EXEMPLAR-BASED JOINT CHANNEL AND NOISE COMPENSATION

Jort F. Gemmeke[⋆]*, Tuomas Virtanen*[†]*, Kris Demuynck*[‡]

[⋆] KU Leuven, Kasteelpark Arenberg 10, 3001 Heverlee, Belgium
[†] Tampere University of Technology, P.O. Box 553, FI-33101 Tampere, Finland
[‡] Ghent University, MultimediaLab, Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium

jgemmeke@amadana.nl,tuomas.virtanen@tut.fi,kris.demuynck@ugent.be

ABSTRACT

In this paper two models for channel estimation in exemplar-based noise robust speech recognition are proposed. Building on a compositional model that models noisy speech and a combination of noise and speech atoms, the first model iteratively estimates a filter to best compensate the mismatch with the observed noisy speech. The second model estimates separate filters for the noise and speech atoms. We show that both models enable noise-robust ASR even if the channel characteristics of the noisy speech do not match those of the exemplars in the dictionary. Moreover, the second model, which is able to estimate separate filters for speech and noise, is shown to be robust even in the presence of bandwidth-limited sources.

Index Terms— Speech recognition, source separation, matrix factorization, noise robustness, channel compensation

1. INTRODUCTION

Compositional models for speech, i.e. models which describe the magnitude spectra of complex sounds as being composed of a purely additive (no negative components) combinations of spectral atoms, have proven to be adept at separating the target speech from interfering sounds such as noise [1, 2], other speakers [3, 4], music [5, 6, 7] and even reverberation [8]. For noise-robust automatic speech recognition (ASR), such compositional models really excel when the atoms also have some temporal extent [9, 10]. Given that speech is primarily a dynamic process, the importance of modeling the temporal dynamics is not unexpected and has been observed with other techniques such as deep belief networks [11] as well.

In this paper, we incorporate a channel estimate into our compositional speech model, leading to a comprehensive framework that copes with both additive and convolutive noise while still being able to capitalize on the dynamic nature of speech and noise by using long spectral-temporal atoms. The method can also provide robustness if the spectral-temporal patches are zeroed out by the channel (bandwidth limitation) or were completely masked by noise. The proposed modifications renders the compositional model better adept at handling the diverse conditions encountered in real-world applications.

In ASR, the most frequently used technique to handle channel mismatches is cepstral mean subtraction, which operates independently from techniques used to cope with the additive noise. In multipass systems, adaptation schemes such as (feature space) maximum likelihood linear regression in the subsequent stages help to overcome some of the inaccuracies introduced by not having an integrated noise model. Predictive schemes such as vector Taylor series [12] compensation on the other hand, do model the joint effect that additive and convolutive noise have on the acoustic model. This allows them to calculate proper model transformations based on a very compact set of parameters that must be estimated. The inherent complexity of both adaptive and predictive model compensation however, makes it very difficult for these techniques to exploit the temporal structure in both speech and noise.

The compositional model proposed in this paper not only has the built-in capability to leverage both spectral and temporal information, it can also estimate both the speech and the noise channel and can be configured to fill in missing spectral components without requiring multiple recognition passes. The techniques needed to build a noise-robust system based on the compositional model have been explored in previous work. In [2], we proposed the use of speech and noise exemplars as dictionary atoms and showed that it is possible to directly map the sparse weights of the clean speech atoms to state posteriors. In a feature enhancement method [9], the compositional model is used to obtain clean speech and noise estimates which are in turn used to define a Wiener filter. Finally, in the hybrid approach [13, 14], both schemes are combined. In [15], we show how these techniques can be successfully applied to create a noise-robust large vocabulary speech recognition system.

The remainder of this paper is organized as follows. First, the compositional model without channel components and the different strategies to employ this model for noise-robust ASR are presented. Next, the model is extended to include channel components and we propose adapted decoding. The system is evaluated on the AURORA-2 database, including a modified testset C which allows us to evaluate the capabilities of the models to fill in missing spectral components. Finally, we draw conclusions and lay out future work.

