Robust speech recognition in multi-source noise environments using convolutive non-negative matrix factorization

Ravichander Vipperla, Simon Bozonnet, Dong Wang and Nicholas Evans

Multimedia Communications Department, EURECOM, Sophia Antipolis, France

{vipperla, bozonnet, wangd, evans}@eurecom.fr

Abstract

Convolutive non-negative matrix factorization (CNMF) is an effective approach for supervised audio source separation. It relies on the availability of sufficient training data to learn a set of bases for each acoustic source. For automatic speech recognition (ASR) in a multi-source noise environment, the varied nature of background noise makes it a challenging task to learn the noise bases and thereby to suppress it from the speech signal using CNMF. A large amount of training data is required to reliably capture noise variation, but this generally leads to an unacceptable computational burden. Here, we address this problem by learning the noise bases using a computationally efficient, online CNMF approach. By learning the noise bases from several hours of ambient noise data and over a few seconds of local acoustic context, we show that background noise can be effectively attenuated from noisy speech. ASR accuracies on the CHiME corpus with the denoised speech show relative improvements in the range of 42.3% for -6 dB signal-to-noise ratio (SNR) to 2.5% for 9 dB SNR.

Index Terms: Convolutive non-negative matrix factorization, online CNMF, speech separation, automatic speech recognition

1. Introduction

Automatic speech recognition (ASR) performance is known to deteriorate in the presence of additive, background noise. While humans are able to dissociate a speaker of interest from a mixture of multiple concurrent sound sources with little or no loss in intelligibility, ASR systems perform poorly, especially when the noise is related to concurrent speech from interfering speakers, i.e. to so-called cocktail party scenario. This paper addresses the problem of recognising the speech of a target speaker under typical ambient noise conditions recorded in a home environment with television sound, music, competing background speech, and short non-stationary noises etc.

The problem is traditionally approached either from an statistical modeling perspective or from a signal enhancement perspective. In this work, we take the latter approach, since statistical models are based on the assumption that noise can be described by an underlying distribution which can be tricky in multisource noise environments. Among existing signal enhancement approaches, we can distinguish systems which use a chain of successive filters [1] in order to separate the mixed speech from the systems based on pattern recognition which first learn a model of the noise and the speech in order to separate the two. For example, [2, 3] use a factorial hidden Markov model (HMM) to separate mixed speech; [4] presents an independent component analysis (ICA) based algorithm for dictionary learning and sparse coding. Non-negative matrix factorization (NMF) and its sparse version (SNMF) have also been used successfully to separate audio stream components [5, 6]. A more sophisticated approach known as convolutive non-negative matrix factorization (CNMF) involves the sharing of decompositions among a set of bases with a time shift and has been shown to perform well in speech enhancement [7] and source separation [8] applications under supervised conditions.

In this paper we report the application of CNMF to denoise a speech signal in multi-source noise environments. The main challenge relates to the efficient learning of noise bases and is the main focus in this paper. We propose an online CNMF algorithm which is able to learn noise bases from several hours of data. This is unfeasible with a traditional CNMF approach due to the enormous computational requirements.

In the following section we first present an overview of CNMF and then discuss the online CNMF algorithm in Section 3. The method employed to denoise a speech signal is described in Section 4. In Section 5 we report our experimental setup and results. A discussion in Section 6 highlights some ideas to extend this work.

2. Convolutive non negative matrix factorization

Non-negative matrix factorization [9] attempts to decompose a non-negative matrix $D \in \Re^{\geq 0}_{M \times N}$ into two matrices W and H with the constraint that elements of the decomposed matrices are non-negative ($W \in \Re^{\geq 0}_{M \times R}$ and $H \in \Re^{\geq 0}_{R \times N}$).

$$D \approx WH \tag{1}$$

Such an approximation is non-unique. Various update rules such as multiplicative, gradient descent, and alternating non-least squares have been proposed [10] to iteratively estimate W and H. All these methods attempt to minimise a cost function of general form:

$$L(W, H) = \arg\min_{W,H} \|D - WH\|_F^2 \tag{2}$$

or a slight variation of it which captures the distance between the original matrix and its approximation. Here, $\|.\|_F$ denotes the Frobenius norm.

