ADAPTIVE NOISE DICTIONARY DESIGN FOR NOISE ROBUST EXEMPLAR MATCHING OF SPEECH

Emre Yılmaz and Hugo Van hamme

Dept. ESAT-PSI
KU Leuven, Belgium

Jort F. Gemmeke

Audience Inc.
Mountain View, CA, USA

ABSTRACT

This paper investigates an adaptive noise dictionary design approach to achieve an effective and computationally feasible noise modeling for the noise robust exemplar matching (N-REM) framework. N-REM approximates noisy speech segments as a linear combination of multiple length exemplars in a sparse representation (SR) formulation. Compared to the previous SR techniques with a single overcomplete dictionary, N-REM uses smaller dictionaries containing considerably fewer noise exemplars. Hence, the noise exemplars have to be selected with care to accurately model the spectrotemporal content of the actual noise conditions. For this purpose, in a previous work, we introduced a noise exemplar selection stage before performing recognition which extracts noise exemplars from a few noise-only training sequences chosen for each target noisy utterance. In this work, we explore the impact of the several design parameters on the recognition accuracy by evaluating the system performance on the CHIME-2 and AURORA-2 databases.

Index Terms— template matching, noise-robustness, automatic speech recognition, sparse representations, exemplar selection

1. INTRODUCTION

Using exemplars in a sparse representation (SR) formulation to model noisy speech has provided major improvements in the automatic speech recognition (ASR) performance compared to conventional approaches such as hidden Markov models (HMM) under adverse conditions [1]. Previously, we have proposed an ASR system that performs noise robust exemplar matching (N-REM) [2] using exemplars of multiple lengths, each associated with a single speech unit such as phones, syllables, half-words or words similar to [3]. Exemplars of different length are organized in separate dictionaries based on the associated speech unit (class) and length unlike the previous SR-based systems [4–6] using a single dictionary with fixed-length exemplars. Using separate dictionaries for each class provides better classification as input speech segments are approximated as a linear combination of exemplars belonging to the same class only [7].

This work has been supported by the KU Leuven research grant OT/09/028 (VASI) and IWT-SBO Project 100049 (ALADIN).

The N-REM dictionaries are substantially less populated compared to a single overcomplete dictionary, as the speech exemplars are associated with a single speech unit and their length distribution is class-dependent which results in unevenly populated speech dictionaries. Unlike the speech exemplars, noise exemplars are extracted from noise-only training sequences for any arbitrary length. As a result, while there are a large number of available noise exemplars for each exemplar length, only the ones that match the actual test noise conditions will be essential for accurate recognition. Thus, noise dictionary design mainly focuses on accurate modeling of the background noise using the smallest possible number of noise exemplars. Previous experiments have shown that rudimentary noise modeling approaches, e.g. using *fixed* noise dictionaries, provide very poor estimation of the noise source [7]. Using much smaller noise dictionaries due to computational restrictions compared to the previous SR-based recognizers with fixed-length exemplars results in inferior performance especially at lower SNR levels. For this reason, we have proposed an adaptive noise exemplar selection technique which chooses the best matching noise-only training sequences from a noise repository using a selection dictionary and extracts the noise exemplars that are used during the recognition from these sequences [2]. In this paper, we further explore the impact of several design parameters, e.g. size of the noise repository and the amount selected noise exemplars, on the recognition performance to reach a compromise between the noise robustness of the recognizer and the computational complexity.

2. NOISE ROBUST EXEMPLAR MATCHING

Training frame sequences representing various noise-free speech units (speech exemplars), each comprised of D mel bands and spanning l frames, are extracted from the alignments obtained with an HMM-based recognizer and 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. Similarly, a single noise dictionary $\mathbf{N}_l$ for each length l is formed by reshaping 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 $(D \cdot l) \times M_{c,l}$ where $M_{c,l}$ is the total number of speech

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

and noise exemplars.

An observed noisy speech segment of length T frames is also reshaped into vectors by applying a sliding window approach [4] with window length of l frames and stored in an observation matrix $\mathbf{Y}_l = [\mathbf{y}_l^1, \mathbf{y}_l^2 ..., \mathbf{y}_l^{(T-l+1)}]$ of dimensionality $(D \cdot l) \times (T - l + 1)$ for $l_{\text{min}} \leq l \leq l_{\text{max}}$ where l_{min} and l_{max} are the smallest and largest speech exemplar lengths respectively. 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, $\mathbf{y}_l \approx \mathbf{A}_{c,l}\mathbf{x}_{c,l}$ for $x_{c,l}^m \geq 0$ where $\mathbf{x}_{c,l}$ is an $M_{c,l}$-dimensional non-negative weight vector. The exemplar weights 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 $\mathbf{\Lambda}$ is an $M_{c,l}$-dimensional vector which contains non-negative values and controls how sparse the resulting vector $\mathbf{x}$ is. The generalized Kullback-Leibler divergence (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 [8]. The generalized KLD is defined as $d(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{k=1}^{K} y_k \log \frac{y_k}{\hat{y}_k} - y_k + \hat{y}_k$.

