HYBRID INPUT SPACES FOR EXEMPLAR-BASED NOISE ROBUST SPEECH RECOGNITION USING COUPLED DICTIONARIES

Deepak Baby and Hugo Van hamme

Department ESAT, KU Leuven, Belgium

{Deepak.Baby, Hugo.Vanhamme}@esat.kuleuven.be

ABSTRACT

Exemplar-based feature enhancement successfully exploits a wide temporal signal context. We extend this technique with hybrid input spaces that are chosen for a more effective separation of speech from background noise. This work investigates the use of two different hybrid input spaces which are formed by incorporating the full-resolution and modulation envelope spectral representations with the Mel features. A coupled output dictionary containing Mel exemplars, which are jointly extracted with the hybrid space exemplars, is used to reconstruct the enhanced Mel features for the ASR back-end. When compared to the system which uses Mel features only as input exemplars, these hybrid input spaces are found to yield improved word error rates on the AURORA-2 database especially with unseen noise cases.

Index Terms: coupled dictionaries, automatic speech recognition, modulation envelope, non-negative matrix factorization

1. INTRODUCTION

One of the biggest issues the current state-of-the-art automatic speech recognition (ASR) systems face is the degradation in performance due to added background noise. So in order to improve noise robustness, most of the ASR systems employ some mechanism which attempts to enhance the speech features by removing these artefacts. Most of these mechanisms, like spectral subtraction [1], vector Taylor series [2], etc., work on spectro-temporal representations spanning a few tens of milli-seconds of the speech recording. In this work, we focus on feature enhancement using non-negative matrix factorization (NMF) using "exemplars" which span hundreds of milli-seconds of the recorded data.

Spectral factorization methods based on NMF attempt to decompose the features extracted from a noisy recording as the weighted sum of speech and noise dictionary atoms or exemplars, and are found to be useful for noise-robust ASR [3–5]. Most of the conventional exemplar-based ASR systems use exemplars extracted from feature spaces like the Mel [6], Gabor [7], DFT (refers to the magnitude of the discrete-Fourier transform throughout this paper) [8] etc., to obtain the compositional model and enhance the corresponding features. These enhanced features are then used to find the enhanced Mel-frequency cepstral coefficients (MFCCs) to be fed to the ASR back-end.

The efficiency of an exemplar-based NMF approach depends on the ability of the chosen exemplar space in differentiating features originating from speech and noise, and it is found that different exemplar spaces yield different performance depending on the type of added noise and signal-to-noise ratio (SNR) levels [9]. It is also noticed that, apart from increasing the computational complexity, using higher dimensional exemplars derived from feature spaces like the DFT [8], or modulation envelope spectra (MS) [9, 10], etc. will result in too detailed modelling of the seen noise cases to generalise well for the unseen noise cases.

In order to address the issues faced by the higher dimensional features and to combine the speech and noise separation properties of different feature spaces, we propose the use of hybrid input spaces to obtain the decomposition. To reconstruct the Mel estimates from this, a variant of the coupled dictionary approach described in [9] is used. In this setup, the exemplars for the coupled hybrid input and the Mel output dictionaries are extracted from the same piece of training data. Then for evaluation, the underlying Mel features are reconstructed using the coupled Mel dictionary, following the decomposition in the hybrid input space.

To obtain a hybrid input space, two feature spaces are chosen first which are called as *primary* and *secondary* feature spaces. A hybrid exemplar is then obtained by concatenating the exemplars belonging to these feature spaces that are extracted from the same piece of training data. In this work, the Mel space is chosen as the primary feature space for its reduced dimensionality and good separation capabilities [9, 11] with the DFT or MS representation as the secondary feature space.

To address the "curse" of large dimensionality of the chosen secondary spaces, we propose to use a *trimmed* secondary exemplar space to be concatenated with the full length primary space exemplar. The trimmed exemplar is obtained by reshaping only a subset of the feature frames belonging to the secondary feature space. The decomposition obtained with such a hybrid space will thus rely mainly on the primary feature space with the trimmed secondary space acting as a cue to regularise the separation.

