2012 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 23–26, 2012, SANTANDER, SPAIN

NOISE-ROBUST DIGIT RECOGNITION WITH EXEMPLAR-BASED SPARSE REPRESENTATIONS OF VARIABLE LENGTH

Emre Yılmaz, Jort F. Gemmeke, Dirk Van Compernolle, and Hugo Van hamme

Dept. ESAT, KU Leuven, Leuven, Belgium

ABSTRACT

This paper introduces an exemplar-based noise-robust digit recognition system in which noisy speech is modeled as a sparse linear combination of clean speech and noise exemplars. Exemplars are rigid long speech units of different lengths, i.e. no warping mechanism is used for exemplar matching to avoid poor time alignments that would otherwise be provoked by the noise and the natural duration distribution of each unit in the training data is preserved. Speech and noise separation is performed by applying non-negative sparse coding using a separate exemplar dictionary for each labeled unit (in this case half-digits) rather than a single dictionary of all units. This approach does not only provide better classification of speech units but also models the temporal structure of speech and noise more accurately. The system performance is evaluated on the AURORA-2 database. The results show that the proposed system performs significantly better than a comparable system using a single dictionary at positive SNR levels.

Index Terms— Exemplar-based recognition, noise robustness, non-negative sparse coding, multiple dictionaries

1. INTRODUCTION

Noise-robust speech recognition has been intensively researched for several decades [1]. Hidden Markov Models (HMM) which have been used as the standard acoustic modeling tools so far are known to perform poorly in case of non-stationary noise due to mismatches between the training and testing conditions. Several approaches such as robust feature extraction, signal and feature enhancement, model compensation and missing data techniques have been proposed to overcome the degradation in recognition accuracy, e.g. [2, 3]. Moreover, as an alternative to noise-robust HMM-based systems, a number of exemplar-based approaches have been proposed in the last years [4, 5, 6].

Exemplar-based recognition recently regained popularity due to the significant increase in computational power and development of fast template matching and search algorithms. In this approach, exemplars, which are labeled speech segments that have occurred in the training data, are compared with the input speech signals using some distance measure. One can trivially classify the segment as the label of the closest exemplar, or by a voting scheme on the set of K nearest neighbors [7, 8]. More recently, sparse linear combinations of exemplars have been used successfully for classification [9, 10].

In [7], the input speech signal is compared with a properly selected subset of a large collection of exemplars and the exemplar sequence with the minimal distance is searched. On the other hand, in [9], a linear combination of the same length exemplars is obtained to jointly approximate the input speech. The latter approach is further applied to noisy speech yielding promising results especially at lower SNRs [11].

In this paper, we propose an exemplar-based noise-robust recognition technique that combines the aforementioned techniques. The proposed technique approximates the input speech segments by linearly combining rigid long exemplars providing a better model for the temporal structure of speech and noise. It uses a separate dictionary for each length of each speech unit. The units are chosen to correspond to speech segments from which words are composed, e.g. phones, demi-syllables, syllables, subwords or even full words.

The idea of using multiple dictionaries to improve the naive non-negative spectrogram factorization is first mentioned in [12]. This approach differs from the proposed one as speech is modeled using a probabilistic model, i.e. non-negative hidden Markov models (N-HMM), and learned from training data whereas our approach uses actual speech segments extracted from training data. In this sense, the proposed approach performs pure exemplar-based recognition without any time warping to increase noise robustness. Furthermore, the natural duration distribution of each unit in the training data is preserved which is different from [11].

Non-negative sparse coding (NSC) is applied to determine the weights of the linear combination similar to [11]. From the geometrical interpretation of NSC-based source separation, it is known that the farther the convex hull of the basis vectors of different sources (speech and noise exemplars in this case) are, the better the separation is. This motivates us to use separate dictionaries for different units.

After obtaining the exemplar weights for each dictionary using a multiplicative update rule to minimize the regularized Kullback-Leibler divergence, the noisy speech is decoded by a search algorithm looking for the digit sequence with the minimum reconstruction error between the approximated and noisy segments.

The rest of the paper is organized as follows. The approach for finding the weights of exemplars is explained in Section 2. The implementation details and evaluation setup are discussed in Section 3. Section 4 presents the recognition results and comments on the system performance. In Section 5, the conclusions and thoughts for future work are discussed .

