# Parallel and Hierarchical Decision Making for Sparse Coding in Speech Recognition

*Dong Wang, Ravichander Vipperla and Nicholas Evans*

Multimedia Communications Department, EURECOM
F-06560 Sophia-Antipolis, France
{dong.wang, vipperla@eurecom.fr, evans}@eurecom.fr

## Abstract

Sparse coding exhibits promising performance in speech processing, mainly due to the large number of bases that can be used to represent speech signals. However, the high demand for computational power represents a major obstacle in the case of large datasets, as does the difficulty in utilising information scattered sparsely in high dimensional features. This paper reports the use of an online dictionary learning technique, proposed recently by the machine learning community, to learn large scale bases efficiently, and proposes a new parallel and hierarchical architecture to make use of the sparse information in high dimensional features. The approach uses multilayer perceptrons (MLPs) to model sparse feature subspaces and make local decisions accordingly; the latter are integrated by additional MLPs in a hierarchical way for making global decisions. Experiments on the WSJ database show that the proposed approach not only solves the problem of prohibitive computation with large-dimensional sparse features, but also provides better performance in a frame-level phone prediction task.

**Index Terms**: sparse coding, feature extraction, posterior feature, speech recognition

## 1. Introduction

Inspired by the discovery in neural science that the brain represents information with only a small number of active neurons [1], sparse coding has been developed as an efficient subspace approach, and has been demonstrated to be a powerful tool in a wide range of research fields, including signal processing, image processing and information representation [2,3]. In the speech community, it has been successfully applied to a number of applications such as blind source separation [4, 5], de-noising [6] and speech enhancement [7]. For speech recognition, a number of results have been published on digit recognition tasks based on TIMIT or TIDIGITS datasets [8,9].

Compared to conventional representations based on short-term spectral analysis such as Mel frequency cepstral coefficients (MFCCs), sparse coding represents speech signals in terms of speech patterns which cover a longer time span and which therefore capture spectro-temporal information. This is a particularly beneficial attribute when used for speech recognition with hidden Markov models (HMM) where temporal information is otherwise generally ignored due to the assumption of independent observations. In contrast to conventional spectro-temporal pattern learning approaches such as principle component analysis (PCA) and non-negative matrix factorisation (NMF), sparse coding enables unambiguous encoding using overcomplete bases, i.e. the number of bases is larger than the feature dimension. This is a fundamental advantage as it supports speech representation with a very large number of bases (even redundant) without introducing ambiguity.

In spite of the promising potential the application of sparse coding to speech recognition is far from being straight forward, particularly in the case of complex tasks that involve tens of thousands of words and hundreds of hours of training data. The first challenge comes from the prohibitive computing required for the learning of a large set of bases; the second relates to making efficient use of the information scattered in high dimensional sparse features. For these reasons, the application of sparse coding in speech recognition is still limited to simple tasks and small databases. Even then, base learning on entire datasets is still prohibitive and manually or randomly selected subsets are often used [8, 9]. This is obviously sub-optimal for complex recognition tasks where a large number of bases are necessary and have to be learnt with large volumes of training data.

In this paper we attempt to tackle the problems associated with large-scale sparse coding from two directions. First we employ an online dictionary learning technique [10], recently proposed by the machine learning community, to speed up base learning. In contrast to conventional batch mode learning, the online approach updates learnt bases frame-by-frame. Although convergence is not guaranteed for small datasets, online learning quickly approaches the optimum when data are plentiful. This is particularly beneficial in the case of complex recognition tasks where a large database is available for base learning. Second, we propose a parallel and hierarchical decision approach to manage high-dimensional sparse features so that the sparsely coded information can be utilised efficiently. Specifically, we build a particular discriminant, e.g. a multilayer perceptron (MLP), to model subset dimensions of high-dimensional sparse features and to make local decisions accordingly; these local decisions are then integrated by additional discriminants in a hierarchical way for making a global decision.

In the following sections we first briefly outline the online base learning algorithm and then introduce the parallel and hierarchical decision approach. Experiments are presented in Section 4 and the paper is concluded in Section 5.

