Low-Rank Representation of Nearest Neighbor Posterior Probabilities to Enhance DNN Based Acoustic Modeling

Gil Luyet[1,3]*, 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
[3]University of Fribourg, Switzerland
{gluyet,pranayd,aasaei,bourlard}@idiap.ch

Abstract

We hypothesize that optimal deep neural networks (DNN) class-conditional posterior probabilities live in a union of low-dimensional subspaces. In real test conditions, DNN posteriors encode uncertainties which can be regarded as a superposition of unstructured sparse noise over the optimal posteriors. We aim to investigate different ways to structure the DNN outputs by exploiting low-rank representation (LRR) techniques. Using a large number of training posterior vectors, the underlying low-dimensional subspace of a test posterior is identified through nearest neighbor analysis, and low-rank decomposition enables separation of the "optimal" posteriors from the spurious uncertainties at the DNN output. Experiments demonstrate that by processing subsets of posteriors which possess strong subspace similarity, low-rank representation enables enhancement of posterior probabilities, and leads to higher speech recognition accuracy based on the hybrid DNN-hidden Markov model (HMM) system.

Index Terms: Deep neural network (DNN), posterior probability, low-rank representation (LRR), k-nearest neighbor (kNN) search, automatic speech recognition (ASR).

1. Introduction

Speech is produced through activation of a few highly constrained articulatory mechanisms leading to various phonetic components and sub-phonetic attributes living in an union of low-dimensional subspaces [1, 2, 3]. Recent developments in speech processing research have recognized DNN as the best computational method to estimate sub-word class-conditional posterior probabilities from the input acoustic features. However, unlike the one-hot output class labels used for DNN training, the real posteriors obtained after the forward pass encode high variability along many dimensions. It has been demonstrated that the actual information learned by DNN is embedded in low-dimensional subspaces [4, 5]. The goal of this paper is to provide further investigations of these structure underlying DNN posteriors.

The fact that DNN estimates of posterior probabilities encode large uncertainties can be quantified through evaluation of the entropy. We group the posteriors in two categories depending on whether the maximum a posteriori (MAP) class label is the correct label associated to that posterior vector or not. The histogram of the distribution of entropy in both groups is depicted in Figure 1 (see Section 3.1 for details of the experimental setup). It is evident that increase in the posterior uncertainties as measured in terms of entropy leads to less accuracy in acoustic modeling. This problem is critical in DNN-hidden Markov model (HMM) automatic speech recognition (ASR) system as it leads to several competitive path likelihoods under Viterbi algorithm. In this paper, we study these uncertainties with regard to their structure exhibited in a space of large number of posterior vectors. More specifically, if the uncertainties are re-occurring patterns, they encode variability pertained to common linguistic features. Otherwise, they introduce unstructured sparse noise. Constructing a matrix by stacking similar posterior vectors, the re-occurring structures result in a low-rank matrix. Sparse error can thus be separated from the structured part through low-rank and sparse decomposition methods [5, 6].

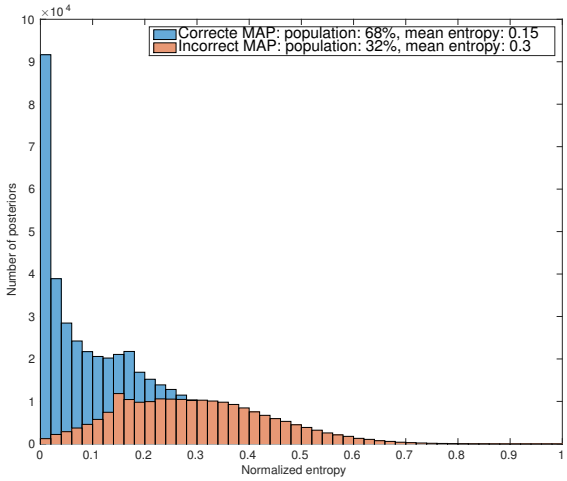

Figure 1: *Distribution of normalized entropies of DNN context-dependent phone posteriors for correctly/incorrectly classified sets based on maximum a posteriori classification.*

