



# Noise Robust Exemplar Matching for Speech Enhancement: Applications to Automatic Speech Recognition

Emre Yilmaz, Deepak Baby and Hugo Van hamme

Dept. ESAT-PSI, KU Leuven, Belgium

{emre.yilmaz, deepak.baby, hugo.vanhamme}@esat.kuleuven.be

## Abstract

We present a novel automatic speech recognition (ASR) scheme which uses the recently proposed noise robust exemplar matching framework for speech enhancement in the front-end. The proposed system employs a GMM-HMM back-end to recognize the enhanced speech signals unlike the prior work focusing on template matching only. Speech enhancement is achieved using multiple dictionaries containing speech exemplars representing a single speech unit and several noise exemplars of the same length. These combined dictionaries are used to approximate the noisy segments and the speech component is obtained as a linear combination of the speech exemplars in the combined dictionaries yielding the minimum total reconstruction error. The performance of the proposed system is evaluated on the small vocabulary track of the 2<sup>nd</sup> CHiME Challenge and the AURORA-2 database and the results have shown the effectiveness of the proposed approach in improving the noise robustness of a conventional ASR system.

**Index Terms:** exemplar matching, noise robustness, speech enhancement, front-end denoising, automatic speech recognition

## 1. Introduction

Speech enhancement techniques, aiming to suppress the background noise degrading the speech signals recorded by a microphone, are often combined with automatic speech recognition (ASR) systems for improved noise robustness [1–3]. These techniques reduce the mismatch between the statistical acoustic models, e.g. hidden Markov models (HMM), trained under noise-free conditions and the target speech by preprocessing the noisy speech and/or features to enhance the noise corrupted spectrotemporal structure and recover the speech component as accurately as possible. Numerous enhancement techniques have been combined with Gaussian mixture model (GMM)-HMM [4–11] and deep neural network (DNN)-HMM [12–16] ASR systems and reported to provide considerable improvements in the recognition accuracy.

This paper presents a novel noise robust ASR system which incorporates an exemplar-based speech enhancement approach, dubbed *noise robust exemplar matching* (N-REM), for denoising the target utterance using the actual occurrences of speech and noise extracted from training data. Unlike previous exemplar-based speech enhancement systems using fixed-length exemplars in a single overcomplete dictionary [17–20], the proposed approach uses exemplars of multiple lengths, each associated with a single speech unit such as phones, syllables, half-words or words [21–23]. These exemplars are organized

in multiple dictionaries based on their length and class (associated speech unit). Using separate dictionaries for different speech units is motivated by the geometrical interpretation of SR-based source separation. It is known that the farther the convex hull of the basis vectors belonging to speech and noise sources are, the better the separation is [24]. Hence, the use of separate dictionaries for each speech unit provides a more precise representation in the high-dimensional feature space.

Previously, the N-REM framework has been successfully applied on small vocabulary ASR tasks [25, 26]. In previous work, the recognizer performs exemplar matching using the mel-scaled spectral representations of the exemplars and noisy speech and relies on a reconstruction error-based back-end to find the most likely hypothesis. However, in the proposed work, N-REM enhances the noisy speech and the enhanced speech represented in mel frequency cepstral coefficient (MFCC) domain is recognized using a conventional GMM-HMM back-end. This system is expected to remedy the poor recognition accuracy at higher SNR levels thanks to the better discrimination of GMMs trained on MFCC features rather than the suboptimal divergence metric used for exemplar matching. Moreover, on account of the more precise representations of the speech units, the proposed front-end is expected to provide better enhancement and recognition than the FE approach [27, 28] which is an alternative exemplar-based sparse representations approach performing enhancement using fixed-length exemplars in a single overcomplete dictionary. We have performed experiments on both the AURORA-2 database and the small vocabulary track of the 2<sup>nd</sup> CHiME Challenge to investigate the performance of the proposed approach under different noise and training conditions and compare the performance with other noise robust recognition systems.

