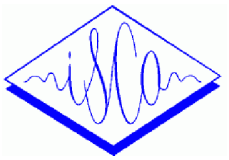
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

# Heterogeneous Convolutive Non-Negative Sparse Coding

*Dong Wang*

Center for Speech and Language Technologies
Tsinghua University, China
wangdong99@mails.tsinghua.edu.cn

*Javier Tejedor*

Human Computer Technology Laboratory
Universidad Autónoma de Madrid, Spain
javier.tejedor@uam.es

## Abstract

Convolutive non-negative matrix factorization (CNMF) and its sparse version, convolutive non-negative sparse coding (CNSC), exhibit great success in speech processing. A particular limitation of the current CNMF/CNSC approaches is that the convolution ranges of the bases in learning are identical, resulting in patterns covering the same time-span. This is obvious unideal as most of sequential signals, for example speech, involve patterns with a multitude of time spans. This paper extends the CNMF/CNSC algorithm and presents a heterogeneous learning approach which can learn bases with non-uniformed convolution ranges. The validity of this extension is demonstrated with a simple speech separation task.

**Index Terms**: non-negative matrix factorization, sparse coding, speech processing

## 1. Introduction

Non-negative matrix factorization (NMF) has been successfully applied in a multitude of applications, due to its capability of learning partial patterns [1, 2]. In speech research, convolutive NMF (CNMF) has been proposed to learn temporal patterns of speech signals [3, 4], and sparse NMF is introduced to learn rich spectral patterns [5, 6, 7]. These two extensions are further combined, resulting in a powerful learning approach referred to as convolutive non-negative sparse coding (CNSC) [8, 9, 10]. An online algorithm has been introduced to deal with the high memory and computation demand of CNSC, which enables CNSC in large scale tasks, for instance large vocabulary speech recognition [11, 12].

The success of CNSC in real applications is largely influenced by the quality of the learned bases, which is in turn determined by the choice of learning configurations, such as the convolution range. In [13] we have shown that for speech signals, bases learned with different convolution ranges reflect patterns in different temporal resolutions. Since the convolution range is uniformed in the conventional CNSC algorithm, i.e., all the bases share the same range, we learned multiple sets of bases with different ranges and then combined them to obtain better signal representation [13]. In this paper, we consider a more general approach: instead of combining bases that are learned with different but uniformed convolution ranges, we learn non-uniformed bases directly. In other words, bases in learning may keep different convolution ranges and hence are 'heterogeneous'. This extension is particularly suited to signals such as speech where patterns (e.g., phones) are highly diverse in time-span. We refer the new learning algorithm to as 'heterogeneous CNSC', and accordingly the conventional approach as 'homogeneous CNSC'.

The new learning algorithm is presented in the next section. In Section 3, we propose a naive incremental search scheme to discover appropriate base distributions. A simple speech separation task is reported in Section 4 to demonstrate the value of the new method, and the paper is concluded in Section 5. The matlab code is available online.[1]

## 2. Heterogeneous CNSC

The heterogeneous CNSC is implemented based on the online CNSC algorithm [11]. Following the formulation in [14], CNSC aims at minimizing the following objective function:

$$L(W,H) \quad = \quad |X - \hat{X}(W)|_2^2 + \lambda |H|_1 \qquad (1)$$

where $\lambda$ is a factor controlling the sparsity of $H$, and $|.|_l$ denotes the $l$-norm, which is equivalent to the sum of squares of the matrix elements when $l = 2$ or to the sum of absolute values when $l = 1$. $X \in R_{0,+}^{M \times N}$ represents the original signal of length $N$ in the $M$-dimensional non-negative space, $W \in R_{0,+}^{M \times R \times P}$ represents $R$ bases with convolution range $P$, and $H \in R_{0,+}^{R \times N}$ represents the coefficients. $\hat{X}$ is the approximate reconstruction of $X$ and has the form:

$$\hat{X} = \sum_{p=0}^{P-1} W(p) \overset{p\rightarrow}{H} \qquad (2)$$

where $\overset{p\rightarrow}{H}$ shifts $H$ by $p$ columns to the right and where $W(p) \in R_{0,+}^{M \times R}$ represents the corresponding bases of $\overset{p\rightarrow}{H}$.