2. EXEMPLAR-BASED SPARSE REPRESENTATIONS

2.1. Noisy speech model

The noisy speech is modeled as a summation of speech and noise exemplars which are acoustic feature vectors with a known label. These exemplars are extracted from a training corpus and represented in linear mel-scaled magnitude spectra domain in order to ensure additivity of speech and noise exemplars.

Speech exemplars consisting of l frames are reshaped as a single column vector and collected in a dictionary $\mathbf{S}_{c,l}$ for each class c and

978-1-4673-1026-0/12/$31.00

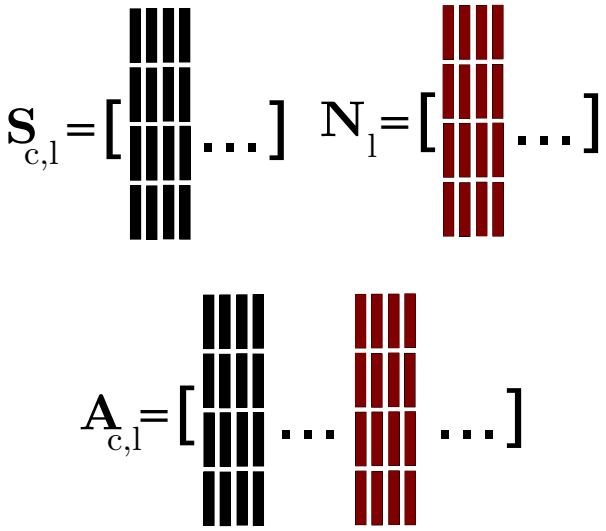

Fig. 1. A unique speech dictionary is formed for each length l of each unit c. Exemplars of different noise types are collected in a single dictionary for each length l. Each speech dictionary is concatenated with this noise dictionary.

each length l as shown in Figure 1. 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 single dictionary $\mathbf{A}_{c,l} = [\mathbf{S}_{c,l}\ \mathbf{N}_l]$ of dimensionality $Dl \times M_{c,l}$ where D is the number of frequency bands and $M_{c,l}$ is the total number of available speech and noise exemplars.

For any class c, a reshaped noisy speech vector $\mathbf{y}_l$ of length l is expressed as a linear combination of the exemplars with non-negative weights:

$$\mathbf{y}_l \approx \sum_{m=1}^{M_{c,l}} x_{c,l}^m \mathbf{a}_{c,l}^m = \mathbf{A}_{c,l}\mathbf{x}_{c,l} \qquad \text{s.t.} \qquad x_{c,l}^m \geq 0 \qquad (1)$$

where $\mathbf{x_{c,l}}$ is an $M_{c,l}$-dimensional sparse weight vector. Sparsity of the weight matrix implies that the noisy speech is approximated by a few speech and noise exemplars which is found to be crucial [6].

2.2. Obtaining the exemplar weights

The exemplar weights are obtained by minimizing the cost function,

$$d(\mathbf{y}_l, \mathbf{A}_{c,l}\mathbf{x}_{c,l}) + \sum_{m=1}^{M_{c,l}} x_{c,l}^m \Lambda_m \qquad \text{s.t.} \qquad x_{c,l}^m \geq 0 \qquad (2)$$

where $\mathbf{\Lambda}$ is an $M_{c,l}$-dimensional vector. The first term is the divergence between the noisy speech vector and its approximation. A regularization term is added in order to limit the l_1-norm of the weight vector. Here, $\mathbf{\Lambda}$ controls how sparse the resulting vector $\mathbf{x}$ is. Defining $\mathbf{\Lambda}$ as a vector, the amount of sparsity enforced on different types of exemplars can be adjusted. The generalized Kullback-Leibler divergence (KLD) is used for d:

$$d(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{k=1}^{K} y_k \log \frac{y_k}{\hat{y}_k} - y_k + \hat{y}_k \qquad (3)$$

which is commonly used in source separation problems and shown to produce better results than Euclidean distance when used in conjunction with linear mel-scaled spectra [13].