## 2. Noise Robust Exemplar Matching

Training frame sequences representing various speech units (speech exemplars) are extracted based on the state-level alignments obtained using an HMM-based recognizer. Speech exemplars, each comprised of  $D$  mel frequency bands and spanning  $l$  frames, are reshaped into a single vector and stored in the columns of a speech dictionary  $\mathbf{S}_{c,l}$ : one for each class  $c$  and each length  $l$ . Each dictionary is of dimensionality  $Dl \times N_{c,l}$  where  $N_{c,l}$  is the number of available speech exemplars of class  $c$  and length  $l$ . Similarly, a single noise dictionary  $\mathbf{N}_l$  for each length  $l$  is formed by reshaping the noise exemplars. Each speech dictionary is concatenated with the noise dictionary of the same length to form a combined dictionary  $\mathbf{A}_{c,l} = [\mathbf{S}_{c,l} \mathbf{N}_l]$  of dimensionality  $Dl \times M_{c,l}$  where  $M_{c,l}$  is the total number of available speech and noise exemplars.

Every noisy speech segment of frame length  $T$  is also reshaped into vectors by applying a sliding window approach

10.21437/Interspeech.2015-241

This research was funded by the KU Leuven research grant GOA/14/005 (CAMETRON) and the European Commission under Contract FP7-PEOPLE-2011-290000 (INSPIRE).

[27] with window length of  $l$  frames and stored in an observation matrix  $\mathbf{Y}_l = [\mathbf{y}_l^1, \mathbf{y}_l^2, \dots, \mathbf{y}_l^{(T-l+1)}]$  of dimensionality  $Dl \times (T-l+1)$ . Due to multiple-length exemplars, the window length  $l$  is varied between the minimum exemplar length  $l_{\min}$  and maximum exemplar length  $l_{\max}$  yielding observation matrices  $\mathbf{Y}_l$  for  $l_{\min} \leq l \leq l_{\max}$ . For every class  $c$ , each observation vector  $\mathbf{y}_l$  is expressed as a linear combination of the exemplars that are stored in the dictionaries of the same length:  $\mathbf{y}_l \approx \sum_{m=1}^{M_{c,l}} x_{c,l}^m \mathbf{a}_{c,l}^m = \mathbf{A}_{c,l} \mathbf{x}_{c,l}$  for  $x_{c,l}^m \geq 0$  where  $\mathbf{x}_{c,l}$  is an  $M_{c,l}$ -dimensional non-negative weight vector. The sparse solutions of  $\mathbf{x}_{c,l}$  yield more realistic approximation of the observed segments without overfitting and have been shown to provide better recognition results [29,30]. The combined dictionaries consisting of speech and noise exemplars are presumed to model all acoustic variability in the observed signal due to pronunciation variation, background noise and so forth. This model can also cope with reverberation by storing reverberated speech exemplars rather than clean speech exemplars.

The non-negative exemplar weights  $\mathbf{x}_{c,l}$  are obtained by minimizing the cost function,  $d(\mathbf{y}_l, \mathbf{A}_{c,l} \mathbf{x}_{c,l}) + \sum_{m=1}^{M_{c,l}} x_{c,l}^m \Lambda_m$  for  $x_{c,l}^m \geq 0$  where  $\Lambda$  is an  $M_{c,l}$ -dimensional vector. The first term is the divergence between the observation vector and its approximation. The second term is a regularization term which penalizes the  $l_1$ -norm of the weight vector to produce a sparse solution.  $\Lambda$  contains non-negative values and controls how sparse the resulting vector  $\mathbf{x}$  is. Defining  $\Lambda$  as a vector, the amount of sparsity enforced on different types of exemplars can be adjusted. In this work, the regularized optimization problem with the aforementioned cost function is solved by applying non-negative sparse coding (NSC) [31]. The generalized KLD is used for  $d$  which is commonly used in source separation problems and shown to produce better results than Euclidean distance when used in conjunction with mel-scaled spectral features [30],  $d(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{k=1}^K y_k \log \frac{y_k}{\hat{y}_k} - y_k + \hat{y}_k$ .

All observation matrices  $\mathbf{Y}_l$  for  $l_{\min} \leq l \leq l_{\max}$  are approximated using the combined dictionaries  $\mathbf{A}_{c,l}$  of the same length by applying the multiplicative update rule given in [25]. To quantify the approximation quality, we use the reconstruction error between the noisy speech segments and their approximations. After a fixed number of iterations for all dictionaries, the reconstruction errors between the observation matrix  $\mathbf{Y}_l$  and its approximations  $\mathbf{A}_{c,l} \mathbf{x}_{c,l}$  are calculated for  $l_{\min} \leq l \leq l_{\max}$ . As the label of each dictionary is known, decoding is performed by applying dynamic programming to find the class sequence that minimizes the reconstruction error to find the best approximation of the target utterance.