The columns of W are the bases vectors that capture the prominent patterns in the data whereas the rows of H are the base activation weights.

The underlying assumption in NMF is that the data represented in each column is independent. However for signals such as speech which display strong spectro-temporal correlation, it is desirable to exploit this dependency while learning the bases. Smaragdis [11] developed a convolutive extension to NMF to address this issue. The decomposition takes the form:

$$\hat{D} \approx \sum_{p=0}^{P-1} W_p \overset{p\rightarrow}{H} \tag{3}$$

where P is the convolution range. The operators $\overset{p\rightarrow}{\cdot}$ and $\overset{\leftarrow p}{\cdot}$ are column shift operators that shift p columns of the matrix to the right and left respectively. Columns vacated after the shift are filled with zeros. Under such a formulation, a sequence of P vectors corresponding to the i^{th} columns of W_p can be seen as base dimensions which capture prominent spectro-temporal patterns.

CNMF update equations which aim to optimise Equation (2) take the following form:

$$W_p = W_p \odot \frac{D \overset{p\rightarrow T}{H}}{\hat{D} \overset{p\rightarrow T}{H}} \tag{4}$$

$$H(p) = H \odot \frac{w_p^T \overset{\leftarrow p}{D}}{w_p^T \overset{\leftarrow p}{\hat{D}}} \tag{5}$$

$$H = \frac{1}{P} \sum_{p=0}^{P-1} H(p) \tag{6}$$

where, $\odot$ is an element wise multiplication operator and where the division is also element-wise. Bases W learnt on speech spectra have been shown to capture efficiently the phonetic patterns [11, 12].

3. Online CNMF

Scrutinizing the update Equations (4) and (5), we notice that the update of H can be implemented by segmenting D into a number of pieces and by computing H on each piece, thereby improving computational efficiency. However, improving the efficiency of the update operations of W is more challenging – here we have to wait for all the data D to be processed in order to complete one iteration. This means that if we have a set of bases trained already, decomposing a large amount of speech with these bases is not a serious problem with parallel computing, however learning bases itself places a high demand on both processing power and memory. To address this problem, we recently proposed an on-line base learning approach for CNMF, which processes the input matrix piece-by-piece and which updates the set of bases using the accumulated sufficient statistics [13]. With very few iterations for each piece of speech, learned patterns quickly converge to local minima of the objective function thereby facilitating its application to large scale tasks.

Simple rearrangement of Equation 4 results in the following basis learning approach:

$$W_p \leftarrow W_p \odot \frac{\sum_u B(p;u)}{\sum_q W_q \sum_u A(q,p;u)} \tag{7}$$

where

$$A(q,p;u) = \overset{q\rightarrow}{H}(u) \overset{p\rightarrow T}{H}(u)$$

and

$$B(p;u) = D(u) \overset{p\rightarrow T}{H}(u)$$

where u is the piece index. The piece length in the segmentation is somewhat arbitrary. For speech signals, a segmentation according to sentence boundaries avoids the splitting of voiced patterns and thus forms a natural choice. For each piece, the bases are updated iteratively using equations (5) and (7). An important aspect of the piecewise iteration is that the computation of $A(q,p;u)$ and $B(p;u)$ only correspond to the current piece being processed and that the contribution of the pieces processed previously can be 'memorised' using two auxiliary variables

$$A(q,p) = \sum_u A(q,p;u)$$

and

$$B(p) = \sum_u B(p;u)$$

This leads to the online pattern learning approach for CNMF, as shown in Algorithm 1, where the flag $activeW$ indicates if the bases should be updated when updating the coefficients. Algorithm 2 illustrates the pattern update process (equation 7). Matlab code for these algorithms is available online[1].

