EXEMPLAR-BASED SPEECH ENHANCEMENT FOR DEEP NEURAL NETWORK BASED AUTOMATIC SPEECH RECOGNITION

Deepak Baby[⋆] *Jort F. Gemmeke*[⋆] *Tuomas Virtanen*[†] *Hugo Van hamme*[⋆]

[⋆]Department ESAT, KU Leuven, Belgium
[†]Department of Signal Processing, Tampere University of Technology, Finland
{Deepak.Baby, Hugo.Vanhamme}@esat.kuleuven.be, jgemmeke@amadana.nl, Tuomas.Virtanen@tut.fi

ABSTRACT

Deep neural network (DNN) based acoustic modelling has been successfully used for a variety of automatic speech recognition (ASR) tasks, thanks to its ability to learn higher-level information using multiple hidden layers. This paper investigates the recently proposed exemplar-based speech enhancement technique using coupled dictionaries as a pre-processing stage for DNN-based systems. In this setting, the noisy speech is decomposed as a weighted sum of atoms in an input dictionary containing exemplars sampled from a domain of choice, and the resulting weights are applied to a coupled output dictionary containing exemplars sampled in the short-time Fourier transform (STFT) domain to directly obtain the speech and noise estimates for speech enhancement. In this work, settings using input dictionary of exemplars sampled from the STFT, Mel-integrated magnitude STFT and modulation envelope spectra are evaluated. Experiments performed on the AURORA-4 database revealed that these pre-processing stages can improve the performance of the DNN-HMM-based ASR systems with both clean and multi-condition training.

***Index Terms*—** deep neural networks, non-negative matrix factorisation, coupled dictionaries, speech enhancement, modulation envelope

1. INTRODUCTION

Automatic speech recognition (ASR) in realistic conditions, where the acoustic data is mixed with a variety of noises and channel variations, is still a major research challenge. Most of the traditional ASR systems, with acoustic modelling based on Gaussian mixture models (GMMs), make use of some speech/feature enhancement mechanism as a pre-processing stage to improve the system robustness. Monaural signal separation techniques like non-negative factorisation (NMF) [1], which exploit a long temporal context, have been successfully used as a speech enhancement front-end for improving the GMM-HMM-based ASR performance [2,3].

Recently, acoustic modelling using deep neural networks (DNNs), dubbed DNN-HMM-hybrid systems, gained popularity over the GMMs due to their improved robustness in realistic conditions [4]. State-of-the-art DNN-based systems contain multiple hidden layers, which enable the setting to learn higher-level information in the acoustic data, together with an output layer that are trained to provide pseudo-likelihoods for the states of an HMM [5,6].

In this work, we investigate the influence of exemplar-based speech enhancement using NMF on the performance of a DNN-HMM-hybrid setting. NMF-based speech enhancement systems work by decomposing the noisy speech as a sparse non-negative linear combination of speech and noise exemplars stored as atoms in a dictionary. The resulting speech and noise estimates are then used to generate a time-varying filter to enhance the noisy STFT and the enhanced speech is obtained using overlap-add method with the enhanced STFT [7,8].

It has been found advantageous to use exemplars from a feature space other than the STFT domain [2,9]. In this case, the mapping of speech and noise estimates to the STFT space may be a low-rank approximation (e.g., Mel feature space [8]) or even non-linear and non-unique (e.g., modulation spectrogram (MS) domain [10]). An approach using coupled dictionaries, where an output dictionary containing exemplars sampled from the STFT domain is used to directly obtain the speech and noise estimates in the STFT space following the decomposition using the input dictionary (containing exemplars from the Mel or MS space), has successfully been used to overcome these issues [3,9].

There exist some studies that investigate the application of a speech enhancement front-end for DNN-based ASR setting. The study presented in [11] shows that the performance of a DNN-based setting can be improved by using a front-end based on the DOLPHIN speech enhancement algorithm [12] which makes use of spectral and locational characteristics of speech and noise for noise reduction. A feature enhancement front-end based on Cepstral-domain minimum mean squared error (C-MMSE) criterion [13] was investigated in [14] which yielded only marginal improvements with a DNN trained on enhanced noisy data.

