Exemplar-based speech enhancement and its application to noise-robust automatic speech recognition

Jort F. Gemmeke[1]*, Tuomas Virtanen*[2]*, Antti Hurmalainen*[2]

[1]Department ESAT, Katholieke Universiteit Leuven, Belgium
[2]Department of Signal Processing, Tampere University of Technology, Finland
jgemmeke@amadana.nl, tuomas.virtanen@tut.fi, antti.hurmalainen@tut.fi

Abstract

In this work an exemplar-based technique for speech enhancement of noisy speech is proposed. The technique works by finding a sparse representation of the noisy speech in a dictionary containing both speech and noise exemplars, and uses the activated dictionary atoms to create a time-varying filter to enhance the noisy speech. The speech enhancement algorithm is evaluated using measured signal to noise ratio (SNR) improvements as well as by using automatic speech recognition. Experiments on the PASCAL CHiME challenge corpus, which contains speech corrupted by both reverberation and authentic living room noise at varying SNRs ranging from 9 to -6 dB, confirm the validity of the proposed technique. Examples of enhanced signals are available at `http://www.cs.tut.fi/~tuomasv/`.

Index Terms: speech enhancement, exemplar-based, noise robustness, sparse representations

1. Introduction

Recognizing speech — either by humans or by machines — in noisy environments remains a difficult problem. A large number of algorithms have been proposed in the literature to address the problem of mitigating the effect of noisy environments on the speech signal. Many of these methods, however, are only effective when speech is corrupted by stationary noise [1, 2], or rely on statistical models of the corrupting noise [3, 4]. Some methods achieve impressive performance when a detailed model of the noise is available [6], but their performance on modelling unseen noise environments is limited.

In this work, we present a non-parametric, speech enhancement method for noisy speech that models noisy speech as a combination of speech and noise. As such, it is very similar to the source separation approaches based on Probabilistic Latent Component Analysis (PLCA) or Non-negative Matrix Factorisation (NMF), largely pioneered by Smaragdis et al. [5]. In this work, however, we build on earlier work on template-based sparse representations [7, 8, 9, 10] and we represent Mel-spectral representations of noisy speech as a sparse, non-negative linear combination of *exemplars*: spectrographic representations of speech spanning 20 frames (200 ms). First a linear combination is found by finding the smallest number of exemplars in a very large collection of speech and noise exemplars (a *dictionary*) that *jointly* approximates the observed speech signal. After obtaining this sparse representation, Mel-spectral reconstructions of the underlying speech and noise sources are created using the weights of the linear combination of exemplars. These are then used to create a time-varying filter to enhance the noisy speech signal.

In comparison to conventional speech enhancement methods, the proposed method has some interesting characteristics. First, since the weights of the speech and noise exemplars are estimated based on a noisy utterance, it does not require an external noise estimator, instead, the speech and noise characteristics are estimated jointly from the noisy signal. Second, since the weights are estimated separately for each 200 ms segment, the method allows modelling non-stationary noises. Third, as the method is exemplar-based both the speech dictionary and the noise dictionary can accurately model many different speakers and noises, to the limit of computational feasibility. At the same time, its exemplar-based nature means the dictionaries can be changed on-the-fly: for example, if more knowledge on the corrupting noise becomes available, these noise frames can be added to the noise dictionary.

We explore the effectiveness of the speech enhancement method using the PASCAL CHiME challenge data. CHiME contains speech corrupted by both reverberation and authentic living room noise at varying signal to noise ratios (SNRs) ranging from 9 to -6 dB. As the speaker identities are assumed to known in the challenge, the speech dictionary will be speaker-dependent. We will use two different noise dictionaries: a fixed dictionary that contains random noise exemplars from the provided background noise data, and an adaptive dictionary that is created on the fly from the background noise surrounding of each noisy utterance to be processed. The method will be evaluated using both SNR measurements of the clean and noisy reconstructed speech as well as by using an automatic speech recognition (ASR) system. Out of interest for ASR applications, we investigate to what extent speech recognition accuracy can improve using re-training and multi-condition methods.