Previous work on exploiting the low-dimensional structures for posterior estimation includes approaches that either constrain the DNN architecture, or transform the DNN outputs. Approaches in the former category include low-rank decomposition of the neural network's weight matrices [7, 8, 9, 10, 11] to reduce DNN complexity. The goal is often smaller footprint for portable devices, such as ASR on smart phones, without compromising the performance. Another technique [12] applies manifold based regularization within DNN architecture to preserve low-dimensional relationships in speech features. Later approaches include enhancement of posterior probabilities by employing HMM topological constraints [13], or by hierarchical application of two neural networks to obtain more accurate posterior estimates [14, 15].

Recently, we considered sparse representation to character-

ize the low-dimensional space of DNN outputs through dictionary learning [16]. It is demonstrated that sparse reconstruction leads to projection on the space of training data dictionary which significantly reduces the rank of class-specific test posteriors. Further analysis verified the correlation between rank reduction and ASR accuracy [5]. This paper builds on our recent findings, and aims to exploit the underlying low-rank representation of DNN posteriors using low-rank representations. Previous work on speech enhancement using low-rank representations consider spectral features [17, 18]. To the best of authors' knowledge, low-rank representation for enhancement of posterior probabilities has not been investigated before.

In the rest of the paper, we explain low-rank representation in Section 2. Experimental analysis is presented in Section 3, and finally, the conclusions are drawn in Section 4.

2. Low-Rank and Sparse Decomposition

In this section, we discuss how to construct matrices corresponding to posterior subspaces on which low-rank representation (LRR) [6] algorithm can be applied.

2.1. Neighboring Posterior Matrices

DNNs compute a vector of class-conditional posterior probabilities $\boldsymbol{z}_t = [p(q_1|\boldsymbol{x}_t) \dots p(q_k|\boldsymbol{x}_t) \dots p(q_K|\boldsymbol{x}_t)]^\top$ for a context appended input acoustic feature vector $\boldsymbol{x_t}$ at time t where q_k denotes the k_{th} class and $.^\top$ stands is transpose operator. DNN training relies on one-hot posterior outputs where the probability of the class that the acoustic feature is associated is 1 and other classes are all 0. The hard labels are obtained using HMM forced alignment with the ground truth speech transcription. Once DNN is trained, test posterior probabilities are estimated through forward pass, and they may exhibit high uncertainty.

We group the posteriors according to their nearest neighbor based similarities, and stack a collection of n neighboring posterior vectors as a matrix $\boldsymbol{M} = [\boldsymbol{z}_{t_1} \dots \boldsymbol{z}_{t_n}]$. We propose that $\boldsymbol{M}$ can be decomposed as superposition of a low-rank part and a sparse error. The low-rank component encapsulates the enhanced posteriors, and the DNN uncertainties present in the form of sparse unstructured noise is separated.

Due to the skewed distribution of the posteriors, LRR based analysis presented in this work is performed on logarithms of posterior probabilities.

2.2. Low-Rank Representation Algorithm

We consider low-rank representation (LRR) algorithm [6] for decomposition of a noisy low-rank matrix $\boldsymbol{M}$ into data lying on union of multiple low-rank subspaces and sparse noise as expressed in

$$\min_{\boldsymbol{Z},\boldsymbol{E}} \|\boldsymbol{Z}\|_* + \lambda\|\boldsymbol{E}\|_1 \text{ , s.t. } \boldsymbol{M} = \boldsymbol{D}\boldsymbol{Z} + \boldsymbol{E} \qquad (1)$$

The rank is quantified by the (relaxed) nuclear norm function denoted by $\|.\|_*$ to obtain a convex cost function where the nuclear norm is the sum of matrix singular values. Dictionary $\boldsymbol{D}$ characterizes the (multiple) subspaces underlying the data. $\boldsymbol{Z}$ is a low-rank matrix such that the product $\boldsymbol{DZ}$ captures the true low-rank component of $\boldsymbol{M}$. λ is the regularization parameter. $\|.\|_1$ denotes the ℓ_1 norm which is defined as the sum of absolute values of the vector elements. It encourages sparsity of the error term. Another approach, namely robust principal component analysis (RPCA) [19]), used earlier for rank analysis in our previous work [5] will be compared with LRR in Section 3.3.1.