NMF speech enhancement using Mel features for DNNs was previously explored in [8] and the setting was found to improve the ASR performance. However, the setting used a pseudo-inverse to map the speech and noise estimates in the Mel space to the STFT space. Using a pseudo-inverse will always result in a low-rank approximation [9] which may be detrimental for a large vocabulary task. In this work, we consider this setting as one of the baseline settings and further explore using exemplar-based speech enhancement for various choices of exemplars together with coupled dictionaries for DNN-based ASR systems. We also investigate whether exemplar-based techniques can further mitigate speaker variability in a DNN-based setting, as observed in a GMM-based ASR setting [15].

This work mainly investigates the following aspects of a DNN-based ASR decoder: how much an exemplar-based speech enhancement front-end with Mel, STFT and MS exemplars can benefit a DNN trained on clean data? Can the performance of DNNs trained

This work has been funded with support from the European Commission under Contract FP7-PEOPLE-2011-290000 (INSPIRE) and IWT-SBO Project 100049 (ALADIN).

on multi-condition data be further improved using an exemplar-based pre-processing stage? Will the low-rank approximation while mapping the Mel estimates to the STFT space have any detrimental effects on the ASR task? How much a DNN-based ASR setting can benefit from enhancement using the MS features together with a coupled STFT dictionary?

In section 2, we describe the exemplar-based speech enhancement technique using coupled dictionaries and the DNN-based back-end for ASR. Section 3 details the various settings evaluated in this work. The experimental setup for evaluation on the AURORA-4 database is explained in Section 4 followed by results and discussion in Section 5. Section 6 concludes the paper.

2. METHODOLOGY

2.1. Speech enhancement using coupled dictionaries

The speech enhancement technique using coupled dictionaries is explained in detail in [9]. Here we only summarize the main steps in the algorithm.

For exemplar-based speech enhancement using coupled dictionaries, the NMF-based decomposition is done using an input dictionary $\mathbf{A}^{\text{in}} = [\mathbf{A}_{\mathbf{s}}^{\text{in}}\ \mathbf{A}_{\mathbf{n}}^{\text{in}}]$, where $\mathbf{A}_{\mathbf{s}}^{\text{in}}$ and $\mathbf{A}_{\mathbf{n}}^{\text{in}}$ are speech and noise dictionaries containing exemplars sampled from speech and noise, respectively. The exemplars can be from an additive and non-negative feature of choice extracted from random segments of training data spanning T frames (which are reshaped to a vector) for temporal continuity. We refer to this exemplar space, where the NMF-based decomposition is obtained, as the *input exemplar space* and is denoted using the superscript ‘in’.

For testing, the noisy data is first converted to the time-frequency representation used to create the exemplars, and a sliding window of length T frames is moved along its frame-axis with a hop size of 1 frame. The features belonging to each of these windows are reshaped and stored as columns in the observation matrix $\mathbf{\Psi}^{\text{in}}$ which is decomposed to obtain the activations $\mathbf{X}^{\text{in}}$ as:

$$\mathbf{\Psi}^{\text{in}} \approx \begin{bmatrix} \mathbf{A}_{\mathbf{s}}^{\text{in}} & \vdots & \mathbf{A}_{\mathbf{n}}^{\text{in}} \end{bmatrix} \begin{bmatrix} \mathbf{X}_{\mathbf{s}}^{\text{in}} \\ \cdots \\ \mathbf{X}_{\mathbf{n}}^{\text{in}} \end{bmatrix} = \mathbf{A}^{\text{in}}\mathbf{X}^{\text{in}} \quad s.t. \quad \mathbf{X}^{\text{in}} \geq 0. \qquad (1)$$

The approximation is done to minimize the Kullback-Leibler divergence between $\mathbf{\Psi}^{\text{in}}$ and $\mathbf{A}^{\text{in}}\mathbf{X}^{\text{in}}$ with an additional sparsity constraint on $\mathbf{X}^{\text{in}}$ [2].

