NOISE-ROBUST SPEECH RECOGNITION WITH EXEMPLAR-BASED SPARSE REPRESENTATIONS USING ALPHA-BETA DIVERGENCE

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

Dept. ESAT, KU Leuven, Leuven, Belgium

ABSTRACT

In this paper, we investigate the performance of a noise-robust sparse representations (SR)-based recognizer using the Alpha-Beta (AB)-divergence to compare the noisy speech segments and exemplars. The baseline recognizer, which approximates noisy speech segments as a linear combination of speech and noise exemplars of variable length, uses the generalized Kullback-Leibler divergence to quantify the approximation quality. Incorporating a reconstruction error-based back-end, the recognition performance highly depends on the congruence of the divergence measure and used speech features. Having two tuning parameters, namely α and β, the AB-divergence provides improved robustness against background noise and outliers. These parameters can be adjusted for better performance depending on the distribution of speech and noise exemplars in the high-dimensional feature space. Moreover, various well-known distance/divergence measures such as the Euclidean distance, generalized Kullback-Leibler divergence, Itakura-Saito divergence and Hellinger distance are special cases of the AB-divergence for different (α, β) values. The goal of this work is to investigate the optimal divergence for mel-scaled magnitude spectral features by performing recognition experiments at several SNR levels using different (α, β) pairs. The results demonstrate the effectiveness of the AB-divergence compared to the generalized Kullback-Leibler divergence especially at the lower SNR levels.

Index Terms— exemplar-based speech recognition, sparse representations, alpha-beta divergence, noise-robustness

1. INTRODUCTION

The performance of automatic speech recognition systems is hindered by non-stationary noise and reverberation in everyday applications. A considerable amount of research has been devoted to tackle the reduced recognition accuracies due to non-stationary noise resulting in a number of approaches which can mainly be classified under robust feature extraction [1], signal and feature enhancement [2], model compensation [3] and missing data techniques [4–6]. Moreover, several front-end approaches, e.g. linear filtering, feature and spectrum enhancement and back-end approaches, e.g. hidden Markov model (HMM) adaptation and acoustic context-dependent likelihood evaluation, have been proposed to mitigate the adverse effect of reverberation on the speech recognizers [7]. All of these techniques are used together with HMM-based speech recognizers which are known to perform poorly in case of mismatches between the training and testing conditions.

As a viable alternative to HMM-based recognizers, exemplar-based (or template-based) speech recognition techniques recently regained popularity due to the significant increase in the available computational power and the development of fast template matching and search algorithms [8–10]. Several hybrid recognition systems combining this approach with statistical models are also proposed [8, 11–13]. Exemplars are labeled speech segments such as phones or syllables, possibly of different length, that have occurred in the training data. Exemplars are compared with the input speech with respect to a distance metric, e.g. Euclidean distance, using dynamic time warping (DTW).

An alternative framework in exemplar-based speech recognition, namely exemplar-based sparse representations (SR), models the spectrogram of input speech segments as a sparse linear combination of exemplars of the same length. SR-based techniques have been successfully used for speech enhancement [14], feature extraction [15], clean [16] and noise-robust speech recognition [17, 18]. We have recently proposed an SR-based speech recognition system which uses exemplars of different length organized in separate dictionaries on the basis of their class and length [19]. Compared to a system using fixed-length exemplars stored in a single dictionary, 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. We have also shown that this system performs reasonably well under noisy conditions in [20].

In SR-based recognition systems, exemplar weights are obtained by solving a regularized convex optimization problem with a cost function comprised of a distance/divergence measure to quantify the approximation quality. The choice of the distance/divergence measure depends on the distribution of the speech and noise sources in the high-dimensional feature space. Magnitude spectral features have been often used in conjunction with the generalized Kullback-Leibler divergence (KLD) in SR-based noise-robust speech recognition, blind source separation and polyphonic music transcription tasks [17, 18, 21]. King et al. investigated the optimal parameter of the beta-divergence as a cost function for non-negative matrix factorization-based speech separation and music interpolation in [22].

