EXEMPLAR SELECTION TECHNIQUES FOR SPARSE REPRESENTATIONS OF SPEECH USING MULTIPLE DICTIONARIES

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

Dept. ESAT, KU Leuven, Leuven, Belgium

ABSTRACT

This paper describes and analyzes several exemplar selection techniques to reduce the number of exemplars that are used in a recently proposed sparse representations-based speech recognition system. Exemplars are labeled acoustic realizations of different durations which are extracted from the training data. For practical reasons, they are organized in multiple undercomplete dictionaries, each containing exemplars of a certain speech unit. Using these dictionaries, the input speech segments are modeled as a sparse linear combination of exemplars. The improved recognition accuracy with respect to a system using fixed-length exemplars in a single dictionary comes with a heavy computational burden. Due to this fact, we investigate the performance of various exemplar selection techniques that reduce the number of exemplars according to different criteria and discuss the links between the salience of the exemplars and the data geometry. The pruned dictionaries using only 30% of the exemplars have been shown to achieve comparable recognition accuracies to what can be obtained with the complete dictionaries.

***Index Terms*—** Exemplar selection, exemplar-based speech recognition, sparse representations, dictionary pruning

1. INTRODUCTION

The success of recently proposed speech recognition systems based on template matching attracted considerable interest in exemplar-based acoustic modeling as a viable alternative to its statistical counterparts [1, 2]. Exemplars are labeled speech segments such as phones, syllables or words, possibly of different length, that are extracted from the training data. Each exemplar is tagged with meta-information including speaker and environmental characteristics. An input speech segment can simply be classified by evaluating the labels of the spatially closest exemplars. Inconsistent exemplar sequences, e.g. sequences with different gender exemplars, can be penalized based on the tagged meta-information.

Although exemplars provide better duration and trajectory modeling compared to the Hidden Markov Models, large amounts of data are required to handle the acoustic variation among different utterances [1]. In order to reduce high memory and computational power requirements, several exemplar selection algorithms are proposed in [3, 4]. The main goal of these techniques is to remove less informative exemplars that are hardly used or whose presence result in inaccurate recognition and achieve comparable recognition accuracies using only a portion of the exemplars.

Another framework in exemplar-based techniques, namely exemplar-based sparse representations (SR), models the spectrogram of input speech segments as a sparse linear combination of exemplars rather than comparing with each individual exemplar. SR-based techniques have been successfully used for speech enhancement [5], feature extraction [6] and clean [7] and noisy [8, 9, 10] speech recognition. In these approaches, fixed-size exemplars are stored in the columns of an overcomplete dictionary which has much higher number of columns (exemplars) than rows (time-frequency cells). We have recently proposed an SR-based speech recognition system which uses exemplars of different length organized in separate dictionaries and which approximates the input speech as a linear combination of the exemplars in each dictionary [11]. Most of these dictionaries are undercomplete having less exemplars than the number of time-frequency cells. We have also shown that this system performs reasonably well under noisy conditions in [12].

Reducing the dimensions of large datasets stored in overcomplete dictionaries has been investigated in different fields and several matrix decompositions such as the singular value decomposition (SVD), rank revealing QR decomposition, CUR matrix decomposition, interpolative decomposition (ID) have been used to obtain a low-rank matrix approximation of the complete data matrix [13]. Though the SVD is known to provide the best rank-k approximation, interpretation of the principal components is difficult in data analysis [14]. Therefore, several CUR matrix decompositions have been proposed in which a matrix is decomposed as a product of three matrices $\mathbf{C}$, $\mathbf{U}$, $\mathbf{R}$ and the matrices $\mathbf{C}$ and $\mathbf{R}$ consist of a subset of the actual columns and rows respectively [15, 16, 17]. Moreover, a probabilistic ID technique which automatically handles the model selection is introduced and applied to polyphonic music transcription task using an overcomplete dictionary containing exemplars of different musical notes in [18].

The exemplar selection techniques proposed in this paper differ from previous work as the dictionaries, which only contain exemplars of the same length and label, are undercomplete due to insufficient training data. Compared to the overcomplete dictionaries with a large number of data points, the redundancy in undercomplete dictionaries is quite limited. Therefore, removing a few highly relevant data points may already result in significant decreases in the recognition accuracy. The use of real exemplars tagged with meta-information is another requirement which prevents applying the SVD or any clustering technique. To the best of our knowledge, there is no prior work on selecting the most salient columns of an undercomplete dictionary. In this paper, we propose various techniques for selecting the most informative columns of the undercomplete dictionaries and analyze the selection problem elaborating on the geometrical structure of the data.