To directly obtain the magnitude STFT estimates (denoted with superscript ‘dft’) of speech and noise, we use a coupled output STFT dictionary $\mathbf{A}^{\text{dft}} = [\mathbf{A}_{\mathbf{s}}^{\text{dft}}\ \mathbf{A}_{\mathbf{n}}^{\text{dft}}]$, where $\mathbf{A}_{\mathbf{s}}^{\text{dft}}$ and $\mathbf{A}_{\mathbf{n}}^{\text{dft}}$ contains exemplars extracted from the same random pieces of training data used to create $\mathbf{A}_{\mathbf{s}}^{\text{in}}$ and $\mathbf{A}_{\mathbf{n}}^{\text{in}}$, respectively. The windowed magnitude STFT estimates for speech and noise are obtained respectively as $\hat{\mathbf{s}}_{\mathbf{w}} = \mathbf{A}_{\mathbf{s}}^{\text{dft}}\mathbf{X}_{\mathbf{s}}^{\text{in}}$ and $\hat{\mathbf{n}}_{\mathbf{w}} = \mathbf{A}_{\mathbf{n}}^{\text{dft}}\mathbf{X}_{\mathbf{n}}^{\text{in}}$. Notice that there are multiple instances of the same frame appearing over multiple columns in this windowed estimate. The frame-level speech and noise estimates in the magnitude STFT domain, $\hat{\mathbf{s}}$ and $\hat{\mathbf{n}}$, are then obtained by averaging out the frames belonging to multiple overlapping windows as explained in [2]. Let this operation be deonted as $[\cdot]^*$, i.e. for example $\hat{\mathbf{s}} = [\hat{\mathbf{s}}_{\mathbf{w}}]^*$.

The noisy STFT $\mathbf{Y}$ is enhanced as $\mathbf{Y}_{\text{enh}} = \mathbf{Y} \odot \hat{\mathbf{s}} \oslash (\hat{\mathbf{s}} + \hat{\mathbf{n}})$, where $\odot$ and $\oslash$ denote the element-wise multiplication and division, respectively. The enhanced speech is then obtained using the overlap-add method. The processing chain to directly obtain the windowed STFT estimates are summarized in Figure 1. Notice that the corresponding exemplars for both the input and output dictionaries are extracted from the same piece of training data which enables the algorithm to directly use the activations $\mathbf{X}^{\text{in}}$ to reliably reconstruct the underlying STFT estimates.

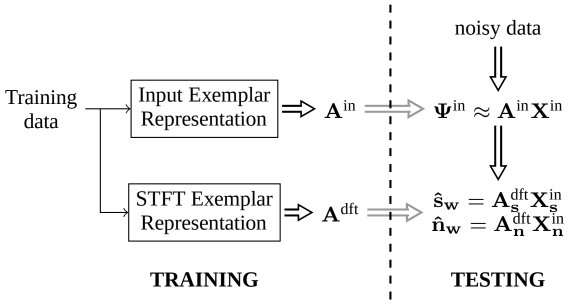
noisy data
Training data
Input Exemplar Representation
$\mathbf{A}^{\text{in}}$
$\mathbf{\Psi}^{\text{in}} \approx \mathbf{A}^{\text{in}}\mathbf{X}^{\text{in}}$
STFT Exemplar Representation
$\mathbf{A}^{\text{dft}}$
$\hat{\mathbf{s}}_{\mathbf{w}} = \mathbf{A}_{\mathbf{s}}^{\text{dft}}\mathbf{X}_{\mathbf{s}}^{\text{in}}$
$\hat{\mathbf{n}}_{\mathbf{w}} = \mathbf{A}_{\mathbf{n}}^{\text{dft}}\mathbf{X}_{\mathbf{n}}^{\text{in}}$
TRAINING
TESTING

Fig. 1. *Block digram overview of the proposed system to directly obtain the STFT estimates using coupled dictionaries.*

2.2. ASR evaluation using DNNs

The evaluations are done on the AURORA-4 database using the “recipe” DNN-HMM-based recognizer in the Kaldi toolkit [16]. A DNN is simply a multi-layer perceptron with multiple hidden layers between its inputs and outputs. Performing back-propagation training on such a network can result in a poor local optimum with a randomly initialized network weights. To circumvent this, a pre-training is done first by considering each pair of adjacent layers as restricted Boltzmann machines (RBM) [17] and then a back propagation training is done over the entire network such that it provides posterior probability estimates for the HMM states [5].