3. Experimental Analysis

We aim to study the appropriate choice of low-rank representation parameters in posterior probability space. In addition, we integrate neighborhood clustering and classification methods with LRR algorithm for unsupervised and supervised frameworks of posterior enhancement. We present an example use case for speech recognition.

3.1. Databases and DNN Setup

The experiments are conducted on digits subset of Numbers'95 [20]. Separate DNNs are trained to estimate context-independent (monophones) and context-dependent tied triphone state (senone) posterior probabilities with three hidden layers and 1024 nodes in each layer. Input features of the DNN consist of Mel-frequency cepstral coefficients (MFCC) concatenated with Δ and $\Delta\Delta$ features by making use of a context of 9 frames where features are computed at every 10ms. We obtain 27 context-independent and 557 senone posteriors.

Kaldi toolkit is used for DNN-HMM hybrid architecture for speech recognition [21] and LRSLibrary [22] is used for LRR.

3.2. Low-rank Representation Parameters

There are two parameters affecting the LRR performance, namely (1) choice of dictionary $\boldsymbol{D}$ and (2) regularization parameter λ.

3.2.1. Dictionary

LRR algorithm presented in (1) relies on dictionary $\boldsymbol{D}$ for characterizing the underlying subspaces [6]. Assuming the posterior subspaces corresponding to the individual classes are linear, we use the left eigenvectors of singular value decomposition (SVD) of class-specific posterior matrices $\boldsymbol{M}$ as the basis vectors to form class-specific dictionaries for low-rank representation. Number of eigenvectors in $\boldsymbol{D}$ is selected so as to preserve 95% variability after truncated SVD reconstruction.

From test data, we create matrix $\boldsymbol{M}$ by randomly choosing senone posterior vectors from multiple classes and apply LRR over it. The choice of dictionary is either concatenation of their 95%-variability preserving eigenvectors (as explained above) or the matrix $\boldsymbol{M}$ itself ($\boldsymbol{D} = \boldsymbol{M}$). In case of SVD, if class-specific dictionaries of correct classes are concatenated to form the dictionary $\boldsymbol{D}$, the dictionary is referred to as "correct-SVD" otherwise, all SVD class-specific dictionaries are concatenated without any knowledge of the correct underlying subspaces/classes. Size of $\boldsymbol{M}$ in all scenarios is 557×1000, but the number of posteriors from each class is arbitrary. Results are listed in Table 1, and represent MAP accuracy after applying LRR for different number of posterior classes present in $\boldsymbol{M}$. The enhanced posteriors $\hat{\boldsymbol{M}}$ are the low-rank component, $\hat{\boldsymbol{M}} = \boldsymbol{DZ}$.

We can see that if the underlying subspaces are known, "correct-SVD" is the best choice for LRR dictionary. This ob-

Table 1: *MAP accuracy after LRR using either data itself or SVD dictionaries for different number of class combination.*

#Classes	Data	Correct-SVD	All SVD
10	71.2%	89.9%	53.8%
50	61.5%	78.2%	51.1%
100	57.3%	68.9%	49.2%
250	52.1%	58.1%	50.8%