The rest of the paper is organized as follows. A brief description of the sparse representations-based speech recognition system is given in Section 2. The proposed exemplar selection techniques are discussed in Section 3. Section 4 explains the experimental setup and implementation details. In Section 5, we present the recognition results and a general discussion on the proposed techniques is given. The conclusions and thoughts for future work are discussed in Section 6.

2. SYSTEM DESCRIPTION

The recognition system that is described in [11] uses a sparse linear combination of the exemplars to model the input speech segments. Each exemplar is associated with a certain speech unit and the duration of each speech unit in the training data is preserved yielding exemplars of different lengths.

Exemplars spanning l frames are reshaped into a single vector and stored in the columns of the dictionary $\mathbf{S}_{c,l}$: one for each speech unit c and each length l. Each dictionary is of dimensionality $Dl \times N_{c,l}$ where D is the number of frequency bands in a frame and $N_{c,l}$ is the number of available exemplars of length l and class c. For any class c, a reshaped input speech vector $\mathbf{y}_l$ of length Dl is expressed as a linear combination of the exemplars with non-negative weights:

$$\mathbf{y}_l \approx \sum_{m=1}^{N_{c,l}} x_{c,l}^m \mathbf{s}_{c,l}^m = \mathbf{S}_{c,l}\mathbf{x}_{c,l} \qquad \text{s.t.} \qquad x_{c,l}^m \geq 0 \tag{1}$$

where $\mathbf{x_{c,l}}$ is an $N_{c,l}$-dimensional sparse weight vector. Sparsity of the weight matrix implies that the input speech is approximated by a small number of exemplars. The exemplar weights are obtained by minimizing the cost function,

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

where Λ is a scalar which controls how sparse the resulting vector $\mathbf{x}$ is. The first term is the divergence measure between the input speech 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. The generalized Kullback-Leibler divergence (KLD) is used for d:

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

The regularized convex optimization problem can be solved using various methods including non-negative sparse coding (NSC). For NSC, the multiplicative update rule to minimize the cost function (2) is derived in [19] and is given by

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

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 sparser and the reconstruction error between the input speech vector and its approximation decreases monotonically.

The first term of Equation (2) expresses the reconstruction error between a speech segment of length l and a class c. Every speech segment of each available exemplar length is approximated as a linear combination of exemplars. This is achieved by applying the sliding window [19] to the input utterance for each available exemplar length and iteratively applying equation (4) using the dictionaries of the corresponding length. After a fixed number of iterations, the reconstruction error is calculated. As the label of each dictionary is known, decoding is performed by finding the class sequence that minimizes the reconstruction error using dynamic programming.

3. EXEMPLAR SELECTION TECHNIQUES

The computational bottleneck of the system described above is the evaluation of Equation (4). The computational complexity per iteration is linearly proportional to the number of exemplars and it can be reduced by removing the less informative and redundant exemplars that are either not used or result in misclassifications. The baseline column selection technique is the randomized column selection algorithm which is proposed as a part of the CUR matrix decomposition in [14]. This algorithm randomly selects a subset of the columns of a data matrix with respect to the probability distribution computed as the normalized statistical leverage scores. Preferably selecting high-statistical leverage columns will, with high probability, lead to a reduced dictionary which approximates the original one almost as well as an SVD-based rank reduction scheme [14].

In this section, we propose several exemplar selection techniques that reduce the number of exemplars stored in the dictionaries discussed in Section 2. These techniques are classified into three categories, namely reconstruction error-based, distance-based and activation-based according to their exemplar selection criteria.

3.1. Reconstruction error-based techniques

The system described in Section 2 approximates input segments as a linear combination of exemplars. Since the approximation quality is measured using the divergence measure in Equation (3), the approximation of an exemplar either using other exemplars in the same-class dictionary or the ones in different-class dictionaries of the same length provides useful information about its salience.

3.1.1. Collinearity reduction (CR)

Exemplars that are well approximated by the other exemplars from the same-class dictionary contain less information compared to the ones with higher reconstruction errors. Therefore, the collinearity reduction technique removes the exemplars that are well approximated as a linear combination of the other exemplars in the same-class dictionary. This idea is applied iteratively by removing the exemplar that is approximated with the minimum reconstruction error at each iteration until the minimum number of exemplars requirement in a dictionary is met.

