Posterior-based Sparse Representation for Automatic Speech Recognition

Sara Bahaadini[1,2], *Afsaneh Asaei*[1], *David Imseng*[1], *Hervé Bourlard*[1,2]

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

Abstract

Posterior features have been shown to yield very good performance in multiple contexts including speech recognition, spoken term detection, and template matching. These days, posterior features are usually estimated at the output of a neural network. More recently, sparse representation has also been shown to potentially provide additional advantages to improve discrimination and robustness. One possible instance of this, is referred to as exemplar-based sparse representation.

The present work investigates how to exploit sparse modelling together with posterior space properties to further improve speech recognition features. In that context, we leverage exemplar-based sparse representation, and propose a novel approach to project phone posterior features into a new, high-dimensional, sparse feature space. In fact, exploiting the properties of posterior spaces, we generate, new, high-dimensional, linguistically inspired (sub-phone and words), posterior distributions. Validation experiments are performed on the Phonebook (isolated words) and HIWIRE (continuous speech) databases, which support the effectiveness of the proposed approach for speech recognition tasks.

Index Terms: speech recognition, posterior feature, sparse representation, hidden variable, exemplar-based modelling.

1. Introduction

Hidden Markov Model (HMM) based modeling and template (exemplar) based techniques are the two main approaches towards Automatic Speech Recognition (ASR). In the last three decades though, HMM-based approaches have been dominant because of their flexibility and their ability to be trained and generalize to unseen data. However, with the always increasing amount of training data, as well as the growing computational and memory resources, the potential of exemplar-based approaches are currently being revisited [1, 2, 3, 4]. These techniques use labeled speech segments, called exemplars or templates, such as phones, syllables or words. In theory, assuming an "infinite" amount of such examples, as well as the "right" representation space and the "right" distance measure, "optimal" recognizers could be sought [5].

One of the emerging approaches in exemplar-based ASR is exemplar-based sparse representation. Recently, sparse representations have gained great attention in signal processing [6] and speech recognition [7]. In [8], sparse reconstruction is shown to improve speech recognition in the presence of overlapping speech interferences. Exemplar-based sparse representation is used in noise robust ASR [3] and Continuous Speech Recognition (CSR) [2]. In [3], the ability of sparse reconstruction in source separation has been used to separate noise from the speech exemplars. In addition, sparse representation is exploited in providing exemplar-based feature vectors for an HMM-based ASR system [2].

To the best of our knowledge, all of the proposed algorithms which benefit from exemplar-based sparse representation use spectral-based features [2, 3, 9, 10]. On the other hand, posterior-based features have shown to yield promising results in exemplar-based approaches in speech recognition. For instance, it has often been shown that Dynamic Time Wrapping (DTW) on Multilayer Perception (MLP) based posterior features yields better performance than a hybrid HMM/MLP system [11]. Deep belief network based posterior [12] features have also shown promising performance in query-by-examples spoken term detection task. Of course, such approaches could also benefit from Deep Neural Networks (DNN) [13] to estimate better phone or sub-phone posteriors.

In the present study, we investigate a new approach in exemplar-based sparse representation, exploiting the properties of posterior features, to further enhance speech recognition features, and improve the discriminant properties (and hierarchical structure potential) of exemplar-based ASR. More specifically, we use sparse recovery to project phone posterior features estimated by an MLP (or a DNN) into a new high-dimensional sparse posterior space (Section 2). This projection is automatically designed to maximize sparsity and discriminant properties of the new posterior space, which may also represent a linguistic level different than phone. This level can be higher (syllable or word) or lower (sub-phone or HMM-state) than phone. To do so, a set of representative exemplars of the new space are collected in a pool of exemplars which is called "dictionary" in compressive sensing literature. The rational behind our approach is that the occurrence frequency of a specific token, e.g. word, syllable, or even sub-phone, among all tokens, is a sparse event. Hence, we impose sparsity during recovery of an exemplar from the pool of all exemplars.

