HMM-REGULARIZATION FOR NMF-BASED NOISE ROBUST ASR

Jort F. Gemmeke[1]*, Antti Hurmalainen*[2]*, Tuomas Virtanen*[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 we extend a previously proposed NMF-based technique for speech enhancement of noisy speech to exploit a Hidden Markov Model (HMM). The NMF-based technique works by finding a sparse representation of specrogram segments of 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. In order to take into account larger temporal context and constrain the representation by the grammar of a speech recognizer, we propose to regularize the optimization problem by additionally minimizing the distance between state emission probabilities derived from the speech exemplar activations, and a posteriori state probabilities derived by applying the Forward-Backward algorithm to the emission probabilities. Experiments on Track 1 of the 2nd CHiME Challenge, which contains small vocabulary 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.
Index Terms: speech enhancement, exemplar-based, noise robustness, Non-Negative Matrix Factorization, Hidden Markov Models

1. INTRODUCTION

These days, there is an increasing attention for Automatic Speech Recognition (ASR), in no small part thanks to applications such as voice search on smart-phones, navigation and home automation. Consequentially, there is a growing demand for *robust* ASR: ASR which functions adequately even in noisy, reverberant environments. Unfortunately, the performance of conventional ASR systems degrades rapidly when the speech signal is corrupted by noise.

The reason why ASR performance drops with increasing levels of noise, is that the observed acoustic features no longer match the acoustic models learned during training. All robust ASR approaches proposed over the last few decades aim to resolve this mismatch. 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 [5], but their performance on modeling unseen noise environments is limited. For an overview, we refer the reader to [6].

A relatively new method based on compositional models for speech, i.e. models which describe the magnitude spectra of complex sounds as being composed of a purely additive combinations of spectral atoms, has proven to be adept at separating the target speech from interfering sounds such as noise [7, 8], other speakers [9, 10], music [11–13] and even reverberation [14]. For noise-robust automatic speech recognition (ASR), such compositional models really excel when the atoms also model temporal context [15, 16]. Due to its close relation to Non-negative Matrix Factorization, we will refer to this class of models as 'NMF'. In NMF, we obtain the weights with which the dictionary atoms combine by minimizing the distance between the representation and the weighted sum of dictionary atoms.

In this work, we propose using even more knowledge of the target speech, in the form of Hidden Markov Models (HMMs). Although HMMs are the cornerstone of virtually every speech recognition system, its use in noise robust ASR techniques based on speech enhancement is limited. For the compositional model, recently a model called the Non-negative HMM (NHMM) was proposed [17]. In this model, there are multiple dictionaries, each corresponding to a specific HMM state, and an observation is modeled as a weighted sum of dictionary reconstructions – each of which is in turn modeled by a weighted sum of dictionary atoms. A HMM transition matrix governs the sequence of dictionaries which can be activated.

In the NHMM, the HMM is learned from data, and for applications such as source separation a factorial HMM is used. While shown to be effective a speaker separation task [17], the approach has some drawbacks for an application to noise robust ASR: First, the use of a factorial HMM leads to an exponential increase in computational complexity which requires solutions such as approximate inference [18]. Moreover, the use of a factorial HMM requires detailed knowledge of the corrupting noise - which is often not available, leading to poor generalization. Finally, as pointed out already in [17], finding the weights of the HMM state posteriors converges much faster than finding the weights with which dictionary atoms combine. As a result, NMF-based methods, which typically need hundreds of iterations to converge, tend to converge to sub-optimal state weights too quickly.

As an alternative perhaps more suited to noise robust ASR, we propose the use of HMM-regularized NMF, an extension of NMF in which the cost function to be minimized is augmented with a term that describes the distance between state emission probabilities, and the a posteriori estimates obtained after applying the Forward-Backward algorithm. Building on our earlier work on exemplar-based speech recognition, we use dictionaries with speech and noise atoms extracted directly from the training data (the exemplars). The speech exemplars are associated with HMM states through a forced alignment of the reverberated training data, and the dictionary atom activations serve as evidence for the states corresponding with the atoms. We then use the HMM of the back-end speech recognizer to obtain a posteriori state estimates.