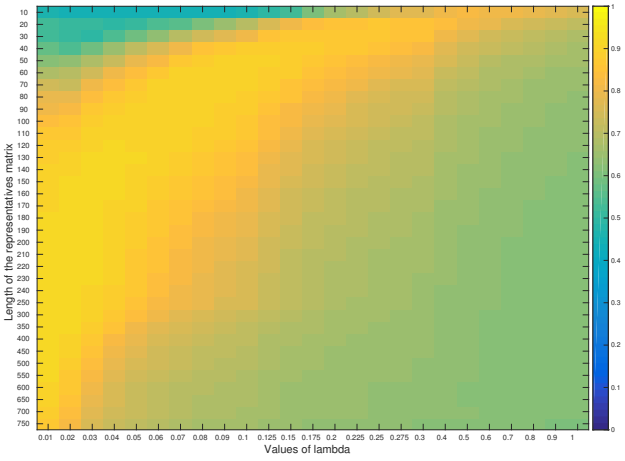

Figure 2: *MAP accuracy of senone posteriors after LRR using data itself as the dictionary for different choices of matrix size and regularization parameter λ. Yellow: 100%, blue: 0%.*

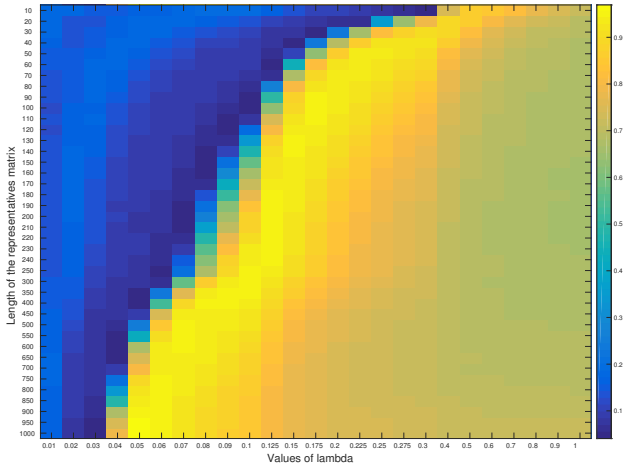

Figure 3: *MAP accuracy of senone posteriors after LRR using correct-SVD as dictionary for different choices of matrix size and regularization parameter λ. Yellow: 100%, blue: 0%.*

servation suggests that the underlying posterior subspaces of each individual class can be well characterized by using SVD eigenvectors. In real case scenario when classes are unknown, using data directly as the dictionary is a reasonable trade-off against the aggregation of all class-specific SVD based dictionaries. Indeed the choice of data as dictionary can exploit the self expressiveness property of the data lying on union of low-dimensional subspace [6], and the empirical observation presented in Table 1 demonstrate its efficiency.

Furthermore, we can see that LRR can effectively exploit the multi-subspace structure of the data through the use of dictionary $\boldsymbol{D}$ for low-rank representation. ASR performance degrades with increasing number of subspaces in $\boldsymbol{M}$. Although the number of posteriors per class in $\boldsymbol{M}$ is an arbitrary number, if the distribution of posteriors among different classes is too unbalanced, and a subspace is not well represented through enough number of posterior vectors, LRR can lead to subspace displacement where some of the posterior vectors from that subspace are structured as per some other dominant subspace. Example of a dominant subspace is the silence class due to which neighboring posteriors are forced to be the silence. To have fewer subspaces in $\boldsymbol{M}$ as well as prevent this issue of over-dominance of popular classes, we are motivated to apply clustering and classification techniques (Section 3.3) before application of LRR. This ensures having enough representatives for all posterior classes in $\boldsymbol{M}$. Nevertheless, LRR is theoretically applicable on large collection of multiple subspaces.

3.2.2. Regularization Parameter

The best value for the regularization parameter λ yields the highest MAP accuracy of enhanced posteriors. This value is proportional to the size of matrix $\boldsymbol{M}$. We conduct further investigations on the appropriate choice of λ. The evaluation principle is to compute the MAP accuracy after applying LRR algorithm for various lengths (between 10 and 1000) of matrix $\boldsymbol{M}$ constructed from the senone posteriors for different values of λs (between 0.01 to 1).