We evaluate the transformation of phone posteriors into word and HMM-state space in isolated word recognition and CSR tasks respectively. For this, we introduce "sparse word posterior" (Section 3) and "sparse HMM-state posterior" (Section 4) feature spaces by sparse recovery from phone posterior features. Moreover, in the case of isolated word recognition, we make use of the exemplar-based approach flexibility [4, 14] in exploiting long-term and short-term contexts to have linguistically richer sparse word posterior features. A context adaptive technique is proposed to exploit various acoustic context spans through fusion of different sparse spaces. Finally, we investigate "hybridization" of DTW scores, to fully exploit temporal properties, and the sparse word posteriors, to maximize discrimination potential. In CSR, the probabilistic form of sparse HMM-state posterior, motivated us to use Kullback-Leibler HMM (KL-HMM) [15] as acoustic model.

2. Sparse representation

2.1. Background

A vector is defined as *sparse* if very few of its components have nonzero values. Given an observation $z \in \mathbb{R}^K$, and an over-complete dictionary matrix $D \in \mathbb{R}^{K \times L}$ where $L \gg K$, the sparse representation α is obtained through the optimization problem stated as:

$$\text{Minimize } \|\alpha\|_0 \quad \text{such that} \quad z = D \times \alpha \tag{1}$$

where the counting function $\|.\|_0 : \mathbb{R}^L \longrightarrow \mathbb{N}$ returns the number of non-zero components in its argument. The non-convex objective of (1) is often relaxed to ℓ_1 norm optimization which can be solved in polynomial time; the ℓ_1 norm, $\|\alpha\|_1$ is defined as sum of the absolute values of the components of α. Further developments consider alternative data reconstruction metrics tailored for a specific application such as classification.

In the present paper, z denotes the phone posterior vector or a short sequence of posterior vectors. The dictionary D is constructed from a large set of posterior-based exemplars or atoms. The large dimensional sparse representation α is then estimated using sparse recovery algorithm expressed as:

$$\hat{\alpha} = \arg\min_{\alpha} \{\lambda \|\alpha\|_1 + \text{KL}(z, D \times \alpha)\} \tag{2}$$

where λ denotes the regularization parameter to control the level of sparsity, and KL is the Kullback-Leibler divergence function.

2.2. Posterior-based sparse representation

A posterior probability $p(q_k|x_t)$, estimated at the MLP/DNN output q_k (associated with a phone), given an input acoustic vector x_t at time t, can be marginalized over L hidden variables r_l as follows:

$$p(q_k|x_t) = \sum_{l=1}^{L} p(q_k, r_l|x_t) = \sum_{l=1}^{L} p(q_k|r_l, x_t)p(r_l|x_t) \tag{3}$$

Considering the observation z_t consisting of the phone posterior features as $z_t = [p(q_1|x_t), \cdots, p(q_K|x_t)]$, an over-complete dictionary D constructed from the exemplars obtained by conditioning the phone posteriors on a different linguistic unit r_l, and exploiting (2) and (3), the sparse posterior-based representation α_t takes the following probabilistic form:

$$\underbrace{\begin{bmatrix} p(q_1|x_t) \\ p(q_2|x_t) \\ \vdots \\ p(q_K|x_t) \end{bmatrix}}_{z_t} = \underbrace{\begin{bmatrix} p(q_1|x_t, r_1) \cdots p(q_1|x_t, r_L) \\ p(q_2|x_t, r_1) \cdots p(q_2|x_t, r_L) \\ \vdots \quad\quad\quad\quad \vdots \\ p(q_K|x_t, r_1) \cdots p(q_K|x_t, r_L) \end{bmatrix}}_{\text{Dictionary matrix:} D} \times \underbrace{\begin{bmatrix} p(r_1|x_t) \\ p(r_2|x_t) \\ \vdots \\ p(r_L|x_t) \end{bmatrix}}_{\alpha_t} \tag{4}$$

The hidden variable r_l can be interpreted based on the atoms which are deputed in the dictionary. For example, if we put some representative of HMM-state probability vectors in the dictionary as the atoms, r_l can be considered as a hidden variable indicating HMM-state, similar to what is done in standard HMM. Then, $p(r_l|x_t)$ will be the HMM-state posterior probability, and the (norm one) normalized α_t will be the sparse HMM-state posterior representation. According to the above formulation, a posterior feature vector $z_t = [p(q_1|x_t), \cdots, p(q_K|x_t)]$ yields a sparse posterior feature vector in a different linguistically meaningful posterior space, $\alpha_t = [p(r_1|x_t), \cdots, p(r_L|x_t)]$, using a dictionary constructed from *appropriate* exemplars representative of the associated labels or hidden variables.