The regularized optimization problem with the aforementioned cost function is solved with non-negative sparse coding (NSC) [9]. For NSC, we apply the multiplicative update rule given in [2] to obtain the exemplar weights. In practice, all observation matrices $\mathbf{Y}_l$ for $l_{\text{min}} \leq l \leq l_{\text{max}}$ are approximated using the combined dictionaries $\mathbf{A}_{c,l}$ of the corresponding length by applying the multiplicative update rule. After a fixed number of iterations, the reconstruction errors between each observation matrix $\mathbf{Y}_l$ and its approximation are calculated. 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.

3. SELECTION DICTIONARY DESIGN

The accuracy of the noise modeling depends on the congruence of the spectrotemporal content of the noise exemplars and actual noise conditions contaminating the target utterance. Therefore, for each noisy utterance, a few noise-only training sequences that are able to model the background noise are selected on the fly and the noise exemplars in $\mathbf{N}_l$ are extracted from these sequences. The selection is performed by applying NSC using a selection dictionary $\mathbf{A}_{L_s}^* = [\mathbf{S}_{L_s}^* \mathbf{N}_{L_s}^*]$ containing speech exemplars from all classes and noise exemplars from different noise-only training sequences. The superscript * marks the dictionaries used in the noise exemplar selection. The speech dictionary $\mathbf{S}_{L_s}^*$ is obtained by concatenating an equal number of speech exemplars of the same length from each class. The length L_{s} can be set to any exemplar length containing abundant speech exemplars from each class.

For the noise dictionary $\mathbf{N}_{L_s}^*$, a noise repository of F noise-only training sequences is created and G noise exemplars are extracted from each noise-only training sequence with an equal frame shift. In total, $\mathbf{N}_{L_s}^*$ contains $F \cdot G$ noise exemplars. Once the selection dictionary $\mathbf{A}_{L_s}^*$ is formed, the observation matrix $\mathbf{Y}_{L_s}$ of length L_{s} is approximated as a linear combination of the exemplars in the selection dictionary $\mathbf{Y}_{L_s} \approx \mathbf{A}_{L_s}^* \mathbf{x}_{L_s}$ for $\mathbf{x}_{L_s} \geq 0$. By accumulating the weights of all noise exemplars extracted from the same training sequence, a total weight for each training sequence is obtained. Evidently, the training sequences having higher weights are expected to model the spectrotemporal properties of the background noise [10]. Hence, the noise dictionaries $\mathbf{N}_l$ for $l_{\text{min}} \leq l \leq l_{\text{max}}$ that are used during the recognition contain noise exemplars extracted from X training sequences with the highest weights.

Noise sniffing [11] is also applied for acquiring noise exemplars on the fly from the immediate neighborhood of the target utterance. The extracted noise exemplars are contained in the noise dictionaries, i.e., $\mathbf{N}_{L_s}^*$ as a part of the selection dictionary. Shifted copies of these frame sequences are also included to provide some degree of shift-invariance [12].

4. EXPERIMENTAL SETUP

4.1. Databases

The training material of AURORA-2 [13] 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. To reduce the simulation times, we subsampled the test sets by a factor of 4 (1000 utterances per SNR).

The small vocabulary track of the 2nd CHiME Challenge [14] 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.

4.2. Exemplar extraction and implementation details

The speech exemplars are extracted from the clean training set of AURORA-2 data. Acoustic feature vectors are represented in mel-scaled magnitude spectra with 23 mel bands. The speech exemplars representing half-digits are segmented by a conventional HMM-based system. There are in total 52,305 speech exemplars excluding 990 silence exemplars. 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 noise-only training sequences are obtained by removing speech from the noisy utterances in the multi-condition training set. The *fixed*

noise dictionaries are extracted from the 16 longest noise-only training sequences with shifts of 4 frames. Consequently, the fixed dictionaries contain between 547-589 noise exemplars depending on the exemplar length. The selection dictionary contains noise exemplars that are extracted from the longest noise-only training sequences. The amount of noise exemplars in the selection dictionaries depends on the chosen F and G value. The selection dictionary also contains 2200 speech exemplars. It uses speech and noise exemplars containing 15 frames. For AURORA-2, an SNR-dependent X value is used as it provides an improved recognition accuracy and reduced computational load at higher SNR levels by using less noise exemplars. The number of noise exemplars extracted from each sequence varies between 77 and 170. The further details of the SNR-dependent noise modeling is given in [2]. The word error rate is used to quantify the recognition accuracy on AURORA-2 data.