This approach offers several advantages. Since in this framework the noise exemplars are not constrained by an HMM, we avoid the exponential complexity of the factorial HMM approach. Also, the method extends to atoms modeling multiple states provided the calculation of state emissions is defined accordingly. Finally, since

The research of Jort F. Gemmeke was funded by IWT-SBO project ALADIN contract 100049. Tuomas Virtanen has been funded by the Academy of Finland, grant #258708.

the HMM contribution is a expressed as a regularizer, we can gracefully tune its influence using a regularization weight. In comparison to another recently introduced method that regularizes NMF by an HMM [19], we propose a simpler solution which allows formulating the algorithm as matrix operations, yielding a more efficient implementation.

We evaluate the performance of the proposed speech enhancement method using Track 1 of the 2nd CHiME challenge [20]. Rather than trying to achieve the best possible recognition accuracies, we aim at exploring to what extend the use of HMM regularization can improve a speech enhancement approach to noise robust ASR working with unadapted acoustic models.

The rest of the paper is organized as follows. The NMF-based speech enhancement method is described in Section 2. The proposed HMM-regularization approach is outlined in Section 3. The experimental setup, such as a description of the 2nd CHiME challenge data, the implementation details of the speech enhancement technique and the speech recognition system are described in Section 4. Automatic speech recognition results with different distance measures, exemplar sizes, and regularization weights are presented in Section 5. We finish with a general discussion and our conclusions in Section 6.

2. NMF-BASED NOISE ROBUST ASR

In this section we briefly describe the baseline system of [21], in order to make the paper self-contained and to introduce the necessary notation. For brevity, and due to its relation to Non-negative Matrix Factorization (NMF), we will refer to this method as 'NMF'.

2.1. Exemplar-based representation of noisy speech

The noise robust ASR technique operates on fixed-size magnitude Mel-spectra, each a $B \times T$ noisy speech spectrogram matrix, with B Mel-frequency bands and T time frames. We assume noisy speech is a linear addition of underlying clean speech and noise magnitude spectrograms. To simplify the notation, the time frames (columns) of each spectrogram are stacked into the noisy speech, clean speech and noise speech 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 modeled 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 decode utterances of arbitrary lengths, we adopt a sliding time window. In this approach, we represent a noisy utterance as W fixed-size, overlapping speech segments, each of length T. We write the analog of (4) as the NMF problem $\mathbf{Y} \approx \mathbf{A}\mathbf{X}$, with $\mathbf{Y}$ a matrix with the multiple $\mathbf{y}$ vectors as it columns and $\mathbf{X}$ containing the corresponding $\mathbf{x}$ sparse representations as its columns.

2.2. Finding exemplar weights

In order to obtain $\mathbf{X}$, we minimize the cost function:

$$d_{KL}(\mathbf{Y}, \mathbf{A}\mathbf{X}) + ||\mathbf{\Lambda} \otimes \mathbf{X}||_1 \quad \text{s.t.,} \quad \mathbf{X} \geq 0 \tag{5}$$

where d_{KL} 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 $\otimes$) with the sparsity penalty matrix $\mathbf{\Lambda}$, defined for each activation entry. In this work, speech and noise exemplars are weighted differently, but the weights are otherwise the same for all exemplars and observations.

The cost function (5) is minimized using a multiplicative updates routine:

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \frac{\mathbf{A}^{\mathrm{T}}(\frac{\mathbf{Y}}{\mathbf{A}\mathbf{X}})}{\mathbf{A}^{\mathrm{T}}\mathbf{1} + \mathbf{\Lambda}}. \tag{6}$$

Here, matrix divisions are element-wise. $\mathbf{1}$ is an all-ones matrix of dimensions $D \times W$.