The simulation results obtained on the AURORA-2 database revealed that, even with the secondary space trimmed down to a single frame, both the hybrid input spaces yield improved performances in terms of word error rate (WER) over the baseline system which uses Mel features only. The computational complexity of the proposed approach is also found to be comparable to that of the baseline system as trimmed secondary spaces are used.

2. METHOD

2.1. Feature enhancement using NMF

NMF-based compositional models attempt to decompose the features extracted from a noisy recording as a sparse non-negative weighted sum of speech and noise atoms or exemplars stored as columns in a speech and noise dictionary denoted as $\mathbf{A}_s$ and $\mathbf{A}_n$, respectively. Exemplars are extracted from training data spanning

This project has been funded with support from the European Commission under Contract FP7-PEOPLE-2011-290000.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

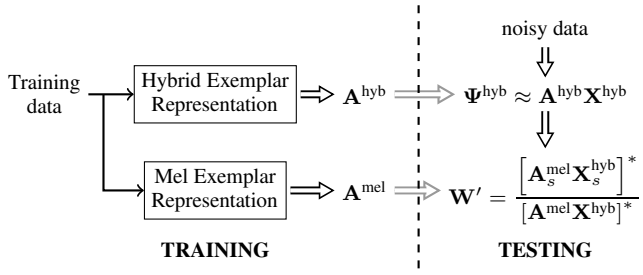

Fig. 1. *Block digram overview of the proposed system using hybrid input spaces and coupled dictionaries for Mel feature enhancement.*

multiple, say T, frames to capture temporal dynamics, followed by reshaping to form a vector. The representation for the noisy utterance in the exemplar space, $\mathbf{\Psi}$, the columns of which are obtained by reshaping sliding windows of length T frames along the length of the utterance [11], is decomposed to get the activations, $\mathbf{X}$, as:

$$\mathbf{\Psi} \approx \begin{bmatrix}\mathbf{A}_s & \mathbf{A}_n\end{bmatrix} \begin{bmatrix}\mathbf{X}_s \\ \mathbf{X}_n\end{bmatrix} = \mathbf{A}\mathbf{X} \quad s.t. \quad \mathbf{X} \geq 0. \tag{1}$$

The approximation is done such that it minimizes the cost function,

$$\mathcal{C} = D_{KLD}(\mathbf{\Psi} \| \mathbf{A}\mathbf{X}) + \mathbf{\Lambda} \odot \mathbf{X} \tag{2}$$

where, D_{KLD} is the element-wise Kullback-Leibler divergence

$$D_{KLD}(x\|y) = x\log(x/y) - x + y \tag{3}$$

and $\mathbf{\Lambda}$ is the sparsity penalty on the activations $\mathbf{X}$ [6]. $\odot$ denotes element-wise multiplication. The frame-wise speech and noise estimates, $\hat{\mathbf{s}}$ and $\hat{\mathbf{n}}$ are then obtained after removing the windowing effect by adding the frames belonging to the overlapping windows in the windowed estimates $\mathbf{A}_s\mathbf{X}_s$ and $\mathbf{A}_n\mathbf{X}_n$, respectively. A frame-level Wiener-like filter is then obtained after element-wise division as, $\mathbf{W} = \hat{\mathbf{s}} \oslash (\hat{\mathbf{s}} + \hat{\mathbf{n}})$, which when applied to the noisy features yields enhanced features.

2.2. Proposed method using hybrid input spaces

In the proposed approach, the activations $\mathbf{X}^{\text{hyb}}$ are obtained using the dictionary $\mathbf{A}^{\text{hyb}} = \begin{bmatrix}\mathbf{A}_s^{\text{hyb}} & \mathbf{A}_n^{\text{hyb}}\end{bmatrix}$, which contains exemplars belonging to a hybrid input space, using the NMF approach explained in Section 2.1. The windowed Mel speech and noise estimates are then reconstructed using the coupled Mel dictionary, which contains coupled exemplars belonging to the Mel feature space, as $\mathbf{A}_s^{\text{mel}}\mathbf{X}_s^{\text{hyb}}$ and $\mathbf{A}_n^{\text{mel}}\mathbf{X}_n^{\text{hyb}}$, respectively. Notice that the corresponding atoms in the coupled dictionaries, $\mathbf{A}^{\text{hyb}}$ and $\mathbf{A}^{\text{mel}}$, are extracted from the same piece of training data which guarantees a reliable reconstruction of the underlying speech and noise estimates in the Mel domain [9,12].