## 2. Online base learning

Given a set of bases $W$, a signal $X$ can be approximately reconstructed through multiplication with a coefficient vector $H$ so that $X \approx WH$. The sparse coding technique optimises $H$ by minimising the following objective function:

$$L(H; W, X) = |X - WH|_2^2 + \lambda |H|_1 \qquad (1)$$

where $|.|_l$ represents the Frobenius $l$-norm. The second term on the right hand side of Equation 1 drives most of the elements in $H$ to zero, thus leading to sparse features. The factor $\lambda$ controls the sparsity (larger values lead to more sparsity), and bases in $W$ are usually constrained to have unit length to ensure a defi-

**Algorithm 1** Online dictionary learning for sparse coding

1: $A :\leftarrow 0$
2: $B :\leftarrow 0$
3: **for** i:=1 to N **do**
4: $\quad H \leftarrow \arg\min_H L(H; W, X_i)$
5: $\quad A \leftarrow A + H \times H^T$
6: $\quad B \leftarrow B + X_i \times H^T$
7: $\quad W \leftarrow update(W, A, B)$
8: **end for**

nite solution for $H$.

An important advantage of sparse coding is that the solution for $H$ remains definite even though the number of bases in $W$ is greater than the feature dimension. This allows a highly accurate representation of signals with a large set of bases. A major problem accompanying the use of large base sets, however, relates to the often prohibitive computation demand associated with base learning. Traditional base learning approaches employ iterative optimisation of the base matrix $W$ and the feature vector $H$, e.g. [8, 11]. Having to read and process all the training data twice in each iteration, these 'batch' learning approaches are highly demanding in both memory resources and computation when the number of bases to be learnt is large and the training data are plentiful. A recent alternative approach, presented in [10] and referred to as online dictionary learning (ODL), processes frames one-by-one (or in small batches) and updates the bases after each is processed. The ODL approach is illustrated in Algorithm 1 where $N$ is the number of frames in the training data. The intermediate matrices $A$ and $B$ are introduced to accumulate sufficient statistics for historical frames and $update(W, A, B)$ is a one-step quadratic optimisation for $W$ given $A$ and $B$.

By updating the bases once for each frame, online learning converges much more quickly than conventional batch learning. In [10] ODL is furthermore shown to converge to optimal bases given sufficient data. When applied to speech signals a potential limitation of ODL relates to the assumption of independent frames and thus it is not suitable for learning temporal patterns. A simple solution is to compose 'temporal features' by concatenating a window of neighbouring frames centred on the current frame so that temporal patterns are captured and learnt through ODL.

Online base learning enables the learning of a larger number of bases with a large database. This is highly appealing for speech recognition: on the one hand the large number of bases provide for more accurate representations of spectro-temporal speech patterns; on the other hand learning based on a large amount data ensures that resulting bases are more representative of speech content (e.g. phones) instead of noise or speaker-dependent characteristics. Therefore, the sparse features $H$, which are derived from large sets of bases learnt with large databases, tend to be more accurate and robust for speech recognition than features based on short-time spectral analysis and other factorisation techniques.

## 3. Parallel and hierarchical decision

Online base learning facilitates the discovery of a large set of spectro-temporal bases to represent speech signals in a reliable and robust way; however, this also means that speech signal representations involve high-dimensional sparse features in which most of the dimensions are uninformative (zero values). How to make efficient use of the information embedded in such high dimensional sparse features is a challenging problem, especially for tasks such as large vocabulary speech recognition where the dimension of sparse features is usually in the order of several hundred.

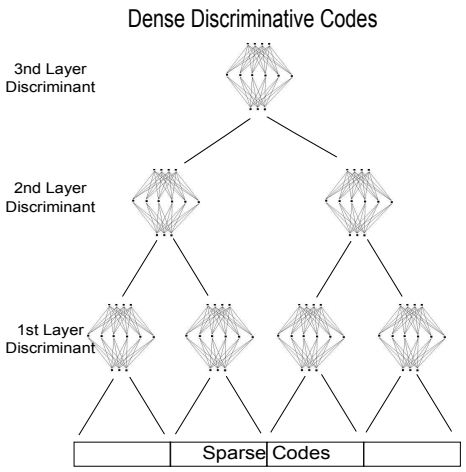


Figure 1: The architecture of parallel and hierarchical decision for sparse features.

One possible solution involves dimension reduction techniques (usually based on linear transforms), e.g. the Kullback Leibler (KL) transform [12]. The problem with this simple approach, though, is that the representation power associated with sparse coding might be largely lost. A better approach is to map the sparse features to certain dense features while retaining the representation power for the task in hand. For speech recognition a natural choice of dense features is phone posterior probabilities, which are discriminative for phone recognition and directly relate to large vocabulary speech recognition. Therefore the problem associated with high-dimensional sparse coding can be largely solved if we can find an efficient method to map sparse features to phone posterior probabilities.