In practice, construction of the dictionary as described in (4) requires an online adaptation for each acoustic observation x_t. Hence, we use training data of the set of all possible units r_l for construction of D and z_t is approximated by a linear combination of training exemplars. In the following, starting from phone posterior features, where q_k is a phone variable, we provide two scenarios for extracting the sparse posterior probabilities, where r_l is either (1) a variable for sub-word in Section 3, or (2) a variable for HMM-state in Section 4.

3. Sparse word posterior

In this section, we introduce the notion of word posterior space by designing a dictionary of sub-word exemplars. The word posterior sparse representation is suitable for the task of isolated word recognition.

3.1. Dictionary design

The dictionary designed to represent the word posterior space is consisted of sub-word exemplars as atoms. The sub-word exemplars are obtained from phone posterior vector along with c left and right neighborhood frames stack into a vector, where c denotes the context size. Hence, the sub-word exemplar W_f corresponding to the f^{th} frame of the acoustic observation of training data is obtained as:

$$\begin{aligned} w_f &= \big[p(q_1|y_f, r_l), \ldots, p(q_K|y_f, r_l)\big]_{1 \times K} \\ W_f &= \big[w_{f-c}, \ldots, w_f, \ldots, w_{f+c}\big]^T_{K(2c+1) \times 1} \end{aligned} \tag{5}$$

where $.^T$ denotes the transpose operator and r_l is the variable for sub-word. As the words span variable number of frames, we denote the number of frames representing word ω by τ_ω; the dictionary thus consists of group exemplars corresponding to each word. For word w, the sub-words r_l are indexed as $l \in \{\sum_{i=1}^{w-1} \tau_i + 1, \ldots, \sum_{i=1}^{w-1} \tau_i + \tau_\omega\}$.

We define the block of τ_ω exemplars representing the ω^{th} word as $\mathcal{W}_\omega$, and the dictionary is obtained as:

$$\begin{aligned} \mathcal{W}_\omega &= [W_1, \ldots, W_{\tau_\omega}]_{K(2c+1) \times \tau_\omega} \\ D &= [\mathcal{W}_1, \ldots, \mathcal{W}_\omega, \ldots, \mathcal{W}_\varpi]_{K(2c+1) \times L} \end{aligned} \tag{6}$$

where ϖ denotes the total number of words and $L = \sum_{\omega=1}^{\varpi} \tau_\omega$.

The dictionary constructed as such exhibits a group dependency structure underlying the components of the sparse word posterior representation. In Section 3.2, the procedure of mapping the sub-word posteriors to the word posterior representation is elaborated. The block structure can be further investigated in the context of model-based sparse recovery [7] which is out of the scope of this paper.

3.2. Word posterior sparse representation

Given the dictionary of sub-word exemplars expressed in (6), and the observation vector of a test sample $Z_t = [z_{t-c}, \ldots, z_t, \ldots, z_{t+c}]$, the activations of sub-word exemplars are estimated using sparse recovery and normalized to yield the sparse sub-word posteriors. To estimate the sparse word posteriors, all the coefficients (in sparse sub-word posterior) corresponding to each word are averaged to form a word level representation.

The simplest way of using the resulting word posterior features for recognition is by direct decoding based on the maximum word posterior probability. To this end, the word posterior

sparse representation is obtained for each frame of a test sample. The word posteriors are then averaged across all frames to yield a probabilistic score for each word in the dictionary. The test sample is then recognized based on the maximum word probability. This simple decoding approach, however has the disadvantage of overlooking the inter-segment dependency (sequencing) between word segments. Hence, as further explain in Section 5.1.5, the word posterior scores integrated with a sequence matching approach such as DTW.