The multiplicative update rule to minimize the cost function (2) is derived in [6] and found as

$$\mathbf{x}_{c,l} \leftarrow \mathbf{x}_{c,l} \odot (\mathbf{A}_{c,l}^T(\mathbf{y}_l \oslash (\mathbf{A}_{c,l}\mathbf{x}_{c,l}))) \oslash (\mathbf{A}_{c,l}^T\mathbf{1} + \mathbf{\Lambda}) \qquad (4)$$

with $\odot$ and $\oslash$ denoting element-wise multiplication and division respectively. $\mathbf{1}$ is a Dl-dimensional vector with all elements equal to unity. Applying this update rule iteratively, the weight vector becomes sparse and the reconstruction error between the noisy speech vector and its approximation decreases monotonically.

2.3. Decoding the noisy speech

Equation (3) expresses the reconstruction error between a speech segment of length l and a class c. It satisfies the conditions to apply dynamic programming, so it is trivial to find the sequence of classes that best matches a sequence of input vectors.

Every noisy frame sequence of each available exemplar length is approximated as a linear combination of exemplars by iteratively applying the update formula. For each class and exemplar length, the approximation is performed separately using the dictionaries discussed in Section 2.1. After a certain number of iterations, the reconstruction error is calculated using Equation (3). As every dictionary contains exemplars with known labels, the entire noisy utterance is searched to find the digit sequence yielding the minimum reconstruction error.

2.4. Scoring silence dictionaries

A known problem of sparse combination approaches working on magnitude spectra is that the silence exemplars are hard to recognize: perfect silence is modeled with zero weights of all exemplars [6]. In a practical noisy mixture, it is well-approximated by combining speech and noise exemplars with small weights, so all classes will score equally well. To overcome this problem, reconstruction errors for the class representing silence have to be compensated. For this purpose, we use the exemplar representation framework to construct a voice activity detector (VAD) in order to predict whether a noisy segment contains speech. Choosing an exemplar length L_s containing abundant samples from each class, we form a single dictionary by concatenating all speech exemplars from different classes plus noise exemplars. After obtaining the exemplar weights for every noisy segment of length L_s, we reconstruct the speech and noise components. For each frequency bin, the ratio of speech and noise magnitude is calculated and averaged over the frequency bins yielding a number used as the voice activity estimate ($0 \leq \text{VAD} \leq 1$).

Since the decoding is done based on the reconstruction error, its dynamic range obtained for each class is highly SNR dependent. This is expected as the same noise dictionary is concatenated to all speech dictionaries of the same length. For high SNRs, the weights are mostly distributed among the speech exemplars yielding a high dynamic range in the reconstruction error. On the contrary, for low SNRs, the weights are mostly distributed among the noise exemplars leading to very close approximations. To avoid overcompensation of the reconstruction errors at lower SNRs, we propose an SNR-dependent compensation factor. SNR estimation is also performed by the single dictionary system used for voice activity detection. The SNR is estimated as the ratio of total speech weights to total noise

weights [6].

$$SNR = \frac{\sum_{w=1}^{W} \sum_{m=1}^{J} x_{L_s}^{w,m}}{\sum_{w=1}^{W} \sum_{m=J+1}^{M_s} x_{L_s}^{w,m}} \quad (5)$$

where $\mathbf{x}_{L_s}^{w}$ is the sparse weight vector corresponding to w^{th} of W noisy segments of length L_s. J is the number of speech exemplars and M_s is number of all exemplars.

The reconstruction errors corresponding to the silence dictionaries are reduced by a value CF depending on the voice activity value assigned to the middle frame of the corresponding noisy segment, the SNR estimate and the reconstruction error itself,

$$CF = d(\mathbf{y}_l, \mathbf{A}_{sil,l}\mathbf{x}_{sil,l}) \cdot min(max(SNR \cdot \alpha, \beta), \gamma) \cdot VAD \quad (6)$$

where α is a scale factor, β and γ are lower and upper limits of the SNR estimate respectively and VAD is the voice activity estimate. It should be noted that including the reconstruction error itself compensates for length differences.

3. IMPLEMENTATION DETAILS AND EXPERIMENTAL SETUP

3.1. Exemplar extraction and dictionary creation

The exemplars used in this system are speech segments extracted from the clean training set of AURORA-2 database [14] which contains 8440 utterances with one to seven digits in American English. Acoustic feature vectors are represented in mel-scaled magnitude spectra with 19 frequency bands. The speech exemplars representing half-digits are segmented by a conventional HMM-based system. There are in total 50,654 speech exemplars including 1300 silence exemplars. The minimum and maximum exemplar lengths are 5 and 30 frames respectively. Exemplars longer than 30 frames are omitted to limit the number of dictionaries.