To perform ASR using a DNN-HMM-hybrid setting, the state emission likelihoods generated by the GMMs are replaced by the pseudo-likelihoods or scaled-likelihoods generated by the DNN.

3. EVALUATED SETTINGS

In this work, we evaluate the exemplar-based speech enhancement for three different input exemplar spaces: Mel, magnitude STFT (referred to as DFT hereafter) and MS spaces, also denoted using the superscripts ‘mel’, ‘dft’ and ‘MS’, respectively. Each of these settings are detailed in this section.

3.1. DFT-DFT Setting

In this setting, DFT exemplar space is chosen as the input exemplar space to obtain the NMF-based decomposition. To obtain DFT exemplars to create the input dictionary $\mathbf{A}^{\text{dft}}$, a random segment of acoustic data spanning T frames (or T_t seconds in time domain) is taken and its full-resolution magnitude STFT of size $F \times T$ is considered for non-negativity, where F is the number of frequency bins used to obtain the STFT. This is then reshaped to a vector of length $F \cdot T$ to obtain its DFT exemplar representation.

During testing, the noisy data is converted to its equivalent DFT exemplar space representation $\mathbf{\Psi}^{\text{dft}}$ as in Section 2. $\mathbf{\Psi}^{\text{dft}}$ is decomposed using $\mathbf{A}^{\text{dft}}$ to obtain the activations $\mathbf{X}^{\text{dft}}$. Notice that in this setting, both the input and output dictionaries are same ($\mathbf{A}^{\text{in}} = \mathbf{A}^{\text{dft}}$). The speech and noise estimates are obtained as $\hat{\mathbf{s}} = [\mathbf{A}_{\mathbf{s}}^{\text{dft}}\mathbf{X}_{\mathbf{s}}^{\text{dft}}]^*$ and $\hat{\mathbf{n}} = [\mathbf{A}_{\mathbf{n}}^{\text{dft}}\mathbf{X}_{\mathbf{n}}^{\text{dft}}]^*$, respectively. These estimates are used to obtain the enhanced STFT as $\mathbf{Y}_{\text{enh}} = \mathbf{Y} \odot \hat{\mathbf{s}} \oslash (\hat{\mathbf{s}} + \hat{\mathbf{n}})$.

3.2. Mel-Mel† and Mel-DFT Settings

Here, the NMF-based decomposition is done using the *Mel dictionary* $\mathbf{A}^{\text{mel}}$ containing Mel exemplars as its columns. A Mel exemplar is obtained by pre-multiplying a magnitude STFT of size $F \times T$ with the STFT-to-Mel matrix $\mathbf{M}$ which contains the magnitude response of B Mel bands along its rows. i.e., $\mathbf{M}$ is of size $B \times F$. The resulting Mel-integrated magnitude STFT is reshaped to obtain a Mel exemplar of length $B \cdot T$. During testing, the noisy speech expressed in the Mel exemplar space $\mathbf{\Psi}^{\text{mel}}$ is decomposed using $\mathbf{A}^{\text{in}} = \mathbf{A}^{\text{mel}} = [\mathbf{A}_{\mathbf{s}}^{\text{mel}}\ \mathbf{A}_{\mathbf{n}}^{\text{mel}}]$ to obtain the activations $\mathbf{X}^{\text{mel}}$. These activations are used to evaluate two settings.

First, as a baseline system, the setting which uses a pseudo-inverse to obtain the STFT is evaluated. For this, we obtain the frame-level speech and noise estimates in the Mel domain $\hat{\mathbf{s}}' = [\mathbf{A}_{\mathbf{s}}^{\text{mel}}\mathbf{X}_{\mathbf{s}}^{\text{mel}}]^*$ and $\hat{\mathbf{n}}' = [\mathbf{A}_{\mathbf{n}}^{\text{mel}}\mathbf{X}_{\mathbf{n}}^{\text{mel}}]^*$ and multiply these with $\mathbf{M}^\dagger = \mathbf{M}^\intercal(\mathbf{M}\mathbf{M}^\intercal)^{-1}$ to map these estimates to the STFT domain. Here, $\intercal$ denotes matrix transpose. The enhanced STFT in this setting is obtained as $\mathbf{Y}_{\text{enh}} = \mathbf{Y} \odot \left\{\mathbf{M}^\dagger\hat{\mathbf{s}}'\right\} \oslash \left\{\mathbf{M}^\dagger(\hat{\mathbf{s}}' + \hat{\mathbf{n}}')\right\}$ [8]. This setting is referred to as *Mel-Mel*† *setting*.