The rest of the paper is organised as follows. The exemplar-based speech enhancement method is described in Section 2. The experimental setup, such as the CHiME database, the implementation details of the speech enhancement technique and the speech recognition system are described in Section 3. The SNR measurement and speech recognition results are presented in Section 4 and discussed in Section 5. Conclusions and suggestions for future work are given in Section 6.

2. Method

The speech enhancement technique employed in this paper is based on representing the signal using magnitudes in the Mel-spectral domain. The input signal is windowed into frames, and a discrete Fourier transform (DFT) of each frame is taken. The absolute values of the DFT in each frame are stored into a vector. The magnitudes in Mel-frequency bands are obtained by multiplying the vector by matrix $\mathbf{B}$, where each row of the ma-

trix is the magnitude response of a single Mel band for the DFT frequencies. We use triangular responses which overlap 50%. The above processing is applied in each frame, and the resulting values are stored into a $B \times T$ noisy speech spectrogram matrix $\mathbf{Y}$ (with B Mel-frequency bands and T time frames).

2.1. Exemplar-based representation of noisy speech

We assume $\mathbf{Y}$ is a linear addition of underlying clean speech $\mathbf{S}$ and noise $\mathbf{N}$ magnitude spectrograms. To simplify the notation, the columns of each matrix are stacked into the vectors $\mathbf{y}$, $\mathbf{s}$ and $\mathbf{n}$, respectively, each of length $D = B \cdot T$

We model $\mathbf{s}$ as a sparse, non-negative linear combination of example speech spectrograms *exemplars*, which are extracted from the training data. The exemplars are denoted as $\mathbf{a}^{\mathrm{s}}_j$, with $j = 1, \ldots, J$ denoting the exemplar index. Accordingly, the noise spectrogram is modelled using K noise exemplars: $\mathbf{a}^{\mathrm{n}}_k$, with $k = 1, \ldots, K$.

We then write:

$$\mathbf{y} \approx \mathbf{s} + \mathbf{n} \tag{1}$$

$$\approx \sum_{j=1}^{J} x^s_j \mathbf{a}^{\mathrm{s}}_j + \sum_{k=1}^{K} x^{\mathrm{n}}_k \mathbf{a}^{\mathrm{n}}_k \tag{2}$$

$$= [\mathbf{A}^{\mathrm{s}} \mathbf{A}^{\mathrm{n}}] \begin{bmatrix} \mathbf{x}^{\mathrm{s}} \\ \mathbf{x}^{\mathrm{n}} \end{bmatrix} \tag{3}$$

$$= \mathbf{A}\mathbf{x} \quad \text{s.t.} \quad \mathbf{x}^{\mathrm{s}}, \mathbf{x}^{\mathrm{n}}, \mathbf{x} \geq 0 \tag{4}$$

with $\mathbf{x}^{\mathrm{s}}$ and $\mathbf{x}^{\mathrm{n}}$ sparse representations of the underlying speech and noise, respectively. In order to obtain $\mathbf{x}$, we minimize the cost function:

$$d(\mathbf{y}, \mathbf{A}\mathbf{x}) + ||\boldsymbol{\lambda} .* \mathbf{x}||_1 \quad \text{s.t.,} \quad \mathbf{x} \geq 0 \tag{5}$$

where d is the generalized Kullback-Leibler (KL) divergence and the second term a sparsity inducing L-1 norm of the activations weighted by element-wise multiplication (operator $.*$) with vector $\boldsymbol{\lambda} = [\lambda_1 \;\; \lambda_2 \ldots \lambda_E]$. The cost function (5) is minimized using a multiplicative updates routine as in [8].

2.2. Speech enhancement

Let us denote speech exemplar j spectrum and noise exemplar k spectrum in frame t as $\mathbf{a}^{\mathrm{s}}_{j,t}$ and $\mathbf{a}^{\mathrm{n}}_{k,t}$, respectively.