It may be noted that the dictionary used for word posterior representation is consisted of exemplars corresponding of word segments. We found this sub-word representation a convenient mean to tackle the variability in the number of frames representing different words. In Section 3.3, we discuss the context adaptivity obtained by this approach and in Section 5.1.4, we provide empirical insights into the "optimal" sub-word exemplars for word recognition.

3.3. Context adaptation

Due to the variability in word length, a constant c may not be efficient to obtain the sparse word posterior representation. While long-span exemplars model intra-segment dependencies more effectively, they are not in favor of short words; a large c may render the number of sub-word exemplars insufficient for representing the short words. To address the issue of appropriate context size, we propose to obtain the sparse representation using dictionaries of different context sizes. The word posterior vectors resulted from these dictionaries are averaged to yield a word posterior representation richer in modeling the intra-segment dependencies.

4. Sparse HMM-state posterior

In this section, we introduce the sparse HMM-state (sub-phone) posterior features applicable for continuous speech recognition.

4.1. Dictionary design

The dictionary designed to represent the HMM-state posterior space consists of sub-phone exemplars denoted as:

$$\rho_l = [p(q_1|y_f, r_l), \ldots, p(q_K|y_f, r_l)]_{1\times K} \tag{7}$$

where r_l indicates an HMM-state. The sub-phone or HMM-state exemplars do not need extra contextual information. The dictionary is thus formed as:

$$D = \left[\rho_1^T, \ldots, \rho_L^T\right]_{K\times L} \tag{8}$$

The state labels of exemplars are extracted from a pre-trained HMM setup. The number of samples per state is very high and it is impractical to construct a massive dictionary from all of them. To achieve a set of sample representatives with a reasonable size, an agglomerative clustering algorithm proposed in [16] is used. The centers of the clusters, which can be considered as the representatives of HMM-states, are used as atoms to populate the dictionary.

4.2. HMM-state posterior sparse representation

Given the dictionary of sub-phone exemplars and a phone posterior feature $z_t \in \mathbb{R}^K$, the sparse HMM-state posterior representation $\alpha \in \mathbb{R}^L$ is estimated by solving (2). To properly exploit the new set of features for CSR in a principled way, KL-HMM modeling [15] provides a suitable recognition back-end considering sparse state emission probabilities for direct modeling of the posterior features. The KL-HMM is trained using the set of new sparse posterior features. The details of the experimental analysis are elaborated in the following Section 5.

5. Experimental analysis

This section is dedicated to the evaluation of the proposed posterior-based sparse representations for isolated word recognition and continuous speech recognition.

5.1. Isolated word recognition

5.1.1. Database

The Phonebook corpus [17] is used for isolated word recognition. The test part of this database is used for evaluations with a similar setup as described in [18]. The test set contains 8 word lists. Each word list consists of 75 unique words. The words are pronounced by approximately 12 different speakers. For each word, one of the samples is randomly selected to construct the dictionary and the rest are kept for evaluations. The average performance over all 8 subsets is evaluated.

5.1.2. Phoneme posterior features

The initial phone posteriors are produced by a 3-layer MLP with 5,000 hidden units and 45 output units. It is trained on 232 hours of Conversational Telephone Speech data [19]. The Perceptual Linear Prediction (PLP) features are extracted from each 10 ms frame of speech and concatenated with the first and second order dynamic features to form a spectral feature vector. The spectral representation of each frame joined with the four adjacent frames both sides are used as the MLP input to extract the phone posterior features [11].

5.1.3. Sparse word posterior features

The dictionary for word posterior features is constructed using only *one sample* for each word. In contrast to the spectral-based approaches [20] where considerably more training samples are demanded, this emphasizes the merit of our posterior-based approach. Given the dictionary, the KL sparse recovery algorithm [3] is used to estimate the sparse word posterior features. The regularization parameter λ is set to 0.8 as it is shown to yield reasonable results. Alternative Euclidean distance based solver has also been investigated, but the recognition accuracy drops by more than 2%. This observation is in line with the prior evidence on suitability of the Kullback-Leibler divergence as a distance metric in posterior feature space [21].

