# Dictionary Learning for Sparse Representation of Neural Network Exemplars in Speech Recognition

Pranay Dighe*†, Afsaneh Asaei*, and Hervé Bourlard*†
*Idiap Research Institute, Martigny, Switzerland, †École Polytechnique Fédérale de Lausanne, Switzerland
Emails: {pranay.dighe, afsaneh.asaei, herve.bourlard}@idiap.ch

***Abstract*—Conventional speech recognition systems relying on exemplar-based sparse representation require huge size exemplars collection to represent the linguistic units. Recent work demonstrates that despite of consistent improvement in automatic speech recognition performance, increasing the size of exemplars collection after a certain (very large) dimension leads to only minor improvements [1]. This observation suggests the need for a better procedure to find a limited size collection of exemplars that can be used for sparse representation. In the present study, the exemplars are neural network sub-word conditional posterior probabilities. In this context, we study the application of dictionary learning for sparse modeling. We demonstrate that the posterior exemplars live in a low-dimensional manifold that can be modeled as a union of subspaces. Furthermore, we evaluate the performance of dictionary learning for exemplar-based speech recognition to compare and contrast it with the traditional exemplars collection approach.**

## I. Union of Subspaces Model

Dictionary learning for sparse representation relies on the assumption that the data can be modeled as a union of subspaces. In this section, we provide supporting evidence that the neural netwrok exemplars conform to this model.

We perform a simple experiment of template matching using dynamic time warping (DTW) for 75 words-vocabulary set of Phonebook database [2]. Exemplars here are in form of (deep) neural network based phone posterior probabilities [3]. Out of 11 utterances for each word, we keep 4 utterances as training templates and use the rest for testing. The 4 utterances in the training set were used to create 15 combinatorial, 1 to 4-sparse templates for DTW matching by averaging after alignment.

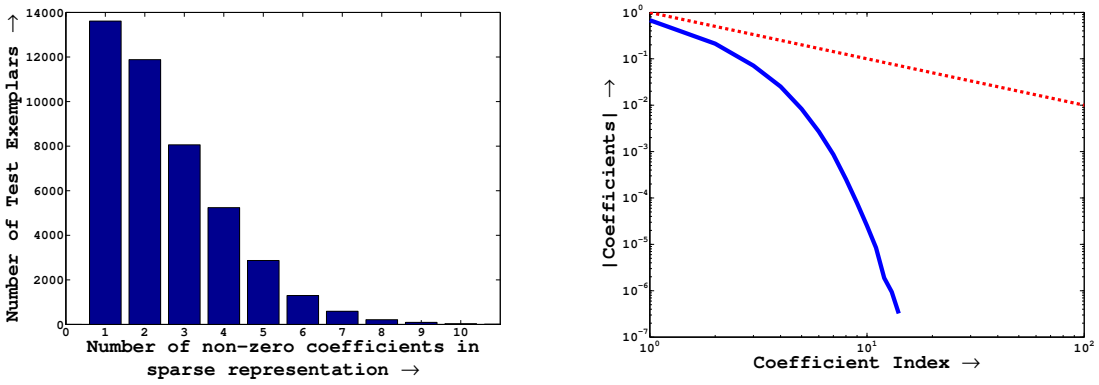
$$\begin{aligned}
&\text{1-sparse templates :}\{T_{U_1}, T_{U_2}, T_{U_3}, T_{U_4}\}\\
&\text{2-sparse templates :}\{T_{U_1U_2}, T_{U_1U_3}, T_{U_1U_4}, T_{U_2U_3}, T_{U_3U_4}\}\\
&\text{3-sparse templates :}\{T_{U_1U_2U_3}, T_{U_2U_3U_4}, T_{U_1U_2U_4}, T_{U_1U_3U_4}\}\\
&\text{4-sparse templates :}\{T_{U_1U_2U_3U_4}\}
\end{aligned} \tag{1}$$

We then quantify the DTW distance of the test utterances with the new constructed templates. The weighted symmetric Kullback-Leibler (KL) divergence is used as the distance measure as it was shown to be an "optimal" metric for neural network exemplars [4]. The smaller distance indicates better characterization of the test templates using the training data. This experiment is run for all test data and the results are listed in Table I. We observe that only 4.9% of test utterances have the least characterization error using a single closest template (DTW assumption). Moreover, only 9.7% are best characterized by the template obtained from averaging the full training set (KL-hidden Markov model - HMM assumption [3]). On the other hand, all remaining 85.4% of the utterances have the least characterization error using the templates which are obtained as a combination of a few (2 or 3) training exemplars. This observation confirm the hypothesis of the effectiveness of the union of subspace approach to model the neural network exemplars.