Noise exemplars are obtained by removing the speech from 16 noisy utterances in the multi-condition training set. 120 exemplars per noise type (in total 480 noise exemplars) are randomly extracted for all possible speech exemplar lengths. Noise modeling is kept modest to avoid long simulation times. After concatenating the different speech and noise dictionaries, the system ends up containing 508 dictionaries of 23 different classes (half-digits plus silence).

3.2. Implementation details

The whole system is implemented in MATLAB and we used GPUs to accelerate the evaluation of Equation (4). The exemplar-based VAD system employs exemplars with $L_s = 17$. In that system, the single dictionary is constructed by concatenating 3489 speech exemplars and 3200 noise exemplars which are extracted from the same 16 noisy utterances. The multiplicative update rule is iterated 200 times to find the exemplar weights. Elements of $\mathbf{\Lambda}$ corresponding to speech exemplars are set to 1.8, and the ones corresponding to noise exemplars are set to 1.6. These values were tuned manually by comparing the voice activity estimates on a randomly selected set of clean utterances.

Recognition is performed using all 508 dictionaries. The reconstruction error shows enough discrimination among classes after only 50 iterations. This is an advantage over e.g. enhancement approaches where the backend models require more accuracy of the reconstructed spectra, and hence more iterations, in order to perform well. Dictionaries are iteratively normalized so that the l_2-norm of the columns are equal to unity and the l_2-norm of the rows are approximately equal. For each length, the same row scaling is applied to the reshaped noisy speech vectors. Elements of $\mathbf{\Lambda}$ corresponding to speech exemplars are set to 0.45 and the ones corresponding to noise exemplars are set to 0.3. α, β and γ values are set to 0.25, 0.02 and 0.3 respectively. These values were tuned on a randomly selected subset of multi-condition training set for maximum recognition accuracy.

Table 1. Word error rates obtained on test set A using the proposed method and the best results of the baseline system using exemplars of length $L = 10$ and $L = 30$

SNR(dB)	20	15	10	5	0	-5	Av_{20-0}
Subway	3.4	4.1	6.8	11.4	28.1	57.1	10.8
Babble	2.4	3.6	6.1	13.4	35.3	70.8	12.2
Car	2.7	3.5	5.8	8.9	28.4	62.3	9.9
Exhibition	3.1	4.9	7.2	13.6	30.1	61.5	11.8
Av.	**2.9**	**4.0**	**6.5**	**11.8**	**30.5**	**62.9**	**11.2**
L = 10	**6.2**	**7.3**	**9.8**	**16.2**	30.5	59.0	**14.0**
L = 30	11.6	12.0	14.5	17.4	**25.1**	**44.2**	16.1

Table 2. Word error rates obtained on test set B using the proposed method and the best results of the baseline system using exemplars of length $L = 10$ and $L = 30$

SNR(dB)	20	15	10	5	0	-5	Av_{20-0}
Restaurant	3.7	4.6	8.8	19.6	46.3	77.2	16.6
Street	2.4	4.7	9.2	21.2	45.0	79.8	16.5
Airport	3.0	4.1	7.1	17.7	42.0	75.6	14.8
Station	3.6	4.7	9.6	20.1	43.1	73.1	16.2
Av.	**3.2**	**4.5**	**8.7**	**19.7**	**44.1**	**76.4**	**16.0**
L = 10	**6.3**	**9.6**	**15.4**	**26.5**	49.4	78.8	**21.5**
L = 30	12.8	14.8	19.6	28.2	**45.2**	**67.6**	24.1

3.3. Experimental setup

The recognition accuracy of the proposed method is evaluated on the test set A and B of the same database. Test set A consists of 4 clean and 24 noisy datasets with four noise types (subway, babble, car and exhibition). The noise types of this test set match the multi-condition training. Test set B has the same number of test sets with four different noise types (restaurant, street, airport, station). To reduce the simulation times, we subsampled the test sets by a factor of 4 (250 utterances per test set).