After finding the best approximation, the denoising is performed by reconstructing the frame-wise speech and noise estimates,  $\hat{s}_{c,l}$  and  $\hat{n}_{c,l}$ , that are obtained after removing the windowing effect by adding the components belonging to overlapping windows from the estimates  $S_{c,l} X_{c,l}^s$  and  $N_l X_{c,l}^n$  respectively. Here,  $X_{c,l}^s$  refers to the exemplar weights of the speech exemplars and  $X_{c,l}^n$  refers to the exemplar weights of the noise exemplars. The frame-level Wiener-like filter is then obtained as in [19],  $W = \mathbf{C}^T \hat{s}_{c,l} \oslash (\mathbf{C}^T (\hat{s}_{c,l} + \hat{n}_{c,l}))$  where  $\mathbf{C}$  is the short-time Fourier transform (STFT)-to-mel matrix containing triangular shaped filter-banks.

### 3. Experimental Setup

#### 3.1. Databases

**AURORA-2:** The training material of AURORA-2 [37] database consists of a clean and a multi-condition training set, each containing 8440 utterances. The multi-condition training set was

constructed by mixing the clean utterances with noise at SNR levels of 20, 15, 10 and 5 dB. Test set A and B consists of 4 clean and 24 noisy datasets at six SNR levels between -5 and 20 dB. The noise types of test set A match the multi-condition training set. Each subset contains 1001 utterances with one to seven digits 0-9 or oh. We use the complete test sets to be able to compare the results with other systems.

**CHIME-2:** The small vocabulary track of the 2<sup>nd</sup> CHiME Challenge [38] addresses the problem of recognizing commands in a noisy and reverberant living room. The clean utterances contain utterances from 34 speakers reading 6-word sequences of the form *command-color-preposition-letter-digit-adverb*. There are 25 different letters, 10 different digits and 4 different alternatives for each of the other classes. The recognition accuracy of a system is calculated based on the correctly recognized letter and digit keywords.

#### 3.2. Dictionary Creation and Implementation Details

**AURORA-2:** The speech exemplars are extracted from the clean training set. Acoustic feature vectors used during speech enhancement are represented in mel-scaled magnitude spectra with 23 frequency bands. There are in total 52,305 speech exemplars representing half-digits. The minimum and maximum exemplar lengths are 8 and 40 frames respectively. Exemplars longer than 40 frames are omitted to limit the number of dictionaries. The recognizer uses in total 675 dictionaries of 23 different classes (half-digits plus silence). The noise dictionaries are created by performing active noise exemplar selection and noise sniffing which are detailed in [25]. The combined dictionaries and observation matrices are  $l_3$ -normalized for all SNR levels. The multiplicative update rule is iterated 100 times for convergence of all frame lengths. Further details can be found in [25].

The enhanced speech is input to a GMM-HMM recognizer employing an HMM topology with 16 states describing each digit and 3 states for silence leading to a total of 179 states. The GMM model is trained on MFCC with 13 static features along with their delta and delta-delta time differences resulting in a 39 dimensional feature space. The emission probabilities of each HMM state is modeled using a GMM of 32 Gaussians with diagonal covariance. For the Viterbi decoder, an HMM topology where all the words have the same word entrance penalties was used. We trained acoustic models on the clean and multicondition training set. To evaluate the impact of retraining on the recognition accuracy, we further train an acoustic model on the enhanced waveforms of the multicondition training set.

**CHIME-2:** The N-REM system for speech enhancement uses exemplars and noisy speech segments that are represented as mel-scaled magnitude spectral features extracted with a 26 channel mel-scaled filter bank ( $D = 26$ ). The frame length is 25 ms and the frame shift is 10 ms. The binaural data is averaged in the spectral domain to obtain 26-dimensional feature vectors. The exemplars are extracted from the *reverberated* utterances in the training set according to the state-based segmentations obtained using the acoustic models in the toolkit provided with the database. Exemplars belonging to each speaker are organized in separate dictionary sets for speaker-dependent modeling yielding 34 different dictionary sets. Based on the availability of the exemplars, the minimum and maximum exemplar lengths are 4 and 45 frames respectively.