The exemplars and noisy speech segments of CHIME-2 data 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. Half-word 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 40 frames respectively. The baseline system performs recognition using noise dictionaries containing 400 sniffed noise exemplars. Each embedded utterance in the development and test set is segmented into noise-only sequences by removing all target utterances. G=5 noise exemplars of 25 frames are extracted from each noise-only sequence and stored in the single noise dictionary. The single noise dictionary size vary depending on the number of available noise-only sequences for each embedded recording. The adaptive noise modeling only evaluates the noise-only sequences that are extracted from the embedded recording which contains the target utterance. The number of noise exemplars extracted from each sequence varies between 95 and 195. The single speech dictionary contains 2354 full-word exemplars (maximum 50 exemplars from 51 classes) of 25 frames. The full-word exemplars are used in the single speech dictionary, as there is no exemplar length L_s containing a vast number of samples from each half-word class. The keyword recognition accuracy is used to evaluate the system performance on the CHIME-2 data.

5. RESULTS AND DISCUSSION

The recognition experiments performed on AURORA-2 data investigates the influence of the selection dictionary size, i.e. the noise repository size F and the number of exemplar extracted from each training sequence in the repository G, on the recognition performance. Choosing an SNR-dependent X best matching training sequences for the recognition is kept in the AURORA-2 experiments [2]. For CHIME-2 data, the selection dictionary is extracted from the noise-only segments of each embedded sequence which results in a fixed value of F. Hence, the CHIME-2 experiments investigates different settings of forming the noise dictionaries using the adaptive noise modeling approach and/or noise sniffing by varying X value. For this purpose, we compare the baseline recognizer using only sniffed exemplars with novel systems adopting adaptive noise modeling with and without the sniffed exemplars.

The performance of the adaptive noise modeling has been evaluated on both test sets of AURORA-2 data at the SNRs of -5, 0 and 5 dB and the results are presented in Table 1. The best results of the proposed setup are given in bold. The details of the other recognition systems can be found in [12]. In these recognition experiments, we compare the word error rates (WER) obtained using *adaptive* and *fixed* noise dictionaries. The experiments with adaptive dictionaries are performed varying F between 160 to 1200 and G between 5 to 15 exemplars per sequence. The results are given at the lower panel of Table 1a and 1b. In Table 1a, the recognition results obtained on test set A are shown. The baseline system using fixed dictionaries provides WERs of 47.1%, 21.2% and 9.3% at SNR level of -5, 0 and 5 dB respectively. The proposed adaptive noise modeling scheme with F=160 and G=5 training sequences reduces the WERs dramatically to 25.2%, 10.9% and 6.2% at the same SNR levels. For G=10, the WER reduces to 24.1% at -5 dB. G=10 is a reasonable choice as increasing G further brings no significant improvement. At SNR levels of 0 and 5 dB, G has a less noticeable impact on the recognition accuracy. Increasing F provides further improvements on the recognition accuracy with WERs of 20.3%, 17.9% and 17.5% for F equal to 480, 800 and 1200 at SNR of -5 dB. The recognition results follow a similar trend at SNRs of 0 and 5 dB. The lower panel of Table 1b presents the recognition results for test set B. The baseline system using fixed dictionaries provides WERs of 57.5%, 23.8% and 8.8% at SNR level of -5, 0 and 5 dB respectively. For the mismatched noise case, the selection technique still provides some improvement for any G and F which is explained by the increased spectral diversity of the available noise exemplars. Unlike the matched case, increasing G or F do not have a considerable impact on the recognition accuracy.

