NOISE ROBUST EXEMPLAR MATCHING WITH COUPLED DICTIONARIES FOR SINGLE-CHANNEL SPEECH ENHANCEMENT

Emre Yılmaz, Deepak Baby, and Hugo Van hamme

Dept. ESAT, KU Leuven, Belgium

ABSTRACT

In this paper, we propose a single-channel speech enhancement system based on the noise robust exemplar matching (N-REM) framework using coupled dictionaries. N-REM approximates noisy speech segments as a sparse linear combination of speech and noise exemplars that are stored in multiple dictionaries based on their length and associated speech unit. The dictionaries providing the best approximation of the noisy mixtures are used to estimate the speech component. We further employ a coupled dictionary approach that performs the approximation in the lower dimensional mel domain to benefit from the reduced computational load and better generalization, and the enhancement in the short-time Fourier transform (STFT) domain for higher spectral resolution. The proposed enhancement system is shown to have superior performance compared to the exemplar-based sparse representations approach using fixed-length exemplars in a single overcomplete dictionary.

Index Terms— speech enhancement, exemplar matching, coupled dictionaries, non-negative sparse coding

1. INTRODUCTION

Single-channel speech enhancement approaches aim to reduce the amount of background noise in speech signals recorded by a microphone and improve the speech intelligibility and quality. These techniques can also be used in the front end of other speech processing tasks such as automatic speech recognition (ASR) to alleviate the degradation due to the background noise. Denoising of monaural speech data is still a rather challenging problem even after the intensive research over several decades [1]. Numerous statistical and data-driven approaches have been proposed to tackle the problem [2–9] (and references therein).

This paper presents a novel exemplar-based speech enhancement approach, dubbed *noise robust exemplar matching* (N-REM), which performs denoising using the actual occurrences of speech and noise extracted from training data. Unlike previous exemplar-based sparse representations (SR) of speech using fixed-length exemplars in a single overcomplete dictionary [10–16], the proposed approach uses exemplars of multiple lengths, each associated with a single speech unit such as phones, syllables, half-words or words [17–19]. These exemplars are organized in multiple dictionaries based on the their length and class (associated speech unit). Using separate dictionaries for different speech units is motivated by the geometrical interpretation of SR-based source separation. It is known that the farther the convex hull of the basis vectors belonging to speech and noise sources are, the better the separation is [20]. Hence, the use of separate dictionaries for each speech unit provides a more precise representation in the high-dimensional feature space.

Previously, the N-REM framework has been shown to perform reasonably well on small vocabulary ASR tasks [21]. This paper describes the initial efforts towards an N-REM based speech enhancement approach. In addition, we incorporate a coupled dictionaries approach [15] which uses a front-end dictionary containing lower dimensional features to obtain the decomposition, and a back-end dictionary containing the full-resolution spectral representations to reconstruct the speech and noise sources. In this way, the proposed approach benefits from the advantages of the lower dimensional features like better generalization and lower computational complexity during the decomposition and higher spectral resolution during the reconstruction of the speech component. For a reliable reconstruction, the mapping between the corresponding exemplars in both the dictionaries should be one-to-one which is realized by extracting the corresponding exemplars of the coupled dictionaries jointly from the same piece of training data.

2. NOISE ROBUST EXEMPLAR MATCHING

2.1. Exemplar extraction and dictionary creation

Training frame sequences associated with a single speech unit (speech exemplars) are extracted based on the state-level alignments obtained using a conventional HMM-based recognizer. Every speech exemplar is represented both in the full-resolution spectral domain (henceforth STFT exemplars) with K frequency bins and lower dimensional mel domain (henceforth mel exemplars) with D mel frequency bands.

This work has been funded with support from the European Commission under Contract FP7-PEOPLE-2011-290000 (INSPIRE) and IWT-SBO Project 100049 (ALADIN)

For the transformation between these domains, we use a STFT-to-mel matrix, $\mathbf{C}$, of dimensionality $D \times K$.

