an attempt is made to recover clean speech spectra $\boldsymbol{o}_{\boldsymbol{x}}$ given noisy speech spectra $\boldsymbol{o}_{\boldsymbol{y}}$. Posterior distribution for dereverberated speech $p(\boldsymbol{o}_{\boldsymbol{x}}|\boldsymbol{o}_{\boldsymbol{y}})$ is then

$$p(\boldsymbol{o}_{\boldsymbol{x}}|\boldsymbol{o}_{\boldsymbol{y}}) \;\propto\; p(\boldsymbol{o}_{\boldsymbol{x}})p(\boldsymbol{o}_{\boldsymbol{y}}|\boldsymbol{o}_{\boldsymbol{x}}), \tag{1}$$

where $p(\boldsymbol{o}_{\boldsymbol{x}})$ is the clean speech prior and $p(\boldsymbol{o}_{\boldsymbol{y}}|\boldsymbol{o}_{\boldsymbol{x}})$ represents the reverberant observation.

In signal processing terms, reverberation can be considered as convolutive interference by a reasonable accuracy. The convolution $\mathsf{b}(t) = (\mathsf{o} * \mathsf{h})(t)$ of a time domain speech signal $\mathsf{o}(t)$ and a FIR filter $\mathsf{h}(t)$ can be expressed as the matrix operation

$$\mathbf{b} = \mathbf{Ho} \tag{2}$$

where $\mathbf{H}$ is the Toeplitz matrix that represents the filter h and $\mathbf{b}$ is the resulting reverberated signal.

Similarly, we can express convolution in the feature domain using linear transformations. First, a supervector $\boldsymbol{s}(t) = [\boldsymbol{o}(t)^\top \;\ldots\; \boldsymbol{o}(t+T-1)^\top]^\top$ is formed from concatenation of T consecutive frames of spectral feature vectors $\boldsymbol{o}(t)$, where T is chosen large enough considering the length of the room impulse response. The dimensionality of the supervector $\boldsymbol{s}(t)$ is $N = TK$, where K is the dimensionality of the original features $\boldsymbol{o}(t)$. Dropping the time index t to keep the notation simpler, we denote by $\boldsymbol{s}_{\boldsymbol{x}}$ and $\boldsymbol{s}_{\boldsymbol{y}}$ the supervectors corresponding to clean speech spectra $\boldsymbol{o}_{\boldsymbol{x}}$ and reverberant spectra $\boldsymbol{o}_{\boldsymbol{y}}$, respectively. The speech features $\boldsymbol{s}_{\boldsymbol{y}}$ that are affected by convolution can be approximated as

$$\boldsymbol{s}_{\boldsymbol{y}} \approx \boldsymbol{H}_{\boldsymbol{y}}\boldsymbol{s}_{\boldsymbol{x}}, \tag{3}$$

where the filter matrix $\boldsymbol{H}_{\boldsymbol{y}}$ performs convolution operation in each frequency channel with the samples included in the supervector $\boldsymbol{s}_{\boldsymbol{x}}$.

For a transformation matrix $\boldsymbol{H}$ corresponding to an arbitrary impulse response, a linear transformation $\boldsymbol{D}$, such as principal component analysis (PCA), can be applied to decorrelate the elements of transformed supervectors $\boldsymbol{Hs}$ as

$$\boldsymbol{c} = \boldsymbol{DHs}, \tag{4}$$

which allows treating the elements of $\boldsymbol{c}$ one-by-one. Given a single supervector $\boldsymbol{s}$ observed under two different convolutive distortions represented by matrices $\boldsymbol{H}_i$ and $\boldsymbol{H}_j$, we denote $\boldsymbol{c}_i = \boldsymbol{DH}_i\boldsymbol{s}$ and $\boldsymbol{c}_j = \boldsymbol{DH}_j\boldsymbol{s}$. As the system is linear, we can write $\boldsymbol{c}_i = \boldsymbol{A}_{\boldsymbol{ij}}\boldsymbol{c}_j$. Under the assumption that matrix $\boldsymbol{D}$ successfully decorrelates both $\boldsymbol{c}_i$ and $\boldsymbol{c}_j$, matrix $\boldsymbol{A}_{\boldsymbol{ij}}$ is diagonal. In this case, we can represent the mapping as an element-wise multiplication of $\boldsymbol{c}_j$ by the diagonal elements of $\boldsymbol{A}_{\boldsymbol{ij}}$,

$$c_i(n) \approx [\boldsymbol{A}_{\boldsymbol{ij}}]_{nn}c_j(n), \tag{5}$$

where $n = \{1, \ldots, N\}$ indexes the elements of the supervectors and diagonal elements of A_{ij}.