2.3. Speech enhancement

For maximum independence of the speech recognizer, we will use a *speech enhancement* approach. 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

$$\tilde{\mathbf{s}}_t = \sum_{j=1}^{J} x^s_j \mathbf{a}^{\mathrm{s}}_{j,t}. \tag{7}$$

$$\tilde{\mathbf{n}}_t = \sum_{k=1}^{K} x^{\mathrm{n}}_k \mathbf{a}^{\mathrm{n}}_{k,t}. \tag{8}$$

For each noisy speech segment (column of matrix $\mathbf{Y}$), 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.

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

$$\mathbf{h}_t = \frac{\mathbf{B}^{\dagger}\hat{\mathbf{s}}_t}{\mathbf{B}^{\dagger}\hat{\mathbf{s}}_t + \mathbf{B}^{\dagger}\hat{\mathbf{n}}_t}, \tag{9}$$

with the matrix division acting element-wise. $\mathbf{B}^{\dagger}$ denotes the MoorePenrose pseudo-inverse of the Mel-matrix $\mathbf{B}$, which maps the Mel magnitude vectors to the DFT domain

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 together with the original phases. Frames are combined using weighted overlap-add to obtain the whole enhanced signal.

3. HMM-REGULARIZED NMF

In order to constrain the speech exemplar activations to those more likely to conform to a Hidden Markov Model (HMM) of speech, we propose regularizing the NMF cost function (5) with the distance between state emission probabilities corresponding to the representation and a posteriori state estimates, obtained by using the Forward-Backward algorithm. The Forward-Backward produces non-zero probabilities only for those state sequences that the HMM is able to generate, and higher probabilities for more likely state sequences.

3.1. State probability calculation

We create a $Q \times W$ *emission probability matrix* $\mathbf{P}$, where each entry $\mathbf{P}_{qw}$ denotes the emission probability of HMM-state q $(1 \dots Q)$ in window w $(1 \dots W)$. The emission probabilities are being generated by the linear model [15]

$$\mathbf{P} = [\mathbf{M}^{\mathrm{s}}\mathbf{M}^{\mathrm{n}}]\,\mathbf{X} = \mathbf{M}\mathbf{X} \tag{10}$$

with $\mathbf{M}^{\mathrm{s}}$ and $\mathbf{M}^{\mathrm{n}}$ mapping speech and noise exemplars to states, respectively. The application of (10) is followed by normalizing the sum over states within each window to unity.

In this work, $\mathbf{M}^{\mathrm{n}}$ is an all-zero matrix of dimension $Q \times K$. $\mathbf{M}^{\mathrm{n}}$ is created by counting for each exemplar in the dictionary, in how many frames each state is activated according to a forced alignment with a GMM/HMM-based recognizer. Note that this constitutes a departure from our recent work, in which we modeled the state activations of frames within exemplars, and a return to the original formulation in [8]. This is necessary in order to have a time dimensionality which matches the spectral representation that is being optimized.

The a posteriori state matrix $\hat{\mathbf{P}}$ is then constructed by applying the Forward-Backward algorithm to the emission probabilities. The initial state and transition probabilities are equal to the word-based HMM speech recognizer back-end. Given the strict grammar of the CHiME/GRID task (cf. Section 4), a single large transition matrix was constructed which encodes state transitions both within words and between words. In order to simplify the calculations, the columns of $\hat{\mathbf{P}}$ are normalized to sum to unity.

3.2. Finding exemplar weights

In order to obtain $\mathbf{X}$, we propose minimizing the cost function:

$$d_{KL}(\mathbf{Y}, \mathbf{A}\mathbf{X}) + ||\mathbf{\Lambda} \otimes \mathbf{X}||_1 + d(\hat{\mathbf{P}}, \mathbf{P}) \quad \text{s.t.,} \quad \mathbf{X} \geq 0 \tag{11}$$

with $d(\hat{\mathbf{P}}, \mathbf{P})$ either the KL divergence or the Euclidean distance between the state emission and a posterior probabilities.