4. RESULTS AND DISCUSSION

To accurately evaluate the improvement gained due to the multiple dictionary framework, the recognition results obtained with the proposed system are compared with the results published in [11]. This baseline system uses a dictionary of speech and noise exemplars of fixed length to produce HMM state likelihoods. The noisy speech is then decoded from these likelihoods using Viterbi decoding. The proposed method differs from this baseline in the use of exemplars of variable length organized in multiple dictionaries. Also, the proposed method does not require an HMM assumption but uses the exemplars directly to match the noisy data. Although more recently,

Table 3. Word error rates obtained on clean test set using the proposed method and the best results of the baseline system using exemplars of length $L = 10$ and $L = 30$

	Clean
Proposed method	**2.9**
L = 10	**4.5**
L = 30	10.5

advances to the framework presented in [11] have been reported on the same database, the results reported in [11] most closely resembles the proposed approach, in part due to the similar approach towards silence scoring.

Table 1 presents the word error rates (WER) obtained with the proposed technique on test set A and the best results obtained with the baseline system. For lower SNRs, the baseline system using exemplars of length L=30 performs better than L=10. The best results at each SNR level are given in bold. It can be seen that the new approach has significantly lower WERs at SNRs larger than 0 dB. There is a significant increase in the WER at SNRs 0 dB and -5 dB. This can be explained with the poor performance of the VAD at lower SNRs and the use of limited amount of noise exemplars. The proposed method has an average WER of 11.2% over the SNR range 20-0 dB, which is better than the baseline results. A similar improvement at positive SNRs is also observed for the mismatched noise case in Table 2.

Table 3 gives the results on clean test set. We see that the new approach achieves the same WER on clean speech and at SNR 20 dB. This is due to the SNR thresholding in the silence compensation factor which does not often distinguish between the SNR estimates of clean utterances and noisy utterances at SNR 20 dB. The proposed approach still achieves a better recognition accuracy with a WER of 2.9% compared to 4.5%.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed an exemplar-based noise-robust digit recognizer which models noisy speech as a sparse linear combination of speech and noise exemplars. These exemplars are rigid long speech units of different lengths, i.e. no time warping is applied to avoid erroneous time alignments due to background noise. Furthermore, the natural length distribution of each speech unit in the training data is preserved to have a better temporal structure modeling.

The exemplar weights are obtained by applying non-negative sparse coding using separate dictionaries for each unit. After obtaining the sparse weights, we apply exemplar matching to find the digit sequence yielding the minimum reconstruction error. The proposed system achieved better recognition accuracy at positive SNRs compared to a system using exemplars of the same length in a single dictionary. Future work includes developing a better silence scoring algorithm, using adaptive dictionaries and investigation of different distance measures in other feature domains.

6. ACKNOWLEDGEMENTS

Emre Yılmaz is also affiliated with IBBT (Interdisciplinary Institute for Broadband Technology). This work is funded by the IMPact program (BATS) and IWT-SBO project 100049 (ALADIN).

7. REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, no. 3, pp. 261 – 291, 1995.

[2] M.J.F. Gales and S.J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 352 –359, Sept. 1996.

[3] B. Raj and R.M. Stern, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101 – 116, Sept. 2005.

[4] J. Ming, R. Srinivasan, and D. Crookes, "A corpus-based approach to speech enhancement from nonstationary noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 822 –836, May 2011.

[5] A. Hurmalainen, J.F. Gemmeke, and T. Virtanen, "Non-negative matrix deconvolution in noise robust speech recognition," in *Proc. ICASSP*, May 2011, pp. 4588 –4591.

[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] M. De Wachter, M. Matton, K. Demuynck, P. Wambacq, R. Cools, and D. Van Compernolle, "Template-based continuous speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1377 –1390, May 2007.

[8] 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.

[9] J.F. Gemmeke, L. ten Bosch, L. Boves, and B. Cranen, "Using sparse representations for exemplar based continuous digit recognition," in *Proc. EUSIPCO*, Glasgow, Scotland, August 24–28 2009, pp. 1755–1759.

[10] 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.

[11] J.F. Gemmeke and T. Virtanen, "Noise robust exemplar-based connected digit recognition," in *Proc. ICASSP*, March 2010, pp. 4546 –4549.

[12] G.J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *Proc. ICASSP*, May 2011, pp. 17 –20.

[13] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066 –1074, March 2007.

[14] H. 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.