For speech data, however, PCA is usually applied after a logarithmic non-linearity,

$$\boldsymbol{c'} = \boldsymbol{D'} \log \boldsymbol{Hs}, \tag{6}$$

where the $\log$ operation is computed element-wise. In this non-linear case, we can write the corresponding transformation between a pair of supervectors as $\boldsymbol{c'}_i = F_{ij}(\boldsymbol{c'}_j)$, assuming that there is a bijective non-linear transformation F_{ij}. By again assuming that matrix $\boldsymbol{D'}$ decorrelates both $\boldsymbol{c'}_i$ and $\boldsymbol{c'}_j$, we can use a set of element-wise mappings $F_{ij}^{(n)}$ so that

$$c'_i(n) \approx F_{ij}^{(n)}(c'_j(n)). \tag{7}$$

In order to simplify the notation, from now on we operate on individual components of the decorrelated supervectors, and drop all indices n. To develop suitable mapping functions F, we apply a distribution matching method similar to [7, 10]. Empirical cumulative distribution $\Phi(c')$ of a variable c' can be approximated by

$$\Phi(c') = \frac{1}{L}\sum_{k=1}^{L}\theta(c' - c'_k), \tag{8}$$

where θ is step function over L samples of form c'_k drawn from the distribution. It follows that simply sorting and scaling data gives an approximation of its inverse cumulative distribution function (ICDF). For dereverberation under a particular recording condition, we use a mapping derived from the empirical distributions of clean speech samples c'_x and reverberant samples c'_y. We denote the ICDFs of the clean and reverberant speech samples by Φ_x^{-1} and Φ_y^{-1}, respectively. The mapping function F_{xy} approximating the transformation from reverberant to clean speech is implemented by constructing a lookup table $\Phi_y^{-1} \underset{F}{\longrightarrow} \Phi_x^{-1}$ using Matlab `interp1` with piecewise cubic interpolation. In terms of Equation (1), this can be seen as approximating the reverberant speech posterior $p(c'_y \mid c'_x)$ and clean speech prior $p(c'_x)$ by samples of corresponding data.

After the lookup tables are defined, a full decorrelated reverberant supervector $\boldsymbol{c'}_y$ can be transformed to a dereverberated log-spectral supervector $\tilde{\boldsymbol{s}}$ estimate by

$$\tilde{\boldsymbol{s}}' = \boldsymbol{D'}^{-1}F_{xy}(\boldsymbol{c'}_y), \tag{9}$$

where mapping F_{xy} is defined by applying individual $F_{xy}^{(n)}$ lookup tables element by element to $\boldsymbol{c'}_y$, and $\boldsymbol{D'}^{-1}$ inverts PCA to get back to the log spectral-temporal domain. Then the supervector representation is dismantled to get estimates of the dereverberated speech log mel-spectrograms $\tilde{\boldsymbol{o}_{\boldsymbol{x}}}'$. Each vector $\tilde{\boldsymbol{o}_{\boldsymbol{x}}}'$ is obtained by taking the average of overlapping samples from all supervectors that contain data for its time frame.

