IDIAP RESEARCH REPORT

SPARSE HIDDEN MARKOV MODELS FOR EXEMPLAR-BASED SPEECH RECOGNITION USING DEEP NEURAL NETWORK POSTERIOR FEATURES

Pranay Dighe Afsaneh Asaei Hervé Bourlard

Idiap-RR-19-2016

AUGUST 2016

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Sparse Hidden Markov Models for Exemplar-based Speech Recognition Using Deep Neural Network Posterior Features

Pranay Dighe[1,2]*, Afsaneh Asaei*[1]*, Hervé Bourlard*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{pranay.dighe,afsaneh.asaei,herve.bourlard}@idiap.ch

Abstract

Statistical speech recognition has been cast as a natural realization of the compressive sensing and sparse recovery. The compressed acoustic observations are sub-word posterior probabilities obtained from a deep neural network (DNN). Dictionary learning and sparse recovery are exploited for inference of the high-dimensional sparse word posterior probabilities. This formulation amounts to realization of a *sparse* hidden Markov model where each state is characterized by a dictionary learned from training exemplars and the emission probabilities are obtained from sparse representations of test exemplars. This new dictionary-based speech processing paradigm alleviates the need for a huge collection of exemplars as required in the conventional exemplar-based methods. We study the performance of the proposed approach for continuous speech recognition using Phonebook and Numbers'95 database.

Index Terms: Speech recognition, Deep neural network posterior features, Compressive sensing, Dictionary learning, Sparse modeling, Hidden Markov models.

1. Introduction

Exemplar-based (template-based) methods and stochastic approaches relying on hidden Markov model (HMM) are often considered as two distinct approaches to automatic speech recognition (ASR). While later enjoys many years of extensive development, exemplar-based methods have recently regained more serious attention [1, 2, 3, 4]. In theory, having an "infinite" amount of exemplars, and the "right" distance measure, "optimal" recognizers could be sought [5]. In practice however, development of an "optimal" exemplar-based system indicates massive memory and computational requirements.

A typical exemplar-based system uses spectral representation of speech signal and a collection of such training exemplars for ASR [2, 3, 6, 7, 8]. It has also been shown that constructing a dictionary of clean exemplars from the training data enables robust ASR if the noisy test exemplars are reconstructed as a sparse linear combination of the training exemplars. Enforcing the sparsity structure has shown to discriminate the subspaces of noise and clean speech leading to some de-noising or separation effect [3, 9]. The exemplar-based sparse representation is also exploited to provide new features to a HMM-based ASR system [2]. In this paper, we focus on exemplar-based sparse representation with two main distinctions: (1) Unlike the previous spectral exemplars, we use sub-word posterior probabilities estimated by a deep neural network (DNN) as exemplars. (2) Instead of taking the collection of training data for sparse representation, we exploit a principled way of dictionary learning for sparse representation.

The core assumption of this approach is that any possible realization of a linguistic unit (eg. a word) lies on a *low-dimensional* surface (a non-linear manifold) modeled by union of subspaces spanned by exemplars already seen in the training set. As a result, the representation of a word in a high-dimensional word posterior space is sparser than the dense sub-word posterior probability space. Using the posterior exemplars, we demonstrate that the statistical speech recognition can be cast as a compressive sensing problem where posterior exemplars are the compressed acoustic observations and higher level word inference requires a high-dimensional sparse recovery. This linguistic compressive mechanism has to be learned from the training exemplars.

Furthermore, the sequential sparse recovery enables us to process the temporal evolution of the word (traversing a path through the non-linear word manifold). We discuss rigorously that this approach develops naturally into a *sparse*-HMM configuration (as in [10]) for ASR where each hidden state is associated with a dictionary. The state models a point on the word manifold using sparse linear combinations of dictionary atoms and the emission likelihoods of the data point are obtained from the weights in the sparse representations.

This paper is organized as follow: Section 2 explains the view on ASR using the CS principles relying on dictionary learning and sparse recovery. We explain how this approach amounts to realization of a sparse modeling HMM in Section 3. The experimental analysis is presented in Section 4.3 and the conclusions are drawn in Section 5.

2. Compressive Sensing Perspective

Each speech utterance is composed of a few words. If we consider the vector representation of an utterance in a linguistic space where each component corresponds to a unique word, this representation is high-dimensional whereas the informative components are highly sparse. The key idea is to exploit the fact that the input features of the ASR system are compressed observations of this naturally high-dimensional representation problem.