2. BACKGROUND

2.1. Noisy speech

Convolutive noise (i.e., channel mismatch) can be approximated by point-wise multiplication in spectral domain, and additive noise can be approximate as the summation of speech and noise magnitude spectra. In this work, we assume the following linear model for Mel-magnitude representations of (possibly filtered) speech corrupted with additive noise:

The research of Jort F. Gemmeke was funded by IWT-SBO project ALADIN contract 100049. The research of T. Virtanen is funded by the Academy of Finland, grant 258708. We acknowledge Hugo Van hamme for helpful discussions.

$$\mathbf{Y} = \mathbf{L}_{\mathrm{s}}\mathbf{S} + \mathbf{L}_{\mathrm{n}}\mathbf{N} \quad (1)$$

with $\mathbf{Y}$ the noisy speech, $\mathbf{S}$ the underlying clean speech and $\mathbf{N}$ the underlying noise — these matrices are of dimension $F \times T$, where F is the number of Mel bands and T is the number of consecutive time frames. $\mathbf{L}_{\mathrm{s}}$ and $\mathbf{L}_{\mathrm{n}}$ represent the Mel-spectral channel filters of the speech and noise source, respectively. In this paper, we restrict ourselves to linear transfer functions (no intermodulation), so the filters are expressed as diagonal matrices of dimension $F \times F$ with the frequency filter coefficients on the diagonal.

2.2. Compositional model for noise-robust ASR

The compositional model for noise-robust ASR is based on representing the observed noisy speech as a linear combination of speech and noise *atoms*. The atoms used for modeling noisy observations are $F \times T_{\mathrm{W}}$ magnitude spectrogram segments, with T_{W} the number of consecutive time frames in an atom. The collection of speech and noise atoms form a *dictionary*. This model is also referred to as Non-negative Matrix Factorization (NMF).

In this work speech and noise atoms are formed by *exemplars*, spectrogram segments extracted from a set of training utterances. For now, we will assume that the speech and noise in the dictionary are not filtered - e.g., captured by a microphone with a flat response. Although in our experiments we use $T_{\mathrm{W}} = 30$, in the remainder of this section we proceed with $T_{\mathrm{W}} = 1$ to simplify the notation. We refer to [9] for a discussion of the use of $T_{\mathrm{W}} > 1$. Consider the model

$$\mathbf{Y} \approx \mathbf{\Psi} = \hat{\mathbf{S}} + \hat{\mathbf{N}} = [\mathbf{A}_{\mathrm{s}}\ \mathbf{A}_{\mathrm{n}}] \begin{bmatrix} \mathbf{X}_{\mathrm{s}} \\ \mathbf{X}_{\mathrm{n}} \end{bmatrix} = \mathbf{A}\mathbf{X} \quad \text{s.t.} \quad \mathbf{X} \geq 0 \quad (2)$$

with the columns of $\mathbf{X}_{\mathrm{s}}$ and $\mathbf{X}_{\mathrm{n}}$ representing sparse linear combinations of the speech and noise dictionaries $\mathbf{A}_{\mathrm{s}}$ and $\mathbf{A}_{\mathrm{n}}$, respectively. The dictionary consists of $J = L + K$ atoms, with L the number of speech exemplars and K the number of noise exemplars. Accordingly, the speech and noise weights $\mathbf{X}_{\mathrm{s}}$ and $\mathbf{X}_{\mathrm{n}}$ are of dimensions $L \times T$ and $K \times T$, respectively.

The weights are obtained by minimizing the Kullback-Leibler (KL) divergence between $\mathbf{Y}$ and $\mathbf{A}\mathbf{X}$. To promote sparsity in $\mathbf{X}$, its entries are weighted element-wise by a penalty $\mathbf{\Lambda}$ [9]. The cost function is minimized by first initializing the activations $\mathbf{X}$ to unity, and then iteratively applying the update rule [9, 16]

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \frac{\mathbf{A}^{\mathrm{T}}\frac{\mathbf{Y}}{\mathbf{\Psi}}}{\mathbf{A}^{\mathrm{T}}\mathbf{1} + \mathbf{\Lambda}} \quad (3)$$

with $\mathbf{1}$ an all-one matrix having dimensions $F \times T$ and $\mathbf{\Lambda}$ a matrix of dimensions $J \times T$ containing the element-wise sparsity penalties. Both the multiplication $\otimes$ and division $\div$ operate element-wise.