The proposed approach is summarised in Figure 1. The notations used to explain the test phase are: $\mathbf{\Psi}^{\text{hyb}}$ for the noisy speech represented in the hybrid exemplar domain and $[\mathbf{Y}]^*$ denotes the matrix obtained after removing the effect of overlapping windows in the windowed observation $\mathbf{Y}$. All matrix divisions should be considered element-wise.

To obtain the hybrid input exemplars, the primary and secondary exemplars are created first from the same piece of training data spanning T frames. Let $\mathcal{T}_S$ be the trimming operator which trims an exemplar spanning T frames down to an exemplar spanning a subset $S \subseteq \{1, 2, \ldots, T\}$ of the T frames. Thus, from an exemplar with frames indexed from 1 through T, the trimming operator $\mathcal{T}_S$ selects only the frames with index contained in S, and reshapes them into a vector.

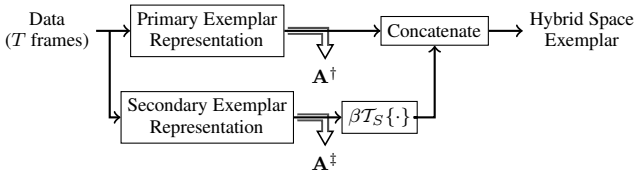

Fig. 2. *Block diagram overview of the processing chain used to obtain the proposed hybrid exemplar representation.*

The trimmed secondary exemplars are obtained by applying $\mathcal{T}_S$ on the secondary exemplars, which are also scaled with β to balance its contribution on obtaining the separation. These trimmed and scaled secondary exemplar is then concatenated with the corresponding primary exemplar to get the hybrid representation. Thus, the hybrid exemplar representation for noisy speech $\mathbf{\Psi}^{\text{hyb}}$ and the hybrid dictionary can be expressed as : Notice that, the proposed approach is equivalent to minimizing the cost function

$$\mathbf{\Psi}^{\text{hyb}} = \begin{bmatrix}\mathbf{\Psi}^{\dagger} \\ \beta\mathcal{T}_S\mathbf{\Psi}^{\ddagger}\end{bmatrix} \quad \text{and} \quad \mathbf{A}^{\text{hyb}} = \begin{bmatrix}\mathbf{A}^{\dagger} \\ \beta\mathcal{T}_S\mathbf{A}^{\ddagger}\end{bmatrix} \tag{4}$$

where, the superscripts $\dagger$ and $\ddagger$ denote the primary and secondary exemplar spaces, respectively. The cost function in this setting thus can be expressed as:

$$\begin{aligned}\mathcal{C}' &= D_{KLD}(\mathbf{\Psi}^{\text{hyb}} \| \mathbf{A}^{\text{hyb}}\mathbf{X}^{\text{hyb}}) + \mathbf{\Lambda} \odot \mathbf{X}^{\text{hyb}} \\ &= D_{KLD}\left(\begin{bmatrix}\mathbf{\Psi}^{\dagger} \\ \beta\mathcal{T}_S\mathbf{\Psi}^{\ddagger}\end{bmatrix} \middle\| \begin{bmatrix}\mathbf{A}^{\dagger} \\ \beta\mathcal{T}_S\mathbf{A}^{\ddagger}\end{bmatrix} \mathbf{X}^{\text{hyb}}\right) + \mathbf{\Lambda} \odot \mathbf{X}^{\text{hyb}} \\ &= D_{KLD}(\mathbf{\Psi}^{\dagger} \| \mathbf{A}^{\dagger}\mathbf{X}^{\text{hyb}}) + \beta D_{KLD}(\mathcal{T}_S\mathbf{\Psi}^{\ddagger} \| \mathcal{T}_S\mathbf{A}^{\ddagger}\mathbf{X}^{\text{hyb}}) + \mathbf{\Lambda} \odot \mathbf{X}^{\text{hyb}}\end{aligned}$$

using (4) and since the cost function being element-wise. It can thus be seen that the secondary space in effect acts as a regularisation to obtain the activations and β acts as the regularisation weight.