Half-word exemplars seemed to generalize sufficiently to unseen data. Dictionary sizes vary with different classes and speakers. *Prewarping* [39] is applied to boost the modeling capabilities of the underpopulated speech dictionaries. The num-

Table 1: Word error rates in % obtained on test set A and B of AURORA-2 data

SNR(dB)	-5	0	5	10	15	20	0-20	clean
N-REM	20.1	10.0	6.3	4.6	3.5	2.7	5.4	1.8
N-REM-SE (clean)	24.4	10.1	5.3	3.4	2.1	1.4	4.4	<b>0.3</b>
N-REM-SE (multi)	<b>17.9</b>	<b>8.2</b>	<b>4.5</b>	<b>2.8</b>	<b>1.9</b>	<b>1.0</b>	<b>3.6</b>	0.8
N-REM-SE (retrain)	19.5	8.6	4.7	3.2	<b>1.9</b>	<b>1.0</b>	3.9	0.5

(a) Test set A

SNR(dB)	-5	0	5	10	15	20	0-20	clean
N-REM	56.9	25.6	10.4	5.7	3.8	3.2	9.7	1.8
N-REM-SE (clean)	56.3	24.1	9.5	4.2	2.3	1.3	8.3	<b>0.3</b>
N-REM-SE (multi)	<b>55.0</b>	<b>23.9</b>	<b>8.9</b>	<b>3.9</b>	<b>1.9</b>	<b>1.2</b>	<b>8.0</b>	0.8
N-REM-SE (retrain)	55.7	24.1	9.3	4.1	2.1	<b>1.2</b>	8.2	0.5

(b) Test set B

Table 2: Comparison of NREM-SE with other recognition systems on AURORA-2 data

Technique	<i>test set A</i>		<i>test set B</i>	
	-5	0-20	-5	0-20
GMM-HMM [28]	77.2	16.9	77.1	15.9
AFE [32]	56.5	7.7	57.7	8.2
NAT [33]	57.7	6.3	58.1	6.3
SC [28]	35.7	7.2	49.8	9.3
FE [28]	30.4	3.6	50.8	6.1
SC+FE [28]	25.6	3.1	<b>43.7</b>	5.0
ESSEM-MCM [34]	-	4.4	-	<b>4.7</b>
RBM-DNN [35]	-	4.5	-	5.1
MASK-RBM-DNN [14]	-	3.8	-	5.0
MS-CD [36]	21.1	<b>2.4</b>	62.4	7.5
FE+MS-CD [36]	20.6	<b>2.4</b>	54.2	6.1
N-REM [25]	20.1	5.4	56.9	9.7
N-REM-SE	<b>17.9</b>	3.6	55.0	8.0

ber of exemplars in each dictionary after prewarping is limited to 50. The noise dictionaries used for the recognition phase contain 200 noise exemplars that are acquired on the fly from the immediate neighborhood of the target utterance in both directions until the frames belonging to other target utterances. In addition to these sniffed noise exemplars, 200-300 noise exemplars are extracted from the most active 2 noise-only sequences selected by adaptive noise exemplar selection technique [25]. The multiplicative update rule is iterated 25 times to obtain the exemplar weights. The columns of the combined dictionaries and observation matrices are  $l_2$ -normalized. Further details can be found in [25].

The enhanced speech is recognized using the baseline HMM structure provided by the challenge organizers at the back-end [38]. The provided acoustic models use 4-10 state word-level HMMs and the emission probabilities of each HMM state is modeled using a GMM of 7 Gaussians. The speech features are standard 39-dimensional MFCCs applied with cepstral mean normalization. We first use the default acoustic models trained on clean, reverberated and noisy training utterances. Similar to the AURORA-2 experiments, we also retrain a new acoustic model using the enhanced isolated noisy training utterances.

### 3.3. Evaluation Metrics

We have opted for the metrics which have been traditionally used for the evaluation of the databases described in Section 3.1 for comparability with the previous literature. The word error rate (WER) has been used to quantify the recognition accuracy for the AURORA-2 digit recognition task. The keyword recognition accuracy (RA) is used to evaluate the system performance on the CHIME-2 data.

## 4. Results

We perform recognition experiments using the proposed system (N-REM-SE) on test set A and B of AURORA-2 and development and test sets of CHIME-2 data. For both datasets, we first compare the performance of N-REM-SE with the exemplar

matching version (N-REM) [25] using the same combined dictionaries and divergence measure. Then, the proposed system is compared with other noise robust ASR systems to evaluate the overall performance of N-REM-SE.