In this paper, the compressed acoustic observations are sub-word conditional posterior probabilities obtained from a deep neural network (DNN). Let $\{q_k\}_{k=1}^{K}$ denote the sub-word (e.g. phonetic) classes. Given an input feature vector x_t at time t, the posterior probability vector

$$z_t = [p(q_1|x_t)p(q_2|x_t)\dots p(q_K|x_t)]^\top$$

where $.^\top$ denotes the transpose operator, is estimated using DNN. According to the marginalization rule of probabilities,

the following relation holds, assuming $p(q|w,x) = p(q|w)$ [8]:

$$\underbrace{\begin{bmatrix} p(q_1|x_t) \\ p(q_2|x_t) \\ \vdots \\ p(q_K|x_t) \end{bmatrix}}_{z_t} = \underbrace{\begin{bmatrix} p(q_1|w_1) & \cdots & p(q_1|w_L) \\ p(q_2|w_1) & \cdots & p(q_2|w_L) \\ \vdots & \vdots & \\ p(q_K|w_1) & \cdots & p(q_K|w_L) \end{bmatrix}}_{\text{Dictionary: } \mathbf{D}=[d_1 \ldots d_l \ldots d_L]} \underbrace{\begin{bmatrix} p(w_1|x_t) \\ \vdots \\ p(w_l|x_t) \\ \vdots \\ p(w_L|x_t) \end{bmatrix}}_{\alpha_t} \tag{1}$$

To model the space of phonetic representation for each word, we assume that the posterior exemplars lie on a low-dimensional (non-linear) manifold which can be characterized as a union of subspaces (UoS). Hence, we model each column (atom) d_l of the dictionary as

$$\underbrace{\begin{bmatrix} p(q_1|w_l) \\ p(q_2|w_l) \\ \vdots \\ p(q_K|w_l) \end{bmatrix}}_{d_l} = \underbrace{\begin{bmatrix} p(q_1|sw_1^{w_l}) \ldots p(q_1|sw_{S_{w_l}}^{w_l}) \\ p(q_2|sw_1^{w_l}) \ldots p(q_2|sw_{S_{w_l}}^{w_l}) \\ \vdots \\ p(q_K|sw_1^{w_l}) \ldots p(q_K|sw_{S_{w_l}}^{w_l}) \end{bmatrix}}_{\text{Word manifold dictionary: } \mathbf{D}_{w_l}} \underbrace{\begin{bmatrix} p(sw_1^{w_l}|w_l) \\ \vdots \\ p(sw_s^{w_l}|w_l) \\ \vdots \\ p(sw_{S_{w_l}}^{w_l}|w_l) \end{bmatrix}}_{a_{w_l}} \tag{2}$$

where $sw_s^{w_l}$ denotes the s^{th} subspace underlying the word w_l, S_{w_l} represents the total number of (over-complete) "bases" to model the sub-space of word w_l and

$$\alpha_t = [a_{w_1}^\top p(w_1|x_t)\ a_{w_2}^\top p(w_2|x_t)\ \ldots a_{w_L}^\top p(w_L|x_t)]^\top$$

Equations (1) and (2) lead us to a very intuitive and natural representation for continuous speech in terms of posterior features and word-to-subword hierarchical dictionaries obtained as

$$\mathbf{D} = [\mathbf{D}_{w_1} \cdots \mathbf{D}_{w_l} \cdots \mathbf{D}_{w_L}]$$

The dictionary $\mathbf{D}$, has an internal partitioning defined by the boundaries of individual sub-dictionaries $\mathbf{D}_{w_l}$. In addition, a sequence of posterior features $\mathbf{Z} = [z_1, \ldots z_t]$, extracted from an utterance of word w_l, will have a hierarchical group structure underlying the individual sparse representation α_t where all the coefficients tend to collaborate to activate a higher level group corresponding to w_l. The sparse representation of $\mathbf{Z}$ yields a matrix $\mathbf{A} = [\alpha_1, .., \alpha_t]$ where the support of the sparse coefficients hold a blocks structure as depicted in Figure 1.

Based on the above formulation, ASR problem is an instance of a compressive sensing problem with the two key components:

1. Dictionary learning for sparse representation of the posterior exemplars.
2. Structured sparse recovery for high-dimensional inference of word probabilities.

In Section 3, we explain how this formulation meets the hidden Markov model framework.