5.1.4. Context adaptation

The word recognition results using the maximum word posterior probability decoding (Section 3.2) is depicted in Figure 1-(a) for various dictionary atom sizes, $(2c + 1)$. We can see that exploiting larger context improves the performance as the intra-segment dependency between frames is better modeled. However, by increasing the atom size beyond 60, the performance remains almost constant and after a certain size, it starts to drop. This effect can be explained as the minimum length of the test samples is around 60 frames (Figure 1-(b)) which indicates that even though using larger contexts is better to model the intra-segment dependency, it is undesirable for shorter words due to the very few number of representatives in the dictionary. The optimal atom size can be justified from its linguistic interpretation. In [17], it is reported that the average syllable per word in Phonebook is 2.7. In our experiment, the average word length is 144 frames. Dividing it by 60 frames yields 2.4, an accept-

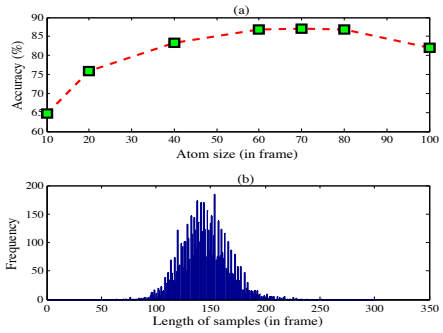

Figure 1: *(a) Word recognition performance vs. the dictionary atom size. (b) Distribution of the size of test samples.*

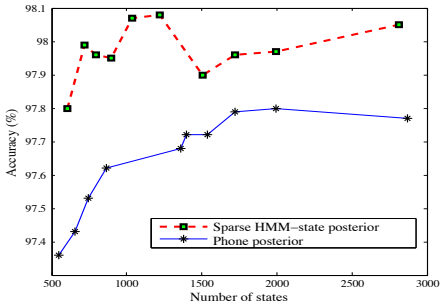

Figure 2: *Performance of the KL-HMM based recognition system with regular posterior and sparse HMM-state posterior features for varying number of states on HIWIRE corpus. Accuracy of conventional HMM/GMM system is 97.3% [22].*

able syllable per word ratio. Hence, setting the atom size equal to the *average syllable length* seems to yield an appropriate size for word recognition.

To evaluate our context adaptation scheme, the word posterior representations obtained from $c = \{10, 40\}$ dictionaries are integrated. The recognition performance using context adapted features is 87.6% which is better than both of the individual word posterior features; which yields 75.9% and 86.7% accuracy respectively (see Figure 1). We conclude that the context adapted space is able to model intra-segment dependency thus achieve a higher discrimination while enough representatives for small and medium length words are preserved.

5.1.5. Hybridization with DTW

To the best of our knowledge, DTW using phone posterior features currently achieves the best results reported on Phonebook for isolated word recognition [18]. To exploit the additional discrimination provided by sparse word posterior representation, the normalized word scores obtained from word posteriors are added to the DTW scores and used for recognition. The word recognition accuracy of the hybrid approach is 93.5% whereas DTW performs 92.2%. Hence, incorporating the sparse word posterior improves the relative performance by 17%. Although DTW is a strong technique for a sequence matching problem, it uses a local distance measure which may not take into account the full discriminative properties of the feature space. By exploiting the sparse word posteriors, complementary evidence are provided resulting in an improvement in recognition performance.

5.2. Continuous speech recognition

The CSR experiments are conducted on HIWIRE corpus [22]. This database contains about 8100 English utterances with non-native speakers. We use the same set-up as [23].

5.2.1. Phoneme posterior features

A 3-layer MLP is used to obtain the phone posterior features. The targets are 117 universal phonemes and the input is 9 frames of 39 dimensional PLP features. More details about the MLP can be found in [23].