Another experiment was conducted on Numbers database [5], where we have huge amount of training data. Instead of k-sparse template matching using DTW as in Phonebook experiment, here we learn dictionaries of size of order $\sim$1000 columns from training data. We perform sparse recovery of test data using these dictionaries and analyse the support-size (number of non-zeros coefficients) of sparse representation. The results are illustrated in Figure 1. We observe that 31% of the test exemplars are represented by one dictionary column whereas 69% are characterized by a linear combination of very small number of columns corresponding to their sparse representation. For dictionary learning and sparse recovery, online dictionary learning algorithm [6] and lasso solver [7] were used respectively.

## II. Dictionary Learning

Once we confirm that the union of subspace model holds for neural network exemplars, we demonstrate experimentally that dictionary learning improves characterisation of the feature space as compared to a simple collection of all exemplars of the training set while its cardinality is still far smaller than the collection size. In isolated word recognition experiment on Phonebook 75-vocabulary dataset, a single exemplar is used as a warm start for dictionary initializing. The remaining 3 exemplars in the training set are then used for updating the dictionary columns using online dictionary learning algorithm [6]. Alternatively, 4 training exemplars are concatenated to form a dictionary for sparse representation. A similar comparison was done for connected digit recognition on Numbers database, where we can either learn word-specific dictionaries or we can directly represent each word using all training exemplars [8]. The results are listed in Table II. We can see that the dictionary learning procedure is quite effective; it can benefit from the abundance of the training data, while it enables us to keep the dimensionality of the exemplar space small and at the same time improve the performance. This observation confirms that dictionary learning is a more efficient way for sparse representation than exemplar collection.




(a) (b)

Fig. 1: Analysis of Sparsity: (a) gives a plot depicting the number of test exemplars versus the number of non-zero coefficients in their sparse representation and (b) shows the fast decay of the non-zero coefficients (blue solid line) as compared to the power-law decay (red dash line).

| | Winning Hypotheses | Average DTW Matching Cost |
|---|---|---|
| 1-sparse | 23/464 | 95.60 |
| 2-sparse | 177/464 | 84.77 |
| 3-sparse | 219/464 | 84.14 |
| 4-sparse | 45/464 | 86.56 |

TABLE I: Comparison of k-sparse templates for characterisation of the test exemplars using the collection of training exemplars. The total number of test exemplars is 464, out of which 23 test exemplars has the least characterization error if a single training exemplar is used. Similarly, 177, 219 and 45 of them have the least characterization error if 2, 3 and 4 of the training exemplars is used for representation respectively. This observation is inline with the union of subspace model of neural network exemplars.

| Task | Dictionary | Collection of Exemplars |
|---|---|---|
| Phonebook | **97.2** | 97.0 |
| Numbers | **85.4** | 78.6 |

TABLE II: Comparing the speech recognition accuracy (%) on Phonebook (isolated word recognition) and (connected word recognition using dictionary learning versus collection of exemplars. Accuracies in case of Connected Digit are given by (100 - WER), where WER is word error rate obtained by Levenshtein distance. The size of training data in Phonebook is small. In this case the dimension of dictionary exemplars (number of learned atoms) is 25% of the full training set. The size of training data in Numbers corpus is large. In this case the dimension of dictionary exemplars (number of learned atoms) is $\sim$3% of the full training set.

## ACKNOWLEDGMENT

The research leading to these results has received funding from by SNSF project on "Parsimonious Hierarchical Automatic Speech Recognition (PHASER)" grant agreement number 200021-153507. The authors would like to acknowledge Dr. David Imseng for his assistance with speech recognition experiments.

## REFERENCES

[1] J. Gemmeke, L. Ten Bosch, L. Boves, and B. Cranen, "Using sparse representations for exemplar based continuous digit recognition," in *Proc. EUSIPCO*. Citeseer, 2009, pp. 24–28.

[2] J. Pitrelli, C. Fong, S. Wong, J. Spitz, and H. Leung, "Phonebook: a phonetically-rich isolated-word telephone-speech database," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1, May 1995, pp. 101–104 vol.1.

[3] G. Aradilla, H. Bourlard, and M. Magimai-Doss, "Using KL-based acoustic models in a large vocabulary recognition task." in *INTERSPEECH*, 2008, pp. 928–931.

[4] S. Soldo, M. Magimai-Doss, J. Pinto, and H. Bourlard, "Posterior features for template-based ASR," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4864–4867.

[5] R. A. Cole, M. Noel, T. Lander, and T. Durham, "New telephone speech corpora at cslu," 1995.

[6] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research (JMLR)*, vol. 11, pp. 19–60, 2010.

[7] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani *et al.*, "Least angle regression," *The Annals of statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[8] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2067–2080, 2011.