In this work, we use the Alpha-Beta (AB)-divergence [23] to quantify the approximation error. AB-divergence is a family of divergences with two parameters, namely α and β. For different values of these parameters, the AB-divergence connects various well-known distance/divergence measures such as Euclidean distance, Hellinger distance, Itakura-Saito divergence and generalized KLD. The higher degree of freedom offered by the AB-divergence has been shown to enable better robustness against noise and outliers [23]. The goal of this work is to investigate to what extent the use of the AB-divergence can improve SR-based noise-robust speech recognition.

The rest of the paper is organized as follows. The exemplar-based sparse representations system using the AB divergence is explained in Section 2. The evaluation setup and implementation details are discussed in Section 3. Section 4 presents the recognition results and comments on the parameters providing the maximum

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

recognition accuracy. In Section 5, the conclusions and thoughts for future work are discussed.

2. SPARSE REPRESENTATION MODEL OF SPEECH WITH EXEMPLARS OF DIFFERENT LENGTH

2.1. Model for noisy speech

The noise-robust recognizer described in [20] models noisy speech segments as a sparse linear combination of speech and noise exemplars that are stored in multiple dictionaries. Speech exemplars, each comprised of D (Mel) frequency channels 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 length l and class c. 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 single 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. For any class c, a reshaped noisy speech vector $\mathbf{y}_l$ of length Dl is expressed as a linear combination of the exemplars:

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

where $\mathbf{x}_{c,l}$ is an $M_{c,l}$-dimensional non-negative weight vector. The combination is hence supposed to model all variability in the signal due to pronunciation variation, reverberation, noise and so forth.

2.2. Obtaining the exemplar weights

The exemplar weights are obtained by minimizing the cost function,

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

where $\mathbf{\Lambda}$ is an $M_{c,l}$-dimensional vector. The first term is the divergence between the noisy speech vector and its approximation. The second term is a regularization term to produce a sparse solution. A sparse weight vector implies that the noisy speech is approximated by a few exemplars from the speech and/or noise dictionaries. $\mathbf{\Lambda}$ contains non-negative values and 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.

Depending on the type of the used speech features, a divergence/distance measure is adopted to compare how well the approximation fits the noisy speech vector. In source separation problems, the generalized KLD has been found to yield better results when it is used in conjunction with magnitude spectral features than for instance the Euclidean distance [24]. Recently, several families of divergences have been proposed and their applications as a cost function for non-negative matrix factorization have been investigated [23, 25]. We compare the baseline setup in [20] with the proposed setup using the AB-divergence to quantify the approximation quality.

2.2.1. Multiplicative update rule for the baseline setup

The baseline system adopts a regularization term which penalizes the l_1-norm of the weight vector to produce a sparse solution

$$r(\mathbf{x}_{c,l}) = \sum_{m=1}^{M_{c,l}} x_{c,l}^m \tag{3}$$

and uses the generalized Kullback-Leibler divergence for d:

$$d(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{k=1}^{K} y_k \log \frac{y_k}{\hat{y}_k} - y_k + \hat{y}_k. \tag{4}$$

The regularized convex optimization problem can be solved by applying non-negative sparse coding (NSC). For NSC, the multiplicative update rule to minimize the cost function (2) is derived in [17] and is given by

$$\mathbf{x}_{c,l} \leftarrow \mathbf{x}_{c,l} \odot (\mathbf{A}_{c,l}^T(\mathbf{y}_l \oslash (\mathbf{A}_{c,l}\mathbf{x}_{c,l}))) \oslash (\mathbf{A}_{c,l}^T\mathbf{1} + \mathbf{\Lambda}) \tag{5}$$

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 noisy speech vector and its approximation decreases monotonically.