2.3. Noise-robust ASR

Once a noise-robust sparse representation is obtained, several options exist to do noise-robust ASR. The Sparse Classification (SC) method [9], associates each frame in the speech dictionary with the corresponding HMM-state found with a forced Viterbi alignment on the training data from which the exemplars were drawn. The weights assigned to the speech exemplars are then used directly to estimate the HMM-state posterior probabilities.

Alternatively, the compositional model can form the base of a feature enhancement (FE) scheme. Given the weights $\mathbf{X}_{\mathrm{s}}$, an initial clean speech estimate can be made: $\hat{\mathbf{S}} = \mathbf{A}_{\mathrm{s}}\mathbf{X}_{\mathrm{s}}$. The estimate $\hat{\mathbf{S}}$ can be improved by Wiener filtering the noisy speech using the following equation, [4],

$$\tilde{\mathbf{S}} = \frac{\hat{\mathbf{S}}}{\hat{\mathbf{S}} + \hat{\mathbf{N}}} \otimes \mathbf{Y} \quad (4)$$

A last option we investigate in this paper is a hybrid combination of the posteriors estimated by SC and those obtained by applying a conventional Gaussian Mixture Model (GMM) to the enhanced features $\tilde{\mathbf{S}}$. The posteriors are combined by multiplication using the method proposed in [13].

3. CHANNEL COMPENSATION

In this section we propose two models to add channel compensation to our compositional model. The first model estimates a single Mel-frequency filter which compensates the noisy observation, while the second method estimates two filters, for the speech and noise part of the exemplar dictionary, respectively.

3.1. Model one: filtered observation

Under the assumption that $\mathbf{L}_{\mathrm{s}} = \mathbf{L}_{\mathrm{n}}$ in (1), we can model a channel mismatch between the speech and noise sources in the dictionary $\mathbf{A}$ and the observed noisy speech $\mathbf{Y}$, as

$$\mathbf{H}_{\mathrm{o}}\mathbf{Y} \approx \mathbf{A}\mathbf{X} \quad \text{s.t.} \quad \mathbf{X} \geq 0 \quad (5)$$

with $\mathbf{H}_{\mathrm{o}}$ again a diagonal, square filter matrix. We estimate $\mathbf{H}_{\mathrm{o}}$ by minimizing the KL-divergence between $\mathbf{H}_{\mathrm{o}}\mathbf{Y}$ and the reconstruction $\mathbf{A}\mathbf{X}$, using the current estimate of $\mathbf{X}$:

$$\mathbf{H}_{\mathrm{o}} = \mathbf{I} \otimes \exp\left(\frac{\log(\frac{\mathbf{\Psi}}{\mathbf{Y}})\mathbf{Y}^{\mathrm{T}}}{\mathbf{Y}^{\mathrm{T}}\mathbf{1}}\right) \quad (6)$$

with $\mathbf{I}$ an $F \times F$ dimensional identity matrix and $\mathbf{1}$ an all-one matrix having dimensions $T \times F$. Equation (6) is derived by setting the gradient of the KL-divergence w.r.t. $\mathbf{H}_{\mathrm{o}}$ to zero. We alternate the update (6) with the update of $\mathbf{X}$ in (3). In the case of using atoms spanning multiple time frames, we average the second term of (6) over the consecutive time frames within segments.

Because alternatively updating (3) and (6) can result in arbitrary scale factors, we constrain the filter weights to average to one in each iteration:

$$\mathbf{H}_{\mathrm{o}} \leftarrow \mathbf{H}_{\mathrm{o}}\frac{F}{||\operatorname{diag}(\mathbf{H}_{\mathrm{o}})||_1} \quad (7)$$

with diag extracting the diagonal elements and $||\ \ ||_1$ representing the L_1 norm.