This online approach is inspired by the online dictionary learning (ODL) [14]. While the ODL approach assumes independent signals, our approach handles convolution. Another property that distinguishes our approach from ODL is that we do not pursue an optimal H for each signal piece; instead, we apply a small number of iterations to obtain a sub-optimal H and assume that it is sufficient for accumulating statistics. This

Algorithm 1 Online CNMF pattern learning

U: number of pieces
K: iteration
$A(i,j) \leftarrow 0, \quad \forall i,j \; ; A(i,j) \in \Re_{R\times R}^{\geq 0}, \quad 0 < i,j < P$
$B(i) \leftarrow 0, \quad \forall i \; ; B(i) \in R_{M\times R}^{\geq 0}, \quad 0 < i < P$
for u := 0 to U-1 **do**
  randomize(H)
  for k := 0 to K-1 **do**
    if activeW **then**
      $W = updateW(A, B, D(u), W, H)$
    end if
    $H = updateH(D, W, H)(Eq.5)$
  end for
  $[\dot{A}, \dot{B}, W] = updateW(A, B, D(u), W, H)$
  $A(i,j) \leftarrow A(i,j) + \dot{A}(i,j)$
  $B(i) \leftarrow B(i) + \dot{B}(i)$
end for

Algorithm 2 CNMF pattern update

Require: A, B, D, W, H
$\dot{A}(i,j) = \overset{i\rightarrow}{H} \overset{j\rightarrow T}{H} \; ; \dot{A}(i,j) \in \Re_{R\times R}^{\geq 0}, \quad 0 < i,j < P$
$\dot{B}(i) = D \overset{i\rightarrow T}{H} \; ; \dot{B}(i) \in \Re_{M\times R}^{\geq 0}, \quad 0 < i < P$
$A = A + \dot{A}$
$B = B + \dot{B}$
for p :=0 to P-1 **do**
  $F \leftarrow 0$
  for q :=0 to P-1 **do**
    $F = F + W_q A(q,p)$
  end for
  $\dot{W}_p = W_p \odot \frac{B(p)}{F}$
end for
$W_p = \frac{\dot{W}_p}{|\dot{W}_p|_2^2} \quad \forall p \; s.f. \; W_p \in \Re_{M\times R}^{\geq 0}$
return $[\dot{A}, \dot{B}, W]$

[1] http://audio.eurecom.fr/software

may lead to a sub-optimal solution for a particular dataset but can remarkably speed up the computation. This form is largely inherited from the conventional CNMF update. We show in [13] that online learning cannot be worse than conventional batch-mode learning and, with a suitable selection of the piece length and number of iterations, it substantially outperforms batch learning with much faster convergence speeds. Similar to ODL, learned patterns tend to be more and more accurate as the quantity of data increases; with increasing iterations, resulting patterns approach the optimal.

4. Purification of speech using CNMF

Smaragdis [11] proposes an elegant approach for audio source separation in a supervised manner. We use a similar methodology as described below:

1. From the training data available for the target speaker and for the background noise, compute the magnitude spectrum and learn bases for the speaker $W^{(Sp)}$ and noise $W^{(Bg)}$ separately.
2. Concatenate the obtained bases to form a larger global set of bases $W^{(Global)} = [W^{(Sp)}\ W^{(Bg)}]$.
3. For a speech signal containing the target speaker and additive noise, decompose the magnitude spectrum using the global bases set generated above to estimate the activations of the bases (using equation 5).
4. From the activations corresponding to the target speaker bases, recompute the magnitude spectrum $Z = \sum_{p=0}^{P-1} W_p^{(Sp)}\ \overset{p\rightarrow}{H}^{(Sp)}$
5. Modulate the magnitude spectrum using the phase spectrum of the mixed signal, to obtain the spectrum for the speaker.
6. Re-synthesize the denoised speech waveform from the above generated spectrogram using the inverse short time Fourier transform.

The real challenge in speech purification in the CHiME challenge is that there are multiple sources of background noise including speech, voices from a television, music and a host of other ambient noises typically encountered in a home environment. As a result there is large variation in the amount and type of background noises corrupting each utterance. Hence appropriate training data to learn the noise bases for each utterance is not readily available. We have investigated a couple of ways to overcome this problem.

5. Experiments

In this section, we describe our experimental setup used for our submission to the CHiME challenge [15] where the task is to recognise the speech utterances in a home environment under six different signal-to-noise ratio (SNR) conditions.

5.1. ASR experimental setup

All the audio provided for the task is recorded from a binaural microphone array. The location of the speaker is also specified to be directly 2 meters directly in front of the microphone array, while the location of each noise source and their distance from the microphones are unknown. We use a simple addition of the two channels with zero delay to obtain a mono-channel audio signal for further processing.