An MLP-based approach can be adopted from the HMM/ANN hybrid architecture [13] and the tandem framework [14], where MLPs with speech feature inputs and phone class label outputs are used to map speech features to phone posteriors. This can be applied to sparse features directly, as in [9] and works well for sparse features with reasonable dimensions; however, if the feature dimension is high, the computation and resources required for MLP training and prediction becomes unaffordable.

In order to solve this problem, we resort to the locality and redundancy of sparse coding. On the one hand useful information for a particular task usually resides in very few dimensions of a sparse feature; on the other hand, information residing in different dimensions might be similar. This motivates a divide-and-conquer approach to dealing with very high dimensional sparse features. Specifically, sparse features can be divided into a number of dimension subsets and, for each subset, a particular discriminant is trained and used to make local decisions based on information contained within the dimensions in question. Local decisions for each subset are then merged by additional discriminants to make broader decisions. The decision process continues until a global decision is reached. This leads to a hierarchical decision approach as illustrated in Figure 1, which shows MLPs as discriminants, although any other suitable discriminant may also be used.

With such a hierarchical approach, each discriminative model operates on a subspace of the sparse feature and thus the computation and resources become manageable; furthermore,

training and prediction of the discriminants within the same layer can be conducted in parallel, leading to an inherent solution to the computational burden. More importantly, the properties of locality and redundancy ensure that local discriminants can be well trained with data in subset dimensions. Although discriminants attached to different dimension subsets may be better at recognising different phones, the hierarchical discriminants can learn this bias and make optimal global decisions. We will see in the next section that this hierarchical decision approach not only solves the computational burden, but that it also improves decision quality by mitigating over-fitting.

# 4. Experiments

We assessed the proposed parallel and hierarchical approach with a frame-level phone recognition task based on the wall street journal (WSJ) speech corpus which consists of 7861 utterances from 92 speakers for training and 742 utterances from 20 speakers for evaluation. Both training and evaluation speech utterances are forced-aligned to their phone transcripts, and each frame is labelled according to the alignment and phone class. Labelled frames in the training utterances are used to train and cross validate the MLP models (6500 utterances for training and the rest for cross validation), and the resulting models are used to predict frame-level phone classes for all utterances in the evaluation set.

The baseline representation involves conventional 13-dimensional MFCCs plus the first and second derivatives (39 dimensions in total) which are computed from the magnitude spectrum of 25ms-windowed speech using 26 Mel-scaled filter banks. For sparse coding, the same 26 Mel-scaled filter banks are used to generate short-time spectral features, from which temporal features are composed by concatenating $T$ neighbouring spectral features. The temporal features are used to learn $R$ bases with the ODL algorithm, and the learnt bases are used to obtain sparse features by decomposing temporal features using the Lasso approach [10]. Different sparsity leads to different bases and features; setting $\lambda = 0.1$ provides good performance and therefore the value is used in all following experiments. The HTK toolkit is used to conduct the forced-alignment and ICSI's *QuickNet* tool is used for manipulating MLPs.

### 4.1. Basic coding

In the first experiment we investigate the quality of sparse features derived with various configurations. We concentrate on the size of learnt bases ($R$) and the number of neighbouring spectral frames covered by a temporal feature. The latter is referred to as the time span and denoted by $T$. Intuitively, a longer time span implies greater variety in spectro-temporal patterns and therefore requires a larger number of bases to represent them with the same accuracy. If, however, the base set is overly large, some bases may converge to non-speech patterns, such as noise, and may lead to the over-fitting of training data. We build MLPs with various inputs including MFCCs and sparse features and use the resulting models to conduct frame-level phone prediction. MLP inputs involve a context window of 9 frames, centred on the prediction frame, as in the tandem approach [14]. The outputs are 46 phone classes, and the hidden nodes are fixed to 4000 according to cross validation.