We use iterative algorithms where $\mathbf{X}$ is initialized with ones and iteratively updated. In order to minimize the cost function (11) with a KL-divergence between state estimates, we propose using the multiplicative update:

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \frac{\mathbf{A}^{\mathrm{T}}\left(\frac{\mathbf{Y}}{\mathbf{A}\mathbf{X}}\right) + \kappa\left(\mathbf{M}^{\mathrm{T}}\left(\frac{\hat{\mathbf{P}}}{\mathbf{P}}\right)\right)}{\mathbf{A}^{\mathrm{T}}\mathbf{1} + \mathbf{\Lambda} + \kappa\left(\mathbf{M}^{\mathrm{T}}\mathbf{1}\right)}. \tag{12}$$

with κ a constant weighing the contribution of the HMM regularization. As before, the leftmost $\mathbf{1}$ is an all-ones matrix of dimensions $D \times W$, whereas the rightmost $\mathbf{1}$ is an all-ones matrix of dimensions $Q \times W$. After each iteration $\mathbf{P}$ and $\hat{\mathbf{P}}$ are recalculated as described in Section 3.1. Since $\hat{\mathbf{P}}$ changes every iteration, convergence may not be guaranteed, but in practice the update was stable.

For minimization of the cost function (11) with a Euclidean distance between state estimates, we propose using the multiplicative update:

$$\mathbf{X} \leftarrow \mathbf{X} \otimes \frac{\mathbf{A}^{\mathrm{T}}\left(\frac{\mathbf{Y}}{\mathbf{A}\mathbf{X}}\right) + \kappa\left(\mathbf{M}^{\mathrm{T}}\hat{\mathbf{P}}\right)}{\mathbf{A}^{\mathrm{T}}\mathbf{1} + \mathbf{\Lambda} + \kappa\left(\mathbf{M}^{\mathrm{T}}\mathbf{P}\right)}. \tag{13}$$

The matrix $\mathbf{1}$ is again an all-ones matrix of dimensions $D \times W$. As for the KL-divergence, after each iteration $\mathbf{P}$ and $\hat{\mathbf{P}}$ are recalculated as described in Section 3.1.

The multiplicative update rules (12) and (13) can be derived by multiplying $\mathbf{X}$ with the ratio between the negative and positive parts of the gradient [22].

Table 1: Speech recognition accuracies using the original noisy features and the enhanced speech using the baseline NMF system. The acoustic models of the recognizer are trained on clean speech, reverberated speech, and noisy speech, respectively. The best results for each system and each SNR are highlighted

SNR (dB)		-6	-3	0	3	6	9
Baseline	clean	11.83	12.33	16.50	17.50	21.75	23.50
	reverberated	32.08	36.33	50.33	64.00	75.08	**83.5**0
	noisy	**49.67**	**57.92**	**67.83**	**73.67**	**80.75**	82.67
NMF	clean	16.42	16.58	21.67	23.25	27.92	27.83
	reverberated	**68.00**	**72.25**	**80.92**	**86.75**	**89.08**	**90.50**
	noisy	62.58	68.17	74.58	78.75	81.92	83.25

4. EXPERIMENTAL SETUP

4.1. Database & speech recognition

Track 1 of the 2nd CHiME Speech Separation and Recognition Challenge [20] is based on the small vocabulary GRID corpus [23], 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 recognized 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 modeling/training, there are 500 reverberated utterances per speaker, and six hours of background noise data. The development and test sets consist of 600 utterances at each SNR level. The development, evaluation and training data are all available 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, however, we only consider monaural signal processing by averaging the two channels.

For speech recognition experiments, we used the HTK recognition setup provided by the CHiME challenge organizers. This setup includes three acoustic models, trained on clean, reverberated and noisy data respectively.

Table 2: Speech recognition accuracies as a function of regularization weight κ and distance measure used in HMM-regularisation. The exemplar sizes is $T = 20$. The reported accuracies are averages over SNRs $\{-6, ..., 9\}$ of the development set, using the reverberated baseline acoustic model. The best result for each distance measure size is highlighted. Dashed values were not evaluated.