Mel speech exemplars, each comprised of D mel frequency bands and spanning l frames, are reshaped into a single vector and stored in the columns of a mel speech dictionary $\mathbf{S}^M_{c,l}$: one for each class c and each length l. Each dictionary is of dimensionality $Dl \times R_{c,l}$ where $R_{c,l}$ is the number of available mel speech exemplars of class c and length l. Similarly, a mel noise dictionary $\mathbf{N}^M_l$ for each length l is formed by reshaping the noise exemplars. Each mel speech dictionary is concatenated with the mel noise dictionary of the same length to form a combined mel dictionary $\mathbf{A}^M_{c,l} = [\mathbf{S}^M_{c,l}\,\mathbf{N}^M_l]$ of dimensionality $Dl \times P_{c,l}$ where $P_{c,l}$ is the total number of available speech and noise exemplars. The same procedure is followed using the STFT speech and noise exemplars to obtain the combined STFT dictionaries $\mathbf{A}^F_{c,l} = [\mathbf{S}^F_{c,l}\,\mathbf{N}^F_l]$ of dimensionality $Kl \times P_{c,l}$.

2.2. Decomposition of noisy speech

The decomposition of noisy mixtures into speech and noise components is performed only in the mel domain. Every observed noisy speech segment of length T frames is also reshaped into vectors by applying a sliding window approach [11] with window length of l frames and stored in an observation matrix $\mathbf{Y}_l = [\mathbf{y}^1_l, \mathbf{y}^2_l ..., \mathbf{y}^{(T-l+1)}_l]$ of dimensionality $Dl \times (T-l+1)$. Due to multiple-length exemplars, the window length l is varied between the minimum exemplar length l_{min} and maximum exemplar length l_{max} yielding observation matrices $\mathbf{Y}_l$ for $l_{\text{min}} \leq l \leq l_{\text{max}}$. For every class c, each observation vector $\mathbf{y}_l$ is expressed as a linear combination of the exemplars that are stored in the dictionaries of the same length:

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

where $\mathbf{x}_{c,l}$ is an $P_{c,l}$-dimensional non-negative weight vector. The sparse solutions of $\mathbf{x}_{c,l}$ yield more realistic approximation of the observed segments without overfitting and have been shown to provide better recognition results [5,22].

The combined dictionaries consisting of speech and noise exemplars are presumed to model all acoustic variability in the observed signal due to pronunciation variation, background noise and so forth. This model can also cope with reverberation by storing reverberated speech exemplars rather than clean speech exemplars.

2.3. Obtaining the exemplar weights

The non-negative exemplar weights $\mathbf{x}_{c,l}$ are obtained by minimizing the cost function,

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

where $\mathbf{\Lambda}$ is an $P_{c,l}$-dimensional vector. The first term is the divergence between the observation vector and its approximation. The second term is a regularization term which penalizes the l_1-norm of the weight vector to produce a sparse solution. $\mathbf{\Lambda}$ 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. In this work, the regularized optimization problem with the cost function in Equation (2) is solved by applying non-negative sparse coding (NSC) [23]. The generalized KLD is used for d which is commonly used in source separation problems and shown to produce better results than Euclidean distance when used in conjunction with mel-scaled spectral features [5],

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

All observation matrices $\mathbf{Y}_l$ for $l_{\text{min}} \leq l \leq l_{\text{max}}$ are approximated using the combined mel dictionaries $\mathbf{A}^M_{c,l}$ of the corresponding length by applying the multiplicative update rule given in [21]. To quantify the approximation quality, we use the reconstruction error between the noisy speech segments and their approximations. After a fixed number of iterations for all dictionaries, the reconstruction errors between the observation matrix $\mathbf{Y}_l$ and its approximations $\mathbf{A}^M_{c,l}\mathbf{x}_{c,l}$ are calculated for $l_{\text{min}} \leq l \leq l_{\text{max}}$. As the label of each dictionary is known, decoding is performed by applying dynamic programming [24] to find the class sequence that minimizes the reconstruction error to find the best approximation of the target utterance.

2.4. Speech enhancement

After finding the best matching dictionaries, the denoising is performed in two ways, either reconstructing the speech and noise components in mel or STFT domain. The former approach provides the frame-wise mel speech and noise estimates, $\hat{s}^M_{c,l}$ and $\hat{n}^M_{c,l}$, that are obtained after removing the windowing effect by adding the components belonging to overlapping windows from the estimates $S^M_{c,l} X^s_{c,l}$ and $N^M_l X^n_{c,l}$ respectively. Here, $X^s_{c,l}$ refers to the exemplar weights of the speech exemplars and $X^n_{c,l}$ refers to the exemplar weights of the noise exemplars. The frame-level Wiener-like filter is then obtained as in [15],

$$W = \mathbf{C}^T \hat{s}^M_{c,l} \oslash (\mathbf{C}^T (\hat{s}^M_{c,l} + \hat{n}^M_{c,l})). \qquad (4)$$

Since $\mathbf{C}$ contains triangular shaped filter-banks, this extrapolation is the same as the piece-wise linear interpolation between D points (mel bands) spread across the 1 to K frequency bins. The resulting filters always fall in the D-dimensional subspace defined by the columns of $\mathbf{C}^T$ which cannot account for all the added noise content along the K dimensional DFT space. The enhanced speech obtained after applying this filter on the noisy DFT thus will result in a sub-optimal noise suppression.