For each corresponding linear domain frame $\tilde{\boldsymbol{o}_{\boldsymbol{x}}} = \log \tilde{\boldsymbol{o}_{\boldsymbol{x}}}'$, we construct a frame-specific mel-spectral domain

Wiener filter, in a manner common to many speech enhancement systems [11]. We define the Wiener filter $\boldsymbol{h_w}$ as

$$\boldsymbol{h_w} = \tilde{\boldsymbol{o_x}} \,./\, \tilde{\boldsymbol{o_y}} \tag{10}$$

where $./$ denotes element-wise division and $\tilde{\boldsymbol{o_y}}$ represents reverberant data that has gone through the same PCA transformation $\boldsymbol{D'}$ and dimensionality reduction operations as the dereverberated data. The generation of $\tilde{\boldsymbol{o_y}}$ omits the lookup table mapping step which is the difference in the procedure compared to generation of $\tilde{\boldsymbol{o_x}}$. The final enhanced mel-spectral features are then obtained as

$$\hat{\boldsymbol{o_x}} = \boldsymbol{h_w} \,.*\, \boldsymbol{o_y}. \tag{11}$$

In the logarithmic domain, with $\boldsymbol{o}' = \log \boldsymbol{o}$, this can be written as

$$\hat{\boldsymbol{o_x}}' = \tilde{\boldsymbol{o_x}}' + \boldsymbol{o_y}' - \tilde{\boldsymbol{o_y}}', \tag{12}$$

where it can be seen more clearly that, through the residual term $\boldsymbol{o_y}' - \tilde{\boldsymbol{o_y}}'$, the filter sums back some of the variation in the reverberant signals that is lost in the dereverberated estimate $\tilde{\boldsymbol{o_x}}'$, after it has been smoothed by the low-order PCA.

After obtaining the initial estimate of dereverberated speech $\hat{\boldsymbol{o_x}}$, we apply the same process defined in equations (3) to (12) iteratively, by substituting reverberant observation $\boldsymbol{o_y}$ with the current estimate $\hat{\boldsymbol{o_x}}$. Finally, after two iterations, the resulting estimates $\hat{\boldsymbol{o_x}}$ are used as the source of acoustic features for the speech recognition.

3. EXPERIMENTS

This section describes the experimental evaluation of the proposed system, using the data sets (Sect. 3.1) and the baseline recognizers (Sect. 3.2) provided by the REVERB challenge. Parameters settings of the proposed approach (Sect. 3.2) and finally the results (Sect. 3.3) are also shown.

3.1. Data

The reverberant speech feature enhancement methods described in this work are evaluated on both artificially distorted clean speech ("SimData") and speech recorded in a noisy, reverberant room ("RealData"). Both data sets are provided by the REVERB challenge, and described in detail in [12]. Separate development and evaluation subsets are provided.

For SimData, clean speech utterances from the WSJCAM0 British English continuous speech recognition corpus [13] are first distorted using measured room impulse responses, and then mixed with measured room noise with a fixed signal-to-noise ratio (SNR) of 20 dB. Utterances in six simulated reverberant environments are provided: two speaker-to-microphone distances (near, far) in each of three rooms of varying size (small, medium, large). The near and far microphone distances are 0.5 m and 2.0 m, while T_{60} reverberation times for the small, medium and large rooms were 0.25 s, 0.5 s and 0.7 s, respectively. The total number of utterances is 1484 and 2176 for the development and evaluation subsets, respectively.

The RealData set consists of real recordings of speakers in a reverberant meeting room. Contents of the utterances are based on the prompts of the WSJCAM0 corpus. The set contains two different test conditions, corresponding to near and far microphone distances of 1.0 m and 2.5 m, respectively. There are, respectively, 179 and 372 utterances in the development and evaluation subsets of RealData.

3.2. Speech Recognition System and Settings

The baseline recognizer provided by the REVERB challenge, based on the HTK toolkit [14], is used to evaluate the speech recognition performance of the proposed methods. The recognition system uses 13-dimensional Mel-frequency cepstral coefficients (MFCCs) augmented with first (D) and second time derivatives (DD). Hidden Markov models with 10-component Gaussian mixture emission distributions are used to model the acoustic features. The clean speech training set of the WSJCAM0 corpus [13] is used to train the acoustic models.

Unsupervised constrained MLLR (CMLLR) adaptation is optionally applied during recognition. For each test condition (room and recording distance), adaptation coefficients for 256 regression classes are calculated based on the entire test set. Unadapted recognition results are used to provide transcriptions for the adaptation.