3.1.2. Discriminative dictionaries (DD)

Dictionary elements of a particular class that are well approximated by a dictionary of another class are likely to cause confusion during recognition. Indeed, any data that is close to these elements may be explained as belonging to the wrong class. The discriminative dictionaries technique iteratively removes the exemplars having the smallest ratio between the reconstruction errors that are obtained using the same-class dictionary and dictionary containing the exemplars of the other classes.

3.2. Distance-based techniques

Distance-based techniques perform exemplar selection considering the spatial closeness of the exemplars which provides information about the data geometry. The symmetric KLD is used as a distance metric which is defined as

$$d_{skld}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{2}(d(\mathbf{y}, \hat{\mathbf{y}}) + d(\hat{\mathbf{y}}, \mathbf{y})) \tag{5}$$

where d is defined in Equation (3).

3.2.1. Removing exemplars with the smallest/largest average distance (SAD/LAD)

Removing the same-class exemplars that either lie in the densely or sparsely populated regions in the feature space has been investigated. This technique retains the exemplars having either the smallest or the largest average distance to the other exemplars stored in the same-class dictionary.

3.2.2. Pruning the closest exemplars (CE)

The second distance-based technique aims to reduce the number of exemplars by discarding one of the exemplars that lie close to each other. At each iteration, the two closest exemplars are found and only one of them is retained in the dictionary.

3.3. Activation-based technique

3.3.1. Active exemplars (AE)

A single activation-based technique is proposed which infers the salience of an exemplar by evaluating the average weight it gets on a recognition task. The exemplar weights in the described system are obtained by applying the multiplicative update rule in Equation (4). Obviously, the exemplars often having higher weights are more decisive in the recognition. Thus, less *active* exemplars are rarely used and they can be removed from the dictionary without a significant loss in the recognition accuracy. The training data is used to quantify how active each exemplar is.

4. EXPERIMENTAL SETUP

4.1. Database

The exemplars used in experiments are speech segments extracted from the clean training set of AURORA-2 database [20] which contains 8440 utterances with one to seven digits in American English. The performance of the proposed exemplar selection techniques is evaluated on the clean test sets of the same database. There are 4 clean test sets, each containing 1001 utterances and recognition experiments are performed on these test sets using the pruned dictionaries.

4.2. Baseline System

Exemplars and input speech segments are represented in root-compressed (with magnitude power = 0.66) mel-scaled magnitude spectra with 17 frequency bands. The frame length is 32 ms and the frame shift is 10 ms. The training data is segmented into the exemplars representing half-digits by a conventional HMM-based recognizer. The system uses 508 dictionaries belonging to 23 different classes. The largest number of exemplars in a dictionary is 283. The minimum and maximum exemplar lengths are 5 and 30 frames respectively. Exemplars longer than 30 frames are removed to limit the number of dictionaries. The baseline system uses 50,654 exemplars in total including 1300 silence exemplars. The l_2-norm of each dictionary column and reshaped input speech vectors are normalized to unity. Further details about the baseline system can be found in [11].

4.3. Implementation of the Proposed Techniques

All of the proposed techniques are applied before the recognition experiments to create the pruned dictionaries. Reconstruction error and activation-based techniques require the evaluation of the multiplicative update rule given in Equation (4) in order to obtain the exemplar weights. The CR and DD techniques are applied iteratively discarding a single exemplar at each step. The AE technique, on the other hand, stores the average weight each exemplar gets during the approximation of the speech segments from the training data and the exemplar selection is performed by preserving the required number of exemplars with the highest average weight value. Distance-based techniques use a square and symmetric distance matrix to identify the spatial closeness of the exemplars. The CE technique iteratively reduces the number of exemplars while the DP and SP techniques are applied in a single step. The recognition accuracies presented in the following section are obtained by reducing the number of exemplars in each dictionary by 10% at each step until only 10% of the exemplars remain in each dictionary.

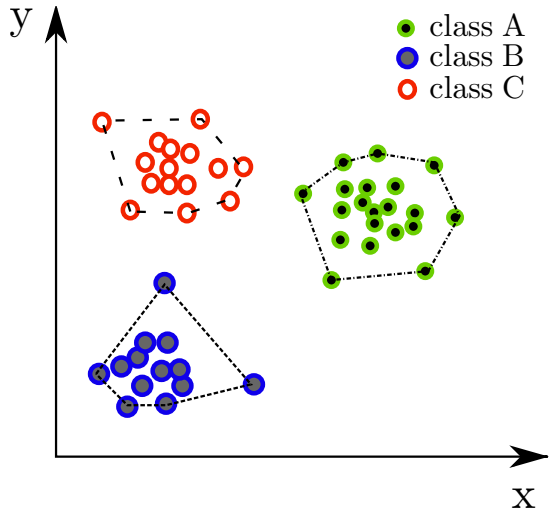

Fig. 1. Illustration of the convex hulls formed by the same class exemplars in two dimensions.