3. DESCRIPTION OF INPUT SPACES

The various input spaces which are chosen to evaluate the proposed approach along with the chosen baseline systems are described in this section.

3.1. Mel, DFT and MS only baselines

For a fair evaluation and completeness, three single-input space baseline systems which uses the Mel, DFT and MS representations respectively are evaluated and compared first. All these systems are evaluated using the coupled Mel output dictionary approach depicted in Figure 1 with the hybrid exemplars replaced by the Mel, DFT and the MS exemplars, respectively.

Mel baseline: This system uses the *Mel exemplars*, which are created by reshaping the Mel-integrated magnitude spectra of acoustic data spanning T frames. The decomposition of the noisy data expressed in the Mel exemplar domain is obtained using the Mel dictionary, $\mathbf{A}^{\text{mel}} = [\mathbf{A}_s^{\text{mel}}\ \mathbf{A}_n^{\text{mel}}]$. The Wiener filter for the noisy Mel enhancement is found using the procedure explained in Section 2.1. Also notice that these dictionaries act as the primary exemplar space dictionaries also for the proposed hybrid approach.

DFT baseline: For this setup, the coupled DFT and the Mel dictionaries are obtained first, with the DFT and Mel exemplars extracted from the same piece of training data. To obtain a DFT exemplar, magnitude spectrogram of a training data spanning T frames is reshaped to a vector. For evaluation, the DFT exemplar representation of the noisy data is decomposed using the *DFT dictionary*,

Experiments	clean	*test set A* 20	15	10	5	0	-5	*test set B* 20	15	10	5	0	-5	*Average Exec. time*
Mel Baseline	0.2	1.5	1.9	3.7	5.0	**11.6**	**27.6**	1.3	**1.6**	**4.4**	**9.0**	**26.7**	**57.9**	5.8 s
DFT Baseline	0.1	0.9	1.9	2.7	7.2	17.2	33.3	0.6	**1.6**	6.3	14.2	35.1	67.8	12.2 s
MS Baseline	**0.0**	**0.7**	**1.3**	**1.9**	**4.4**	12.5	30.5	**0.5**	1.7	5.1	11.2	34.8	69.0	10.8 s

Table 1. *WER in % obtained for various baseline systems as a function of SNR in dB evaluated on a subset of 100 files per test set of the AURORA-2 database. The average execution time per utterance required by the setting is also shown.*

$\mathbf{A}^{\text{dft}} = [\mathbf{A}_s^{\text{dft}}\ \mathbf{A}_n^{\text{dft}}]$. The activations thus obtained, $\mathbf{X}^{\text{dft}}$ are then applied on to the coupled Mel dictionary to get the speech and noise estimates for noisy Mel enhancement (ref. Section 2.2).

MS baseline: 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 [13] which are called modulation envelopes. Let B be the number of frequency bands considered. The MS representation for acoustical data is obtained by taking the short-time Fourier transform (STFT) of the modulation envelopes corresponding to each frequency band [14]. For non-negativity, only the magnitude of the STFT is considered.

Because of the low-pass filtering operation, only very few lower bins of the MS will contain significant energy and it is possible to truncate each of the MS to the lowest b bins [15]. All these truncated MS of size $b \times T$ each are then stacked to obtain a matrix of size $(B \cdot b) \times T$ which are referred to as *MS features* [9]. The MS exemplars are then obtained by reshaping the MS features which are stored in the *MS Dictionary*, $\mathbf{A}^{\text{MS}} = [\mathbf{A}_s^{\text{MS}}\ \mathbf{A}_n^{\text{MS}}]$. The MS baseline system is then evaluated using the coupled dictionary approach explained in Section 2.2 with the decomposition obtained in the MS exemplar space.