We used the standard ASR setup provided for the CHiME challenge without any modification. The setup uses speaker dependent acoustic models trained on Mel frequency cepstral coefficients with energy,1^{st} and 2^{nd} order derivatives. Features are further cepstral mean normalised.

Each test utterance takes the form "<*verb*> <*colour*> <*preposition*> <*letter*> <*digit*> <*coda*>" and only the recognition hypotheses for the letter and digit are scored.

The language model is a simple lattice that covers all possible sequences in the above pattern and the utterances are decoded using HTK [16].

5.2. CNMF experimental setup

As explained in Section 4, our enhancement algorithm involves processing the spectral representation of speech signal in order to suppress noise. In order to choose the optimal values for window size and overlap size to be used for parametrization, we converted the development set utterances to spectral representations, re-synthesized the waveforms and computed the ASR accuracies. A window size of 25 msec and an overlap length of 10 msec were found to give the best accuracies and have been used in all experiments described further.

We learn speaker bases from the training set available for each speaker using a convolutional span of 4 frames. In our setup, this is equivalent to capturing prominent spectro-temporal patterns that span about 70 msec; we expect to capture important sub-phone patterns with such a setup. A set of 100 bases per speaker has been empirically chosen, although in our experience, a set of around 60 bases is generally sufficient to capture minimal phonetic characteristics.

5.3. Using local acoustic context for learning noise bases

The success of CNMF-based approaches depends on learning reliable bases for the sources that need to be separated. However, the type of background noise in each utterance is unknown a-priori. For all experiments reported here we assume that the noise either side of the utterance interval is representative of that within the interval. Since each utterance is only in the order of 1.5 seconds in length, this assumption may be realistic in the case of relatively stationary noise. However, the background environment is dynamic in practice and thus we expect that not all noises occurring within the utterance will be captured from the adjacent acoustic context.

For all experiments reported here, we use up to 2 seconds of audio either side of the utterance interval, subject to availability. This data is used to learn the noise bases for each utterance with the CNMF algorithm. Since the quantity of data in this case is quite low, a set of 20 bases was used to avoid over-fitting and the convolutional range was set to 4 frames in order to match that of speaker bases.

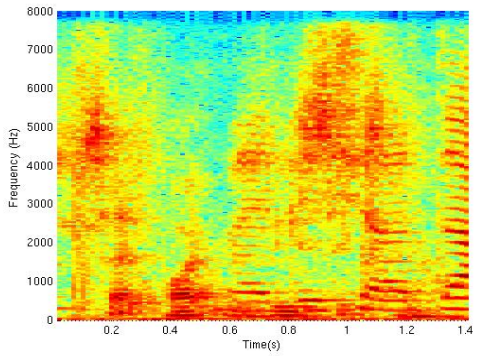
ASR performance is shown in Table 1 against SNR for the development and test sets on which we observe average relative improvements of 9.05% and 9.72% respectively. Consistent improvements are obtained for all conditions but are more prominent for lower SNRs.

Figure 1 shows an example spectrogram of a noisy speech signal with a child speaking in the background (a) before and (b) after denoising. The higher formants that correspond to the child's speech are seen to be effectively suppressed by CNMF denoising.

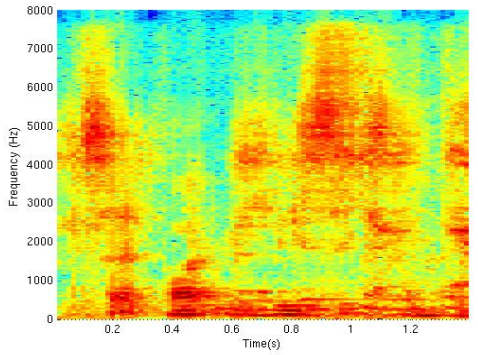
As effective as the approach seems to be, there are a number of difficulties in effectively using local acoustic context:

Table 1: *Accuracies (%) with CNMF using background basis learnt from the local acoustic context*

SNR	Development Set		Test Set	
	Baseline	CNMF	Baseline	CNMF
-6 dB	31.08	38.33	30.33	36.42
-3 dB	36.75	43.92	35.42	43.83
0 dB	49.08	58.17	49.50	58.08
3 dB	64.00	67.83	62.92	66.67
6 dB	73.83	76.50	75.17	78.17
9 dB	83.08	83.67	82.50	85.33



(a) Speech spectrogram with a child speaking in the background



(b) Speech spectrogram after background removal using CNMF

Figure 1: Spectrograms of waveform with a child speaking in the background before and after purification with CNMF.