5.2.2. Sparse HMM-state posterior

The dictionary matrix is constructed from the means of 800 clusters in phone posterior space as it is explained in Section 4. The KL-solver [3] is used for sparse representation in an 800-dimensional sparse HMM-state posterior space using 117-dimensional phone posterior features. The regularization parameter is set to 0.8. The speech recognition is achieved using the KL-HMM framework which is appropriate to model a sparse feature space [15]. The continuous speech recognition accuracy using sparse HMM-state posterior features using KL-HMM modeling is compared with the phone posteriors features in Figure 2. The conventional HMM/GMM [22] system performs 97.3%. The performance using the sparse HMM-state posteriors is 98.1% which yields 30% relative improvement compared to the conventional HMM/GMM, and 10% relative improvement compared to the KL-HMM modeling using regular phone posteriors, which yields 97.8% accuracy. Moreover, as Figure 2 illustrates, HMM-state posterior features require less number of KL-HMM states, reducing the complexity of the model used for speech recognition.

6. Conclusion

In this paper, a novel posterior-based sparse representation was proposed, exploiting exemplar-based sparse representation and the properties of the posterior feature space. The proposed approach resulted in a new type of statistical formalism where the hidden variable can accommodate different types of linguistic unit, resulting in a new way to map posterior features into a different linguistically-inspired feature space. The sparse word posteriors as well as HMM-state posteriors were obtained from phone posteriors and investigated in the context of isolated word recognition and continuous speech recognition respectively. The resulting word posterior scores, which may miss some of the temporal properties of the utterances, can be further enhanced by integrating standard (posterior-based) DTW scores. The numerical evaluations resulted in improved word recognition rate where efficient methodologies were incorporated to tackle the length variability of the words. Furthermore, the sparse HMM-state posterior features outperformed the best state-of-the-art results using the appropriate framework of KL-HMM to model sparse features while reducing the model complexity required for continuous speech recognition.

7. Acknowledgments

This work was sponsored by the Swiss NSF 200020-144281 funding on “Adaptive Multilingual Speech Processing (A-MUSE)”. Afsaneh Asaei acknowledges SNSF 200021-153507 project on “Parsimonious Hierarchical Automatic Speech Recognition (PHASER)”. The authors would like to thank Philip N. Garner from Idiap Research Institute for his fruitful discussions on the statistical sections.

8. 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] 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.

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

[8] A. Asaei, H. Bourlard, and V. Cevher, "Model-based compressive sensing for multi-party distant speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

[9] 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.

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

[11] 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.

[12] Y. Zhang, R. Salakhutdinov, H.-A. Chang, and J. Glass, "Resource configurable spoken query detection using deep boltzmann machines," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 5161–5164.

[13] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[14] G. Heigold, P. Nguyen, M. Weintraub, and V. Vanhoucke, "Investigations on exemplar-based features for speech recognition towards thousands of hours of unsupervised, noisy data," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4437–4440.

[15] 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.

[16] D. Imseng and J. Dines, "Decision tree clustering for KL-HMM," Idiap, Tech. Rep., 2012.

[17] J. Pitrelli, C. Fong, S. H. Wong, J. R. Spitz, and H. C. 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. IEEE, 1995, pp. 101–104.

[18] G. Aradilla, H. Bourlard *et al.*, "Posterior features applied to speech recognition tasks with user-defined vocabulary," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 3809–3812.

[19] T. Hain, V. Wan, L. Burget, M. Karafiat, J. Dines, J. Vepa, G. Garau, and M. Lincoln, "The ami system for the transcription of speech in meetings," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–357.

[20] J. Gemmeke and B. Cranen, "Noise robust digit recognition using sparse representations," *Proceedings of ISCA 2008 ITRW Speech Analysis and Processing for knowledge discovery*, 2008.

[21] A. Asaei, B. Picart, and H. Bourlard, "Analysis of phone posterior feature space exploiting class-specific sparsity and MLP-based similarity measure," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4886–4889.

[22] J. Segura, T. Ehrette, A. Potamianos, D. Fohr, I. Illina, P. Breton, V. Clot, R. Gemello, M. Matassoni, and P. Maragos, "The HIWIRE database, a noisy and non-native english speech corpus for cockpit communication," *Online. http://www. hiwire. org*, 2007.

[23] D. Imseng, R. Rasipuram, and M. Magimai-Doss, "Fast and flexible kullback-leibler divergence based acoustic modeling for non-native speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 348–353.