3.2. Hybrid input spaces: Mel-DFT and Mel-MS spaces

In this work, we investigate the Mel-DFT and Mel-MS hybrid spaces. For this, the Mel, DFT and MS exemplars are created first as explained in Section 3.1 from the same piece of data spanning T frames. The trimmed secondary exemplars are then created, by applying $\mathcal{T}_S$ on the DFT and MS exemplars, which are also scaled with β_1 and β_2, respectively. These are then concatenated with the corresponding Mel exemplar (ref. Section 2.2) to get the hybrid Mel-DFT and Mel-MS exemplar representations, respectively.

During testing, for every sliding window of length T along the length of the noisy utterance, the Mel and the secondary exemplar representations are obtained. The secondary exemplar representation is then trimmed using $\mathcal{T}_S$ and scaled, followed by concatenating with the Mel exemplar representation to be stored as columns in $\mathbf{\Psi}^{\text{hyb}}$.

4. EVALUATION EXPERIMENTS

4.1. Experimental setup

For evaluation, test sets 'A' and 'B' of the AURORA-2 corpus which contains utterances of digits from '0-9' and 'oh' are used. The training set of the corpus is composed of 8440 clean speech utterances and 6768 noisy utterances which are corrupted by four additive noises (subway, babble, car and exhibition hall). Test set A contains 4004 clean utterances which are divided into four equal subsets to obtain the noisy utterances corrupted by the noise types present in the training data at varying SNRs (20, 15, 10, 5, 0 and -5 dB) leading to 24 noisy subsets. Test set B also contains the same number of subsets but corrupted with four other (unseen) noise types (restaurant, train station, street and airport). The WERs obtained after taking the average over the four noise types for clean speech, -5 dB and the combined average of results obtained for SNRs ranging from 20-0 dB are presented.

The noise data required to obtain the noise exemplars are created from the noisy training set using the two step procedure explained in [6]. The clean and the noise samples are pre-processed by removing the dc component and applying pre-emphasis with filter coefficient 0.97. The exemplars for the Mel, and the trimmed DFT and MS spaces are then created using the steps explained in Section 3. To extract the coupled exemplars, random pieces of training data spanning 300 ms were used. No supervision was done to avoid the overlap between the chosen random pieces of data or to avoid silence. Then, for each of the chosen random piece of training data:

1. To obtain the Mel exemplars, the DFT of the chosen random piece of training data was first obtained using a window length and hop size of 25 ms and 10 ms respectively with 128 frequency bins within the Nyquist frequency (4 kHz), leading to a DFT representation of size 128×30. This is then Mel-integrated with $B = 23$ channels. These frame-level Mel features of size 23×30 thus obtained are then reshaped to obtain a Mel exemplar of length 690.

2. To obtain the DFT exemplar, the DFT representation obtained in Step 1 is reshaped to a vector (of length $3,840$).

3. To obtain the MS feature representation, the data is first split across $B = 23$ frequency channels using the equivalent rectangular bandwidth filter banks implemented using Slaney's toolbox [16]. Each of these is then half-wave rectified and low-pass filtered at a 3 dB cut-off frequency of 30 Hz to obtain the modulation envelopes. The modulation spectra for each channel is then found by taking the STFT of each of these envelopes with a window length of 64 ms and hop size 10 ms. With the sampling frequency of 8 kHz and STFT with 128 bins within the Nyquist frequency, each of the spectra was truncated to $b = 5$ bins and are stacked to get the MS features [9] of size 115×30. The MS exemplar representation is then obtained after reshaping the MS features to a vector of length $3,450$.

For evaluation, the coupled dictionaries $\mathbf{A}^{\text{mel}}$, $\mathbf{A}^{\text{dft}}$ and $\mathbf{A}^{\text{MS}}$ were created with 10000 speech and noise exemplars each. The hybrid input space dictionaries were then created as explained in Section 3.2 for different choices of S, β_1 and β_2. During testing, the corresponding exemplar space representations of the noisy data, $\mathbf{\Psi}$, were obtained as explained in Section 3 using the settings given above. The NMF-based decomposition was obtained with 600 multiplicative updates with sparsity constraint. A sparsity penalty of 1.5 for speech and 1 for noise exemplars as in [17] were used for all the evaluated decompositions except for the MS and DFT baselines, both of which used 1.75 and 0.75 respectively as in [9].