Now consider heterogeneous bases where the convolution range $P$ depends on the base index $r$, denoted by

[1] http://homepages.inf.ed.ac.uk/v1dwang2/public/tools/index.html

$P_r$. This means different bases may cover different time spans. The reconstruction is then changed to:

$$\hat{X} = \sum_{r=1}^{R} \sum_{p=0}^{P_r-1} W(r;p) \overset{p\rightarrow}{H}(r) \tag{3}$$

where $W(r;p)$ denotes the $p$-th column of the $r$-th base, and $\overset{p\rightarrow}{H}(r)$ is the coefficient matrix associated with the $r$-th base shifted $p$ columns to the right. Similar to the derivation in [14] the update equations for $W$ and $H$ are given by:

$$W(r;p) \leftarrow W(r;p) \odot \frac{X \overset{p\rightarrow}{H}(r)^T}{\hat{X} \overset{p\rightarrow}{H}(r)^T}$$

$$H_p(r) \leftarrow H(r) \odot \frac{W(r;p)^T \overset{\leftarrow p}{H}}{W(r;p)^T \overset{\leftarrow p}{\hat{X}} + \lambda \Xi}$$

where $\odot$ is the element-wise product and where the division is also element-wise. $\Xi$ is a matrix with all the elements equal to 1. $\overset{\leftarrow p}{\hat{X}}$ is the reconstructed signal shifted $p$ columns to the left, and $\overset{\leftarrow p}{H}$ shifts $H$ by $p$ columns to the left. Note that, for different $p$, the update for $H$ is different and is usually averaged over $p$:

$$H(r) = \frac{1}{P_r} \sum_{p=0}^{P_r-1} H_p(r).$$

The online procedure [11] can be easily applied to the heterogeneous learning, so that a large set of heterogeneous patterns can be learned with large amount of data. For example, we can use this approach to learn phone and sub-phone patterns simultaneously and discover the most representative patterns based on certain constraints such as sparsity.

## 3. Incremental base distribution search

A particular difficulty accompanied with heterogeneous learning is that we need to discover how many bases should be distributed to a specific convolution range. Suppose our task is to learn $R$ heterogeneous bases, and the possible convolution range is from 1 to $T$. Given a cost measure (e.g., reconstruction cost), we can examine all the possible base distributions and find the optimal one that leads to the least cost. However, this exhaustive search is highly ineffective; for a large $T$, we have to resort to some heuristic approach and look for a sub-optimal solution. In this paper, we use a simple *incremental search* which increases the number of bases one by one, and for each newly added base, the convolution range $t$ is chosen to be the value that leads to the least cost with yet existing bases. Algorithm 1 illustrates the steps, where $Learn(R)$ and $Eval(R)$ are learning and evaluation procedures parameterized by $R$ respectively.

**Algorithm 1** Incremental search

```
TN: max convolution range
RN: total number of bases
R(t) := 0   ∀ 1 ≤ t ≤ TN
for r := 1 to RN do
    c' := MAX_FLOAT
    t' := −1
    for t := 1 to TN do
        R(t) := R(t) + 1
        c := Learn(R) & Eval(R)
        if c < c' then
            (c',t')=(c,t)
        end if
        R(t) := R(t) - 1
    end for
    R(t') := R(t') + 1
end for
```

The computation complexity is $T \times R$ times of learning and evaluation, and the resulted solution is sub-optimal. Other methods such as evolutive algorithms are under investigation.

## 4. Experiments

We use a toy experiment to demonstrate the validity of heterogeneous learning. The task is to learn two sets of bases from speech signals of male and female speakers respectively, and then to use the learned bases to separate the male and female voices from a segment of mixed speech. The same experiment was used in [11] to demonstrate the online learning technique.