LRR is applied on class-specific posterior exemplars using either data or "correct-SVD" as the dictionary. The MAP accuracy is averaged over all classes. The results are depicted in Figures 2–3.

The dependency of the best choice of regularization parameter on the matrix size is evident, however the quality of enhanced posteriors is fairly stable in particular in the case of using data as the dictionary. In practice, the regularization parameter can be either learned for different sizes and used through a look-up table for LRR enhancement, or the development set can be used for tuning a reasonable λ.

One interesting observation is that the LRR performance is far more stable (better spread of yellow region in Figure 2 as compared to sharp yellow-blue transition in Figure 3) when data is used as the dictionary for low-rank representation compared to the "correct-SVD" dictionaries. Although higher MAP accuracy can be obtained in the later case, sensitivity to the appropriate selection of the regularization parameter is higher. Hence, in the subsequent tests on neighboring posterior enhancement on test data, the DNN posterior matrix i.e. data itself is chosen as the LRR dictionary.

3.3. Low-rank Representation of Neighboring Posteriors

To enhance posterior estimation of test data, neighboring posteriors are first identified through clustering or classification techniques. This procedure ensures that $\boldsymbol{M}$ has 1) data from fever subspaces and 2) enough number of posteriors are present for each underlying subspace for low-rank representation. However, due to clustering/classification inaccuracies, neighboring posteriors exhibit a multi-subspace structure in $\boldsymbol{M}$. To demonstrate this property, we conduct an experiment using single subspace low-rank representation (RPCA) and compare it to the LRR performance in the following Section 3.3.1.

3.3.1. Multi-Subspace Structure

For this experiment, we apply k-means clustering to group the neighboring posteriors based on the principle of low-rank matrix factorization [23]. Next, RPCA [19] and LRR are applied on each group to separate the underlying low-rank matrix of enhanced posteriors from the sparse errors. RPCA formulation assumes data lying on a single subspace whereas LRR generalizes this idea through the use of dictionary for characterizing the multi-subspace structures. Table 2 lists the results of MAP classification accuracies. Senone posteriors are used for this experiment.

Table 2: *MAP accuracy of low-rank representation posteriors using RPCA and LRR on clusters of similar posteriors obtained from k-means. MAP accuracy of the initial DNN posteriors is* 61.3%.

Low-rank representation	MAP accuracy
RPCA	68.5%
LRR	77.0%

We can see that LRR outperforms RPCA based single subspace low-rank representation. This experiment also confirms that application of LRR on neighboring posteriors can improve MAP classification. The next step is to show an example use case of this method in DNN-HMM speech recognition relying on posteriors for acoustic modeling.

3.3.2. Enhanced Posteriors for DNN-HMM ASR

To speed up the computation, monophone posteriors used for this experiment. Test posteriors are processed in small groups of neighboring posteriors. We consider (1) MAP classification, (2) k-means clustering and (3) k nearest neighbor (kNN) classification method. For k-means, the value of k is set to the maximum number of classes, i.e. dimension of posterior vectors or number of DNN outputs which is 27. The k parameter for kNN is tuned using development set, and it is equal to 1500. Cosine distance is used as the similarity measure in both k-means and kNN.

Once the neighboring test posteriors are identified through either of the methods stated above, subsets of 1000 posterior vectors are used for LRR decomposition. Corresponding vectors in the low-rank component, $\hat{\boldsymbol{M}} = \boldsymbol{DZ}$, are normalized to sum to one, and used as the enhanced posteriors for DNN-HMM ASR on digits subset of Numbers'95 database. Results are presented in Table 3.