For the ASR back-end, a GMM-HMM based decoder using MFCC features was used. Each digit in the HMM topology was described by 16 states together with 3 silence states resulting in a total of 179 states. The GMM models were trained on the MFCCs obtained from the clean training data and enhanced noisy training data using the respective front-ends (referred to as *retraining*), with 13 static features along with their velocity and acceleration coeffi-

S	β_1	clean	*test set A* (20-0)	-5	*test set B* (20-0)	-5
Mel Baseline		0.4	5.2	28.0	9.2	**59.4**
Hybrid Mel-DFT space						
{1}	0.5	0.4	5.0	27.6	9.3	59.9
{1}	0.2	0.4	4.9	27.1	9.2	60.1
{15}	0.2	0.4	5.1	27.8	9.4	60.4
Switching	0.2	**0.3**	**4.7**	**26.6**	**9.0**	59.7

Table 2. *WER in % obtained as a function of SNR in dB on the AURORA-2 database for the hybrid Mel-DFT space approach. The results obtained for various choices of S and β_1 are given.*

cients leading to a 39 dimensional feature space. The GMM for each of the HMM state was modelled using 32 Gaussians with diagonal covariance.

4.2. Comparison between the baseline systems

To reduce the experimentation time, we compare the three chosen baseline systems evaluated on a subset of 100 files per test set which is tabulated in Table 1. It can be seen that for test set A, the Mel baseline performs better at lower SNRs and as the SNR increases, higher dimensional features yield better separation than the Mel features resulting in improved WERs. The higher dimensionality of these features results in poorer modelling of the unseen cases which explains their inferior performance for test set B. Also notice that the MS and DFT baseline settings are computationally expensive which is almost twice that of the Mel baseline setting.

The different baseline streams were also found to yield complementary results which can also benefit the hybrid input space approach. For the remaining part of this paper, the Mel exemplar system is chosen as the baseline for its good performance, lower dimensionality and also being the primary input space for the hybrid setup.

4.3. Parameters for the hybrid Mel-DFT space

With the baseline system chosen as the Mel exemplars only case, which is the same as the primary exemplar space chosen for the proposed hybrid spaces, the effectiveness of the proposed approach relies on the optimal choices of S and β. These are the two parameters which decide on the contribution of the secondary feature spaces on regularising the speech and noise separation resulting from the Mel baseline system. This section details the analysis of the hybrid Mel-DFT space for different choices of S and β_1 which is summarised in Table 2.

As the minimum choice, the effect of using the secondary DFT spaces with $|S|=1$ are investigated. As a pilot experiment, the effect of the first DFT frame i. e., $S=\{1\}$ with $\beta_1=0.5$ is investigated, which resulted in marginal performance improvement over the Mel only system. The optimum value of β_1 to get the best separation was then found to be 0.2 after doing a grid search in the range $[0.05,\ 0.5]$ on a subset of 100 files per noise type . The $S=\{1\}$ system with tuned β_1 is then evaluated over the complete test set which confirmed the effectiveness of the secondary DFT space in significantly improving the recognition results.

With the middle frame more correlated with the other frames in the given temporal context of 30 frames, the choice of $S=\{15\}$ was supposed to be more effective as it can be a better representative of all the DFT frames compared to the first DFT frame. However, on the contrary, the simulation experiments yielded inferior performance when compared to the $S=\{1\}$ case.