### 4.1. AURORA-2 Results

Table 1 and 2 presents the results obtained on AURORA-2 data. The clean speech recognition performance of all systems are given in the last column of Table 1a and Table 1b. The exemplar matching system provides a WER of 1.8% on clean speech which is higher than any setup with a GMM-HMM back-end. N-REM-SE trained on clean and multicondition data has a WER of 0.3 and 0.8 on clean speech respectively. As expected, the GMM-HMM back-end considerably improves the clean speech recognition performance.

N-REM provides WERs of 20.1%, 10.0% and 6.3% at SNR levels of -5 dB, 0 dB and 5 dB. N-REM-SE trained on clean speech performs surprisingly well with WERs of 24.4%, 10.1% and 5.3% at the same SNR levels. Training the acoustic models of N-REM-SE on multicondition data improves the results by an absolute improvements of 6.5%, 1.9% and 1.0% respectively. Unlike the other exemplar-based approaches which use a single fixed noise dictionary, retraining the acoustic models on the enhanced training data does not bring any improvement in case of N-REM-SE. This is due to the adaptive noise modeling adopted in N-REM-SE which selects a different set of noise exemplars for each noisy utterance on the fly. Consequently, retraining does not help the back-end to cope with the artifacts introduced by speech enhancement in this scenario. The models trained on multicondition data yield better or similar recognition performance at all SNR levels.

Using a GMM-HMM back-end reduces the WER in general at higher SNR levels similar to the clean speech performance. Compared to the WERs of 4.6%, 3.5% and 2.7% provided by N-REM at SNRs of 10 dB, 15 dB and 20 dB respectively, N-REM-SE trained on multicondition data has WERs of 2.8%, 1.9% and 1.0% at the same SNRs. Moreover, this system has an average WER (0-20) of 3.6% compared to the 5.4% of N-REM. Multicondition trained N-REM-SE also shows superior performance at all SNR levels of test set B compared to N-REM and other N-REM-SE variants. This system provides an average WER of 8.0% compared to the 9.7% of N-REM and 8.3% of retrained N-REM-SE.

Table 2 lists the recognition results of some other noise robust ASR systems data with state-of-the-art performance on AURORA-2 data. This list is by no means exhaustive and it only includes the recognition results published on the complete test sets for a fair comparison. From these results, it can be concluded that the recognition systems using exemplar-based speech enhancement approaches, e.g. FE variants and N-REM-SE, provide impressive performance in matched noise scenarios. Other exemplar-based systems which do not rely on a statistical model at the back-end, e.g. SC and N-REM, mainly suffer from low recognition accuracies at higher SNR levels resulting in worse average WER results. On the other hand, the ESSEM-MCM and RBM-DNN methods perform almost equally well

Table 3: Keyword recognition accuracies in % obtained on the development and test set of CHIME-2 data

SNR(dB)	-6	-3	0	3	6	9	Avg
N-REM	<b>70.4</b>	<b>77.9</b>	<b>84.8</b>	<b>90.4</b>	<b>92.6</b>	<b>93.8</b>	<b>85.0</b>
N-REM-SE (clean)	21.1	23.4	27.9	30.5	34.4	34.9	28.7
N-REM-SE (reverb)	67.3	74.7	81.7	88.2	89.8	91.5	82.2
N-REM-SE (noisy)	60.8	68.8	74.3	78.1	81.8	83.2	74.5
N-REM-SE (retrain)	69.4	75.9	82.8	87.4	88.6	91.7	82.6

(a) Development Set

SNR(dB)	-6	-3	0	3	6	9	Avg
N-REM	<b>71.0</b>	<b>78.9</b>	<b>85.3</b>	<b>88.7</b>	<b>91.9</b>	<b>92.8</b>	<b>84.8</b>
N-REM-SE (clean)	19.9	22.8	26.8	29.7	34.2	38.1	28.6
N-REM-SE (reverb)	69.8	76.8	84.3	87.3	90.3	91.6	83.4
N-REM-SE (noisy)	60.3	69.1	74.8	78.0	81.4	82.8	74.4
N-REM-SE (retrain)	70.3	76.4	84.5	86.7	89.3	90.6	83.0