Next, the setting which directly obtains the speech and noise estimates in the STFT domain using the coupled dictionary approach is evaluated. The enhanced STFT in this case is obtained as $\mathbf{Y}_{\text{enh}} = \mathbf{Y} \odot \hat{\mathbf{s}} \oslash (\hat{\mathbf{s}} + \hat{\mathbf{n}})$, where $\hat{\mathbf{s}} = [\mathbf{A}_{\mathbf{s}}^{\text{dft}}\mathbf{X}_{\mathbf{s}}^{\text{mel}}]^*$ and $\hat{\mathbf{n}} = [\mathbf{A}_{\mathbf{n}}^{\text{dft}}\mathbf{X}_{\mathbf{n}}^{\text{mel}}]^*$ (refer Section 2). This setting is referred to as *Mel-DFT setting*.

3.3. MS-DFT Setting

This setting makes use of MS exemplars to obtain the compositional model using NMF. The MS representation was proposed as part of a computational model for human hearing which relies on the low frequency amplitude modulations within various frequency bands [18]. To obtain an MS exemplar, T frames of acoustic data are considered and are filtered using a filter-bank having B channels (to have a reliable comparison with the Mel-based settings). The resulting B band-limited signals are half-wave rectified to model non-negative nerve firings and low-pass filtered at a 3 dB cut-off frequency of around 20 Hz to obtain the modulation envelopes. The magnitude STFT of these envelopes yields B modulation spectrograms [10] of size $K \times T$ each, where K is the number of modulation frequency bins used to obtain the STFT.

As there is a low-pass filtering operation, it is possible to truncate each of these modulation spectrograms to their lowest few, say k, bins [3,19], i.e, each modulation spectrogram now has size $k \times T$. To obtain a two-dimensional representation, we stack these modulation spectrograms originating from B channels to a matrix of size $(B \cdot k) \times T$. These are then reshaped to a vector of length $B \cdot k \cdot T$ to obtain the MS exemplar [3]. Let the MS dictionary be $\mathbf{A}^{\text{MS}} = [\mathbf{A}_{\mathbf{s}}^{\text{MS}}\ \mathbf{A}_{\mathbf{n}}^{\text{MS}}]$.

During testing, the noisy data expressed in the MS exemplar domain is decomposed using the MS dictionary to obtain the activations $\mathbf{X}^{\text{MS}}$. The frame-level speech and noise estimates for enhancing the noisy STFT are obtained as $\hat{\mathbf{s}} = [\mathbf{A}_{\mathbf{s}}^{\text{dft}}\mathbf{X}_{\mathbf{s}}^{\text{MS}}]^*$ and $\hat{\mathbf{n}} = [\mathbf{A}_{\mathbf{n}}^{\text{dft}}\mathbf{X}_{\mathbf{n}}^{\text{MS}}]^*$, respectively. Notice that such an approximation will work only if the mapping between the MS and the DFT exemplars are one-to-one. In our previous work, temporal oversampling while obtaining the modulation spectra is successfully used to remedy this [9].

4. EXPERIMENTAL SETUP

4.1. AURORA-4 database

AURORA-4 database is a large vocabulary continuous speech database based on the WSJ0 corpus of read speech. The test set of the corpus is divided as: test A (330 clean utterances), test B (clean utterances in test A added with 6 different noise types, summing to $330 \cdot 6 = 1\ 980$ utterances), test C (330 clean utterances with microphone variation), test D (6 noisy versions of utterances in test C, summing to $1\ 980$ utterances), all from 8 speakers. The six noise types used in both test B and D are car, street, train station, babble, restaurant and airport noises added at varying SNRs between 5 and 15 dB.

For training the acoustic models and creating the dictionaries, clean training set and multi-condition training (MCT) sets are used, both containing 7 137 utterances each from 83 speakers. The MCT contains clean utterances with microphone variation and noisy utterances with artificially added noises present in the test sets at varying SNRs between 10 and 20 dB. The database also contains a development set with the same structure as that of the test sets, from 10 speakers.