### 4.1. Experimental setup

We chose two speakers (one male and one female) from the Wall Street Journal (WSJ) speech database[2]. The individual speech segments were obtained by concatenating all the utterances of each speaker, and the mixed speech was obtained by simply adding the two individual segments using the Sox toolkit[3], with appropriate zero-padding applied to the shorter segment.

The speech signals are windowed into frames of 32ms with a frame shift of 16ms, thereby resulting in a frame rate of 62.5 frames per second. The Fourier transform is applied to each frame and the magnitude spectrum is used as a non-negative representation which is suitable for CNSC. With the individual spectrum, the male and female bases are learned by the CNSC algorithm. In the decoding phase, the spectrum of the mixed speech signal is projected onto the two sets of individual bases, and the male and female spectra are reconstructed with the corresponding bases and projected coefficients. For simplicity, we use batch learning in this study although the online learning can be equally applied. The number of iterations is set to 100 for both learning and decoding, which has shown well balance between speed and convergence [13].

[2] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC94S13A
[3] http://sox.sourceforge.net/

We use the averaged reconstruction cost (square error) of the resulting magnitude spectrum to evaluate the separation, which is simply defined as:

$$\Delta = \frac{1}{N_m}|M_m - \tilde{M}_m|_2^2 + \frac{1}{N_f}|M_f - \tilde{M}_f|_2^2$$

where $M_m$ and $M_f$ are the original magnitude spectrum of male and female speech respectively, and $\tilde{M}_m$ and $\tilde{M}_f$ are the corresponding reconstructed magnitude spectrum. $N_m$ and $N_f$ are the numbers of frames of the original male and female speech segments respectively. Note the reconstructed signals are chopped to meet the length of the original signals.

### 4.2. Homogeneous learning

We first study the behavior of homogeneous learning, which gives us a better understanding where the heterogeneous learning takes its advantage.

#### *4.2.1. Sparsity test*

The first experiment investigates the behavior of homogeneous learning with various sparsity settings. Figure 1 presents the results when the sparsity $\lambda$ is set to different values in learning and decoding. The number of bases $R$ is set to 40 and the convolution range $T$ is set to 5. We first observe that higher sparsity in learning leads to less reconstruction cost, or better quality of separation. This can be explained by the fact that learning with higher sparsity penalizes pattern overlapping in signal representation, which results in patterns that are more representative for characteristics of signals in both the temporary and the spectral dimensions. For decoding, higher sparsity does not provide better separation, partly due to the discrepancy between the sparse-regulated objective function in decoding and the non-sparse metric in evaluation.

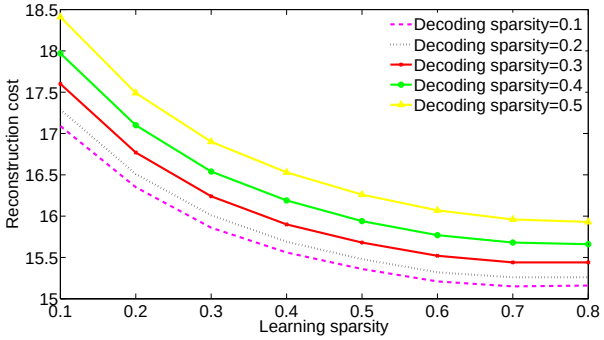



Figure 1: *Homogeneous learning with various sparsity values in learning and decoding.*

#### *4.2.2. Convolution range*

The second experiment studies the impact of the number of bases $R$ and convolution range $T$. The sparsity is set to 0.4 and 0.1 in learning and decoding respectively. The results are shown in Figure 2.