3.2. Model two: filtered dictionary

If $\mathbf{L}_{\mathrm{s}} \neq \mathbf{L}_{\mathrm{n}}$ in (1), or alternatively, if the channel characteristics of the speech and noise sources in the dictionary $\mathbf{A}$ are not the same, we can estimate Mel-magnitude filters for both the speech and noise components of the dictionary:

$$\mathbf{Y} \approx \mathbf{\Psi}_2 = \hat{\mathbf{S}} + \hat{\mathbf{N}} = [\mathbf{H}_{\mathrm{s}}\mathbf{A}_{\mathrm{s}}\ \ \mathbf{H}_{\mathrm{n}}\mathbf{A}_{\mathrm{n}}]\,\mathbf{X} \quad \text{s.t.} \quad \mathbf{X} \geq 0 \quad (8)$$

Table 1: Accuracy obtained on AURORA-2 testset C as a function of SNR for baseline methods. The performance on corresponding data from testset A and B are also shown.

method		*test set C* clean	5	-5	Avg 20-0
MFCC baseline		99.8	78.2	22.9	83.9
NMF baseline	SC	96.6	87.0	40.3	88.0
	FE	99.8	93.8	52.2	94.2
	Hybrid	99.8	94.0	51.4	94.0
		corresponding data set A+B			
MFCC baseline		99.8	78.5	19.7	84.6
NMF baseline	SC	97.5	92.5	61.1	93.1
	FE	99.8	94.8	64.6	94.0
	Hybrid	99.8	95.8	70.1	96.3

with $\mathbf{H}_{\mathrm{s}}$ and $\mathbf{H}_{\mathrm{n}}$ diagonal matrices representing the Mel-spectral filters of the speech and noise source, respectively.

In order to estimate $\mathbf{H}_{\mathrm{s}}$ and $\mathbf{H}_{\mathrm{n}}$, we alternate update rule (3) with the following two update rules

$$\mathbf{H}_{\mathrm{s}} \leftarrow \mathbf{H}_{\mathrm{s}} \otimes \frac{\frac{\mathbf{Y}}{\mathbf{\Psi}_2}\left(\mathbf{A}_{\mathrm{s}}\mathbf{X}_{\mathrm{s}}\right)^{\mathrm{T}}}{\mathbf{1}\left(\mathbf{A}_{\mathrm{s}}\mathbf{X}_{\mathrm{s}}\right)^{\mathrm{T}}} \tag{9}$$

$$\mathbf{H}_{\mathrm{n}} \leftarrow \mathbf{H}_{\mathrm{n}} \otimes \frac{\frac{\mathbf{Y}}{\mathbf{\Psi}_2}\left(\mathbf{A}_{\mathrm{n}}\mathbf{X}_{\mathrm{n}}\right)^{\mathrm{T}}}{\mathbf{1}\left(\mathbf{A}_{\mathrm{n}}\mathbf{X}_{\mathrm{n}}\right)^{\mathrm{T}}} \tag{10}$$

with $\mathbf{1}$ is an all-one matrix having dimensions $T \times F$. The filters $\mathbf{H}_{\mathrm{s}}$ and $\mathbf{H}_{\mathrm{n}}$ are initialized as the identity matrix. As in Section 3.1, the filter updates are followed by a normalization (cf. (7)).

In practice, we obtain better results if we normalize the rows of the filtered dictionary $[\mathbf{H}_{\mathrm{s}}\mathbf{A}_{\mathrm{s}}\ \mathbf{H}_{\mathrm{n}}\mathbf{A}_{\mathrm{n}}]$ to equal Euclidean norm, normalize the rows of $\mathbf{Y}$ by the same factor, and finally normalize the columns of the resulting dictionary to unity Euclidean norm after updating the filters. These are the same normalizations that are carried out in the baseline framework prior to optimization [9, 13].