3. Sparse Modeling HMM

The sparse modeling of posterior exemplars can be expressed through formulation of a novel sparse HMM configuration

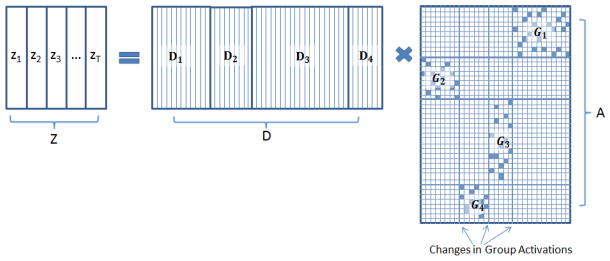

Figure 1: Given a sequence of acoustic features in $\mathbf{Z}$, the sparse representation matrix $\mathbf{A}$ will have a block structure associated to the word-specific dictionaries where the inner block coefficients are sparse. Such a collaborative hierarchical sparsity structure can be exploited using C-HiLasso algorithm [11] for the sparse recovery.

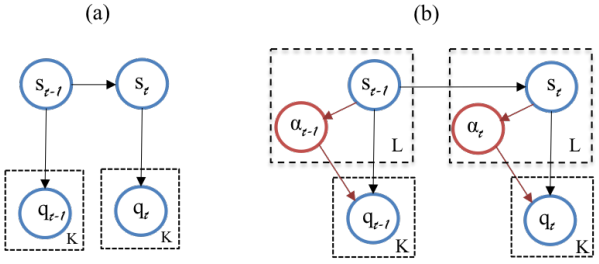

Figure 2: Graphical model for (a) the conventional HMM and (b) sparse modeling HMM. Each state of the conventional HMM is now replaced by a sparse modeling dictionary with an exponentially large number of sub-states. The activation of the sparse sub-states is controlled by the sparse representation vector a.

where the hidden states are characterized with the sparse representation dictionaries; Figure 2 depicts a graphical model of this idea. We explain the details of this new modeling paradigm in this section.

For brevity and simplicity, we assume that each hidden state models the manifold for exactly one word from the vocabulary. The hidden state corresponding to word w_l is thus modeled by the word manifold dictionary $\mathbf{D}_{w_l}$. The goal is to infer the most probable sequence of hidden states for a given sequence of observations. The observation z_t associated to the state $s_t = w_l$ is modeled as $z_t = \mathbf{D}_{w_l} a_{w_l}$ where $\mathbf{D}_{w_l} \in \mathbb{R}^{K \times S_{w_l}}$ is an over-complete dictionary ($K < S_{w_l}$). Sparse representation of the observation z_t using $\mathbf{D}_{w_l}$ gives the sparse sub-word posterior probability vector $a_{w_l} = [p(sw_s^{w_l}|w_l) \ldots p(sw_{S_{w_l}}^{w_l}|w_l)]^\top$.

The dictionary $\mathbf{D}_{w_l}$ characterizes the (non-linear) manifold associated to the state w_l. Sparse recovery of a_{w_l} using this dictionary indicates that the observation z_t lies on a low-dimensional subspace that can be characterized as a linear combination of the subspaces defined by the dictionary atoms. The dictionary defines an exponentially large number of subspaces where the observation z_t can live. Sparse recovery of a_{w_l} using this dictionary leads to the selection of independent subspaces to characterize the low-dimensional subspace of the observation z_t. We refer to this new HMM configuration as sparse modeling HMM.

We assume the model mismatch (noise) for the dictionary-based modeling, i.e. $z_t = \mathbf{D}_{w_l} a_{w_l}$ to be an independent Gaus-

Algorithm 1 EM Algorithm - *Sparse* HMM

Require: : $\mathbf{Z}$, λ (regularization parameter), $\mathbf{D}^0$ (initialization)

1: **for** $i = 1$ `to max-iter` **do**
2: **for** $t = 1$ `to` T **do**
3: Sparse Coding of z_t to determine α_t:

$$\alpha_t = \arg\min_{\alpha} \left\{ \frac{1}{2} \|z_t - \mathbf{D}^{(i-1)}\alpha\|_2^2 + \lambda \|\alpha\|_1 \right\} \quad (4)$$

4: **end for**
5: Updating $\mathbf{D}^{(i)}$ with $\mathbf{D}^{(i-1)}$ as warm restart:

$$\mathbf{D}^{(i)} = \arg\min_{\mathbf{D}} \left\{ \sum_{j=1}^{t} (\frac{1}{2} \|z_j - \mathbf{D}\alpha_j\|_2^2 + \|\alpha_j\|_1) \right\} \quad (5)$$