5. RESULTS AND DISCUSSION

In this section, we present the word error rates (WER) that are obtained on the clean test set of AURORA-2 using the dictionaries pruned with the techniques discussed in Section 3. These results are compared with the recognition accuracies obtained with the complete dictionaries and the dictionaries that are pruned with the randomized column selection algorithm of the CUR matrix decomposition. The recognition experiments on clean data provide information about both the performance of the proposed exemplar selection techniques and the size of the smallest dictionaries that sufficiently represent clean speech as a design parameter. It is worth mentioning that basic HMM/GMM systems provide higher recognition accuracies (about a percent) on the clean test set compared to the baseline recognizer using the complete dictionaries [20]. However, unlike this framework, it is not easy to account for background noise in HMMs [12].

Table 1 presents the WER results. The baseline system with the complete dictionaries has a WER of 1.68%. For each technique, the smallest dictionary size in which the WER has increased less than 10% (i.e. 1.68% * 1.1 = 1.85%) over the baseline is given in bold. The dictionaries pruned with collinearity reduction (CR) and active exemplars (AE) provide results lying in this error bound using 30%

Table 1. Average word error rates obtained on four clean test sets using the complete and pruned dictionaries. First row provides the result obtained using the complete dictionaries.

Removed exemplars (%)	# of exemplars	CR	AE	SAD	CE	DD	LAD	CUR
0	50654	1.68	1.68	1.68	1.68	1.68	1.68	1.68
10	45968	1.66	1.67	1.67	1.72	2.05	2.30	1.67
20	40858	1.73	1.69	1.69	1.76	4.43	2.71	1.57
30	35793	1.79	1.76	1.69	1.78	2.73	3.10	1.51
40	30687	1.75	1.73	1.69	1.81	2.99	3.62	**1.73**
50	25531	1.76	1.75	1.69	**1.78**	3.41	4.15	1.97
60	20533	1.76	1.77	**1.79**	1.92	3.86	4.51	2.01
70	15468	**1.79**	**1.84**	2.08	2.01	4.29	4.90	2.10
80	10362	1.91	2.30	2.19	2.14	5.27	5.92	2.50
90	5293	2.28	4.66	3.80	2.58	6.87	6.77	3.18

Table 2. Average word error rates obtained on four clean test sets using the DD and LAD techniques for outlier removal.

Removed exemplars (%)	# of exemplars	DD	LAD
0	50654	1.68	1.68
1	50568	1.76	1.72
2	50045	1.83	1.82
3	49544	1.85	1.95
4	49018	1.85	2.06

of the exemplars. Removing the exemplars with the smallest average distance (SAD) and pruning the closest exemplars (CE) performs slightly worse than the CR and AE staying in the bound using 40% and 50% of the exemplars respectively. The CUR decomposition gives similar WERs using more than 50% of the exemplars. The simulation times of the final system using 30% of the exemplars are reduced by a factor of 3, varying from 2.8 to 4 seconds depending on the utterance duration.

The hypothetically appealing idea of obtaining more discriminative dictionaries (DD) and removing the exemplars with the largest average distance (LAD) do not work for the intended task. Even after removing 10% of the exemplars, the WER exceeds 2%. The results obtained with these techniques imply that the spatial position of a data point provides some clues about how informative it is in the recognition. Due to the non-negativity of the data, each dictionary forms a convex hull that lies in the positive orthant. There are a few exemplars that lie on or next to the boundaries whereas the center is densely populated. A two dimensional illustration of the ideal (perfectly separable) case with three different classes is given in Figure 1. Considering the exemplar selection criteria of the LAD, it is apparent that it mainly discards exemplars that are further away from the densely populated region in the convex hulls. Similarly, the DD aims to reduce the confusions between the dictionaries and these confusions are mostly due to the exemplar lying on the boundaries in each convex hull. Removing these exemplars results in narrower convex hulls spanned by each dictionary which provides a less accurate description of the cone. On the other hand, other techniques retaining the exemplars lying on the boundaries and preserving the convex hull formed by each dictionary performs significantly better than the DD and LAD. It should be noted that most active exemplars typically lie on the convex hull boundaries which are rather decisive in the recognition.