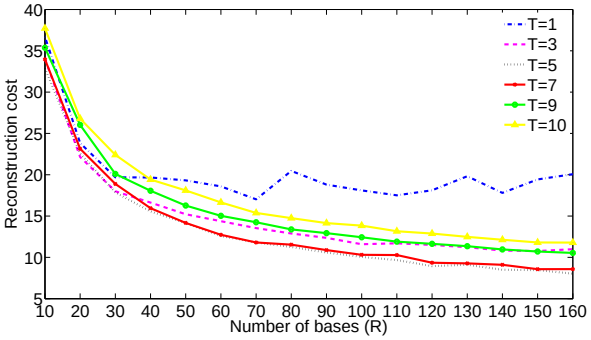



Figure 2: *Homogeneous learning with various numbers of bases ($R$) and convolution range ($T$).*

We observe that the reconstruction cost decreases in general with increasing $R$. This is not surprising as more detailed patterns can be learned with a larger number of bases, which in turn results in better signal representation. The case $T = 1$ is a bit exceptional, where the reconstruction cost is not further decreased with $R$ larger than 70. This may be attributed to the limited number of spectral patterns that might be 'over-learned' if $R$ is too large.

Another observation is that with a fixed $R$, the reconstruction cost reaches the minimum with a median $T$. This on one hand indicates that long time-span patterns are generally representative for speech characteristics, on the other hand, it shows that involving over complex patterns may reduce representation power. This can be attributed to the large number of potential patterns with a large convolution range, which leads to incomplete representation if $R$ is small. This conjecture is supported by the results of $T = 3$ and $T = 9$: with a small number of bases $R$, $T = 9$ provides clearly worse separation than $T = 3$, whilst with $R$ increasing, the performance gap gradually reduces, suggesting that patterns of $T = 9$ are becoming complete; finally when $R = 160$, $T = 9$ reaches the same as or even slightly better performance than $T = 3$.

### 4.3. Heterogeneous learning

From the previous experiment, we see that long time-span patterns are generally helpful in signal representation; the problem with the homogeneous learning is that, when the convolution range is increased, the time-span of all the patterns has to be increased. This results in two issues: on one hand, we have to learn some long time-span patterns that are not essentially necessary, and on the other hand, some highly representative short time-span patterns are lost. The heterogeneous learning solves this problem by allowing patterns with non-uniformed convolution ranges being learned.

Tables 1 and 2 present the results of heterogeneous learning with $\lambda = 0.4$ and $\lambda = 0.7$ respectively. The

| | Homogeneous | Heterogeneous | |
|---|---|---|---|
| $T$ | $\Delta$ | Base distribution | $\Delta$ |
| 1 | 19.67 | [40] | 19.67 |
| 2 | 17.71 | [19 21] | 18.43 |
| 3 | 16.63 | [10 7 23] | 16.93 |
| 4 | 16.40 | [10 7 10 13] | 17.76 |
| 5 | **15.56** | [9 7 6 10 8] | 17.14 |
| 6 | 15.67 | [6 4 4 4 8 14] | 16.01 |
| 7 | 15.95 | [1 7 5 6 4 4 13] | 15.73 |
| 8 | 17.04 | [1 1 5 4 7 9 3 10] | **15.07** |
| 9 | 18.06 | [7 4 3 2 4 5 3 2 10] | 15.30 |
| 10 | 19.42 | [6 1 3 3 4 4 3 6 1 9] | 15.93 |

Table 1: Homogeneous and heterogeneous learning with $R$=40, $\lambda = 0.4$. $\Delta$ is the reconstruction cost.

| | Homogeneous | Heterogeneous | |
|---|---|---|---|
| $T$ | $\Delta$ | Base distribution | $\Delta$ |
| 1 | 18.80 | [40] | 17.83 |
| 2 | 16.89 | [18 22] | 17.83 |
| 3 | 16.22 | [11 7 22] | 16.77 |
| 4 | 15.96 | [10 10 10 10] | 16.71 |
| 5 | **15.15** | [9 8 7 6 10] | 16.56 |
| 6 | 15.32 | [12 5 6 5 3 9] | 16.46 |
| 7 | 15.61 | [8 2 3 4 7 7 9] | 15.35 |
| 8 | 16.62 | [2 6 6 7 3 6 5 5] | 15.15 |
| 9 | 17.66 | [5 5 3 6 3 5 6 4 3] | **14.85** |
| 10 | 19.19 | [7 6 1 7 2 4 3 1 1 8] | 15.72 |