6: **end for**
7: **return** $\mathbf{D}^{\texttt{max-iter}}$

sian noise distributed as $\mathcal{N}(0, \sigma_{w_l}^2 \mathbf{I})$. Thus, the distribution of observation z_t is given by $\mathcal{N}(\mathbf{D}_{w_l} a_{w_l}, \sigma_{w_l}^2 \mathbf{I})$. To have a sparse latent variable a_{w_l}, a probability distribution which generates sparse vectors can be used as a prior. For simplicity, we choose a Laplace distribution with parameter $\lambda_{w_l} > 0$[1]. The estimation of the emission likelihoods would be

$$p(z_t|s_t = w_l) = \int_{a_{w_l}} p(z_t|a_{w_l}, s_t = w_l) p(a_{w_l}|s_t = w_l)\, da_{w_l} \quad (3)$$

To describe the sparse modeling HMM, two sets of parameters must be learned: (1) state-distribution parameters $\Theta_s = (\mathbf{D}_{w_l}, \sigma_{w_l}, \lambda_{w_l})$ and (2) state-transition probabilities. The transition probabilities can be directly computed from the frequency of word transitions. Then, our goal is to learn the state-distribution parameters given some training data for each individual word[2]. Since $p(z_t|s_t)$ depends on the hidden variable a_{w_l}, we use EM algorithm to maximize the log-likelihood of the observations corresponding to each state $s_t = w_l$, $\mathcal{L}_{\Theta_s} = \sum_t \log p_{\Theta_s}(z_t|s_t = w_l)$ w.r.t. the parameters Θ_s. Due to the carefully chosen distributions for model-mismatch and a_{w_l}, the estimation and maximization of log-likelihood $\log p_{\Theta_s}$ reduces to standard l_1-norm minimization and dictionary learning problems of compressive sensing [10]. We summarize the EM procedure as derived in [10] in Algorithm 1.

Algorithm 1 is analogous to the method of optimal directions (MOD) in sparse dictionary learning [15]. It enables us to leverage any combination of dictionary learning algorithm and sparse solver that fulfills the optimization equations 4 and 5 and updates α_t and $\mathbf{D}$ jointly. We refer to the MOD approach to learning the sparse HMM parameters as MOD-HMM. In particular, we use the fast variant of this idea developed as the online dictionary learning by [16]. Furthermore, we study alternative approaches to dictionary learning relying on KSVD [17] as well as sparse NMF [18]. We refer to these two variants of sparse HMM implementation as KSVD-HMM and NMF-HMM accordingly.

[1] Although sampling from a Laplace does not generate sparse vectors [12], we choose this prior as it leads to LASSO sparse recovery during EM parameter learning [13, 14].

[2] Recall that we refer to each state as words. The algorithms are general and applicable to any linguistic unit

It may be noted that the sparse HMM framework is fundamentally different than hybrid exemplar-based/HMM proposed in [3] in the following ways:

1. Exemplars in the proposed framework are posterior probabilities, so the probabilistic relation in (1) can amount to direct estimation of word posterior probabilities.
2. EM derivation of the HMM state parameters requires dictionary learning. This key step has been ignored in all previous exemplar-based methods, to the extent of our knowledge. Learning dictionary essentially finds an over-complete basis set from the training data and alleviates the need of storing a huge collection of exemplars.
3. The earlier proposed exemplar-based sparse representation approach (in [3]) for ASR is hybridised with HMM and exploits the (state) label matrix for each exemplar obtained from a conventional HMM system for decoding. In contrast, we decode obtained sparse recovery based likelihoods directly.

4. Experiments and Analysis

The experiments are designed to evaluate the performance of the proposed approach for continuous speech recognition. In addition, we conduct some intial tests on isolated word recognition to compare different algortihmic approaches for implementation of the sparse HMM.

4.1. Database and Setup

The isolated word recognition experiments are performed on Phonebook speech corpus [19] recorded on single microphone channel at 16KHz. The performance is averaged over 8 different sets of 75 words vocabulary each on Phonebook data. Each word has around 11 utterances, out of which we use 4 for learning dictionaries and the rest for testing. This setup is similar to the experiments in [20]. To capture contextual information in the data, c adjacent posterior frames from both sides of a frame are concatenated to form a long context appended frame. A context-size of 20 frames was found to perform best for Phonebook. For continuous speech recognition, Numbers'95 corpus [21] is used which is recorded over telephone channel at 8KHz for connected digits tasks. A subset (only utterances that involve the 10 digits (*zero* to *nine*) and *oh*) of Numbers'95 corpus are used for the experiments. Context size of 8 frames and 1 frame are used for word-based and state-based experiments respectively in case of Numbers'95 corpus.