κ	0	0.005	0.01	0.05	0.1	0.2	0.5	0.8	1	1.5	2
KL-divergence	81.25	81.43	81.54	**81.72**	81.58	-	80.24	-	76.24	-	-
Euclidean distance	81.25	-	81.26	-	81.43	81.57	**81.87**	81.86	81.85	81.51	81.18

Table 3: Speech recognition accuracies as a function of regularization weight κ, for exemplar sizes $T \in \{1, 5, 10, 20\}$. HMM-regularization uses Euclidean distance. The reported accuracies are averages over SNRs $\{-6, ..., 9\}$ of the development set, using the reverberated baseline acoustic model. The best result at each exemplar size is highlighted.

κ	0	0.01	0.1	0.5	1
$T = 1$	60.27	60.85	65.97	69.90	**70.15**
$T = 5$	73.60	73.61	75.01	**76.36**	75.82
$T = 10$	79.10	79.11	79.61	**80.24**	79.75
$T = 20$	81.25	81.26	81.43	**81.87**	81.85

4.2. Speech enhancement

The speech enhancement setup of the small vocabulary track employed methods first described in the 2011 CHiME workshop [24] and later refined to a form which is also used in this work [21]. A speech basis comprising 5000 exemplars was generated for each speaker by pseudo-random sampling of training data and selective reduction with word frequency equalization. A matching speaker-dependent basis was always used for factorization as the identity of target speakers was known. For a noise model, 5000 exemplars were sampled from the noise context of each test utterance individually. All factorization took place in a 40-band monaural Mel magnitude domain, where the bands were normalized by applying an equalization curve acquired from training speech. Temporally the model used 25 ms frames with 10 ms shift, and exemplar size (window length) of 20 frames. We do not retrain the acoustic models on the enhanced speech signals.

4.3. HMM-regularization

For the HMM-regularization, we used a 'burn-in' period of 20 iterations in which we use update rule (6), followed by 280 iterations of update rule (12) or (13).

5. EXPERIMENTS

5.1. Baseline & acoustic model

In the 2nd CHiME challenge, three baseline acoustic models are provided, corresponding to training the speech recognizer of clean, reverberated and noisy speech features. Since in this work, we do not consider retraining the acoustic model, we first evaluate the performance of the baseline NMF system with each acoustic model. We use an exemplar size of $T = 20$, the best performing setting in the first CHiME challenge [24].

The results in Table 1 show that on the unprocessed noisy features, the best results are obtained using the 'noisy' acoustic model, except for 9 dB SNR, for which the 'reverberated' acoustic model scores best. With the baseline NMF method, a substantial increase in performance is observed both at high and low SNRs. Interestingly, all the best results are now obtained with the reverberated model, due to the effectiveness of the front-end in reducing the mismatch with the reverberated speech used during training. In the remainder of this work, we will use the reverberated acoustic model to evaluate the effectiveness of the proposed NMF techniques.

5.2. HMM-regularization distance measure

In Section 3 we proposed two approaches for HMM-regularization: either measuring the mismatch between posterior and a posteriori state estimates using a KL-divergence, or using the Euclidean distance. In this experiment, we first investigate the effectiveness of the two distance measures for a large number of regularization weights κ.

From the results in Table 2 we can observe that both distance measures serve to improve the average performance when compared to the baseline NMF system (here indicate by $\kappa = 0$). Although the differences are small, it seems the Euclidean distance offers a better performance than the KL-divergence. More importantly, the Euclidean distance measure is far more robust against varying values of κ, which should improve generalization.

An explanation for the poorer performance of the KL-divergence can probably be found in the asymmetry of the KL-divergence: overestimates of small values in $\hat{\mathbf{P}}$ are not penalized as much as underestimates of larger values. With most of the probability mass of $\hat{\mathbf{P}}$ centered on a few states, however, this may not provide enough constraints for consistent improvements. In contrast, the Euclidean distance is symmetric and penalizes both under- and overestimates equally. In the remainder of this work, we will only report on experiments with the Euclidean distance measure.

Another observation we can make from Table 2 is that the overall improvement rather limited - from 81.25% average accuracy to 81.87% for $\kappa = 0.5$ with the Euclidean distance measure. This triggered us to investigate the performance of the proposed technique at other exemplar sizes than $T = 20$, detailed in the next Section.