3.3. Feature enhancement

When using the proposed models, the SC method described in Section 2.3 can be used without any change as it operates directly on the sparse representations $\mathbf{X}_{\mathrm{s}}$. The FE method, however, needs to be modified to account for the estimated filters. Under the assumption that the filter characteristics of the acoustic model used to recognize the enhanced features matches the speech in the exemplar dictionary, a general formulation for the use of dictionary or observation filtering is

$$\tilde{\mathbf{S}} = \mathbf{H}_{\mathrm{s}}^{-1}\frac{\mathbf{H}_{\mathrm{s}}\mathbf{A}_{\mathrm{s}}\mathbf{X}_{\mathrm{s}}}{[\mathbf{H}_{\mathrm{s}}\mathbf{A}_{\mathrm{s}}\ \ \mathbf{H}_{\mathrm{n}}\mathbf{A}_{\mathrm{n}}]\,\mathbf{X}} \otimes \mathbf{H}_{\mathrm{o}}\mathbf{Y} = \frac{\hat{\mathbf{S}}}{\hat{\mathbf{S}}+\hat{\mathbf{N}}} \otimes \mathbf{H}_{\mathrm{o}}\mathbf{Y} \tag{11}$$

with either $\mathbf{H}_{\mathrm{o}}$ or $\mathbf{H}_{\mathrm{s}}$ and $\mathbf{H}_{\mathrm{n}}$ the identity matrix when not estimated.

4. EXPERIMENTAL SETUP

4.1. Noisy speech material

We used the AURORA-2 database to evaluate our approach. AURORA-2 provides three testsets, described in detail in [17]. In short, testsets A and B contain four different kinds of noise, while testset C (a subset of both A&B) is filtered with a different channel characteristic than the A & B testsets and the training sets. AURORA-2 provides a clean and a multi-condition training set: we use the clean training set to train the acoustic models of the HMM-based speech recognizer and we use the speech and noise samples underlying the multi-condition training set to populate our clean and noise dictionaries, respectively.

Table 2: Accuracy obtained on AURORA-2 testset C as a function of SNR with the proposed methods.

method		SNR clean	5	-5	Avg 20-0
FOBS oracle	SC	97.3	92.0	59.0	91.8
	FE	99.8	94.0	61.6	94.9
	Hybrid	99.8	95.7	67.0	95.7
FDICT oracle	SC	97.2	91.7	59.2	91.9
	FE	99.8	94.8	61.6	94.9
	Hybrid	99.8	95.4	66.1	95.5
FOBS estimated	SC	98.0	93.0	55.9	93.5
	FE	99.8	95.1	62.1	95.5
	Hybrid	99.8	96.6	67.5	96.1
FDICT estimated	SC	98.5	93.4	58.9	93.6
	FE	99.8	93.1	53.1	93.8
	Hybrid	99.8	96.1	62.8	95.7

Our evaluation focuses on testset C. We use the same random, representative subset of 10% of the utterances (i.e. 200 utterances per SNR level) as used in [18, 13]. Additionally, we created a dataset with a bandwidth limiting filtering in the clean speech component of the testset C noisy speech by filtering the clean speech using a 10-point Butterworth low-pass filter with a cutoff frequency of 1.6Khz before adding the noise. In practice, this results in filtering away the top-5 Mel frequency bands of the clean speech, a scenario similar to some of the material in the AURORA-4 database. For all testsets, the results of the four noise types are averaged and we display the results for clean speech, 5 dB, -5 dB, and the average over the 20-0 dB range.

4.2. Speech recognition

The acoustic feature vectors used in the compositional model consisted of Mel frequency magnitude spectrograms, spanning $F = 23$ bands with a frame length of 25 ms and a frame shift of 10 ms. The dictionary consists of $L = 10000$ speech exemplars and $K = 4000$ noise exemplars, each spanning $T_{\mathrm{W}} = 30$ frames. Digits were described by 16 HMM states with an additional 3-state silence word, resulting in a 179-dimensional state-space. Each frame in each exemplar was annotated using an HMM-state label obtained through forced alignment with the canonical transcription obtained with the GMM-based recognizer also employed for use with FE.