(b) Test set

Table 4: Comparison of NREM-SE with other recognition systems on CHIME-2 data

Technique	Baseline AM		Retrained AM	
	Dev Avg	Test Avg	Dev Avg	Test Avg
GMM [38]	68.7	68.8	-	-
SCSS [40]	76.0	77.7	-	-
FE [41]	81.2	82.2	-	-
HMM-FE [41]	81.9	82.7	-	-
BSE [42]	81.8	82.0	81.5	83.2
FASST [43]	<b>82.9</b>	<b>84.2</b>	<b>84.7</b>	<b>85.7</b>
N-REM-SE	82.2	83.4	82.6	83.0

under matched and mismatched noise conditions.

The hybrid SC+FE system appears to be a nice compromise with a remarkable -5 dB performance and one of the lowest average WERs on both test sets. Compared to this system and other exemplar-based systems, the gap between the matched and mismatched noise is larger for the proposed system due to the smaller amount of noise exemplars in the class- and length-dependent dictionaries. This results in poor generalization against unseen noise types. In the case of matched noise, N-REM-SE has a better -5 dB performance, which is actually the best among all systems, and a comparable average WER on test set A.

#### 4.2. CHIME-2 Results

Table 3 and 4 presents the results obtained on CHIME-2 data. We first focus on Table 3 presenting the recognition accuracies obtained on the development and test sets to compare the performance of the exemplar matching-based system and the proposed recognizer. The results on both sets follow a similar trend; hence the results on the test set are discussed only. N-REM provides RAs of 71.0%, 78.9% and 85.3% at -6 dB, -3 dB and 0 dB. N-REM-SE trained on reverberated data has RAs of 69.8%, 76.8% and 84.3% at the same SNR levels. These results are slightly worse than the exemplar matching system. Retraining the acoustic models does not improve the recognition performance similar to the results obtained on AURORA-2 data. The proposed setup with the acoustic models trained on clean and noisy speech provides inferior results.

The overall performance of N-REM is also mildly better with an average RA (Avg) of 84.8% compared to the proposed recognizer trained on the reverberated data with 83.4%. The CHIME-2 results favor the exemplar matching system over N-REM-SE unlike the AURORA-2 experiments. The same observation holds for the single dictionary counterparts, FE and SC [27], considering the recognition results reported in [25]. We discuss the differences between AURORA-2 and CHIME-2 databases to get more insight for the reduced performance of the exemplar-based front-end denoising systems (N-REM-SE and FE) on CHIME-2 task compared to the other systems (N-REM and SC) which do not rely on a statistical model at the back-end. Firstly, the variation in both speech and noise components in the noisy mixtures are more significant in CHIME-2 com-

pared to AURORA-2. The former is due to the smearing effect of reverberation degrading the spectrotemporal content of the speech exemplars and the latter is an outcome of the highly non-stationary room noise. Secondly, there are only few exemplars available in training data, especially for *letters*, to obtain accurate representations of each speech unit in the high-dimensional feature space. Hence, the speech enhancement quality provided by the combined dictionaries with increased variation is less effective in compensating for the mismatch between the target speech and the acoustic models trained on neither reverberated nor noisy speech. Under such a scenario, adopting a GMM-HMM back-end is less favorable compared to the N-REM and SC systems which either relies on the reconstruction error or estimates state likelihoods directly from the exemplar weights at the back end respectively.

To be able to evaluate the enhancement performance of the N-REM front-end, we present some results obtained using other speech or feature enhancement-based recognition systems in Table 4. The recognition results of the best performing GMM-HMM baseline trained reverberated data is also provided as a reference. The best performing FASST system does not only benefit from spectral enhancement, but also from spatial enhancement using spatial full-rank covariance matrices [43]. All other systems benefiting only from either spectral or spatial enhancement and using the standard acoustic models provided as a part of the CHIME-2 challenge perform moderately compared to the more sophisticated approaches with some speaker and environment adaptation techniques. N-REM-SE provides a reasonable performance outperforming the other exemplar-based sparse representation systems, FE and HMM-FE, which use exemplars of fixed length in a single overcomplete dictionary for feature enhancement.