Although the importance of the exemplars lying on the boundaries for the recognition accuracy has been shown, it can still be claimed that some of these exemplars can be outliers resulting in misclassifications. A discussion on the misclassifications due to the outliers in a convex hull can be found in [21]. To analyze the impact of the outliers on the recognition accuracy, we further apply the DD and LAD to remove a few percent of the exemplars. From the results in Table 2, it is not evident that these techniques work for outlier removal either. This can be either due to the non-existence of outliers in most dictionaries or their negligible impact on the recognition accuracy.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed several exemplar selection techniques for undercomplete dictionaries and analyzed which exemplars these techniques tend to select considering the geometrical structure formed by the data points in the feature space. Techniques based on the collinearity reduction (CR) and selecting the active exemplars (AE) provided the best results by achieving recognition accuracies that are in the 10% error bound of the baseline results using only 30% of the exemplars. The distance-based techniques, namely removing exemplars with the smallest average distance (SAD) and pruning the closest exemplars (CE), perform slightly worse than the CR and AE. All of these techniques outperform the CUR decomposition which has been successfully used for reducing the size of overcomplete dictionaries.

Discriminative dictionaries (DD) and removing the exemplars with the largest average distance (LAD) provides inferior results revealing the connection between the spatial position of an exemplar and its salience in the recognition. The DD and LAD mostly discard exemplars lying on the boundaries of the convex hulls resulting in a less accurate description of the cone. On the other hand, the SAD and CE explicitly remove the exemplars lying in the densely populated region of the convex hulls without deforming the boundaries and provide much better results than the DD and LAD. Hence, it can be concluded that the exemplars lying on the boundaries of the convex hulls are highly informative and discarding these exemplars

results in high recognition accuracy loss.

Future work includes reducing the redundancy in rows (frequency bands) and designing an efficient implementation of the described system providing faster recognition.

7. ACKNOWLEDGEMENTS

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

8. REFERENCES

[1] 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.

[2] 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.

[3] D. Seppi and D. Van Compernolle, "Data pruning for template-based automatic speech recognition," in *Proc. INTERSPEECH*, Makuhari, Chiba, Japan, Sept. 2010, pp. 985–988.

[4] X. Sun and Y. Zhao, "New methods for template selection and compression in continuous speech recognition," in *Proc. INTERSPEECH*, Florence, Italy, Aug. 2011, pp. 985–988.

[5] J. F. Gemmeke, T. Virtanen, and A. Hurmaleinen, "Exemplar-based speech enhancement and its application to noise-robust automatic speech recognition," in *International Workshop on Machine Listening in Multisource Environments*, Sept. 2011, pp. 53–75.

[6] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, and A. Sethy, "Sparse representations features for speech recognition," in *Proc. INTERSPEECH*, Sept. 2010, pp. 2254–2257.

[7] 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.

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

[9] 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.

[10] Q. F. Tan and S. S. Narayanan, "Novel variations of group sparse regularization techniques with applications to noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1337 –1346, May 2012.

[11] E. Yılmaz, D. Van Compernolle, and H. Van hamme, "Combining exemplar-based matching and exemplar-based sparse representations of speech," in *Symposium on Machine Learning in Speech and Language Processing (MLSLP)*, Portland, OR, USA, Sept. 2012.

[12] 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.

[13] N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Review*, vol. 53, no. 2, pp. 217–288, 2011.

[14] M. W. Mahoney and P. Drineas, "CUR matrix decompositions for improved data analysis," *Proceedings of the National Academy of Sciences*, vol. 106, no. 3, pp. 697–702, 2009.

[15] S. A. Goreinov, E. E. Tyrtyshnikov, and N. L. Zamarashkin, "A theory of pseudoskeleton approximations," *Linear Algebra and its Applications*, vol. 261, no. 13, pp. 1–21, 1997.

[16] A. Frieze, R. Kannan, and S. Vempala, "Fast Monte-Carlo algorithms for finding low-rank approximations," *J. ACM*, vol. 51, no. 6, pp. 1025–1041, Nov. 2004.

[17] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, "Relative-error CUR matrix decompositions," *SIAM J. Matrix Anal. Appl.*, vol. 30, no. 2, pp. 844–881, Sept. 2008.

[18] I. Ari, A. T. Cemgil, and L. Akarun, "Probabilistic interpolative decomposition," in *IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, Santander, Spain, Sept. 2012.

[19] 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.

[20] 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.

[21] T. N. Sainath, D. Nahamoo, D. Kanevsky, B. Ramabhadran, and P. Shah, "A convex hull approach to sparse representations for exemplar-based speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Hawaii, USA, Dec. 2011, pp. 59 –64.