

EXPLOITING LOW-DIMENSIONAL STRUCTURES TO ENHANCE DNN BASED ACOUSTIC MODELING IN SPEECH RECOGNITION

Pranay Dighe^{*◦}

Gil Luyet^{†*}

Afsaneh Asaei^{*}

Hervé Bourlard^{*◦}

^{*}Idiap Research Institute, Martigny, Switzerland

[◦]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland

[†]University of Fribourg, Switzerland

{pranay.dighe, afsaneh.asaei, herve.bourlard}@idiap.ch, gil.luyet@unifr.ch

ABSTRACT

We propose to model the acoustic space of deep neural network (DNN) class-conditional posterior probabilities as a union of low-dimensional subspaces. To that end, the training posteriors are used for dictionary learning and sparse coding. Sparse representation of the test posteriors using this dictionary enables projection to the space of training data. Relying on the fact that the intrinsic dimensions of the posterior subspaces are indeed very small and the matrix of all posteriors belonging to a class has a very low rank, we demonstrate how low-dimensional structures enable further enhancement of the posteriors and rectify the spurious errors due to mismatch conditions. The enhanced acoustic modeling method leads to improvements in continuous speech recognition task using hybrid DNN-HMM (hidden Markov model) framework in both clean and noisy conditions, where upto 15.4% relative reduction in word error rate (WER) is achieved.

Index Terms— Sparse coding, Dictionary learning, Deep neural network, Union of Low Dimensional Subspaces, Acoustic modeling.

1. INTRODUCTION

A need for sparse representations for better acoustic modeling of speech has been advocated consistently for better characterization of the underlying low-dimensional and parsimonious structure of speech [1, 2, 3, 4]. Two major emerging trends, namely deep neural networks (DNN) and exemplar-based sparse modeling, are different approaches of exploiting sparsity in speech representations to achieve invariance, discrimination and noise separation [5, 4, 6].

On the other hand, speech utterances are formed as a union of words which in turn consist of phonetic components and sub-phonetic attributes. Each linguistic component is produced through activation of a few highly constrained articulatory mechanisms leading to generation of speech data in union of low-dimensional subspaces [7, 8, 9]. However, most existing speech classification and acoustic modeling methods do not explicitly take into account the multi-subspace structure of the data.

The present study focuses on exploiting the multi-subspace low-dimensional structure of speech learned from the training data to enhance DNN based acoustic modeling of unseen test data. Hence, this also has the potential to enable domain adaptation and handling mismatch in the framework of DNN based acoustic modeling.

1.1. Prior Works

Sparse representation has been proven powerful as features used for acoustic modeling. As argued in [2], if data is projected

into high-dimensional space, the underlying structures are disentangled. These structures form a union of low-dimensional subspaces which models the non-linear manifold where speech data resides. Prior work on sparse representation includes exemplar-based methods [3, 10] where sparse representation, learned using spectral features achieve promising performance in automatic speech recognition (ASR) specially due to their robustness in handling noise and corruption.

Recent advancement in DNN based acoustic modeling relies on estimation of highly sparse sub-word class-conditional posterior probabilities. While the conventional Gaussian mixture models (GMM) are statistically inefficient in modeling data lying on or near non-linear manifolds [11, 7, 8], DNNs achieve accurate sparse acoustic modeling through multiple layers of non-linear transformations [12]. The hidden layers of DNN successively learn underlying structures at different levels and express them as highly invariant and discriminative representations towards deeper layers. While enforcing sparsity constraints during DNN training is mostly employed for the purpose of regularization to prevent overfitting, various studies have shown that sparsity in DNN architectures directly contributes towards simpler networks and superior performance in ASR. Successful application of sparse activity [13] (very few neurons being active), sparse connectivity [14] (very few non-zero weights) as well as better performance of sparsity inducing techniques like dropout neural network training [15] confirm the belief that ‘*sparser*’ is better for acoustic modeling in ASR.

1.2. Motivation and Contributions

We point out two issues with respect to the state-of-the-art DNN based acoustic models which motivate further consideration of sparse modeling:

- Q1. Previous studies [16, 17] have found sparse activations in DNNs by showing how individual neurons in hidden layers learn being selectively active in different ways towards distinct phone patterns. Since this sparsification learned by hidden layers is not explicitly hand-crafted, we ask upto what extent the union of low-dimensional subspaces structure for speech is actually being exploited by DNNs ?
- Q2. Despite of being effective in seen conditions, DNNs are found highly sensitive to unseen variations in data [12]. The mismatch condition causes erroneous estimates of posterior probabilities which is exhibited as spurious noises in the output posterior probabilities. Can we correct these errors through a low-dimensional model to improve acoustic modeling in noisy conditions ?