1. Sufficient data may not be readily available in practical situations. A speaker diarization system may be used to detect utterance intervals, but may be ineffective in low SNR conditions.

2. Noise conditions either side of the utterance are not necessarily representative of those within the utterance.

3. The dynamic nature of multi-source noise environments limits the use of longer local acoustic context and thus it is more difficult to learn reasonable base patterns.

5.4. Global background noise bases

In order to overcome the limitations of learning noise bases on local acoustic context as discussed above, ideally they would be learned in a context independent fashion. This is possible using the background training set provided in the CHiME challenge. There is a reasonable amount of noise recordings of about 7 hours in this set which captures most background scenarios encountered.

With such a large data set, however, it is infeasible to learn the bases with a classical algorithm due to excessive memory and computational requirements to store and process such large matrices. Hence, in this case, we adopt the online pattern learning algorithm described in Section 3.

Background training data is provided in segments of 5 minutes each. We use the same partition structure as pieces in the online training algorithm. We learn background bases with 100, 150 and 200 dimensions respectively, all with a convolutional span of 4 frames. ASR performance for the development and test sets are shown in Tables 2 and 3 respectively, for each of the noise bases.

Table 2: ***Development set**: Accuracies (%) with CNMF using global background bases. Notation used: Bg 100 - Background noise bases with a dimension of 100*

SNR	Baseline	Bg 100	Bg 150	Bg 200
-6 dB	31.08	40.92	42.17	41.83
-3 dB	36.75	46.67	48.50	48.17
0 dB	49.08	59.75	60.25	61.67
3 dB	64.00	69.50	70.33	69.83
6 dB	73.83	78.50	77.17	75.42
9 dB	83.08	83.42	83.00	82.67

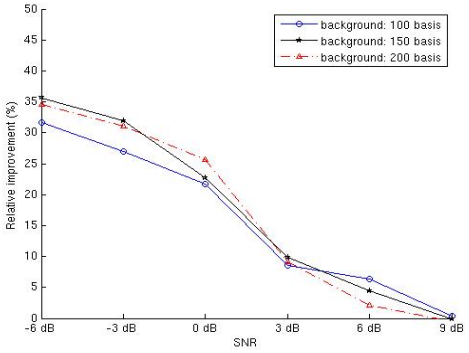
Table 3: ***Test set**: Accuracies (%) with CNMF using global background noise bases. Notation used: Bg 100 - Background noise basis with a dimension of 100*

SNR	Baseline	Bg 100	Bg 150	Bg 200
-6 dB	30.33	41.50	41.58	39.83
-3 dB	35.42	45.67	47.83	49.92
0 dB	49.50	58.50	61.33	59.92
3 dB	62.92	68.17	68.08	67.00
6 dB	75.17	78.83	78.42	77.50
9 dB	82.50	85.75	84.17	83.75

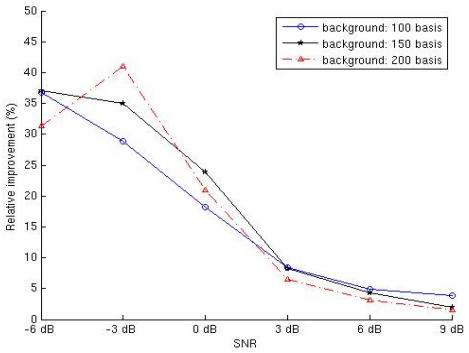
Noise bases learned over a global context are shown to give better ASR performance than those using a short local context. On average, the best relative improvements are obtained using 150 bases with a gain of 12.9% and 13.6% across all SNR conditions for the development and test sets respectively.

From Figure 2 we see that relative improvements for low SNR conditions are much greater than those for higher SNR conditions. We also observe that, as the number of background bases increases, performance for lower SNR conditions improves while there is a marginal degradation for higher SNR conditions. This trend is to be expected. For higher SNR conditions, the amount of noise is significantly lower than that of the speech signal and the use of a larger set of bases for background

noise leads to the projection of some speaker energy onto the noise bases, thus leading to some loss of useful information.