4.2. NMF-based speech enhancement

The dictionaries used by the NMF-based speech enhancement setting was created using the utterances present in the MCT set. The speech dictionaries were created using the clean speech utterances present in the MCT set. The noise data needed for creating the noise dictionaries were obtained by subtracting the original clean speech from the noisy utterance as in [3,8]. To obtain the exemplars for creating the coupled dictionaries, a segment of training speech or noise segment spanning $T = 15$ frames, as used in [8,20] is taken at random and the following operations are used to obtain the coupled exemplars from each domain.

1. The DFT exemplar is obtained by taking the magnitude STFT of the segment with a window-length and hop-size of 25 ms and 10 ms, respectively. $F = 512$ bins were used, which was truncated to the first 256 bins considering only the positive half of the symmetric spectrum. The resulting magnitude STFT of size 256×15 is reshaped to obtain a DFT exemplar of length $3\ 840$.
2. To obtain the Mel exemplar, the magnitude STFT obtained above is pre-multiplied with the STFT-to-Mel matrix $\mathbf{M}$ containing $B = 40$ Mel bands. The resulting Mel-integrated spectra of size 40×15 is reshaped to obtain the Mel exemplar of length 600.
3. To obtain the MS exemplar, the time domain signal is first filtered into $B = 40$ channels using the equivalent-rectangular bandwidth filter-banks implemented using Slaney's toolbox [21]. The modulation envelopes are obtained with a low-pass 3 dB cut-off frequency of 30 Hz (as in [3,9]). The resulting modulation envelopes are then converted to its magnitude modulation spectrogram representation using a window-length of 64 ms with $K = 1024$ frequency bins. Notice that a hop-size of around 25 ms is sufficient for this setting. To make the mapping between the Mel and DFT exemplars as close as one-to-one, a temporal oversampling with a hop size of 10 ms is used [9].

   The values of $K = 1024$ and a low-pass cut-off of 30 Hz results in a value of $k = 5$ bins. Therefore, each of the modulation spectra is truncated to its lowest 5 bins and are stacked to get a representation of size 200×15. The MS exemplar is obtained after reshaping it to a vector of length $3\ 000$.

In this work, speech dictionaries of $10\ 000$ coupled speech exemplars each, extracted by random sampling, are used. The coupled noise dictionaries used are comprised of two parts. A fixed part containing $5\ 000$ exemplars extracted from randomly sampled noise

Setting	test A	test B	test C	test D	Avg.
No Enh.	2.9	45.2	43.6	64.5	50.3
Mel-Mel†	2.8	17.3	39.8	45.1	29.8
DFT-DFT	2.8	24.8	39.9	48.4	34.4
Mel-DFT	2.8	15.9	39.4	42.8	28.1
MS-DFT	**2.7**	**14.8**	**38.9**	**40.8**	**26.8**

Table 1. *Average WERs in % obtained for various settings on the AURORA-4 database with DNN trained on the clean training data. Best scores are highlighted in bold font.*

Setting	test A	test B	test C	test D	Avg.
No Enh.	3.5	7.3	10.3	21.8	13.5
Mel-Mel†	3.6	6.6	10.4	20.7	12.7
DFT-DFT	**3.4**	6.8	**9.7**	20.3	12.5
Mel-DFT	3.6	6.8	10.3	20.8	12.9
MS-DFT	3.5	**6.2**	10.2	**19.4**	**11.9**

Table 2. *Average WERs in % obtained for various settings on the AURORA-4 database with retrained DNNs on enhanced MCT data. Best scores are highlighted in bold font.*

segments, and a small (“sniffed”) noise dictionary extracted from cyclicly shifted versions of the first $T = 15$ frames of the noisy test utterance that is being decoded (resulting in 15 exemplars). This results in a total of $15\,015$ exemplars in all the dictionaries. The fixed part of all the coupled dictionaries are created only once and are kept fixed for all the evaluations in this paper.