FE and hybrid recognition experiments employed the GMM-based recognizer operating on MFCC features, derived from the Mel-spectral features used in the compositional model. Here, we used 13 static MFCC features along with their delta and delta-delta time derivatives resulting in a 39 dimensional feature vector. The MFCC features were mean and variance normalized on a per-utterance basis. The acoustic model consisted of 64 Gaussians with

Table 3: Accuracy obtained on AURORA-2 bandwidth-limited version of testset C as a function of SNR.

method		SNR clean	5	-5	Avg 20-0
MFCC baseline		95.2	56.2	16.8	66.0
NMF	SC	59.5	74.5	30.2	73.2
	FE	92.5	77.0	37.9	80.2
	Hybrid	83.1	79.6	41.5	80.3
FOBS	SC	88.9	79.3	46.5	80.3
	FE	94.0	78.8	49.0	84.2
	Hybrid	97.6	81.9	54.0	85.1
FDICT	SC	93.8	88.0	49.5	88.8
	FE	95.1	84.1	43.9	87.4
	Hybrid	98.1	90.2	55.3	91.2

diagonal covariance per HMM-state. In the hybrid system, the GMM probabilities (based on the FE stream) were raised to the power 0.33 prior to combining with the SC probability estimates as described in [13].

4.3. Implementation

The SC and FE speech decoding systems were implemented in MATLAB; we refer the reader to [9, 18, 13] for further implementation details. For optimization, we used 600 iterations of (3). We used a 'burnin' period of 5 iterations in which we only use update rule (3). Since the filters converge much faster than (3), we stop updating the filters after a 'burnout' period of 200 iterations. Other settings were taken from the best performing system in [13]: a speech and noise sparsity of 1.5 and 1, respectively, and the use of two additional noise dictionaries: an artificial noise dictionary [19] and one based on noise sniffing [18]. Note that the noise sniffing dictionary is based on $\mathbf{Y}$, and as such should not be filtered with $\mathbf{H}_{\mathrm{n}}$.

5. EXPERIMENTS AND RESULTS

In Table 1, we display the performance of the baseline MFCC recognizer operating directly on the noisy speech, as well as the previously proposed exemplar-based NMF framework described in Section 2. We can observe that for the MFCC baseline recognizer the channel characteristic has little impact on the recognition accuracy. However, for the exemplar-based methods, referred to as "NMF", the performance decreases substantially at lower SNRs. The fact that the SC method has the largest decrease in accuracy when going from matched to mismatched channel characteristics, from 61.1% to 40.3% at SNR -5 dB, shows that in the mismatched condition the selected exemplars not only are spectrally dissimilar, but also associated with incorrect acoustic states. Even though on testset A+B the hybrid recognition achieves the best results, on testset C the performance is drops below that of FE due to the low accuracy of SC.

In Table 2, we display the testset C performance of the two proposed models namely filtered observations ("FOBS", (5)) and filtered dictionaries ("FDICT", (8)). The top two panels show 'oracle' results in which the observation or dictionary filters are not updated, but are initialized using the correct filters, i.e, the ratio between the testset C channel characteristic (MIRS) and the train&testset A/B channel characteristic (G.712). In the bottom two panels the performance of the FOBS and FDICT models are shown with estimated filters.

From the oracle results it is clear that both models can compensate most of the channel mismatch introduced in testset C. For example for SC the accuracy increased from 40.3% at SNR -5 dB to 59.2% for FDICT - only slightly lower than the 61.1% obtained on the corresponding testset A+B data. With FOBS the estimated filters achieve similar performance as when using oracle filters, and even better than oracle performance for SC. This is probably due to the filter compensating not only for the channel characteristic mismatch, but also for the train-testing mismatch. When using the FDICT model the oracle performance is not achieved due to a lower FE performance. This might be due to the fact that now two filters need to be estimated rather than one. Still, with both models the estimated filters achieve a substantial increase in noise robustness over the baseline system.

Finally, Table 3 shows results for the modified testset C containing the bandwidth-limited speech. It is clear that the missing frequency bands greatly affect the MFCC baseline recognizer even in the absence of corrupting noise, reducing the accuracy from 99.8% to 95.2% on clean speech. At lower SNRs, the decrease is even more pronounced because the zero-out frequency bands get filled with noise. The NMF baseline is also greatly affected: the clean speech accuracies drop even below the MFCC baseline. The channel estimation methods FOBS and FDICT substantially improve the channel and noise robustness over the NMF baseline. Although both SC and FE yield lower accuracies on clean speech, the hybrid SC/FE systems do outperform the MFCC baseline.