The model for the clean speech spectrum, and model for the noise spectrum are given as

$$\tilde{\mathbf{s}}_t = \sum_{j=1}^{J} x^s_j \mathbf{a}^{\mathrm{s}}_{j,t}. \tag{6}$$

$$\tilde{\mathbf{n}}_t = \sum_{k=1}^{K} x^{\mathrm{n}}_k \mathbf{a}^{\mathrm{n}}_{k,t}. \tag{7}$$

In order to decode utterances of arbitrary lengths, we adopt a sliding time window approach as in [8]. In this approach, we represent a noisy utterance as a number of fixed-size, overlapping speech segments, each of length T. For each segment, we calculate clean speech estimates $\tilde{\mathbf{s}}_t$ and noise estimates $\tilde{\mathbf{n}}_t$ as described above. For the entire utterance, the segment-wise estimates are averaged over the overlapping windows, to get a single clean speech and noise estimate per each frame t. The spectral estimates of speech and noise averaged over windows are denoted with vectors $\hat{\mathbf{s}}_t$ and $\hat{\mathbf{n}}_t$, respectively.

Table 1: SNR measurements on development set. The rows "Enhanced-fixed" and "Enhanced - adaptive" refer to speech enhancement with a fixed or adaptive noise dictionary, respectively.

SNR (dB)	-6	-3	0	3	6	9
Noisy speech	-7.0	-4.8	-2.5	-0.2	2.0	3.8
Enhanced - fixed	-2.2	-1.2	0.2	1.7	3.1	4.2
Enhanced - adaptive	-2.8	-1.6	0.0	1.6	3.1	4.4

Incidentally, the use of a sliding window approach is an alternative for a de-convolution approach [5] in which exemplars can be placed at arbitrary positions in the spectrogram. Theoretically, the use of deconvolution would alleviate the need for time-shifted exemplars in the dictionary [11]. However, pilot tests revealed that in our experimental setup, the use of deconvolution yields the same separation quality at the cost of roughly a factor two in computational effort.

We design a DFT-domain filter magnitude response vector $\mathbf{h}_t$ for each frame t as

$$\mathbf{h}_t = \mathbf{B}^T \hat{\mathbf{s}}_t ./ (\mathbf{B}^T \hat{\mathbf{s}}_t + \mathbf{B}^T \hat{\mathbf{n}}_t), \tag{8}$$

with $./$ denoting element-wise division. The multiplication by $\mathbf{B}^T$ maps the Mel magnitude vectors to the DFT domain. We also tried mapping the Mel magnitude vectors by multiplication with the pseudo-inverse of $\mathbf{B}$. It produced marginally better results in some test scenarios, but was not used in simulation results shown in this paper.

Element-wise multiplication between the complex DFT vector of noisy speech in frame t and the corresponding filter magnitude response above is calculated to obtain an enhanced complex spectrum. The enhanced spectrum is transformed into time domain by taking inverse DFT. Frames are combined using overlap-add to get the whole enhanced signal. Example samples are available at `http://www.cs.tut.fi/~tuomasv/`

3. Experimental setup

3.1. Database

The PASCAL 'CHiME' Speech Separation and Recognition Challenge [1] is designed to address some of the problems occurring in real word noisy speech recognition. The challenge data is based on the GRID corpus [12], in which 34 speakers read simple command sentences. These sentences are of form *verb-colour-preposition-letter-digit-adverb*. There are 25 different 'letter' class words and 10 different digits. Other classes have four word options each. When doing automatic speech recognition, the recognition accuracy is the percentage of correctly recognised letter and digit keywords.

CHiME utterances simulate a scenario, where sentences are spoken in a noisy living room. The original, clean speech utterances are reverberated according to the actual room response, and then mixed to selected noise sections, which produce the desired SNR mixture level for each noisy set. The noisy sets have target SNR levels of 9, 6, 3, 0, -3 and -6 dB.