Posterior features are obtained from a deep neural network with 3 hidden layers. Features (MFCC+Δ + $\Delta\Delta$) with a context of 4 frames are used at the input layer. The 27 dimensional phone posteriors and 83 dimensional state posteriors are computed for word based and state based experiments respectively.

4.2. Algorithmic Approaches for Sparse HMM

We learn a dictionary for each of the words using Algorithm 1. We have already seen in (2) that the columns of this dictionary represent the subspaces of the word manifold. Sparse recovery of a test exemplar using this dictionary yields a sparse vector that essentially denotes the low-dimensional UoS of the word manifold in which the given test exemplar lies. When a sequence of such test exemplars are decoded using a word manifold dictionary, the resulting sparse representation traverses on

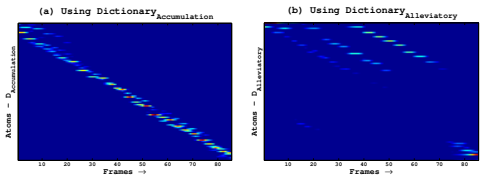

Figure 3: Sparse representation of the word "Accumulation" under sparse recovery with dictionary that corresponds to (a) $\mathbf{D}_{\text{Accumulation}}$ and (b) $\mathbf{D}_{\text{Alleviatory}}$. The sequencing pattern underlying the sparse representation is exhibited when the correct dictionary is used.

the word manifold through different unions of subspaces. Figure 3 shows this phenomenon when a word is decoded using the correct word manifold dictionary versus when a wrong dictionary is used.

Based on Algorithm 1, we employ different dictionary learning algorithms to develop different versions of sparse HMM. The EM derivation of the sparse HMM parameters is analogous to the method of optimal directions in dictionary learning and sparse recovery. We use a fast implementation of this idea referred to as the *online* dictionary learning [16]. Furthermore, we use the K-SVD [17] and NMF [18] algorithms for dictionary learning. Unlike the online dictionary learning, the KSVD and NMF approaches update the dictionary and sparse coefficients jointly. To estimate the sparse representation using NMF algorithmic updates, we use the solver presented in [3] which optimizes the generalized kullback-divergence as opposed to the Euclidean reconstruction error.

In each case, a given test utterance is decoded using all word manifold dictionaries. The sparse recovery with minimum reconstruction error is used as the decision criteria for labeling the test utterance. Results are compared in Table 1 for different dictionary learning approaches. MOD-HMM based on online dictionary learning [16] was found to perform best both in terms of performance and computational speed. The HMM-MLP recognition accuracy reported for this task is 98.8% as reported in [20].

System	Accuracy(in %)
MOD-HMM	97.8
KSVD-HMM	88.9
NMF-HMM	89.0

Table 1: A comparison of different dictionary learning algorithms for sparse HMM implementation evaluated for isolated word recognition on Phonebook 75 words task.

Furthermore, we compare the performance of word-based dictionary learning against conventional approach of using collection of exemplars for sparse representation in Table 2. Although, dimension of the dictionary (number of learned atoms) is only a small fraction of the size of collection of exemplars (25% for Phonebook and ∼3% for Numbers), dictionaries give much better recognition performance.

4.3. Continuous Speech Recognition Results

The performance of sparse HMM is evaluated for connected digit recognition task using Numbers'95 corpus. The word error rate of the HMM-MLP system is 7.2% [22].

We implement a system similar to one used in [6], however *the collection of exemplars is replaced by dictionary learning*. Using the dictionaries, we estimate the posterior-based sparse representation and use them to decode the most likely sequence of digits relying on Viterbi dynamic programming. We only use MOD based dictionary learning in these experiments. Dictionaries can be learned at word or state-level as in [6]. We use word and state alignments on training data for learning word or state-specific dictionaries. A dictionary for *pause* class is also learned from exemplars of pauses in the training data. In case of continuous speech, we can decode an utterance 1) either using manifold dictionary (D_{w_l} in eq.2) and compare reconstruction errors given by each dictionary D_{w_l} or 2) using the complete dictionary (D in eq.1) and rely on group lasso or C-HiLasso [11] for efficient sparse recovery as these variants of Lasso are tailor-made to handle structured sparsity exhibited in figure 1. A comparison of results is shown in Table 3 in terms of word error rate % (WER).