Table 2: Homogeneous and heterogeneous learning with $R$=40, $\lambda = 0.7$. $\Delta$ is the reconstruction cost.

total number of bases $R$ is fixed to $40$, and the (maximal) convolution range $T$ varies from $1$ to $10$. The incremental search is applied to look for the base distribution, and for efficiency, the number of iterations is set to $50$ in the search. We should note that homogeneous learning is a special case of heterogeneous learning where the convolution range of all bases is set to be uniformed. This means that heterogeneous learning should not be worse than homogeneous learning. In practice, however, the incremental search usually cannot find the optimal solution, and so the results with heterogeneous learning are not necessarily better than those obtained with homogeneous learning. Nevertheless, from Tables 1 and 2, we still observe clear advantage with the heterogeneous learning when $T$ is large, and the best results obtained with heterogeneous learning are better than those obtained with homogeneous learning.

Looking at the base distributions over $T$, we find that long time-span bases take a large portion, confirming the importance of long temporal patterns. Interestingly, the short time-span bases are not less important: in most cases, the total number of bases with $T \leq 3$ is rather significant. Comparing Tables 1 and 2, it seems that higher sparsity encourages more median time-span patterns which usually posses good balance between representative power and generality. We finally emphasize that all these observations are based on the sub-optimal search; much work remains to discover properties of the patterns learned with heterogeneous learning and properties of the heterogeneous learning itself.

## 5. Conclusions

We propose a heterogeneous convolutive non-negative sparse coding approach. Allowing non-uniformed convolution ranges, the new approach is able to learn patterns with various time spans. We also propose an incremental search to look for appropriate base distributions. A simple speech separation task demonstrates that the heterogeneous learning can discover more representative pattern sets than the conventional homogeneous learning. This work is still in the preliminary stage and a multitude of research remains for future work, such as efficient search methods for optimal base distributions and analysis of properties of optimal bases.

## 6. References


[1] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 12, pp. 111–126, 1994.

[2] D. D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

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

[4] D. FitzGerald and E. Coyle, "Shifted non-negative matrix factorisation for sound source separation," in *Proc. IEEE Workshop on Statistical Signal Processing*, 2005, pp. 1132–1137.

[5] P. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[6] M. Heiler and C. Schnorr, "Learning sparse representations by non-negative matrix factorization and sequential cone programming," *Machine Learning Research*, vol. 7, pp. 1385–1407, 2006.

[7] H. Kim and H. Park, "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis," *Bioinformatics*, vol. 23, no. 12, pp. 1495–1502, 2007.

[8] P. D. O'Grady and B. A. Pearlmutter, "Convolutive non-negative matrix factorisation with a sparseness constraint," in *Proc. IEEE International Workshop on Machine Learning for Signal Processing*, 2006, pp. 427–432.

[9] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 1066–1074, 2007.

[10] W. Wang, A. Cichocki, and J. A. Chamber, "A multiplicative algorithm for convolutive non-negative matrix factorization based on squared Euclidean distance," *IEEE Trans. on Signal Process.*, vol. 57, no. 5, pp. 2858–2864, 2009.

[11] D. Wang, R. Vipperla, and N. Evans, "Online pattern learning for non-negative convolutive sparse coding," in *Proc. Interspeech*, 2011, pp. 65–68.

[12] R. Vipperla, S. Bozoonet, D. Wang, and N. Evans, "Robust speech recognition in multi-source noise environments using convolutive non-negative matrix factorization," in *Proc. CHiME*, 2011.

[13] D. Wang, R. Vipperla, and N. Evans, "Parallel and hierarchical decision making for sparse coding in speech recognition," in *Proc. Interspeech*, 2011, pp. 2557–2560.

[14] W. Wang, "Convolutive non-negative sparse coding," in *Proc. IJCNN*, 2008, pp. 3681–3684.