The proposed distribution matching (DM) system uses the MFCC+D+DD acoustic features that are, except the feature enhancement step, identical to those of the baseline recognizer. The distribution matching based feature enhancement step (see Section 2) is performed on the $K = 23$ dimensional mel-spectral features computed during the MFCC processing as follows.

First, the normalization method proposed in [4] is used to further reduce the effects of any spectral and gain alteration due to reverberation in the mel-spectral features. The normalization method is based on estimating gains of each frequency channel from the largest energy samples along the time trajectory. This is based on the observation that large energy time-frequency bins are more likely to contain clean speech or early reflections than those of lower energy that may contain energy in reverberation tails. This allows similar normalization based on clean speech alone regardless of the reverberation (or noise) conditions. Then spectral supervector representation of the DM system is constructed from normalized mel-spectral feature vectors and it uses a time context of $T = 20$ frames. The performance is also evaluated with no time context ($T = 1$).

The PCA transformation in Equation (6) is estimated for clean speech data taken from the training part of corpus for over 1000 utterances. For the primary approach, with a time

Table 1. Evaluation test word error rates on clean conditions. The results are shown for the baseline system without (Baseline) and with (Baseline-ada) CMLLR adaptation, and similarly for two versions of the proposed method, without (DM) and with CMLLR-adaptation (DM-ada). The best results are bolded.

Method	Room			Ave.
	1	2	3	
Baseline	12.89	12.64	12.13	12.55
Baseline-ada	**11.78**	**11.42**	**11.21**	**11.47**
DM	12.92	12.67	12.06	12.55
DM-ada	11.84	11.50	11.45	11.59

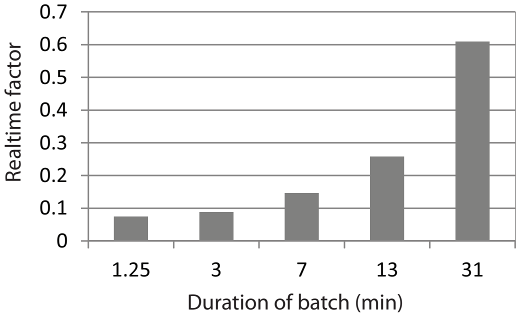

Fig. 1. Real-time factor as function of batch length used for distribution mapping.

context of $T = 20$ frames, experiments are performed using both at the full dimensionality of 460 as well as with a dimensionality reduction step where only the 40 principal components are retained. A dimensionality of 12 is used for the variant that has no time context ($T = 1$). In the recognition phase, unless otherwise noted, we assume full batch processing and always collect the reverberant posterior distribution estimates of Equation (8) based on the whole evaluation or development test condition, and for the corresponding speech prior we use an equal length sample taken from the clean training set.

3.3. Results

Tables 1 and 2 show the word error rate results of the evaluation test on the clean and reverberant data, respectively. For the *reverberant* data (Table 2), the systems can be ranked from worst to best based on overall averages in the following order: baseline without adaptation (Baseline), baseline with adaptation (Baseline-ada), proposed system without adaptation (DM) and the proposed system with adaptation (DM-ada). When applied without adaptation, the proposed method (DM) outperforms the baseline (Baseline) in all reverberant conditions. Similarly, when each system is applied with adaptation, the proposed system (DM-ada) outperforms the adapted baseline (Baseline-ada) in all reverberant conditions. For *clean* data (Table 1), the best performing system is the baseline with CMLLR adaptation (Baseline-ada), with a small margin compared to the second best (DM-ada).

Table 3 demonstrates effects of different parameters in the results, for three different aspects, using the development set data. First, the effect of the time context length is addressed. If the performance is better for contexts considerably longer than one frame, it can be used as a evidence that a long temporal context is required. In the development set results, the system using $T = 20$ time frames outperforms the one using a window length of $T = 1$, equivalent to having no context. Second, we compare our primary approach of using dimensionality reduction to 40 principal components to the full dimensional PCA that retains all components, and observe that better results in all cases are obtained for the 40-component approach. Third, we compare results of our approach with and without the Wiener filter. We notice that without the Wiener filter, the performance drops even below that of the baseline system in several cases. When the parameter settings are contrasted, we notice that having sufficient temporal context is more important than the dimensionality reduction for the PCA (40 vs. full), and that the Wiener filter is beneficial when the dimensionality reduction is applied.