For modelling/training, there are 500 reverberated utterances per speaker (no noise), and six hours of background noise data. The development and test sets consist of 600 utterances at each SNR level, Additionally, noiseless (only reverberation)

[1] http://www.dcs.shef.ac.uk/spandh/chime/challenge.html

development utterances are available. Development and test utterances are both given in a strictly endpointed format, but also as embedded signals within a longer noise context. All data is stereophonic and has a sampling rate of 16 kHz.

3.2. Dictionary selection

In our speech enhancement experiments, we used a dictionary $\mathbf{A}$ consisting of 5000 speech and 5000 noise exemplars with $T = 20$. In exemplar-based sparse classification [9], using 20 adjacent frames to model the temporal modulations was found sufficient to distinguish well between speech and noise by modeling their temporal modulations. On the other hand, very different time-frequency resolutions such as the DFT spectrum in individual 60 ms frames have been found to produce good exemplar-based enhancement for speech recognition [10].

We created two different dictionaries for each speaker. In the first, the 5000 noise exemplars are randomly extracted from the provided background noise data. The second set is different for each test utterance, and there 5000 noise exemplars are selected by sampling the neighbourhood of embedded utterances in both directions with a shift of 4 to 7 frames, excluding locations where other test utterances were embedded.

In both dictionaries, the speech part of the dictionary is formed by constructing an initial speech dictionary for the specific speaker by extracting exemplars from a randomly selected subset of 60% of the noiseless speech training utterances, using a random frame shift of 4 to 8 frames. This resulted in an initial dictionary containing approximately 10000 to 17000 partially overlapping exemplars for each speaker. These were then reduced to a fixed size of 5000 exemplars by selecting exemplars such that there is a maximally flat coverage between words. This is done because in the initial dictionary, words from classes with fewer options are overrepresented.

Finally, each speaker-dependent dictionary was reweighted to have equal Euclidean norm over Mel bands and exemplars. During feature extraction for speech enhancement, the same band weights were applied to the noisy utterances to unify the scale of bands.

3.3. Speech enhancement

The speech enhancement method requires, in addition to the noisy speech signal, mel-spectral magnitude features. These were calculated from partially overlapping 25 ms frames with a shift of 10 ms between frames. We used 26 Mel bands, which matches the number of bands used for the default CHiME MFCC models. Features were extracted separately for both stereo channels, thus effectively doubling the number of feature bands. Using a low-resolution frequency representation (26 Mel bands) makes the exemplars less sensitive to the pitch and allows us to represent a wide range of spectral shapes without the need to model all the phoneme-pitch combinations.

In the sparsity penalty matrix $\boldsymbol{\lambda}$ we use two different values, one for speech exemplars and another for noise exemplars. These were tuned on a random subset of the development set by maximising recognition accuracy using the exemplar-based classification system described in [8]. The tuned values are 2.0 and 1.7 for speech and noise exemplars, respectively. Although the speech enhancement method uses binaural features to find a sparse representation, the resulting time-varying filter is applied to mono waveforms (the average of binaural waveform) because the subsequent analysis does not exploit the binaural nature of the signals.

The SNRs of the enhanced signals, also known as the signal-to-distortion ratio (SDR) [13], were calculated as

$$\mathrm{SNR}_{\mathrm{dB}} = 10 \log_{10} \frac{\sum_n y(n)^2}{\sum_n (y(n) - \hat{y}(n)^2)}. \qquad (9)$$

Above $y(n)$ is the original signal as the function of time n and $\hat{y}(n)$ is the corresponding enhanced signal.

3.4. Speech recognition

For speech recognition experiments, we used the HTK recognition setup provided by the CHiME challenge organisers. In brief, the recogniser works on mean-normalised 12 -band Mel-cepstral features and their delta and delta-delta derivatives. If the speech is binaural, the two channels are averaged prior to feature extraction. The recogniser employs whole-word models ranging from 4 tot 10 states per word, with each state described by 7 Gaussian mixtures. The employed models are speaker-dependent, and trained by performing 4 more iterations of EM training on a speaker-independent model using the speaker-specific training utterances.