2.2.2. Multiplicative update rule for the proposed setup

In the proposed setup, we use the AB-divergence for d, which is equivalent to the generalized KLD for ($\alpha = 1$, $\beta = 0$) [23], to compare the noisy speech vector and its approximation. The AB-divergence is given by:

$$d_{AB}^{(\alpha,\beta)}(\mathbf{y}, \hat{\mathbf{y}}) = -\frac{1}{\alpha\beta}\sum_{k=1}^{K} y_k^\alpha \hat{y}_k^\beta - \frac{\alpha}{\gamma} y_k^\gamma - \frac{\beta}{\gamma}\hat{y}_k^\gamma \qquad \text{for} \quad \alpha, \beta, \gamma \neq 0, \tag{6}$$

where $\gamma = \alpha + \beta$. The extended forms of the AB-divergence at $\alpha = 0, \beta = 0$ or $\gamma = 0$ can be found in [23]. In the initial experiments presented in Section 4, the proposed setup using the AB-divergence adopts the multiplicative update rule derived in [23] which does not take the sparsity inducing regularization term into account, i.e. $\mathbf{\Lambda} = \mathbf{0}$. The multiplicative update rule which minimizes the first term of cost function (2) for $\alpha \neq 0$ is given by

$$\mathbf{x}_{c,l} \leftarrow \mathbf{x}_{c,l} \odot ((\mathbf{A}_{c,l}^T\mathbf{Z}_{c,l}) \oslash (\mathbf{A}_{c,l}^T(\mathbf{A}_{c,l}\mathbf{x}_{c,l})^{\cdot[\gamma-1]}))^{\cdot[1/\alpha]}, \tag{7}$$

where $\mathbf{Z}_{c,l} = \mathbf{y}_l^{\cdot[\alpha]} \odot (\mathbf{A}_{c,l}\mathbf{x}_{c,l})^{\cdot[\beta-1]}$ and $^{\cdot[\,]}$ denotes element-wise exponentiation. Investigation of the impact of induced sparsity on the recognition accuracy with the proposed setup remains as future work.

2.3. Decoding the noisy speech

The first term of Equation (2) expresses the reconstruction error for class c and a noisy speech segment of length l. Every noisy speech segment of each available exemplar length is approximated as a linear combination of exemplars. This is achieved by applying the sliding window approach [17] to the noisy utterance for each available exemplar length and iteratively applying Equation (5) using the dictionaries containing exemplars 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 applying dynamic programming (taking the grammar into account) to find the class sequence that minimizes the reconstruction error.

3. EXPERIMENTAL SETUP AND IMPLEMENTATION DETAILS

3.1. Database

The small vocabulary track of the 2nd 'CHiME' Challenge [26] addresses the problem of recognizing commands in a noisy living

β \ α	-5	-4.5	-4	-3.5	-3	-2.5	-2	-1.5	-1	-0.5	0.1	0.5	1
5	**57.33**	**65.25**	55.83	45.33	38.33	33.67	27.50	24.08	20.67	-	-	-	-
4.5	39.25	57.83	**64.83**	56.00	47.00	39.25	34.50	28.83	25.17	21.42	-	-	-
4	21.92	39.92	58.33	**65.25**	56.83	47.92	40.92	34.42	29.83	26.58	21.33	-	-
3.5	-	22.58	40.25	58.33	**64.67**	56.92	48.17	41.75	34.92	30.33	26.25	22.17	-
3	-	-	22.75	40.83	58.92	**63.50**	56.92	48.92	41.17	36.00	30.75	28.67	22.75
2.5	-	-	-	23.08	41.17	57.50	**62.25**	56.92	50.00	42.00	35.42	32.08	28.25
2	-	-	-	-	23.17	41.67	56.92	**62.50**	56.33	50.00	40.92	36.83	32.33
1.5	-	-	-	-	-	23.33	41.00	57.25	**61.83**	56.50	48.50	42.33	36.08
1	-	-	-	-	-	-	23.25	40.50	55.50	**61.33**	54.58	49.25	42.08
0.5	-	-	-	-	-	-	-	22.92	40.08	54.58	**60.58**	54.25	48.50
0	-	-	-	-	-	-	-	-	22.75	39.00	55.83	**59.00**	53.67
-0.5	-	-	-	-	-	-	-	-	-	22.42	42.25	53.25	**57.50**
-1	-	-	-	-	-	-	-	-	-	-	26.00	37.92	52.42
-1.5	-	-	-	-	-	-	-	-	-	-	-	21.67	37.83
-2	-	-	-	-	-	-	-	-	-	-	-	-	20.83