## 5. Conclusion

This paper presents a novel noise robust ASR system using a single-channel speech enhancement setup that performs noise robust exemplar matching to separate speech and noise sources in the front-end. The exemplars used in this technique are associated with a certain speech unit and organized in separate dictionaries based on the associated speech unit and length. The noisy mixtures are approximated as a sparse linear combination of the speech and noise exemplars in each dictionary. The proposed system has provided comparable performance with the other state-of-the-art ASR systems on two popular small vocabulary recognition tasks, AURORA-2 and CHIME-2.

Future work includes investigating the performance of the hybrid N-REM recognizer which combines the acoustic scores obtained from the statistical models and exemplar matching. Moreover, replacing the mel-scaled magnitude spectral features with perceptually motivated modulation spectrogram features is expected to provide better separation of speech and noise. Finally, an extension of the proposed system working on databases with larger vocabulary using e.g. phone or biphone sized exemplars remains as a future work.

## 6. References

- [1] D. Kolossa and R. Haeb-Umbach, Eds., *Robust Speech Recognition of Uncertain or Missing Data - Theory and Applications*. Springer, 2011.
- [2] T. Virtanen, R. Singh, and B. Raj, *Techniques for Noise Robustness in Automatic Speech Recognition*. Wiley, 2012.
- [3] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM TASLP*, vol. 22, no. 4, pp. 745–777, 2014.
- [4] D. Van Compernelle, "Noise adaptation in a hidden markov model speech recognition system," *Computer Speech & Language*, vol. 3, no. 2, pp. 151–167, 1989.
- [5] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), hidden markov models and the projection, for robust speech recognition in cars," *Speech Communication*, vol. 11, no. 2-3, pp. 215–228, Jun. 1992.
- [6] H. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Proc. ICASSP*, 1995, pp. 153–156.
- [7] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *Proc. ICSLP*, Oct. 2000, pp. 806–809.
- [8] D. Macho, L. Mauuary, B. Noé, Y. M. Cheng, D. Ealey, D. Jouviet, H. Kelleher, D. Pearce, and F. Saadoun, "Evaluation of a noise-robust DSR front-end on aurora databases," in *Proc. INTERSPEECH*, 2002, pp. 17–20.
- [9] C. Breithaupt and R. Martin, "Statistical analysis and performance of DFT domain noise reduction filters for robust speech recognition," in *Proc. INTERSPEECH*, 2006, pp. 365–368.
- [10] ETSI, *ETSI ES 202 050 V1.1.5 (2007-01), Advanced front-end feature extraction algorithm*, January 2007.
- [11] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, "Robust speech recognition using a cepstral minimum-mean-square-error-motivated noise suppressor," *IEEE TASLP*, vol. 16, no. 5, pp. 1061–1070, Jul. 2008.
- [12] A. L. Maas, Q. V. Le, T. O’Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Proc. INTERSPEECH*, 2012, pp. 22–25.
- [13] M. L. Seltzer, Y. Dong, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. ICASSP*, May 2013, pp. 7398–7402.
- [14] B. Li and K. C. Sim, "Improving robustness of deep neural networks via spectral masking for automatic speech recognition," in *Proc. ASRU*, 2013, pp. 279–284.
- [15] A. Narayanan and D. Wang, "Investigation of speech separation as a front-end for noise robust speech recognition," *IEEE/ACM TASLP*, vol. 22, no. 4, pp. 826–835, Apr. 2014.
- [16] D. Baby, J. F. Gemmeke, T. Virtanen, and H. Van hamme, "Exemplar-based speech enhancement for deep neural network based automatic speech recognition," in *Proc. ICASSP*, May 2015.
- [17] P. Smaragdīs, M. Shashanka, and B. Raj, "A sparse non-parametric approach for single channel separation of known sounds," in *NIPS*, 2009, pp. 1705–1713.
- [18] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based speech enhancement and its application to noise-robust automatic speech recognition," in *Proc. CHIME-2011*, Sept. 2011, pp. 53–75.