Figure 1 demonstrates computational time of the method without an ASR back-end, showing the real-time factor against length of the batch used for the lookup table construction. The software implementation in Matlab is run on a single core of an Intel(R) Xeon(R) CPU E3-1230 V2 processor at 3.30 GHz. The rightmost bar in the figure denotes the setting that was used to conduct the ASR simulations in this study. It corresponds to using a full batch of development set Room 3 far-condition data, which is ca. 31 min in duration. The real-time factor in that case was 0.61. Required computation drops when the batch length is reduced, as less data is used for the lookup table construction. We did not pay particular attention to computational efficiency, thus with a little effort it should be possible to implement the approach with reduced computational cost. In the present version we have prioritized the ease of implementation at the cost of some unnecessary computation, such as reconstructing the lookup table for every utterance. These could be simply done once for each batch with a little effort in optimizing the implementation.

4. DISCUSSION

In this study, we addressed speech dereverberation using an unsupervised single channel approach that utilizes a speech prior, but makes only weak assumptions on the properties of reverberation. The assumptions that we make, or that are built in our method, are that a long temporal context is required, reverberation has a convolutive effect, and that we can successfully decorrelate both the clean and reverberant speech, when represented as long-context supervectors of short-term spectral observations, using a PCA transformation learned for clean speech. Our results demonstrate that in the all reverberant cases, we achieve better performance compared to the clean speech trained baseline. Furthermore, we showed that our method can be coupled to CMLLR adaptation with cumulative benefits.

Tests with different parameter settings on the proposed system demonstrate that it is essential to use long context in the supervector representation. In this paper, we demonstrate this by comparing a temporal context of only one frame to the proposed 20 frame context, but our earlier development tests showed that a 20 frame context was better than e.g. a 10 frame context. Showing this and related results was omitted for compactness, and as they are not compatible with the final version of the approach presented in this paper. Our results in comparing the systems with and without the Wiener filter demonstrate, first of all, that spectra originating directly from a low dimensional PCA representation are overly smooth to represent speech accurately. Secondly, comparing the high dimensional PCA to the Wiener filtered low dimensional PCA, we notice that the Wiener filter is better at representing short-term variation than utilizing the high dimensional PCA.

In the present study, we chose to use full batch processing over utterance-wise from the challenge alternatives, because of the need to have a sufficient amount of data available for robust estimation of the distribution of the reverberant speech features. In the development stage, we conducted experiments with a version using a posterior model that is accumulated utterance by utterance, and the full batch version was only marginally better. Systematic investigations on the need of adaptation data are left, however, for future studies.

Regarding the computational cost of the proposed method, it should be straightforward to implement it more efficiently. The first step would be to remove unnecessary computation that was left in the method for ease of implementation (see Sect. 3.3). Secondly, the computation could be reduced through histogram equalization using a more coarse distribution sampling [7]. After collecting sufficient data for the estimation of the reverberant posterior distribution in the corresponding condition, the method can applied with a low latency of only one spectral frame, if the supervector context is taken to represent the past mel-spectra. The question of whether full batch data is actually necessary is out of the scope of this study.

The present study used a standard PCA to decorrelate the spectral supervectors. During the development of the method, we also conducted experiments using more sophisticated approaches, such as the stacked denoising auto-encoder (SDAE) [15], with which we generated bottleneck features of similar dimensionality as the PCA used in this work. With SDAE, we obtained better speech reconstruction accuracy, but coupling

Table 2. Evaluation test word error rates on reverberant conditions. The results are shown for the baseline system without (Baseline) and with (Baseline-ada) CMLLR adaptation, and similarly for two versions of the proposed method, without (DM) and with CMLLR-adaptation (DM-ada). The best results are bolded.