The coupled dictionary approach remedies this problem by using the STFT speech and noise dictionaries to obtain the the frame-wise speech and noise estimates $\hat{s}^F_{c,l}$ and $\hat{n}^F_{c,l}$ from the estimates $S^F_{c,l}X^s_{c,l}$ and $N^F_l X^n_{c,l}$ respectively. The resulting Wiener-like filter can be written as

$$W_{cd} = \hat{s}^F_{c,l} \oslash (\hat{s}^F_{c,l} + \hat{n}^F_{c,l}). \quad (5)$$

3. EXPERIMENTAL SETUP

The enhancement performance of N-REM is evaluated on the test set A and B of the AURORA-2 corpus [25]. The training material of AURORA-2 consists of a clean and a multi-condition training set, each containing 8440 utterances with one to seven digits in American English. 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 consists of 4 clean and 24 noisy datasets with four noise types (subway, babble, car and exhibition) at six SNR levels, 20, 15, 10, 5, 0 and -5 dB. The noise types of this test set match the multi-condition training set. Test set B has the same number of test sets with four different noise types (restaurant, street, airport, station) at the same SNR levels. Each subset contains 1001 utterances. To reduce the simulation times, we subsampled the test sets by a factor of 4 (250 utterances per test set, 1000 utterances per SNR). A different subset with 100 utterances from each test set is used for development purposes. All data has a sampling frequency of 8 kHz.

The speech exemplars are extracted from the clean training set. Acoustic feature vectors are represented in the full-resolution STFT domain with K =129 bins and mel-scaled magnitude spectra with 23 frequency bands. The speech exemplars representing half-digits are segmented by a conventional HMM-based system. The recognizer uses in total 53285 speech exemplars distributed to 675 dictionaries of 23 different classes (half-digits plus silence). The number of noise exemplars varies depending on the duration of the noise-only sequences that are selected in the preprocessing step and the estimated SNR level of the target utterance. On average, the recognizer uses 11355 and 6621 noise exemplars/utterance in total at SNR level of -5 dB and 20 dB respectively. 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 dictionaries are created by performing active noise exemplar selection and noise sniffing [21]. The combined dictionaries and observation matrices are l_2-normalized for all SNR levels. The multiplicative update rule is iterated 100 times for convergence.

The performance of the proposed setup is compared with several baseline speech enhancement systems such as the optimally-modified log-spectral amplitude (OM-LSA) estimator combined with improved minima controlled recursive averaging technique described in [26] and several exemplar-based SR systems described in [15] which use a single over-complete dictionary containing either fixed length full resolution spectral features (SPEC) or mel-scaled spectral features (MEL). Moreover, the SR-based system adopting the coupled dictionary approach (MELCP) is also considered. The dictionary used by SPEC, MEL and MELCP contains 10000 speech and 10000 noise exemplars. Further details about these systems can be found in [15]. Two evaluation metrics have been used for the performance evaluation. Firstly, the signal-to-distortion ratio (SDR) improvements (ΔSDR) are calculated using the BSS Evaluation Toolkit [27]. Then, the perceptual evaluation of speech quality (PESQ) [28] improvements (ΔPESQ) are also presented to compare the perceptual speech quality of the proposed system with the baselines.