The recognition accuracies provided by the baseline and the proposed systems on the development and test set of CHIME-2 data are presented in Table 2a and Table 2b. The results on development and test sets follow a similar pattern, thus, we focus only on the test set results. The baseline system using 400 sniffed exemplars provides 69.3%, 76.8% and 84.5% at SNRs of -6, -3 and 0 dB. The recognition system using only adaptive dictionaries with X=3 provides comparable results with 69.8%, 76.5% and 83.9% at the same SNR levels. The mixed dictionaries obtained by combining 200 sniffed exemplars (SE) with adaptive noise dictionaries hav-

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

SNR(dB)	**-5**			**0**			**5**		
NREM (*Fixed*)	47.1			21.2			9.3		
GMM/HMM	60.8			24.3			7.3		
SC	35.2			13.8			7.4		
FE	30.4			10.7			3.3		
NREM (*Adpt.*)	G = 5	G = 10	G = 15	G = 5	G = 10	G = 15	G = 5	G = 10	G = 15
F = 160	25.2	24.1	23.5	11.0	10.8	10.5	6.2	5.8	6.1
F = 320	23.2	21.2	21.0	9.8	9.5	9.5	5.9	5.9	5.6
F = 480	21.6	20.3	20.0	10.1	9.4	9.8	5.8	5.6	5.5
F = 640	20.2	18.5	18.4	**9.1**	9.2	9.4	5.8	5.6	5.6
F = 800	19.9	17.9	18.0	9.5	8.7	9.3	5.8	5.6	5.6
F = 1200	**19.0**	**17.5**	**17.2**	9.3	**8.4**	**8.9**	**5.6**	**5.5**	**5.3**

(a) Test set A

SNR(dB)	**-5**			**0**			**5**		
NREM (*Fixed*)	57.5			23.8			8.8		
GMM/HMM	64.0			25.9			7.4		
SC	52.4			23.5			11.0		
FE	52.6			20.5			5.7		
NREM (*Adpt.*)	G = 5	G = 10	G = 15	G = 5	G = 10	G = 15	G = 5	G = 10	G = 15
F = 160	57.1	55.8	56.1	23.5	**23.1**	23.4	8.2	**8.0**	**8.2**
F = 320	55.6	56.2	55.9	23.4	**23.1**	23.5	8.2	8.4	8.8
F = 480	55.8	56.2	**55.7**	22.8	23.4	23.1	8.6	8.3	8.4
F = 640	**55.2**	56.4	**55.7**	22.8	**23.1**	23.0	8.2	8.3	8.7
F = 800	56.0	**55.7**	55.8	22.8	23.3	**22.7**	**7.9**	8.3	8.6
F = 1200	55.4	56.1	56.6	**22.1**	23.9	23.2	8.4	8.6	8.7

(b) Test set B

ing X=2 provide the best performance. This system provides 71.2%, 78.9% and 85.3% at SNRs of -6, -3 and 0 dB with an absolute improvement of 1.9%, 2.1% and 0.8% respectively. Another setup that gives promising results is the one using noise dictionaries with 300 SE and X=1. All setups using adaptive noise modeling provide comparable results at higher SNRs. The recognition results with higher X values are not reported as increasing X does not improve the results with an increased computational burden.

From these results, it can be concluded that the preliminary noise sequence selection technique benefits from the larger noise repository with a rather coarse sampling of the noise-only sequences in the repository. For AURORA-2 data, setting G=10 exemplar per sequence captures the within noise-only sequence variation well enough and larger G values do not improve the recognition accuracy. Finally, depending on the available memory, the noise repository size F can be increased further to have better coverage of the variation in background noise and hence improved performance. The experiments on CHIME-2 data shows that combining sniffed exemplars with the exemplars extracted from the selected sequences provides superior noise modeling compared to only sniffing similar amounts of noise exemplars. Furthermore, it has been shown that the best recognition performance at lower SNR levels is achieved using 350-450 mixed noise exemplars per dictionary. Increasing the amount of noise exemplars further does not bring any improvement. This upper bound on the recognition performance is explained by the poor speech modeling provided by the speech dictionaries due to the limited amount of training data.

6. CONCLUSION

This paper investigates the impact of several parameters of an exemplar-based adaptive noise modeling technique on the recognition accuracy. A non-negative sparse coding-based noise exemplar selection technique is described in the previous work that selects noise exemplars on-the-fly to be able to model the spectrotemporal content of the actual noise conditions. Using the optimal parameters, the final system with adaptive noise modeling uses less noise exemplars compared to the system using fixed dictionaries and provides better recognition accuracy on the AURORA-2 data. Moreover, the experiments on CHIME-2 data show that the mixed dictionaries containing sniffed and adaptively selected noise exemplars outperform the baseline using sniffed exemplars only. Overall, the proposed approach appears to be an effective noise dictionary design scheme that can be incorporated in exemplar-based ASR approaches.