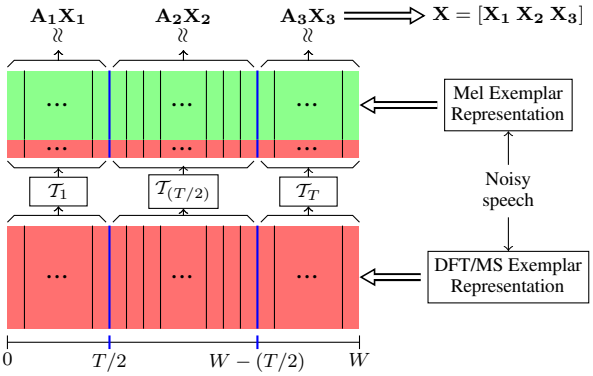

Fig. 3. *Block diagram overview of the proposed switching approach to obtain the activations.*

An analysis of the $S=\{1\}$ and $S=\{15\}$ cases revealed that such a fall in performance can be attributed to the reshaping operation when considering multiple frames (here, $T=30$) to obtain the exemplar space representation of the noisy test utterance $\mathbf{\Psi}^{\text{hyb}}$ (ref. Section 2). For the utterances in which the speech onset happens before the $15th$ frame, $S=\{15\}$ system was found to fail in detecting the speech onset resulting in a substitution or deletion. To address this and to capture the effectiveness of the middle DFT frame, a switching system which chooses the set S adaptively along the length of the utterance is proposed.

The proposed switching approach is depicted in Fig. 3. As explained in Section 2, the noisy utterance is first converted to the primary and secondary exemplar space representations by means of a sliding window spanning T frames along the length of the utterance. Let W be the total number of resulting sliding windows. In the switching approach, for the first and the last $T/2$ sliding windows of the utterance we use the secondary exemplar with $S=\{1\}$ and $S=\{T\}$ respectively, and $S=\{T/2\}$ for all the remaining windows falling in the middle. Thus we need to use three different dictionaries ($\mathbf{A_1}$, $\mathbf{A_2}$ and $\mathbf{A_3}$) depending on the choice of S in this setup, and the resulting activations are concatenated to obtain the overall activations as $\mathbf{X}=[\mathbf{X_1}\ \mathbf{X_2}\ \mathbf{X_3}]$. It can be seen from Table 2 that the assessment of the proposed approach yielded improved WERs over all the other investigated setups.

The performance improvement over the baseline system can be attributed to the inclusion of a secondary feature space which can regularise and improve the speech and noise separation. Also notice that the secondary space is not required to span the entire temporal context considered per exemplar to obtain a significant improvement in separating speech from noise.

4.4. Comparison of Mel-DFT and Mel-MS spaces

A comparison between the systems using the proposed hybrid input spaces is presented in this section. To obtain the Mel-MS results, the switching setup is used with a $\beta_2=0.1$ which was found after a grid search same as in Section 4.3. Table 3 summarizes the evaluated results.

It can be seen that both the proposed approaches yield statistically significant ($p<0.01$) improvement in performances when compared to the Mel baseline system for both seen and unseen noise cases. Also notice that a significant 16% relative WER improvement is obtained on test set B SNR(20-0), suggesting that the proposed approach can mitigate the effects of unseen noise cases as well. Inclusion of the MS space as a secondary space was found be more effective when compared to the DFT space. This can be attributed to the better speech and noise separation properties of the MS fea-

Experiments	clean	*test set A* (20-0)	*test set A* -5	*test set B* (20-0)	*test set B* -5
GMM trained on clean data					
Mel Baseline	0.4	5.2	28.0	9.2	59.4
Mel-DFT space	**0.3**	**4.7**	**26.6**	9.0	59.7
Mel-MS space	**0.3**	**4.7**	27.2	**8.8**	**59.2**
GMM trained on enhanced noisy data					
Mel Baseline	0.8	2.9	23.0	7.4	55.5
Mel-DFT space	0.6	2.8	22.2	6.5	53.3
Mel-MS space	**0.5**	**2.7**	**21.2**	**6.2**	**52.3**

Table 3. *WER in % obtained for various approaches as a function of SNR in dB on the AURORA-2 database.*

tures when compared to the DFT features, with noise having a different modulation frequency content from speech which was observed in [9, 10].