4. RESULTS

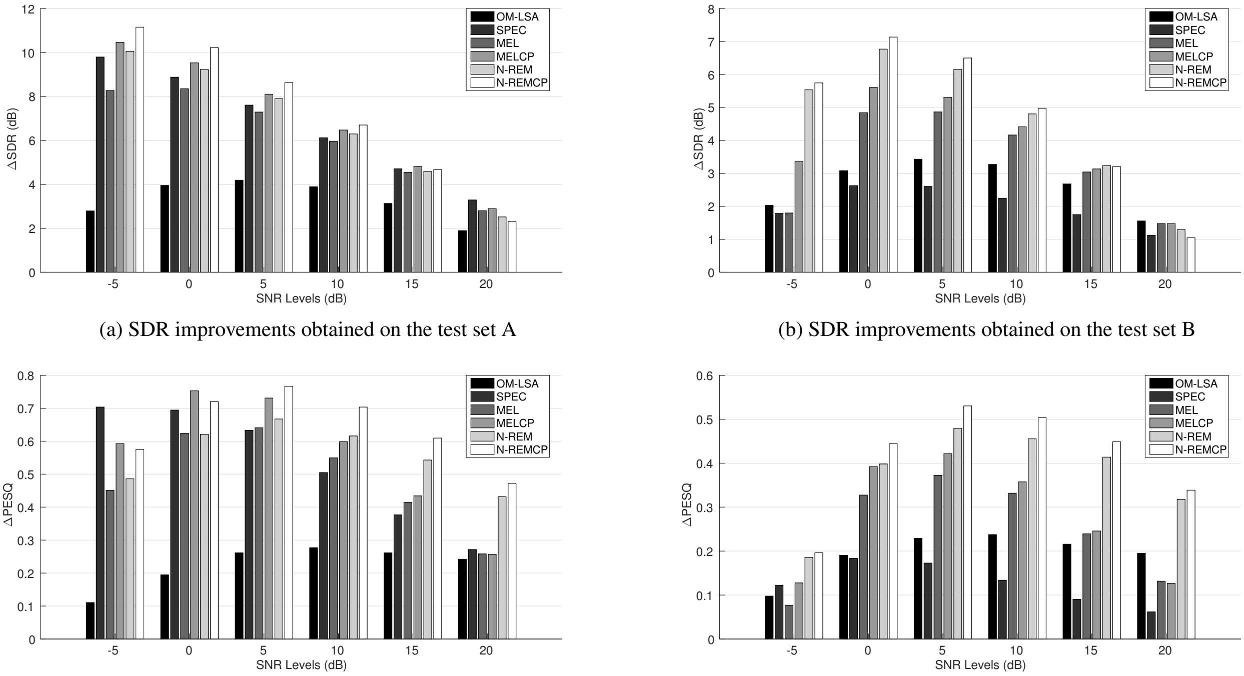
The ΔSDR and ΔPESQ values obtained on both test sets of AURORA-2 data are presented in Figure 1. Figure 1a illustrates the ΔSDR provided on the test set A. The N-REM setup performing the enhancement in mel domain as shown in Equation (4) provides ΔSDRs of 10.1 dB, 9.2 dB and 7.9 dB at SNR levels of -5, 0 and 5 dB respectively. The comparable MEL system yields 8.3 dB, 8.4 dB and 7.3 dB at the same SNR levels.

N-REMCP which performs the enhancement in the STFT domain as shown in Equation (5), achieves better enhancement than N-REM providing 11.2 dB, 10.2 dB, 8.6dB at SNRs of -5, 0 and 5 dB with an absolute improvement of 1.1 dB, 1.0 dB and 0.7dB. For the same SNRs, the baseline MELCP system provides 10.5 dB, 9.5 dB and 8.1 dB. Both N-REM setups outperform their SR-based counterparts with a considerable margin.

At SNR levels of 10 dB and 15 dB, all systems except OM-LSA provide comparable results with ΔSDRs values between 6.0-6.7 dB at SNR of 10 dB and 4.5-4.8 dB at SNR of 15 dB. The SPEC system outperforms the others with a ΔSDR of 3.3 dB at SNR of 20 dB. OM-LSA provides the worst results at all SNR levels.

The ΔSDR obtained on test set B, which are shown in Figure 1b, clearly demonstrates the improved enhancement provided by N-REM systems especially at lower SNR levels. N-REM provides ΔSDRs of 5.5 dB, 6.8 dB and 6.2 dB at SNRs of -5, 0 and 5 dB. These results are significantly higher than 1.8 dB, 4.8 dB and 4.9 dB of the MEL system. The N-REMCP system outperforms MELCP with an absolute improvement of 2.4 dB, 1.5 dB and 1.2 dB at the same SNRs respectively. N-REM based systems still perform better than the baselines at SNR of 10 dB, while they are slightly worse than MEL and MELCP at 20 dB. At this SNR level, OM-LSA provides the best results with a ΔSDR of 1.6 dB. SPEC is the worst performing system at all SNR levels of test set B.