We experimented with several acoustic models. The baseline model is trained using the reverberant 'clean' training signals and is not retrained on noisy signals. We also trained a multi-condition model by doing another 4 iterations of speaker-dependent retraining using both clean training utterances as well as the noisy, speaker-dependent development utterances. In addition, we train acoustic models that are (partly) based on speech processed with the enhancement method. For the 'clean' model, we first apply speech enhancement to those 40% of training utterances that were not used for extracting exemplars in the speech dictionary. As no adaptive noise dictionary can be constructed for these clean utterances, only the fixed noise dictionary was used. The enhanced-clean acoustic model was then constructed using the provided training recipe. For the multi-condition models, we retrained this enhanced-clean model with another 4 iterations of speaker-dependent retraining using both partly-enhanced clean training utterances as well as the enhanced, noisy, speaker-dependent development utterances. This was done both using the fixed noise dictionary and the adaptive noise dictionary.

4. Results

4.1. SNR measurements

We carried out SNR measurements of the enhanced noisy speech using the development set, for which clean speech utterances are provided. This was done both using the fixed noise dictionary as well as the adaptive noise dictionary. For comparison, we also did the SNR measurement on the original noisy utterances. The measured average SNRs are shown in Table 1.

Note that the measured SNR levels of the original noisy utterances do not match the corresponding SNR level designations. This is mainly due to the fact that during the creation of the CHiME database, SNR measurements were done using high-pass filtered noise utterances.

In order to evaluate the effect of the enhancement algorithm on clean speech, we also carried out an SNR measurement of enhanced clean speech using those 40% of training utterances that were not used for extracting exemplars in the speech dictionary. Only the fixed noise dictionary was used. The average SNR on these 6800 utterances was 18.9, with a standard deviation of 2.8.

Table 2: Speech recognition accuracies using the CHiME baseline acoustic model. The rows "Enhanced-fixed" and "Enhanced - adaptive" refer to speech enhancement with a fixed or adaptive noise dictionary, respectively.

(a) Development set

SNR (dB)	-6	-3	0	3	6	9	inf
Noisy speech	31.1	36.8	49.1	64.0	73.8	83.1	94.2
Enhanced - fixed	44.5	54.5	64.8	75.5	82.8	87.7	-
Enhanced - adaptive	37.9	48.4	59.5	71.3	79.2	85.9	-

(b) Test set

SNR (dB)	-6	-3	0	3	6	9
Noisy speech	30.3	35.4	49.5	62.9	75.0	82.4
Enhanced - fixed	45.2	54.3	67.3	75.0	83.8	87.8
Enhanced - adaptive	48.2	58.5	71.8	78.0	85.0	88.5

Table 3: Speech recognition accuracies using an acoustic model trained on enhanced clean speech, where the clean speech is enhanced using the fixed noise dictionary. The rows "Enhanced-fixed" and "Enhanced - adaptive" refer to speech enhancement with a fixed or adaptive noise dictionary, respectively.

(a) Development set

SNR (dB)	-6	-3	0	3	6	9	inf
Noisy speech	31.4	35.9	49.2	61.8	73.3	83.5	94.2
Enhanced - fixed	46.3	54.5	66.3	75.5	83.3	87.2	-
Enhanced - adaptive	39.4	49.1	59.8	71.6	80.6	86.9	-

(b) Test set

SNR (dB)	-6	-3	0	3	6	9
Noisy speech	30.3	34.8	47.9	60.8	74.1	82.9
Enhanced - fixed	44.9	54.7	68.1	76.3	84.5	89.0
Enhanced - adaptive	49.3	58.3	70.5	77.3	85.8	88.9