(a) SNR = -6 dB

β \ α	-5	-4.5	-4	-3.5	-3	-2.5	-2	-1.5	-1	-0.5	0.1	0.5	1
5	**69.75**	**76.58**	69.50	60.17	50.67	43.75	39.00	33.75	30.17	-	-	-	-
4.5	49.25	71.00	**77.42**	70.58	60.83	51.25	44.92	39.58	34.67	31.50	-	-	-
4	28.75	50.17	71.58	**78.08**	71.67	61.50	52.67	44.25	39.33	35.42	30.25	-	-
3.5	-	29.33	51.58	71.83	**78.00**	72.17	62.83	53.33	45.58	40.67	35.25	31.67	-
3	-	-	29.75	51.92	72.00	**77.33**	73.00	63.75	54.08	47.08	40.92	37.92	33.58
2.5	-	-	-	29.33	52.50	72.42	**77.33**	72.58	63.25	54.42	46.33	42.42	39.83
2	-	-	-	-	28.75	52.33	72.33	**76.33**	72.92	63.00	53.33	48.25	44.33
1.5	-	-	-	-	-	28.75	52.67	71.25	**76.50**	72.67	62.42	56.00	49.25
1	-	-	-	-	-	-	29.25	52.67	70.92	**75.50**	70.08	63.00	55.25
0.5	-	-	-	-	-	-	-	29.33	52.50	71.17	**74.92**	70.75	61.83
0	-	-	-	-	-	-	-	-	29.92	51.92	72.17	**73.58**	69.83
-0.5	-	-	-	-	-	-	-	-	-	29.67	55.58	69.00	**73.08**
-1	-	-	-	-	-	-	-	-	-	-	33.50	50.42	67.58
-1.5	-	-	-	-	-	-	-	-	-	-	-	29.42	49.25
-2	-	-	-	-	-	-	-	-	-	-	-	-	29.00

(b) SNR = 0 dB

Table 1: Keyword recognition accuracies evaluated for different (α, β) pairs on the development set at SNR level of 0 and -6 dB. Recognition accuracies obtained using the generalized KLD are marked with gray background. The best result of each column is given in bold.

room. The clean utterances are taken from the GRID corpus [27] which contains 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.

The clean utterances are artificially reverberated using binaural room impulse responses which include speaker head movement effects. Then they are mixed with binaural recordings of genuine room noise at SNR levels of 9, 6, 3, 0 ,-3 and -6 dB. The training set contains 500 utterances per speaker (17,000 utterances in total) with clean, reverberated and noisy versions. Noisy utterances are provided both in isolated or embedded form. The embedded recordings contain 5 seconds of background noise before and after the target utterance. The development and test sets contain 600 utterances from all speakers at each SNR level (3600 utterances in total for each set) both in isolated and embedded from. The immediate noise context of the target utterances are available in the embedded recordings. The development set also contains 600 noise-free reverberated utterances. All data has a sampling frequency of 16 kHz.

3.2. Exemplar extraction and dictionary creation

The exemplars and noisy speech segments 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 speech exemplars representing half words are extracted from the reverberated utterances in the training set according to the state-based segmentations obtained by applying forced alignment using a conventional HMM-based speech recognizer. Exemplars belonging to each speaker are organized in separate dictionary sets for speaker-dependent modeling yielding 34 different dictionary sets. The minimum and maximum exemplar lengths are 2 and 40 frames respectively. Exemplars longer than 40 frames are omitted to limit the number of dictionaries. The usage of very short exemplars is viable due to the existence of a strict grammar. Dictionary sizes vary with class, but are limited to 200. The silences between the words are assumed to be negligible, hence, dictionaries representing

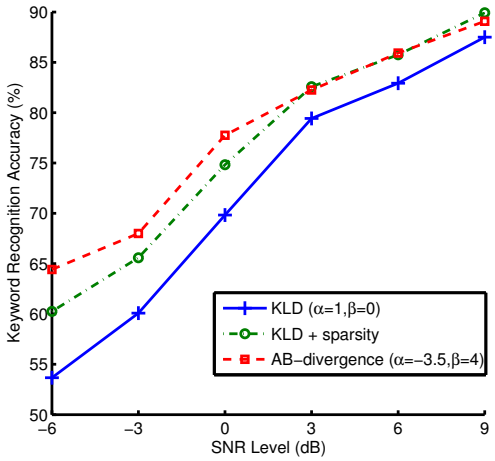

Fig. 1: Keyword recognition accuracies evaluated for the baseline and proposed setups on the test set at SNR levels from -6 dB to 9 dB

a silence class are not used. Further details can be found in [28].