The NMF based decomposition is obtained with 350 iterations of NMF-multiplicative updates with the activations $\mathbf{X}^{\text{in}}$ initialized as $(\mathbf{A}^{\text{in}})^\intercal \mathbf{\Psi}^{\text{in}}$. Sparsity penalties used to obtain the speech activations for the decompositions using the Mel, DFT and MS dictionaries are 1.2, 1.7 and 1.6 respectively, which are tuned using the development set [3,8]. For all settings, the sparsity penalty for noise activations are fixed as 0.5 times the speech sparsity penalty, to avoid extra computational effort while tuning both the speech and noise sparsity penalties in a grid search [8].

4.3. DNN-HMM-based decoder for ASR

In this work, DNNs trained on clean training data and enhanced MCT data, referred to as *clean DNN* and *retrained DNN*, are used for ASR evaluation. To obtain a retrained DNN, the MCT data is first enhanced using the respective speech enhancement front-end and the resulting data is used to train the DNN. The recipe recognizer based on DNN in the Kaldi toolkit is based on the implementation presented in [5]. All DNNs used are comprised of 6 hidden layers with 2 048 sigmoid neurons per layer. The input layer used 40 Mel coefficients with a temporal context of 11 frames, summing to a total of 440 input features. Average word error rates (WER) in $\%$ are used to evaluate and compare the performance of the various settings.

5. RESULTS AND DISCUSSION

5.1. Results on DNN trained on clean data

The results obtained for a DNN trained on the clean training set of the AURORA-4 database are tabulated in Table 1. The first row denotes the results obtained on the AURORA-4 data without any pre-processing. Only the clean part of the development sets were used for cross-validation during the clean DNN training.

It is evident that the noise mismatch (set B and D) and channel mismatch (set C and D) both have a detrimental effect on the accuracy of the system without enhancement. We succeed best in reducing the discrepancy caused by the noise mismatch, because that is what the enhancement is designed for (Wiener-like filtering, but no convolutional channel mismatch model). A relative WER improvement of around 50% is obtained with the speech enhancement front-end using MS exemplars over the setting with unenhanced data. The Mel-DFT setting yielded a better performance when compared to the Mel-Mel† setting even though the NMF-based decompositions in both systems are done using the same the Mel dictionaries, suggesting that the pseudo-inverse mapping in the latter can have detrimental effects on the accuracy of the system.

Also notice that there is a WER improvement for both test A and C clean speech sets as well. It can be attributed to the ability of exemplar-based models to reduce the speaker variability while training and testing, due to the projection onto the clean speech manifold, similar to the observation made for a GMM-based ASR setting in [15].

5.2. Results on retrained DNNs

The ASR results obtained using the retrained DNNs are tabulated in Table 2. It is evident that a DNN trained on the MCT data can yield superior WER improvements over a DNN trained on clean data, thanks to its multiple hidden layers. It can be seen that an exemplar-based speech enhancement front-end together with DNN retraining can further improve the performance of a DNN-based ASR system for all test cases.

Also notice that, unlike the observations made with the clean DNN, the Mel-DFT and Mel-Mel† settings yielded almost similar WERs and the DFT-DFT setting performs better than the Mel based settings with retrained DNNs. These can be attributed to the ability of the DNNs to learn the deformations introduced by the speech enhancement front-end as well while retraining. The MS-DFT setting yielded the best overall performance in this setting with statistically significant WER improvements of $p < 0.0001$ over the no enhancement setting and $p < 0.001$ over the Mel-Mel† baseline setting.

6. CONCLUSIONS AND FUTURE WORK

In this work, we investigated the performance of a DNN-HMM-based ASR system with an exemplar-based speech enhancement front-end. It is observed that the ASR performance on clean speech using clean DNNs can be further improved using the exemplar-based techniques. The WER improvement can be attributed to the ability of the investigated exemplar-based systems in reducing the noise mismatch (by speech enhancement) and speaker mismatch (by projecting the test features to clean speech manifold). A speech enhancement front-end with MS exemplars was also investigated in this work, which yielded statistically significant improvements. It is also observed that retraining the DNNs using the enhanced multi-condition training data can further improve the accuracy of a DNN-based system. The best performing settings in this work yielded an average overall WERs of 26.8% and 11.9% with clean and retrained DNNs, respectively.