Table 4: Speech recognition accuracies using a multi-condition acoustic model. The rows "Enhanced-fixed" and "Enhanced - adaptive" refer to speech enhancement with a fixed or adaptive noise dictionary, respectively. The rows "Noisy speech" are recognised using an acoustic model trained on clean speech and noisy development data. The rows "Enhanced - fixed" are recognised using a model trained on enhanced clean speech (fixed dictionary), which was then refined using development data enhanced with a fixed dictionary. The rows "Enhanced - adaptive" are recognised using a model trained on enhanced clean speech (fixed dictionary), which was then refined using development data enhanced with a adaptive dictionary.

(a) Development set

SNR (dB)	-6	-3	0	3	6	9	inf
Noisy speech	81.8	86.7	93.1	96.8	98.7	99.6	99.75
Enhanced - fixed	89.9	94.5	97.5	99.0	99.8	99.8	-
Enhanced - adaptive	87.3	92.2	97.1	98.7	99.8	99.8	-

(b) Test set

SNR (dB)	-6	-3	0	3	6	9
Noisy speech	34.7	40.9	50.8	61.7	72.0	80.1
Enhanced - fixed	49.3	57.4	68.9	76.6	83.5	86.3
Enhanced - adaptive	52.8	59.4	69.2	76.1	84.1	87.2

4.2. Speech recognition

We carried out speech recognition experiments using three types of acoustic models (cf. Section 3.4). The 'baseline' acoustic model is provided by the CHiME organisers and is trained on clean speech (only reverberated). Recognition results using this acoustic model for the original noisy speech, as well as the enhanced speech with fixed or adaptive noise dictionaries are shown in Table 2. A second acoustic model was created by training on enhanced clean speech, shown in Table 3. Finally, recognition was carried out using multi-condition trained models. For brevity, in Table 4 only the results are shown of the corresponding acoustic model, e.g. the recognition results for the enhanced speech using an adaptive dictionary were obtained using a multi-condition model trained on development data enhanced with an adaptive dictionary.

5. Discussion

5.1. Effectiveness of speech enhancement

The results in Table 1 show that both variants of exemplar-based speech enhancement — using a fixed or adaptive noise dictionary — improve the measured SNR substantially. Also in Tables 2-4 it can be observed that speech recognition on the enhanced speech achieves substantially higher accuracies, both at low and high SNRs.

5.2. Influence of noise dictionary

With the exception of the highest SNR level, speech enhancement using a fixed noise dictionary achieves slightly higher SNR improvements, as shown in Table 1. Likewise, substantially higher recognition accuracies are obtained on the development set using a fixed noise dictionary.

On the test set however, speech enhancement using an adaptive dictionary performs better than speech enhancement using a fixed noise dictionary, although the differences are smaller. In Table 2 it can be observed that the baseline recognition scores differ only slightly between the test and development set, indicating that it is not very likely that the test set is much more difficult than the development set. A more likely explanation is that the chosen sparsity values, that were tuned on development data using a fixed dictionary, are over-trained on the development set and suboptimal for the enhancement with an adaptive dictionary.

5.3. Retraining the acoustic model

When speech enhancement is used as a noise robustness front-end prior to speech recognition, it is common practise to adapt or retrain the acoustic models used by the ASR system to account for the artefacts the speech enhancement process introduces. Often, this improves the results substantially since ASR systems are very sensitive to the artefacts created by speech en-

hancement, such as musical noise.

Comparing the results in Tables 2 and 3, we can observe that retraining the acoustic models on enhanced clean speech generally improves the speech enhancement recognition results, both on the development data as on the test set. As expected, no consistent improvement can be observed on the original noisy speech.

At the same time, the benefit of retraining the acoustic model is only slight. This may indicate that the distortions introduced by the speech enhancement method are not severe, a hypothesis supported by the high SNR on clean speech reported in Section 4.1.