Table 3: *ASR accuracy using LRR on neighboring posteriors for enhanced monophone posteriors estimation.*

Posterior estimation	WER%
DNN	5.9
MAP+LRR	4.9
k-Means+LRR	4.8
kNN+LRR	3.5

It can be seen that kNN outperforms all the other methods and achieves a relative WER improvement of nearly 40% with respect to the baseline where DNN posteriors are used for acoustic modeling. The reason is that kNN exploits the variability encoded in the training posteriors, and yields to accurate classification [24]. On the other hand, if no labeled data is available, unsupervised clustering based on k-means is beneficial to enhance the posteriors through low-rank representation. Figure 4 illustrates the consistent reduction of the rank and WER using kNN along with LRR. The approximate rank (aRank) quantifies the number of singular values to preserve 99% variability. Figure 5 shows a kNN followed by LRR enhanced posteriors for a sample utterance.

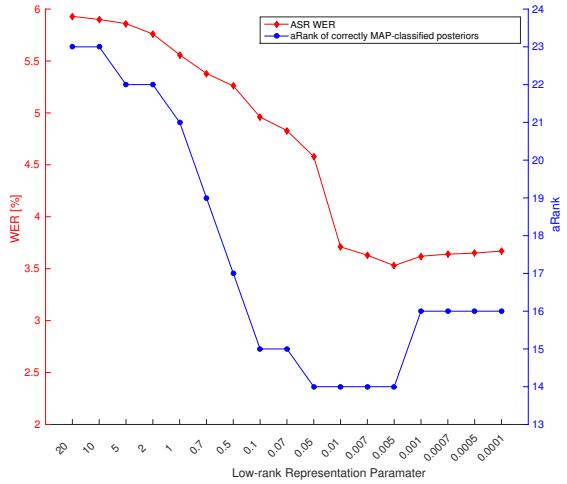

Figure 4: *Variation of the aRank and WER for different regularization parameter.*

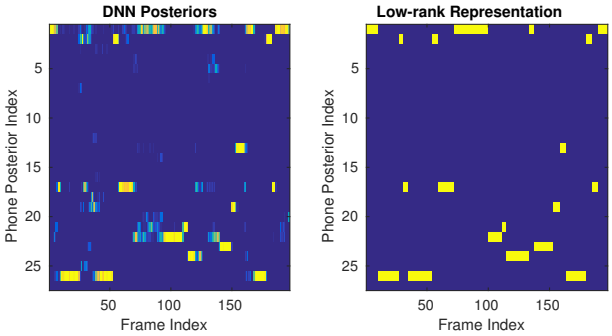

Figure 5: *DNN monophone posteriors vs. their low-rank representation for a sample utterance.*

4. Conclusions and Future Work

This paper investigates the intrinsic low-dimensional structure of high-dimensional space of DNN posterior probabilities. We have devised a simple framework of grouping the neighboring posteriors following by low-rank representation to enhance posterior estimation. This procedure alleviates some of the DNN uncertainties due to mismatch or unseen condition leading to unstructured sparse errors in posterior estimation.

kNN was found the best method to identify the neighboring posteriors through the use of training posterior variabilities. However, if no labeled data is available, unsupervised k-means clustering relying on the principle of low-rank matrix factorization [23] is an effective method as well to enhance posterior estimation and acoustic modeling. Considering the posterior space as a union of low-dimensional subspaces, alternative clustering techniques such as subspace clustering can exploit this property in splitting the space into neighboring vectors according to their subspace similarity [25].

An example use case was shown on DNN-HMM speech recognition. Beyond ASR, other applications that rely on estimation of DNN posteriors can also benefit from the proposed approach, such as spoken query detection [26], parametric speech coding [27] and linguistic parsing [28]. Thorough experiments on large speech corpora for a broad range of applications is planned for future research.

5. Acknowledgments

Research leading to these results was funded by SNSF project on "Parsimonious Hierarchical Automatic Speech Recognition (PHASER)" grant agreement number 200021-153507.

6. References

[1] L. Deng, "Switching dynamic system models for speech articulation and acoustics," in *Mathematical Foundations of Speech and Language Processing*. Springer New York, 2004, pp. 115–133.