One future work is to explore the setting based on the MS features for other choices of low-pass 3 dB cut-off frequencies and exemplar sizes. Since the MS features are motivated from human auditory processing, another research direction is to investigate the performance of a DNN with the MS features at its input layer as an attempt to model human speech recognition.

7. REFERENCES

[1] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, pp. 788–791, October 1999.

[2] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, “Exemplar-based sparse representations for noise robust automatic speech recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067 –2080, Sept. 2011.

[3] Deepak Baby, Tuomas Virtanen, Jort F. Gemmeke, Tom Barker, and Hugo Van hamme, “Exemplar-based noise robust speech recognition using modulation spectrogram features,” in *Spoken Language Technology Workshop (SLT), 2014 IEEE*, South Lake Tahoe, USA, Dec 2014.

[4] G. Hinton, Li Deng, Dong Yu, G.E. Dahl, A Mohamed, N. Jaitly, A Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[5] Karel Veselý, Arnab Ghoshal, Lukás Burget, and Daniel Povey, “Sequence-discriminative training of deep neural networks,” in *INTERSPEECH*, 2013, pp. 2345–2349.

[6] G.E. Dahl, Dong Yu, Li Deng, and A Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, Jan 2012.

[7] Bhiksha Raj, Tuomas Virtanen, Sourish Chaudhuri, and Rita Singh, “Non-negative matrix factorization based compensation of music for automatic speech recognition,” in *INTERSPEECH*, Makuhari, Japan, 2010.

[8] J. T. Geiger, J. F. Gemmeke, B. Schuller, and G. Rigoll, “Investigating NMF speech enhancement for neural network based acoustic models,” in *INTERSPEECH*. 2014, ISCA.

[9] D. Baby, T. Virtanen, T. Barker, and H. Van hamme, “Coupled dictionary training for exemplar-based speech enhancement,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 2883–2887.

[10] S. Greenberg and B.E.D. Kingsbury, “The modulation spectrogram: in pursuit of an invariant representation of speech,” in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, 1997, vol. 3, pp. 1647–1650 vol.3.

[11] Marc Delcroix, Yotaro Kubo, Tomohiro Nakatani, and Atsushi Nakamura, “Is speech enhancement pre-processing still relevant when using deep neural networks for acoustic modeling?,” in *INTERSPEECH*. 2013, pp. 2992–2996, ISCA.

[12] Tomohiro Nakatani, Shoko Araki, Marc Delcroix, Takuya Yoshioka, and Masakiyo Fujimoto, “Reduction of highly non-stationary ambient noise by integrating spectral and locational characteristics of speech and noise for robust asr.,” in *INTERSPEECH*. 2011, pp. 1785–1788, ISCA.

[13] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443–445, Apr 1985.

[14] M.L. Seltzer, Dong Yu, and Yongqiang Wang, “An investigation of deep neural networks for noise robust speech recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 7398–7402.

[15] Tara N. Sainath, Bhuvana Ramabhadran, David Nahamoo, Dimitri Kanevsky, and Abhinav Sethy, “Sparse representation features for speech recognition,” in *INTERSPEECH*, 2010, pp. 2254–2257.

[16] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, “The Kaldi speech recognition toolkit,” in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[17] Geoffrey E. Hinton, “A practical guide to training restricted boltzmann machines,” in *Neural Networks: Tricks of the Trade (2nd ed.)*, pp. 599–619. 2012.

[18] C. J. Plack, *The sense of hearing.*, Lawrence Erlbaum Associates Publishers, 2005.

[19] T. Barker and T. Virtanen, “Non-negative tensor factorization of modulation spectrograms for monaural sound source separation,” in *Proc. INTERSPEECH*, 2013.

[20] Jürgen T. Geiger, Felix Weninger, Antti Hurmalainen, Jort Gemmeke, Martin Wöllmer, Björn Schuller, Gerhard Rigoll, and Tuomas Virtanen, “The TUM+TUT+KUL Approach to the CHiME Challenge 2013: Multi-Stream ASR Exploiting BLSTM Networks and Sparse NMF,” in *Proceedings of the 2nd CHiME workshop*, June 2013, pp. 25–30.

[21] M. Slaney, “Auditory toolbox version 2,” *Interval Research Corporation*, vol. 10, pp. 1998, 1998.