5.3. Exemplar size

In order to test with multiple exemplar sizes, we must first establish whether the optimal regularization weight $\kappa = 0.5$ is valid for other exemplar sizes. To that end, in Table 3 we show the average accuracy as a function of exemplar size $T \in \{1, 5, 10, 20\}$ for various values of κ. From Table 3, we can conclude that HMM-regularization improves the results for all exemplar sizes, and that the optimal value for the regularization weight is fairly constant at $\kappa \approx 0.5$.

Moreover, we can observe that for small exemplar sizes, the benefit of HMM-regularization is far bigger than for larger exemplar sizes. To study this in more detail, we display the full development set results for each exemplar size with $\kappa = 0.5$ in the left panel of Table 4. Here, we can observe that for $T = 1$, the use of HMM-regularization improves the accuracy at -6 db SNR from 38.50% to 49.25%, comparable to the use a noisy acoustic model in the speech recognizer (which yields 49.67% at -6 dB SNR). Additionally, the

Table 4: Speech recognition accuracies using the original noisy features and the enhanced speech, for exemplar sizes $T \in \{1, 5, 10, 20\}$. The left table shows the results on the development set, and the right table the results on the test set. HMM-regularization uses Euclidean distance with a regularization weight of $\kappa = 0.5$.

(a) Development set

SNR (dB)		-6	-3	0	3	6	9
Baseline	reverberated	32.08	36.33	50.33	64.00	75.08	83.50
	noisy	49.67	57.92	67.83	73.67	80.75	82.67
$T = 1$	NMF	38.50	45.17	55.17	64.83	75.75	82.17
	NMF-HMM	49.25	55.42	67.50	76.50	83.75	87.00
$T = 5$	NMF	55.92	60.25	71.25	79.67	85.83	88.67
	NMF-HMM	58.17	65.08	74.67	82.42	88.08	89.75
$T = 10$	NMF	63.50	69.42	77.92	84.83	88.75	90.17
	NMF-HMM	63.92	71.83	79.33	86.08	89.50	90.75
$T = 20$	NMF	68.00	72.25	80.92	86.75	89.08	90.50
	NMF-HMM	69.08	73.58	81.50	87.33	89.42	90.33

(b) Test set

SNR (dB)		-6	-3	0	3	6	9
Baseline	reverberated	32.17	38.33	52.08	62.67	76.08	83.83
	noisy	49.33	58.67	67.50	75.08	78.83	82.92
$T = 1$	NMF	37.17	42.58	52.17	63.25	74.92	82.25
	NMF-HMM	49.67	55.17	66.50	76.08	82.58	88.83
$T = 5$	NMF	53.67	61.25	71.50	79.08	85.17	89.58
	NMF-HMM	58.25	66.42	74.75	83.17	87.50	91.25
$T = 10$	NMF	63.00	72.33	78.42	85.00	90.08	91.08
	NMF-HMM	64.42	73.08	80.42	85.75	90.42	92.25
$T = 20$	NMF	67.25	75.92	81.08	86.42	90.67	92.00
	NMF-HMM	67.00	77.00	81.83	87.00	91.17	92.42

accuracy at high SNRs also improved substantially, from around 83% to 87.00%.

At the same time, we can observe that for larger exemplar sizes the gains are diminishing. Further investigation showed that this is probably because most errors in the CHiME small vocabulary task are due to confusions between words, not between word-class (such as letters, digits, etc.). It can be observed that when using a long temporal context, the majority of the selected exemplars at a particular observation window correspond to the correct class. As such, the potential benefit of HMM-regularization in penalizing unlikely exemplar activation sequences is limited.

5.4. Test set evaluation

As a final experiment, we evaluate the baseline NMF system and the proposed HMM-regularized NMF approach on the test set for the exemplar sizes $T \in \{1, 5, 10, 20\}$. The results in the right panel of Table 4 shows the same trends as for the development set: The use of HMM-regularization virtually always improves the results, but the gain is much larger at smaller exemplar-sizes than for the best performing, large exemplar sizes.