Table 1 illustrates the phone error rate (PER), for phone class prediction on both the training set and the evaluation set, using MFCCs and sparse features with differing values of $R$ and $T$. We first observe that, for sparse coding, both a larger time span ($T$) and a larger base set ($R$) tend to provide better phone accuracy. With a fixed time span, the use of more bases is usually beneficial. For example, when $T$ is set to 15 and 21, 100 bases provides better performance than 50 bases. However, if the time span is small, and hence the number of likely patterns is limited, too many bases may lead to decreased performance on account of over-fitting. This is the case with $T$=9 where the learning of 100 bases leads to significantly better results on the training set but poorer performance on the evaluation set. Finally, with sufficiently large time spans and base sets sparse features tend to outperform MFCCs, confirming our conjecture that sparse coding leads to better representation for speech signals.

| | | PER% | | | |
|---|---|---|---|---|---|
| | | Train | | Eval | |
| | | $R$=50 | $R$=100 | $R$=50 | $R$=100 |
| SPARSE | $T$=9 | 25.89 | 28.17 | 22.62 | 28.36 |
| | $T$=15 | 25.59 | 27.15 | 21.39 | 26.80 |
| | $T$=21 | 24.99 | 27.08 | 21.03 | **26.20** |
| MFCC | | 25.54 | | 27.37 | |

Table 1: Frame-level PERs with various sparse features.

| | MLP Features | PER% | |
|---|---|---|---|
| | | Train | Eval |
| | MFCC | 25.54 | 27.37 |
| | T9 | 25.89 | 28.17 |
| | T15 | 25.59 | 27.15 |
| | T21 | 24.99 | 27.08 |
| Combination | T9+T15 | 22.76 | 26.87 |
| | T9+T15+T21 | 20.90 | 26.23 |
| Hybridisation | MFCC+T9 | 25.05 | 27.15 |
| | MFCC+T15 | 23.05 | 25.68 |
| | MFCC+T21 | 21.61 | **25.06** |

Table 2: Frame-level PERs with code combination and hybridisation. $Tn$ represents sparse features with $R = 50$ and $T = n$.

### 4.2. Combination and hybridisation

In the second experiment, we examine the complementarity of bases learnt with different configurations. We concentrate on varying the time span, and assume that different time spans will lead to the identification of different bases which thus convey different and complementary information. For instance, shorter time spans may lead to phone patterns while longer time spans may lead to word patterns. In order to make use of such complementary information we combine the sparse features derived from different sets of bases and use them to train the MLPs to conduct phone prediction. This approach is referred to as code combination.

In addition, sparse coding with longer time spans is assumed to be complementary to shorter-time analysis and thus it should also be possible to combine sparse features with MFCCs. This combination is referred to as code hybridisation since it merges sparse features and dense features. Table 2 presents the results on the training and evaluation sets with code combination and hybridisation. The three sets of 50 bases are the same as those used in the previous experiment and are combined with each other or hybridised with the 13-dimension MFCCs. We observe that both combination and hybridisation lead to performance improvements in phone prediction.

### 4.3. Hierarchical decision

Finally we examine the proposed parallel and hierarchical approach according to a three-layer architecture. In the first layer

| | MLP Features | PER% | |
|---|---|---|---|
| | | Train | Eval |
| LAYER 1 | (1) R50,T9 | 25.54 | 28.17 |
| | (2) R50,T15 | 25.89 | 27.15 |
| | (3) R100,T21,H1 | 22.49 | 26.89 |
| | (4) R100,T21,H2 | 24.12 | 26.99 |
| LAYER 2 | (1)+(2) | 23.48 | 25.94 |
| | (3)+(4) | 20.51 | 25.25 |
| LAYER 3 | (1)+(2)+(3)+(4) | 20.18 | **24.47** |

Table 3: Frame-level PERs with hierarchical decisions.

decisions are made by four MLPs based on features with four sets of bases:

1. 50 bases spanning 9 frames (R50,T9);
2. 50 bases spanning 15 frames (R50,T15);
3. the first 50 bases of the 100 bases learnt with a time span of 21 frames (R100,T21,H1);
4. the second 50 bases of the 100 bases learnt with a time span of 21 frames (R100,T21,H2).

In the second layer, decisions based on feature sets (1) and (2) are merged by an MLP whose inputs are the outputs of the two MLPs in the first layer, i.e. 96 posterior probabilities, and the outputs are 46 phone classes as before. The size of hidden nodes are fixed to 500. Decisions based on feature sets (3) and (4) are merged in the same way. Finally, in the third layer, the two decisions from the second layer are merged in the same way.