3.3. Noise dictionaries

The noise dictionaries used in the experiments contain noise exemplars which are extracted from the embedded recordings in the training set and from the neighborhood of each target utterance in both directions until the frames belonging to other target utterances. Each dictionary contains 50 noise exemplars extracted from the training set and 150 noise exemplars from the immediate neighborhood of the target utterance.

3.4. Implementation details

The recognition system is implemented in MATLAB and we used GPUs to accelerate the evaluation of Equation (5) and (7). The multiplicative update rules are iterated 50 times to find the exemplar weights. Elements of $\mathbf{\Lambda}$ in Equation (5) are tuned for the highest recognition accuracy on the development data and set to 1.75 and 3 for speech and noise exemplars respectively. The l_2-norm of dictionary columns and reshaped noisy speech vectors are normalized to unity.

4. RESULTS AND DISCUSSION

We have initially performed recognition experiments on the development data with various (α, β) pairs seeking for the pair providing the best recognition accuracy and its underlying relation to the used speech features. The generalized KLD, which is the divergence measure of the baseline system, is a special case of the Alpha-Beta divergence with $(\alpha = 1, \beta = 0)$. The recognition accuracies are obtained for various (α, β) pairs in the range of [-5,1] and [-2,5] respectively with steps of 0.5 applying the multiplicative update rule in Equation (7) and the results are presented in Table 1 for SNR levels of -6 dB and 0 dB. The results that are lower than the 70% of the best results at each SNR level are not presented. As the multiplicative update rule in Equation (7) is not defined for $\alpha = 0$, this value of α is replaced with $\alpha = 0.1$. The recognition accuracies obtained using the generalized KLD are presented in cells marked with gray background and they are equal to 53.67% and 69.83% at SNR levels of -6 dB and 0 dB respectively. The best result obtained for each column is given in bold.

It can be clearly seen from Table 1 that the highest recognition accuracies are obtained for the (α, β) pairs that are on the line $\alpha + \beta = 0.5$. The best results at SNR levels of -6 dB and 0 dB are 65.25% and 78.08% obtained for $(\alpha = -3.5, \beta = 4)$ resulting in an absolute improvement of 11.58% and 8.25% respectively. Firstly, these results imply that the best results are provided when the large values in the approximation $\mathbf{A}_{c,l}\mathbf{x}_{c,l}$ are slightly downweighted compared to the smaller values [23]. Moreover, negative α values provide better results as they suppress the impact of larger $\mathbf{y}_{l,k}/\mathbf{A}_{c,l,k}\mathbf{x}_{c,l,k}$ ratios, i.e. ratios between the k^{th} element of noisy speech vector $\mathbf{y}_l$ and its approximation $\mathbf{A}_{c,l}\mathbf{x}_{c,l}$, on the total reconstruction error [23]. This can be interpreted as putting less emphasis on the spectral peaks that are approximated with a significant error and downweighting these erroneously approximated spectral peaks increases the noise-robustness as expected. It is worth mentioning that the (α, β) values providing the best recognition accuracy cannot be generalized to other recognition tasks due to their dependence on the noise and reverberation characteristics. For different tasks, a search over the (α, β) plane has to be performed for the best performance.