(a) Development Set



(b) Test Set

Figure 2: Relative improvement in accuracies over the baseline for Development and Test sets purified with global background basis learnt using online CNMF

5.5. Using background bases learnt from global and local acoustic context

Noise bases learned only on the global context or local context have limitations. While bases learned on the global context cannot accurately capture finer noise variations, those learned on the local context suffer from lack of data and subsequently problems with modeling unseen patterns. In this set of experiments, we use the background bases learned from both local and global contexts as described in Sections 5.3 and 5.4 respectively.

Each test utterance is denoised using a global noise basis of 100, 150 and 200 bases augmented with 20 noise bases learnt from the local acoustic context. This gives further improvements in ASR performance for both development and test sets as illustrated in Tables 4 and 5 respectively. On an average, for the 170 noise bases case, a relative improvement of 14.7% and 14.9% are achieved on the development set and the test set respectively.

5.6. Voting scheme

Figure 3 shows that the choice of base dimension has a bearing on performance at different levels of SNR. While a larger

Table 4: ***Development set**: Accuracies (%) with CNMF using background bases learnt on global and local acoustic context*

SNR	Baseline	Bg 100 + 20	Bg 150 + 20	Bg 200 + 20
-6 dB	31.08	43.75	43.50	43.67
-3 dB	36.75	48.33	49.67	48.58
0 dB	49.08	61.67	61.75	62.42
3 dB	64.00	71.75	71.67	70.58
6 dB	73.83	79.00	77.75	75.72
9 dB	83.08	83.83	83.08	82.83

Table 5: ***Test Set**: Accuracies (%) with CNMF using background bases learnt on global and local acoustic context*

SNR	Baseline	Bg 100 + 20	Bg 150 + 20	Bg 200 + 20
-6 dB	30.33	42.92	42.92	40.25
-3 dB	35.42	48.33	49.00	51.00
0 dB	49.50	60.17	62.25	60.75
3 dB	62.92	69.42	68.58	68.08
6 dB	75.17	79.25	78.67	76.75
9 dB	82.50	85.58	84.42	83.58

background bases set gives better performance for lower levels of SNR, a smaller bases set is more beneficial for higher levels of SNR. To obtain consistent improvements for all SNRs, we use a voting scheme to combine the recognition hypotheses obtained with 100, 150 and 200 global background bases augmented with 20 bases learnt from the local context. During ties in the voting scheme, recognition hypotheses with 150 bases is selected as default.

Table 6 shows ASR performance with the voting scheme. We obtain an average relative improvement of 15.7% across all SNR conditions for the test set, with a relative improvement of about 42.3% at -6 dB and about 2.5% improvement at 9 dB.

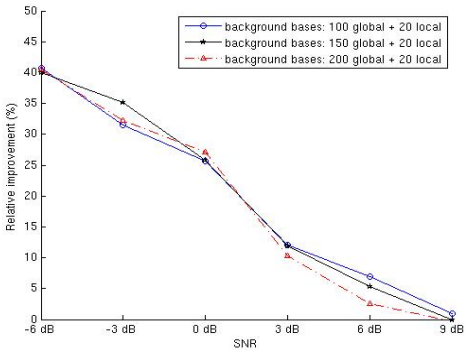
Table 6: *Accuracies (%) with the voting scheme*

SNR	Development Set			Test Set		
	Baseline	Voting	Rel. ↑	Baseline	Voting	Rel. ↑
-6 dB	31.08	43.83	41.02	30.33	43.17	42.33
-3 dB	36.75	49.25	34.01	35.42	50.42	42.35
0 dB	49.08	63.00	28.36	49.50	61.75	24.75
3 dB	64.00	71.83	12.23	62.92	69.42	10.33
6 dB	73.83	78.33	6.10	75.17	79.33	5.53
9 dB	83.08	83.58	0.60	82.50	84.58	2.52

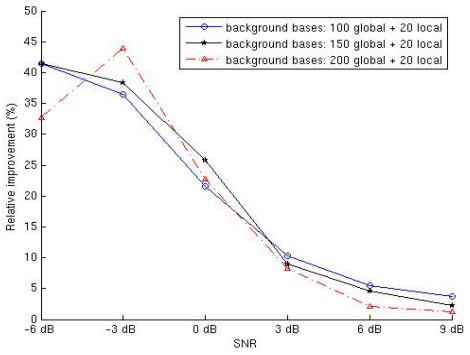
6. Discussion

An online CNMF implementation is extremely beneficial in this scenario, where global noise bases are learnt from over several hours of data. To illustrate, with the conventional offline approach, the training of bases on the 7 hours of noise data used in our experimental setup would require M=201 and $N \approx 1.6$ million. This is infeasible in terms of memory requirements and computational time.