Results shown in Table 3 reveal several interesting characteristics. First, decisions based on half of the 100 bases learnt with time spans of 21 frames approach the accuracy of decisions based on the entire base set (26.20% as shown in Table 1). This observation indicates that the sparse features are indeed local and redundant and that, therefore, reliable decisions can be made using only a fraction of the full feature dimension. Second, we observe that hierarchical decisions tend to give better results than integrative decisions. This can be seen from the results of the second layer, where hierarchical decisions based on the two sets of 50 bases (R50,T9 and R50,T15) give better performance than integrated decisions using code combination (Table 2). Similarly, hierarchical decisions based on the two subsets of 50 bases (R100,T21,H1 and R100,T21,H2) give better performance than decisions based on the entire base set (Table 1). Comparing results for training and evaluation sets, it seems that hierarchical decision making can mitigate the overfitting problem by making decisions based on feature subspaces. This is analogous to sub-band speech recognition [15] and the TRAP approach based on multiple critical bands [16], where the locality and redundancy of the speech spectrum leads to noise-robust recognition through the integration of decisions based on sub-bands.

## 5. Conclusions

This paper investigates the application of large-scale sparse coding to speech recognition. We first report an online learning technique to learn a large set of bases, which leads to high-dimensional sparse representations of speech signals. To overcome the associated computational burden we propose a parallel and hierarchical decision making approach. Here, local decisions are based on feature subsets. Global decisions are made by combining local decisions in a hierarchical way. In a frame-level phone recognition task experiments demonstrate that the hierarchical approach not only solves the computational burden, but also provides better performance compared to the conventional approach based on full dimensions. It is expected that improvements in phone prediction lead to corresponding improvements for continuous speech recognition using the tandem framework or the hybrid approach.

Further work is required to optimise the approach and to thoroughly assess the potential. For instance, in the current work we simply segment the feature dimension linearly and exclusively, while nonlinear and overlapping segmentation should be investigated. Clustering might also help to discover optimal dimension groups for local decisions. Dense features other than phone posteriors should also be investigated.

## 6. Acknowledgements

This work was partially supported by the French Ministry of Industry (Innovative Web call) under contract 09.2.93.0966, 'Collaborative Annotation for Video Accessibility' (ACAV).

## 7. References

[1] P. Földiák and M. Young, "Sparse coding in the primate cortex," in *The Handbook of Brain Theory and Neural Networks*, M. Arbib, Ed. MIT Press, 1995, pp. 895–898.

[2] R. Ribshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, vol. 58, pp. 267–288, 1996.

[3] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, pp. 33–61, 1999.

[4] T. Virtanen, "Separation of sound sources by convolutive sparse coding," in *Proc. SAPA'2004*, 2004.

[5] M. G. Jafari, S. A. Abdallah, M. D. Plumbley, and M. E. Davies, "Sparse coding for convolutive blind audio source separation," in *ICA'06*, 2006, pp. 132–139.

[6] M. Schmidt, J. Larsen, and F.-T. Hsiao, "Wind noise reduction using non-negative sparse coding," in *Proc. IEEE 2007 workshop on Machine Learning for Signal Processing*, 2007, pp. 431–436.

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

[8] W. Smit and E. Barnard, "Continuous speech recognition with sparse coding," *Computer Speech and Language*, vol. 23, no. 2, 2009.

[9] G. Sivaram, S. Nemala, M. Elhilali, T. Tran, and H. Hermansky, "Sparse coding for speech recognition," in *Proc. ICASSP'10*, 2010.

[10] 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, January 2010.

[11] R. Grosse, R. Raina, H. Kwong, and A. Y. Ng, "Shift-invariant sparse coding for audio classification," in *Proc. of Twenty-third Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007.

[12] M. Heckmann, "Supervised vs. unsupervised learning of spectro temporal speech features," in *Proc. SAPA 2010*, 2010.

[13] N. Morgan and H. Bourlard, "Continuous speech recognition: An introduction to the hybrid HMM/connectionist approach," *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 24–42, May 1995.

[14] H. Hermansky, D. P. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. ICASSP'00*, Istanbul, Turkey, June 2000, pp. 1635–1638.

[15] S. Tibrewala and H. Hermansky, "Sub-band based recognition of noisy speech," in *Proc. ICASSP'97*, 1997.

[16] H.Hermansky and S.Sharma, "Temporal patterns (TRAPS) in ASR of noisy speech," in *Proc. ICASSP'99*, Phoenix, Arizona, USA, Mar. 1999.