	SimData							RealData		
	Room 1		Room 2		Room 3		Ave.	Room 1		Ave.
	Near	Far	Near	Far	Near	Far	–	Near	Far	–
Baseline	18.32	25.77	42.71	82.71	53.56	87.97	51.82	90.07	88.01	89.04
Baseline-ada	14.86	19.10	24.59	64.48	34.16	79.34	39.40	82.88	80.49	81.68
DM	18.20	23.01	27.99	53.53	37.47	67.14	37.87	73.14	71.37	72.25
DM-ada	**14.74**	**18.50**	**20.89**	**39.59**	**26.74**	**51.80**	**28.70**	**64.32**	**60.16**	**62.24**

Table 3. Development test word error rate results on reverberant test conditions, all without CMLLR adaptation. Results are shown for the baseline recognizer (Baseline), the three parameter settings of the proposed method — the version used for the final results utilizing a context of $T = 20$ time frames (DM), the version with no time context ($T = 1$), and the version using $T = 20$ time frames and no dimensionality reduction (full dim) — as well as the variant without Wiener filter (no filter). The best results are bolded.

	SimData							RealData		
	Room 1		Room 2		Room 3		Ave.	Room 1		Ave.
	Near	Far	Near	Far	Near	Far	–	Near	Far	–
Baseline	15.29	25.29	43.90	85.80	51.95	88.90	51.81	88.71	88.31	88.51
DM ($T = 20$)	**14.75**	**21.95**	**28.44**	**56.15**	**34.52**	**63.95**	**36.60**	**62.69**	**64.46**	**63.57**
$T = 1$	16.10	24.93	30.88	74.14	39.94	79.23	44.17	70.99	71.09	71.03
full dim	15.46	23.13	29.48	62.53	35.44	67.68	38.92	66.56	68.97	67.75
no filter	32.30	42.65	53.76	78.75	61.72	85.93	59.15	79.23	80.45	79.83

SDAE to the distribution mapping did not perform as well as the simpler PCA. Non-linear independent component analysis was also tried out in earlier development stages. One of the reasons for the superior performance of PCA might be that we have learned mappings only from clean speech. Using data from the distribution of reverberant speech in learning the mapping is certainly possible, but comes with an increase in computational cost. However, using more sophisticated decorrelation methods is certainly among our future interests.

5. REFERENCES

[1] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 29, pp. 254–272, 1981.

[2] B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, pp. 117–132, 1998.

[3] S. Thomas, S. Ganapathy, and H. Hermansky, "Recognition of reverberant speech using frequency domain linear prediction," *IEEE Signal Processing Letters*, vol. 15, pp. 681–684, 2008.

[4] K. J. Palomäki, G .J. Brown, and J.P. Barker, "Techniques for handling convolutional distortion with 'missing data' automatic speech recognition," *Speech Communication*, vol. 42, pp. 123–142, 2004.

[5] K. J. Palomäki, G. J. Brown, and J. Barker, "Recognition of reverberant speech using full cepstral features and spectral missing data," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP-2006)*, Tolouse, France, 2006, vol. 1, pp. 289–292.

[6] M.J.F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Comput. Speech Lang.*, vol. 12, pp. 75–98, 1998.

[7] S. Dharanipragada and M. Padmanabhan, "A non-linear unsupervised adaptation technique for speech recognition," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP-2000)*, Beijing, 2000.

[8] A. Krueger, O. Walter, V. Leutnant, and R. Haeb-Umbach, "Bayesian Feature Enhancement for ASR of Noisy Reverberant Real-World Data," in *Proc. Interspeech*, Portland, USA, Sep. 2012.

[9] C. Evers, J. R. Hopgood, and J. Bell, "Blind speech dereverberation using batch and sequential monte carlo methods," in *IEEE Int. Symposium on Circuits and Systems*, May 2008, pp. 3226–3229.

[10] G. Saon, S. Dharanipragada, and D. Povey, "Feature space gaussianization," in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP-2004)*, 2004, vol. I, pp. 329–332.

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

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

[13] 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. Int. Conf. Acoust. Speech Signal Process. (ICASSP-1995)*, Detroit, MI, USA, 1995, pp. 81–84.

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

[15] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, Dec. 2010.