We further compare the ΔPESQ values to evaluate the perceptual quality of the enhancement systems. The ΔPESQ



(a) SDR improvements obtained on the test set A

(b) SDR improvements obtained on the test set B

(c) PESQ improvements obtained on the test set A

(d) PESQ improvements obtained on the test set B

Fig. 1: SDR and PESQ improvements on the test set A and B of AURORA-2 data

values obtained on test set A are shown in Figure 1c. On test set A at SNR -5 dB, SPEC has the highest ΔPESQ of 0.70 followed by MELCP and N-REMCP with a ΔPESQ of 0.59 and 0.57 respectively. At 0dB, MELCP performs the best with 0.75, while N-REMCP and SPEC yield 0.72 and 0.69 respectively. N-REMCP has the highest ΔPESQ at all SNR levels higher than 0 dB. The performance gap between the N-REM based systems and baselines increases at higher SNR levels. The improved perceptual quality of N-REM and N-REMCP is also apparent from the better ΔPESQ results on test set B at all SNR levels which is shown in Figure 1d.

From these results, it can be concluded that the N-REM based systems in general perform better speech enhancement than the baseline systems on account of the separate speech dictionaries which result in more accurate representations of acoustic units in the high-dimensional feature space. Two prominent advantages of these systems are the superior ΔSDR performance under the mismatched noise scenario and overall improvement in the perceptual speech quality. A final comment about the presented results is that the coupled dictionary approach highly improves the enhancement quality also in the N-REM based speech enhancement especially at the lower SNR levels.

5. CONCLUSION

This paper presents a novel single-channel speech enhancement system that performs noise robust exemplar matching to separate speech and noise sources using exemplars, each associated with a certain speech unit. These exemplars are organized in separate dictionaries based on the associated speech unit and length and unseen noisy mixtures are approximated as a sparse linear combination of the speech and noise exemplars in each dictionary.

We further adopt the coupled dictionary approach which performs the approximation in the lower dimensional mel domain and the enhancement in the full-resolution STFT domain. The ΔSDR and ΔPESQ results demonstrate the improved speech enhancement achieved by the proposed system. Future work includes investigating the speech enhancement performance of N-REM using the flexible alpha-beta divergence which yielded improved speech recognition performance and replacing the mel-scaled magnitude spectral features with perceptually motivated modulation spectrogram features. Moreover, an extension of the proposed system working on databases with larger vocabulary remains as a future work.

REFERENCES

[1] Philipos C. Loizou, *Speech Enhancement: Theory and Practice (Signal Processing and Communications)*, 1 edition, June 2007.

[2] T.V. Sreenivas and P. Kirnapure, “Codebook constrained Wiener filtering for speech enhancement,” *IEEE TSAP*, vol. 4, no. 5, pp. 383–389, 1996.

[3] R. Martin, “Noise power spectral density estimation

based on optimal smoothing and minimum statistics," *IEEE TSAP*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[4] V. Grancharov, J. Samuelsson, and Bastiaan Kleijn, "On causal algorithms for speech enhancement," *IEEE TASLP*, vol. 14, no. 3, pp. 764–773, 2006.

[5] 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.

[6] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE TASLP*, vol. 15, no. 1, pp. 1–12, 2007.

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

[8] J.R. Jensen, J. Benesty, M.G. Christensen, and S.H. Jensen, "Enhancement of single-channel periodic signals in the time-domain," *IEEE TASLP*, vol. 20, no. 7, pp. 1948–1963, 2012.

[9] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE TASLP*, vol. 21, no. 10, pp. 2140–2151, 2013.

[10] P. Smaragdis, M. Shashanka, and B. Raj, "A sparse non-parametric approach for single channel separation of known sounds.," in *NIPS*, 2009, pp. 1705–1713.

[11] 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.

[12] 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.

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

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

[15] D. Baby, T. Virtanen, T. Barker, and H. Van hamme, "Coupled dictionary training for exemplar-based speech enhancement," in *Proc. ICASSP*, May 2014, pp. 2883–2887.

[16] N. Mohammadiha and S. Doclo, "Single-channel dynamic exemplar-based speech enhancement," in *Proc. INTERSPEECH*, Sept. 2014.

[17] 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.

[18] T. Deselaers, G. Heigold, and H. Ney, "Speech recognition with state-based nearest neighbour classifiers," in *Proc. INTERSPEECH*, Antwerp, Belgium, 2007, pp. 2093–2096.

[19] 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.

[20] David Donoho and Victoria Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?," in *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[21] 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.

[22] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, Dec. 2004.

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

[24] H. Ney, "The use of a one-stage dynamic programming algorithm for connected word recognition," *IEEE TASSP*, vol. 32, no. 2, pp. 263–271, Apr. 1984.

[25] 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.

[26] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE TSAP*, vol. 11, no. 5, pp. 466–475, Sept 2003.

[27] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[28] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001, pp. 749–752.