It was also observed in [9] that the MS features can perform well only for the seen noise cases as the MS features lead to more accurate representation of speech and noise, which will not generalise well for the unseen noises. But in the proposed approach, it is found that using trimmed MS exemplars as secondary features can be beneficial for unseen noises also.

The average execution times per utterance are tabulated in Table 4. It can be seen that the hybrid exemplar space yields an improved performance at a comparable computational complexity.

Setting	Mel Baseline	Mel-DFT Space	Mel-MS Space
Exec. time	5.9 s	6.4 s	6.2 s
Size	690	818	805

Table 4. *Average execution time needed in seconds per utterance for the various settings with 20000 exemplars in the dictionary. Size (or length) of the exemplars are also shown.*

5. CONCLUSION AND FUTURE WORK

In this work, we presented an exemplar-based feature enhancement method for ASR using hybrid input spaces and coupled dictionaries. The use of hybrid spaces was found to yield improved recognition accuracies over the baseline system. This paper also presents an effective way of combining multiple input spaces by means of an adaptively trimmed secondary exemplar representation without much increase in the computational complexity. The trimmed representation is also found to be effective in reducing the effects of overtraining to seen noise cases and generalises better to unseen noise cases when compared to full length exemplar representations.

Further, possibly adaptive, feature dimensionality reduction and its effect on reducing overfitting are to be investigated. Another future work is to study the effect of the number of noise exemplars and sparsity penalties in modelling unseen noise cases.

REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, April 1979.

[2] P. Moreno, B. Raj, and R. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Acoustics, Speech, and Signal Processing, 1996 IEEE International Conference on*, May 1996, vol. 2, pp. 733–736.

[3] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The Munich 2011 CHiME challenge contribution: NMF-BLSTM speech enhancement and recognition for reverberated multisource environments," in *CHiME 2011 Workshop on Machine Listening in Multisource Environments*, 2011.

[4] E. Yılmaz, J. F. Gemmeke, D. Van Compernolle, and H. Van hamme, "Noise-robust digit recognition with exemplar-based sparse representations of variable length," in *IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, Santander, Spain, Sept. 2012.

[5] J. T. Geiger, F. Weninger, A. Hurmalainen, J. F. Gemmeke, M. Wöllmer, B. Schuller, G. Rigoll, and T. 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.

[6] 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.

[7] A. Hurmalainen and T. Virtanen, "Modelling spectro-temporal dynamics in factorisation-based noise-robust automatic speech recognition," in *Acoustics, Speech and Signal Processing, 2012 IEEE International Conference on*, 2012, pp. 4113–4116.

[8] T. Virtanen, R. Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*, Wiley, 2012.

[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, 2014 IEEE International Conference on*, May 2014, pp. 2883–2887.

[10] D. Baby, T. Virtanen, J. F. Gemmeke, T. Barker, and H. Van hamme, "Exemplar-based noise robust speech recognition using modulation spectrogram features," in *Spoken Language Technology Workshop, 2014 IEEE*, South Lake Tahoe, USA, December 2014.

[11] J. F. Gemmeke and T. Virtanen, "Noise robust exemplar-based connected digit recognition," in *Acoustics, Speech and Signal Processing, 2010 IEEE International Conference on*, March 2010, pp. 4546 –4549.

[12] N. Mohammadiha and S. Doclo, "Single-channel dynamic exemplar-based speech enhancement," in *INTERSPEECH*. 2014, ISCA.

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

[14] S. Greenberg and B. Kingsbury, "The modulation spectrogram: in pursuit of an invariant representation of speech," in *Acoustics, Speech, and Signal Processing, 1997 IEEE International Conference on*, 1997, vol. 3, pp. 1647–1650.

[15] T. Barker and T. Virtanen, "Non-negative tensor factorization of modulation spectrograms for monaural sound source separation," in *INTERSPEECH*. 2013, ISCA.

[16] M. Slaney, "Auditory toolbox version 2," *Interval Research Corporation*, vol. 10, 1998.

[17] J. F. Gemmeke and H. Van hamme, "Advances in noise robust digit recognition using hybrid exemplar based systems," in *INTERSPEECH*. 2012, ISCA.