[2] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[3] A. Asaei, M. Cernak, and H. Bourlard, "On compressibility of neural network phonological features for low bit rate speech coding," in *INTERSPEECH*, 2015.

[4] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[5] P. Dighe, G. Luyet, A. Asaei, and H. Bourlard, "Exploiting low-dimensional structures to enhance DNN based acoustic modeling in speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[6] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 171–184, 2013.

[7] V. Sindhwani, T. Sainath, and S. Kumar, "Structured transforms for small-footprint deep learning," in *Advances in Neural Information Processing Systems*, 2015, pp. 3070–3078.

[8] P. Nakkiran, R. Alvarez, R. Prabhavalkar, and C. Parada, "Compressing deep neural networks using a rank-constrained topology," in *INTERSPEECH*, 2015.

[9] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks." in *INTERSPEECH*, 2013, pp. 2345–2349.

[10] J. Xue, J. Li, and Y. Gong, "Restructuring of deep neural network acoustic models with singular value decomposition." in *INTERSPEECH*, 2013, pp. 2365–2369.

[11] T. N. Sainath, B. Kingsbury, V. Sindhwani, E. Arisoy, and B. Ramabhadran, "Low-rank matrix factorization for deep neural network training with high-dimensional output targets," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6655–6659.

[12] V. S. Tomar and R. C. Rose, "Manifold regularized deep neural networks," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[13] H. Ketabdar, "Enhancing posterior based speech recognition systems," Ph.D. dissertation, Ecole Polytechnique Fédérale de Lausanne, Lausanne , Switzerland, 0 2008.

[14] H. Ketabdar and H. Bourlard, "Enhanced phone posteriors for improving speech recognition systems," *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 18, no. 6, pp. 1094–1106, 2010.

[15] J. P. Pinto, G. S. V. S. Sivaram, M. Magimai.-Doss, H. Hermansky, and H. Bourlard, "Analysis of mlp based hierarchical phoneme posterior probability estimator," *IEEE Transcations on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 225–241, 0 2011.

[16] P. Dighe, A. Asaei, and H. Bourlard, "Sparse modeling of neural network posterior probabilities for exemplar-based speech recognition," *Speech Communication: Special Issue on Advances in Sparse Modeling and Low-rank Modeling for Speech Processing*, 2015.

[17] M. Gavrilescu, "Noise robust automatic speech recognition system by integrating robust principal component analysis (RPCA) and exemplar-based sparse representation," in *International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, June 2015, pp. S–29–S–34.

[18] Y.-H. Yang, "Low-rank representation of both singing voice and music accompaniment via learned dictionaries," in *Proceedings of the 14th International Society for Music Information Retrieval Conference*, November 4-8 2013, http://www.ppgia.pucpr.br/ismir2013/wp-content/uploads/2013/09/17_Paper.pdf.

[19] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

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

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[22] https://github.com/andrewssobral/lrslibrary.

[23] C. Bauckhage, "k-means clustering is matrix factorization," *arXiv preprint arXiv:1512.07548*, 2015.

[24] A. Asaei, H. Bourlard, and B. Picart, "Investigation of kNN classifier on posterior features towards application in automatic speech recognition," Idiap-RR-11, Tech. Rep., 2010.

[25] R. Vidal, "Subspace clustering," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 52–68, 2011.

[26] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2009, pp. 421–426.

[27] M. Cernak, A. Lazaridis, A. Asaei, and P. Garner, "Composition of Deep and Spiking Neural Networks for Very Low Bit Rate Speech Coding," *EPFL-ARTICLE-217528*, [online] http://infoscience.epfl.ch/record/217528/files/VLBRCoding.pdf.

[28] M. Cernak, A. Asaei, and H. Bourlard, "On structured sparsity of phonological posteriors for linguistic parsing," *arXiv preprint-arXiv:1601.05647*, 2016.