Results show that the use of higher order bases dimension



(a) Development Set



(b) Test Set

Figure 3: Relative improvement in accuracies over baseline for the Development and Test sets purified with background bases learnt from the global and local acoustic context.

for higher SNR conditions leads to poorer performance. In our ongoing work we are investigating automatic noise basis selection and noise diarization in order to learn bases from more homogeneous noise segments.

Finally, all experiments reported here are based on spectral magnitude representations in order to satisfy the non-negative constraint. We have begun to investigate the application of our approach to mel-scaled spectral estimates which also satisfy the non-negative constraint.

7. Conclusions

This paper reports the successful application of convolutive non-negative matrix factorization (CNMF) to improve the performance of automatic speech recognition (ASR) in a multi-source noise environment. The focus of the work presented in this paper relates to the efficient learning of noise bases and its suppression or separation from noise-degraded speech. When used in conjunction with noise bases learnt from local acoustic context, global noise bases learnt using an online CNMF approach are shown to give substantial improvements in ASR accuracies over all noise conditions specified in the CHiME corpus.

8. Acknowledgments

We acknowledge the financial support from '*The Adaptable Ambient Living Assistant (ALIAS)*' project funded through the joint national Ambient Assisted Living (AAL) programme and partial support from the French Ministry of Industry (Innovative Web call) under contract 09.2.93.0966, '*Collaborative Annotation for Video Accessibility (ACAV)*'.

9. References

[1] O. D. Deshmukh and C. Y. Espy-Wilson, "Modified phase opponency based solution to the speech separation challenge," in *Proc. Interspeech*, 2006.

[2] S. T. Roweis, "One microphone source separation," in *In Advances in Neural Information Processing Systems 13*. MIT Press, 2000, pp. 793–799.

[3] T. Virtanen, "Speech recognition using factorial hidden markov models for separation in the feature space," in *Proc. Interspeech*, 2006.

[4] G. jin Jang and T. won Lee, "A maximum likelihood approach to single-channel source separation," *Journal of Machine Learning Research*, vol. 4, pp. 1365–1392, 2003.

[5] P. D. O'Grady and B. A. Pearlmutter, "Discovering convolutive speech phones using sparseness and non-negativity," in *Proceedings of the 7th international conference on Independent component analysis and signal separation*, 2007, pp. 520–527.

[6] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *International Conference on Spoken Language Processing (Interspeech)*, Sep 2006.

[7] C. D. Sigg, T. Dikk, and J. M. Buhmann, "Speech enhancement with sparse coding in learned dictionaries." in *ICASSP'10*, 2010, pp. 4758–4761.

[8] W. Wang, "Convolutive non-negative sparse coding," in *IEEE International Joint Conference on Neural Networks.*, June 2008, pp. 3681 –3684.

[9] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[10] A. Cichocki, R. Zdunek, A. H. Phan, and S. ichi Amari, *Nonnegative Matrix and Tensor Factorizations*. Wiley, 2009.

[11] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing.*, vol. 15, no. 1, pp. 1–12, 2007.

[12] P. O'Grady and B. Pearlmutter, "Convolutive non-negative matrix factorisation with a sparseness constraint," in *Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing.*, Sept 2006, pp. 427–432.

[13] D. Wang, R. Vipperla, and N. Evans, "Online pattern learning for non-negative convolutive sparse coding," in *Proc. Interspeech*, 2011.

[14] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research*, vol. 2010, no. 11, pp. 19–60, 2010.

[15] H. Christensen, J. Barker, N. Ma, and P. Green, "The CHiME corpus: A resource and a challenge for computational hearing in multisource environments," in *Interspeech*, 2010.

[16] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for Hidden Markov Model Toolkit Version 3.4)*, 2006.