5.4. Multi-condition training

As a final approach to improving speech recognition by adapting the acoustic model, we consider multi-condition training. Multi-condition training is a straightforward approach to achieving noise robustness, but is known to suffer from a lack of generalisability to unseen conditions and a reduced performance on high-SNR speech.

Comparing Tables 2 and 4 we can observe that when doing noisy speech recognition using a multi-condition trained acoustic model, the performance on the development set increases enormously, because we are now effectively testing on the training data. At the same time, we observe that on the test set, results only improve at SNRs $<$ 3dB, and the improvement is far smaller. This confirms the conventional wisdom on multi-condition training.

In combination with speech enhancement, a very similar effect can be observed, resulting in a trade-off between low and high SNR accuracy. Interestingly, the benefit of speech enhancement with a fixed noise dictionary using a multi-condition trained model seems slightly larger than speech enhancement using an adaptive dictionary. This may be due to the fact that thanks to the fixed dictionary, the speech enhancement makes more consistent errors which can be learned and compensated by the acoustic model.

6. Conclusions and future work

We proposed an exemplar-based technique for speech enhancement of noisy speech. The technique works by finding a sparse representation of the noisy speech in a dictionary containing both speech and noise exemplars, and uses the activated dictionary atoms to create a time-varying filter to enhance the noisy speech.

The speech enhancement algorithm was evaluated using measured SNR improvements as well as by using automatic speech recognition. Experiments on the PASCAL CHiME challenge corpus, showed substantial improvements, both in measured SNR and in speech recognition accuracy. It was shown that on the unseen test data, using a adaptive noise dictionary performed better than a fixed noise dictionary, and that the speech recognition results can be improved by retraining of the acoustic models.

In future work, we plan on a more through investigation of the impact of feature dimension, amount of time context and sparsity penalty values on speech enhancement quality. Also, we plan on refining the speech enhancement method, for example by performing multiple iterations of speech enhancement.

7. Acknowledgements

The research of Jort F. Gemmeke was funded by IWT-SBO project ALADIN contract 100049. Tuomas Virtanen and Antti Hurmalainen have been funded by the Academy of Finland.

8. References

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, 1979.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, pp. 443–445, 1985.

[3] B. Raj, R. Singh, and R. M. Stern, "On Tracking Noise with Linear Dynamical System Models," in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, Montreal, Canada, 2004.

[4] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, Albuquerque, USA, 1990.

[5] Paris Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Proceedings of the 5th International Symposium on Independent Component Analysis and Blind Signal Separation*, Granada, Spain, September 2004.

[6] John R. Hershey, Steven J. Rennie, Peder A. Olsen, and Trausti T. Kristjansson, "Super-human multi-talker speech recognition: A graphical modeling approach," *Computer Speech and Language*, vol. 24, no. 1, 2010.

[7] Paris Smaragdis, Madhusudana Shashanka, and Bhiksha Raj, "A sparse non-parametric approach for single channel separation of known sounds," in *Proceedings of Neural Information Processing Systems*, Vancouver, Canada, 2009.

[8] Jort F. Gemmeke and Tuomas Virtanen, "Noise robust exemplar-based connected digit recognition," in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, Dallas, USA, 2010.

[9] J.F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, 2011, accepted for publication.

[10] B. Raj, T. Virtanen, S. Chaudhure, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in *Interspeech 2010*, Tokyo, Japan, 2010.

[11] A. Hurmalainen, J.F. Gemmeke, and T. Virtanen, "Non-negative matrix deconvolution in noise robust speech recognition," in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, Prague, Czech Republic, 2011.

[12] M. P. Cooke, J. Barker, S. P. 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, 2006.

[13] Emmanuel Vincent, Hiroshi Sawada, Pau Bofill, Shoji Makino, and Justinian P. Rosca, "First stereo audio source separation evaluation campaign: data, algorithms and results," in *Proceedings of the 7th international conference on Independent component analysis and signal separation*, Berlin, Heidelberg, 2007, ICA'07, pp. 552–559, Springer-Verlag.