After investigating the recognition accuracies over the (α, β) plane using the development data, we present the recognition results on the test data using the baseline setup with and without induced sparsity (the latter is equivalent to the AB-divergence with $(\alpha = 1, \beta = 0)$) and the proposed setup using the AB-divergence with $(\alpha = -3.5, \beta = 4)$ at SNR levels from -6 dB to 9 dB in Figure 1. The proposed setup performs better than the baseline setup with induced sparsity providing 4.17% and 2.42% absolute improvement at SNR levels of -6 dB and -3 dB respectively. From these results, it can be concluded that AB-divergence provides large improvements at lower SNR levels, once the (α, β) pair that couples well with used speech features is obtained.

5. CONCLUSIONS

This paper analyzes the AB-divergence as a metric to compare the exemplars and noisy speech segments represented in magnitude spectral features in the exemplar-based sparse representations framework. Having two tuning parameters, α and β, AB-divergence provides higher flexibility compared to the commonly used Kullback-Leibler divergence. AB-divergence links several well-known divergence/distance measures for different (α, β) pairs and, as a result, the performance analysis of any recognition system using various divergence/distance measures boils down to a grid search over the (α, β) plane.

Using the provided multiplicative update rules in [23], we perform recognition experiments for different (α, β) pairs to find the parameters yielding the highest keyword recognition accuracy on the development data of the 2nd CHIME Challenge. After finding the optimal parameters, the keyword recognition accuracies obtained on the test data are compared with the baseline system using the generalized KLD. These recognition experiments have shown that the proposed setup using the AB-divergence with the tuned parameters provides higher recognition accuracies than the generalized KLD with induced sparsity at lower SNR levels.

The provided multiplicative update rule for AB-divergence does not take the sparsity inducing L_1 penalty into account. Investigation of the impact of induced sparsity with the proposed setup remains as future work.

6. ACKNOWLEDGEMENTS

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

7. REFERENCES

[1] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct 1994.

[2] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. ICASSP*, May 1996, vol. 2, pp. 733 –736 vol. 2.

[3] 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.

[4] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267 – 285, 2001.

[5] 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.

[6] M. Van Segbroeck and H. Van hamme, "Advances in missing feature techniques for robust large-vocabulary continuous speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 123 –137, Jan. 2011.

[7] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114 –126, Nov. 2012.

[8] 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.

[9] 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.

[10] S. Sundaram and J. R. Bellegarda, "Latent perceptual mapping with data-driven variable-length acoustic units for template-based speech recognition," in *ICASSP*, 2012, pp. 4125–4128.

[11] S. Axelrod and B. Maison, "Combination of hidden Markov models with dynamic time warping for speech recognition," in *Proc. ICASSP*, May 2004, vol. 1, pp. 173–176.

[12] G. Aradilla, J. Vepa, and H. Bourlard, "Improving speech recognition using a data-driven approach," in *Proc. INTERSPEECH*, Lisbon, Portugal, 2005, pp. 3333–3336.

[13] X. Sun, X. Chen, and Y. Zhao, "On the effectiveness of statistical modeling based template matching approach for continuous speech recognition," in *INTERSPEECH*, 2011, pp. 453–456.

[14] 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 (CHIME)*, Sept. 2011, pp. 53–75.

[15] 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.

[16] T. N. Sainath, A. Carmi, D. Kanevsky, and B. Ramabhadran, "Bayesian compressive sensing for phonetic classification," in *Proc. ICASSP*, March 2010, pp. 4370–4373.

[17] 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.

[18] 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.

[19] 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.

[20] 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, pp. 1–4.

[21] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, pp. 177–180.

[22] B. King, C. Fevotte, and P. Smaragdis, "Optimal cost function and magnitude power for NMF-based speech separation and music interpolation," in *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2012, pp. 1–6.

[23] A. Cichocki, S. Cruces, and S.-I. Amari, "Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization," *Entropy*, vol. 13, no. 1, pp. 134–170, 2011.

[24] 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.

[25] A. Cichocki, R. Zdunek, and S. Amari, "Csiszár's divergences for non-negative matrix factorization: family of new algorithms," in *Proc. of the 6th International Conference on Independent Component Analysis and Blind Signal Separation*, 2006, pp. 32–39.

[26] 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.

[27] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition.," *Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[28] 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 *2nd International Workshop on Machine Learning in Multisource Environments (CHIME)*, June 2013, pp. 39–43.