- [19] D. Baby, T. Virtanen, T. Barker, and H. Van hamme, "Coupled dictionary training for exemplar-based speech enhancement," in *Proc. ICASSP*, May 2014, pp. 2883–2887.
- [20] N. Mohammadiha and S. Doclo, "Single-channel dynamic exemplar-based speech enhancement," in *Proc. INTERSPEECH*, Sept. 2014, pp. 2690–2694.
- [21] M. De Wachter, K. Demuyck, D. Van Compernelle, and P. Wambacq, "Data driven exemplar based continuous speech recognition," in *Proc. EUROSPEECH*, 2003, pp. 1133–1136.
- [22] T. Deselaers, G. Heigold, and H. Ney, "Speech recognition with state-based nearest neighbour classifiers," in *Proc. INTERSPEECH*, Antwerp, Belgium, 2007, pp. 2093–2096.
- [23] L. Golipour and D. O’Shaughnessy, "Context-independent phoneme recognition using a k-nearest neighbour classification approach," in *Proc. ICASSP*, Apr. 2009, pp. 1341–1344.
- [24] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" in *NIPS*. Cambridge, MA: MIT Press, 2004.
- [25] E. Yılmaz, J. F. Gemmeke, and H. Van hamme, "Noise robust exemplar matching using sparse representations of speech," *IEEE/ACM TASLP*, vol. 22(8), pp. 1306–1319, Aug. 2014.
- [26] —, "Noise robust exemplar matching with alpha-beta divergence," *Submitted to Speech Communication*, 2015.
- [27] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE TASLP*, vol. 19(7), pp. 2067–2080, Sept. 2011.
- [28] J. F. Gemmeke and H. Van hamme, "Advances in noise robust digit recognition using hybrid exemplar-based techniques," in *Proc. INTERSPEECH*, Portland, USA, Sept. 2012, pp. 1–4.
- [29] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, 2004.
- [30] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE TASLP*, vol. 15, no. 3, pp. 1066–1074, March 2007.
- [31] P. Hoyer, "Non-negative sparse coding," in *IEEE Workshop on Neural Networks for Signal Processing*, 2002, pp. 557–565.
- [32] H. G. Hirsch and D. Pearce, "Applying the Advanced ETSI front-end to the Aurora-2 task," Tech. Rep., Sept. 2006, version 1.1.
- [33] O. Kalinli, M. Seltzer, J. Droppo, and A. Acero, "Noise adaptive training for robust automatic speech recognition," *IEEE TASLP*, vol. 18, no. 8, pp. 1889–1901, Nov 2010.
- [34] Y. Tsao, J. Li, C.-H. Lee, and S. Nakamura, "Soft margin estimation on improving environment structures for ensemble speaker and speaking environment modeling," in *Proc. 3rd Int. Universal Communication Sym.*, pp. 404–408.
- [35] B. Li and K. C. Sim, "A spectral masking approach to noise-robust speech recognition using deep neural networks," *IEEE/ACM TASLP*, vol. 22, no. 8, pp. 1296–1305, Aug 2014.
- [36] D. Baby, T. Virtanen, J. F. Gemmeke, T. Barker, and H. Van hamme, "Exemplar-based noise robust automatic speech recognition using modulation spectrogram features," in *IEEE SLT Workshop*, South Lake Tahoe, USA, Dec. 2014.
- [37] H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ASR2000*, Sept. 2000, pp. 181–188.
- [38] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second CHiME speech separation and recognition challenge: Datasets, tasks and baselines," in *Proc. ICASSP*, Vancouver, Canada, May 2013, pp. 126–130.
- [39] E. Yılmaz, J. F. Gemmeke, and H. Van hamme, "Noise-robust automatic speech recognition with exemplar-based sparse representations using multiple length adaptive dictionaries," in *Proc. CHIME-2013*, Vancouver, Canada, June 2013, pp. 39–43.
- [40] P. Mowlae, J. A. Morales-Cordovilla, F. Pernkopf, H. Pessen-thainer, M. Hagmuller, and G. Kubin, "The 2nd CHIME speech separation and recognition challenge: Approaches on single-channel source separation and model-driven speech enhancement," in *Proc. CHIME-2013*, 2013, pp. 59–64.
- [41] J. F. Gemmeke, A. Hurmalainen, and T. Virtanen, "HMM-regularization for NMF-based noise robust ASR," in *Proc. CHIME-2013*, 2013, pp. 47–52.
- [42] F. Nesta, M. Matassoni, and R. F. Astudillo, "A flexible spatial blind source extraction framework for robust speech recognition in noisy environments," in *Proc. CHIME-2013*, 2013, pp. 33–38.
- [43] D. T. Tran, E. Vincent, D. Jouviet, and K. Adiloglu, "Using full-rank spatial covariance models for noise-robust ASR," in *Proc. CHIME-2013*, 2013, pp. 31–32.