Table 2: Recognition accuracies in percentages obtained on development and test set the CHIME-2 data - SE: Sniffed Exemplars

SNR(dB)	**-6**	**-3**	**0**	**3**	**6**	**9**
NREM (400SE)	69.4	76.4	85.0	90.1	92.9	93.3
GMM/HMM	49.3	58.6	67.5	75.0	78.8	82.9
SC	75.5	81.4	87.5	89.9	92.4	92.3
FE	68.0	72.2	80.9	86.7	89.0	90.5
NREM (*Adpt.*)	**-6**	**-3**	**0**	**3**	**6**	**9**
X = 1	64.8	71.3	80.6	86.8	90.9	92.4
X = 2	66.0	73.6	82.0	89.1	92.1	93.2
X = 3	69.1	76.3	84.6	89.3	92.3	93.5
X = 1 + 100SE	68.3	76.4	83.6	90.2	92.3	92.9
X = 2 + 100SE	67.0	73.9	83.0	**90.7**	92.3	93.4
X = 3 + 100SE	67.1	74.6	83.7	90.4	92.5	93.3
X = 1 + 200SE	65.7	73.9	83.6	90.3	92.4	93.3
X = 2 + 200SE	70.6	**78.0**	84.7	90.4	92.6	**93.8**
X = 1 + 300SE	**71.3**	77.8	**85.1**	90.3	**92.8**	93.6

(a) Development Set

SNR(dB)	**-6**	**-3**	**0**	**3**	**6**	**9**
NREM (400SE)	69.3	76.8	84.5	88.8	91.9	93.5
GMM/HMM	49.7	57.9	67.8	73.7	80.8	82.7
SC	76.5	81.3	88.9	90.5	92.7	93.2
FE	67.2	75.9	81.1	86.4	90.7	92.0
NREM (*Adpt.*)	**-6**	**-3**	**0**	**3**	**6**	**9**
X = 1	65.5	72.2	80.8	86.4	89.8	93.1
X = 2	68.4	75.3	83.6	87.8	90.3	92.8
X = 3	69.8	76.5	83.9	87.8	90.5	92.7
X = 1 + 100SE	69.5	74.9	**85.3**	88.7	91.9	93.3
X = 2 + 100SE	68.0	75.3	84.5	87.7	91.8	92.7
X = 3 + 100SE	67.7	75.3	84.0	87.5	91.0	92.6
X = 1 + 200SE	67.2	74.9	**85.3**	87.0	92.4	93.2
X = 2 + 200SE	**71.2**	**78.9**	**85.3**	88.7	91.9	92.8
X = 1 + 300SE	70.6	77.4	**85.3**	**88.8**	**92.6**	**93.4**

(b) Test set

REFERENCES

[1] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, D. Van Compernolle, K. Demuynck, J. F. Gemmeke, J. R. Bellegarda, and S. Sundaram, "Exemplar-based processing for speech recognition: An overview," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 98–113, Nov. 2012.

[2] 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.

[3] M. De Wachter, K. Demuynck, D. Van Compernolle, and P. Wambacq, "Data driven exemplar based continuous speech recognition," in *Proc. European Conf. on Speech Communication and Technology*, 2003, pp. 1133–1136.

[4] J.F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE TASLP*, vol. 19, no. 7, pp. 2067–2080, Sept. 2011.

[5] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in *Proc. INTERSPEECH*, Makuhari, Chiba, Japan, 2010, pp. 717–720.

[6] D. Kanevsky, T. Sainath, B. Ramabhadran, and D. Nahamoo, "An analysis of sparseness and regularization in exemplar-based methods for speech classification," in *Proc. INTERSPEECH*, Makuhari, Chiba, Japan, 2010, pp. 2842–2845.

[7] 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)*, pp. 1–4.

[8] 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.

[9] P.O. Hoyer, "Non-negative sparse coding," in *IEEE Workshop on Neural Networks for Signal Processing*, 2002, pp. 557–565.

[10] E. Yılmaz, J. F. Gemmeke, and H. Van hamme, "Exemplar selection techniques for sparse representations of speech using multiple dictionaries," in *Proc. EUSIPCO*, Marrakesh, Morocco, Sept. 2013, pp. 1–5.

[11] J. F. Gemmeke and T. Virtanen, "Artificial and online acquired noise dictionaries for noise robust ASR," in *Proc. INTERSPEECH*, 2010, pp. 2082–2085.

[12] 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.

[13] H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA Tutorial and Research Workshop ASR2000*, Sept. 2000, pp. 181–188.

[14] 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.