Moreover, the FDICT model, which is able to estimate separate filters for speech and noise performs indeed better than FOBS which only estimates a single observation filter that cannot fully capture the different channel characteristics of the underlying clean speech and noise sources. In fact, inspection of the enhanced features produced by FE showed that at lower SNRs, the model actually performs *bandwidth extension*: it estimates clean speech estimates of the missing frequency bands by filtering the noise energy. At high SNRs, this does not occur due to the multiplicative nature of the Wiener filter in combination with the lack of energy in the zeroed out frequency bands. To use the model for bandwidth extension at higher SNRs we could use the clean speech estimate $\hat{\mathbf{S}}$ without or in combination with Wiener filtering.

6. CONCLUSIONS AND FUTURE WORK

In this paper two models for channel estimation in exemplar-based noise-robust speech recognition were proposed. We showed both models enable noise-robust ASR even if the channel characteristics of the noisy speech do not match those of the exemplars in the dictionary. Moreover, the second model, which is able to estimate separate filters for speech and noise, was shown to be robust if bandwidth-limited speech is combined with fullband noise.

Future work consist of evaluation on tasks like AURORA-4, a noisy speech database which is known to exhibit mismatching channel characteristics. Another topic of future work is joint speech dereverberation and denoising: Due to use of exemplars spanning multiple time frames, the proposed models are able to estimate time-varying filters with minor modifications.

7. REFERENCES

[1] B. Schuller, F. Weninger, M. Wllmer, Y. Sun, and G. Rigoll, "Non-negative matrix factorization as noise-robust feature extractor for speech recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2010.

[2] J. F. Gemmeke and T. Virtanen, "Noise robust exemplar-based connected digit recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2010.

[3] B. Raj and P. Smaragdis, "Latent variable decomposition of spectrograms for single channel speaker separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005.

[4] P. Smaragdis, M. Shashanka, and B. Raj, "A sparse non-parametric approach for single channel separation of known sounds," in *Proc. Neural Information Processing Systems*, 2009.

[5] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, 2007.

[6] 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.

[7] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. M. E. Davies, "Sparse representations in audio & music: from coding tosource separation," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 995–1005, 2009.

[8] K. Kumar, R. Singh, B. Raj, and R. M. Stern, "Gammatone sub-band magnitude-domain dereverberation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2011.

[9] 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, 2011.

[10] T. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, D. Van Compernolle, K. Demuynck, J. Gemmeke, J. Bellegarda, and S. Sundaram, "Exemplar-based processing for speech recognition: An overview," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 98 –113, nov. 2012.

[11] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82 –97, nov. 2012.

[12] A. Acero, L. Deng, T. T. Kristjansson, and J. Zhang, "Hmm adaptation using vector taylor series for noisy speech recognition," in *Proc. ICSLP*, 2000, pp. 869–872.

[13] J. F. Gemmeke and H. Van hamme, "Advances in noise robust digit recognition using hybrid exemplar-based techniques," in *Proc. INTERSPEECH*, 2012.

[14] F. Weninger, M. Wöllmer, J. Geiger, B. Schuller, J. Gemmeke, A. Hurmalainen, T. Virtanen, and G. Rigoll, "Non-negative matrix factorization for highly noise-robust ASR: to enhance or to recognize," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, March 2012.

[15] H. Kallasjoki, U. Remes, J. F. Gemmeke, T. Virtanen, and K. J. Palomäki, "Uncertainty measures for improving exemplar-based source separation," in *Proc. INTERSPEECH*, 2011.

[16] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Neural Information Processing Systems*, April 2001, pp. 556–562.

[17] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. of ISCA ASR2000 Workshop, Paris, France*, 2000, pp. 181–188.

[18] J. F. Gemmeke, A. Hurmalainen, T. Virtanen, and Y. Sun, "Toward a practical implementation of exemplar-based noise robust ASR," in *Proc. EUSIPCO*, 2011, pp. 1490–1494.

[19] J. F. Gemmeke and T. Virtanen, "Artificial and online acquired noise dictionaries for noise robust ASR," in *Proc. INTERSPEECH*, 2010, pp. 2082–2085.