6. DISCUSSION AND CONCLUSIONS

We can conclude that the HMM-regularization that was proposed in this paper is indeed effective, but much more so for small exemplar sizes than for the exemplar-sizes that are generally employed these days for NMF-based noise robust ASR. This once again underlines the importance in modeling long temporal context in noise robust ASR, but also shows that for this task, the long temporal context is adequately handled by the exemplars themselves.

This also puts to the question the effectiveness of related models which model individual time frames combined with a HMM [17, 19]. Given the computational demands of running Forward-Backward at every iteration — not investigated exhaustively in this paper but estimated to be as expensive as a regular NMF iteration with the current parameter settings —, the use of dictionary atoms that span a longer temporal context than a single frame seems an attractive alternative.

Interestingly, while an evaluation on the CHiME dataset *seemed* like a best-case scenario for the HMM-regularization technique, due to its constrained grammar, it may in fact not represent an adequate test case as it turned out most errors were not, in fact, due to selecting the wrong word-classes but due to within-word-class confusions. On more challenging and less constrained tasks, it is possible that HMMs do provide useful constraints, if not at the level of consecutive time frames, then at the level of subsequent exemplars.

In future work, we plan on a thorough comparison with related models such as the NHMM, investigation of HMM structures defined at larger timescales, and evaluation on more complex tasks such as CHiME-WSJ.

7. 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] 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.

[6] Tuomas Virtanen, Rita Singh, and Bhiksha Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*, Wiley, 2012.

[7] Björn Schuller, Felix Weninger, Martin Wöllmer, Yang Sun, and Gerhard Rigoll, "Non-negative matrix factorization as noise-robust feature extractor for speech recognition," in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, Dallas, USA, 2010.

[8] Jort F. Gemmeke and Tuomas Virtanen, "Noise robust exemplar-based connected digit recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4546–4549.

[9] Bhiksha Raj and Paris Smaragdis, "Latent variable decomposition of spectrograms for single channel speaker separation," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, 2005.

[10] 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.

[11] Tuomas 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.

[12] 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.

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

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

[15] Jort F. Gemmeke, Tuomas Virtanen, and Antti 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.

[16] Tara N. Sainath, Bhuvana Ramabhadran, David Nahamoo, Dimitri Kanevsky, Dirk Van Compernolle, Kris Demuynck, Jort F. Gemmeke, Jerome R. Bellegarda, and Shiva Sundaram, "Exemplar-based processing for speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 98–113, nov. 2012.

[17] G. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden markov modeling of audio with application to source separation," *Latent Variable Analysis and Signal Separation*, pp. 140–148, 2010.

[18] Gautham J. Mysore and Maneesh Sahani, "Variational inference in non-negative factorial hidden markov models for efficient audio source separation," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

[19] Emad M. Grais and Hakan Erdogan, "Hidden markov models as priors for regularized nonnegative matrix factorization in single-channel source separation," in *Interspeech 2012*, Portland, Oregon, 2012.

[20] Emmanuel Vincent, Jon Barker, Shinji Watanabe, Jonathan Le Roux, Francesco Nesta, and Marco Matassoni, "The Second 'CHiME' Speech Separation and Recognition Challenge: Datasets, Tasks and Baselines," in *Proceedings of IEEE International Conference on Audio, Speech and Signal Processing*, 2013.

[21] Antti Hurmalainen, Jort F. Gemmeke, and Tuomas Virtanen, "Modelling non-stationary noise with spectral factorisation in automatic speech recognition," *Computer Speech & Language*, 2012.

[22] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. with application to music analysis," *Neural Computation*, vol. 21, no. 3, 2009.

[23] 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.

[24] Antti Hurmalainen, Katariina Mahkonen, Jort F. Gemmeke, and Tuomas Virtanen, "Exemplar-based recognition of speech in highly variable noise," in *International Workshop on Machine Listening in Multisource Environments*, 2011, pp. 1–6.