Task	Dictionary	Collection of Exemplars
Phonebook	**97.2**	97
Numbers	**85.4**	78.6

Table 2: Comparing the speech recognition accuracy (%) using dictionary learning versus collection of exemplars.

System	D_{w_l}+Lasso	D+C-HiLasso
Word Dictionary	14.6	18.5
State Dictionary	14.0	15.9

Table 3: Connected digit recognition word error rate (%) on Numbers'95 database.

We can see that state dictionaries with word based decoding performs the best, whereas a collection of (posterior) exemplars approach has the 21.4% WER for this task. As discussed in section 3, this is the first result on decoding the sparse recovered posteriors *directly* to word utterances, and unlike the previous work [3] the exemplar-based sparse representation scores are not hybridized with HMM for decoding.

5. Conclusions

We have proposed and evaluated a novel compressive sensing approach for speech recognition. Posterior exemplars are central players in development of this new statistical ASR framework that builds on dictionary learning to model a non-linear manifold of word posterior representations and sparse recovery for high-dimensional inference of the sparse word probabilities. Learning a dictionary alleviates the need of huge collection of exemplars needed in previous sparse representation approaches. We found this formulation analogous to realization of a hidden Markov model where the state distributions are characterized by a sparse coding dictionary. Further studies in this direction may lead to development of a novel ASR paradigm based on hierarchical sparse modeling of DNN posterior probabilities that can advance the conventional DNN-HMM framework.

6. Acknowledgments

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.

7. References

[1] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, D. Van Compernolle, K. Demuynck, J. F. Gemmeke, J. R. Bellegarda, and S. Sundaram, "Exemplar-based processing for speech recognition: An overview," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 98–113, 2012.

[2] T. N. Sainath, B. Ramabhadran, M. Picheny, D. Nahamoo, and D. Kanevsky, "Exemplar-based sparse representation features: From TIMIT to LVCSR," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 8, pp. 2598–2613, 2011.

[3] 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.

[4] M. De Wachter, M. Matton, K. Demuynck, P. Wambacq, R. Cools, and D. Van Compernolle, "Template-based continuous speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1377–1390, 2007.

[5] P. A. Devijver and J. Kittler, *Pattern recognition: A statistical approach*. Prentice-Hall London, 1982, vol. 761.

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

[7] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, and A. Sethy, "Sparse representation features for speech recognition." in *INTERSPEECH*, 2010, pp. 2254–2257.

[8] S. Bahaadini, A. Asaei, D. Imseng, and H. Bourlard, "Posterior-based sparse representation for automatic speech recognition," in *Proceeding of Interspeech*, 2014.

[9] A. Asaei, "Model-based sparse component analysis for multiparty distant speech recognition," Ph.D. dissertation, École Polytechnique Fédéral de Lausanne (EPFL), 2013.

[10] L. Tao, E. Elhamifar, S. Khudanpur, G. D. Hager, and R. Vidal, "Sparse hidden markov models for surgical gesture classification and skill evaluation," in *Information Processing in Computer-Assisted Interventions*. Springer Berlin Heidelberg, 2012, pp. 167–177.

[11] P. Sprechmann, I. Ramirez, G. Sapiro, and Y. C. Eldar, "C-hilasso: A collaborative hierarchical sparse modeling framework," *Signal Processing, IEEE Transactions on*, vol. 59, no. 9, pp. 4183–4198, 2011.

[12] V. Cevher, "Learning with compressible priors," in *Advances in Neural Information Processing Systems (NIPS)*, 2009, pp. 261–269.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[14] 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.

[15] K. Engan, S. O. Aase, and J. Hakon Husoy, "Method of optimal directions for frame design," in *ICASSP*, 1999, pp. 2443–2446.

[16] 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.

[17] M. Aharon, M. Elad, and A. Bruckstein, "KSVD: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, 2006.

[18] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems (NIPS)*, 2001, pp. 556–562.

[19] J. Pitrelli, C. Fong, S. H. Wong, J. R. Spitz, and H. C. Leung, "Phonebook: A phonetically-rich isolated-word telephone-speech database," in *ICASSP*, 1995, pp. 101–104.

[20] S. Soldo, M. Magimai-Doss, J. Pinto, and H. Bourlard, "Posterior features for template-based ASR," in *ICASSP*, 2011, pp. 4864–4867.

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

[22] G. Aradilla, "Acoustic models for posterior features in speech recognition," Ph.D. dissertation, Ecole Polytechnique Fédérale de Lausanne, Lausanne , Switzerland, 2008.