In this paper, we address those issues by explicit modeling of the underlying structures in speech using prior knowledge that speech data lives in the union of low-dimensional subspaces. We implement this idea using the principled dictionary learning and sparse coding algorithms over DNN posterior probabilities to recover sparse representations where non-zero values correspond to class-specific subspaces. These subspace sparse representations are then used to *enhance* the original DNN posterior probabilities through dictionary based reconstruction. We build upon compressive sensing and subspace sparse recovery theory to provide theoretical support for validity of our approach. We also elaborate on our choice of features (DNN based posterior probabilities) and algorithms for dictionary learning [18] and structured sparse coding [19] which essentially distinguish our approach from previous exemplar based sparse representation methods [20, 3, 10]. We demonstrate improvements in performance achieved by the proposed enhanced acoustic modeling in hybrid DNN-HMM continuous ASR system using Numbers'95 database [21] and show increased robustness in noisy conditions.

In the rest of the paper, the proposed subspace sparse acoustic modeling method is elaborated in Section 2. The experimental analysis are carried out in Section 3. Section 4 provides the concluding remarks and directions for future work.

2. SUBSPACE SPARSE ACOUSTIC MODELING

In this section, we model the space of DNN class-conditional posterior probabilities as a union of low-dimensional subspaces. Relying on the subspace sparse representation, we show how the posteriors can be enhanced for more accurate class-specific representations.

2.1. Subspace Sparse Representation

Speech features reside on or near non-linear manifolds which can be best characterized by union of low-dimensional subspaces. The proposed approach relies on the fact that a data point in a union of subspaces can be more efficiently reconstructed using a sparse combination of data points from its own subspace than data points from other subspaces, thus resulting in a *subspace-sparse representation* [22].

To state it more precisely, let $\mathbf{S} = \{\mathcal{S}_\ell\}_{\ell=1}^L$ be a set of linear disjoint subspaces associated to L classes in \mathbb{R}^m such that the dimensions of individual subspaces $\{r_\ell\}_{\ell=1}^L$ are smaller than the dimension of the actual space, i.e. $\forall \ell, r_\ell < m$. Speech features z lie in the union $\cup_{\ell=1}^L \mathcal{S}_\ell$ of these low-dimensional subspaces. Let $\mathbf{D}_\ell \in \mathbb{R}^{m \times n_\ell}$ be the class-specific over-complete dictionary for subspace \mathcal{S}_ℓ where n_ℓ is the number of atoms in \mathbf{D}_ℓ and $n_\ell > r_\ell$. Each data point in \mathcal{S}_ℓ can then be represented as a sparse linear combination of the atoms from \mathbf{D}_ℓ .

Defining ℓ_1 -norm of a vector (denoted by $\|\cdot\|_1$) as the sum of the absolute values of its components, the *subspace sparse recovery* (SSR) property [22] for union of disjoint subspaces asserts that ℓ_1 -norm sparse representation of a data point over collection of all class-specific dictionaries $\{\mathbf{D}_\ell\}_{\ell=1}^L$ can lead to separation of the class-specific subspaces by selecting atoms only from the underlying class of the data point for its reconstruction. Thus, the obtained sparse representations have activations only for the atoms corresponding to the actual subspace \mathcal{S}_ℓ where z lives.

Considering a speech utterance as the union of words, phones or sub-phonetic components, the subspaces \mathcal{S}_ℓ can be modeled at different levels (time granularity) corresponding to any of these speech units. Consequently a dictionary \mathbf{D} can be constructed by learning basis sets \mathbf{D}_ℓ for individual classes. In the present study, we focus on context-dependent senones (c.f. Section 2.2.1) for their superior quality in DNN-HMM framework. Nevertheless there is no theoret-

ical/algorithmic impediment in applying it for larger units such as words.

The rigorous proof of SSR property (see Theorem 2 in [22]) requires certain conditions and assumptions on disjoint subspaces. Since we train DNN with binary senone target outputs, the intersection of senone subspaces is expected to be a rare event and suggests disjointness of subspaces. Although further theoretical analysis is beyond the scope of the present work, experiments conducted in Section 3 empirically confirm that SSR property indeed holds for subspace-sparse modeling of senones.

2.2. Class-Specific Dictionary Learning

There are two key considerations for dictionary learning in sparse subspace acoustic modeling. Namely, the choice of features and algorithmic developments.

2.2.1. Senone Posterior Probabilities as Speech Features

A posterior feature z is a vector consisting of class-conditional probabilities at the output layer of DNN. In contrast to spectral features, posterior features are proven highly effective for sparse modeling [23, 24]. They are inherently sparse and invariant to speaker/environmental conditions presented in the DNN training data. Although we choose to work with posterior probabilities at context-dependent senone levels (tied triphone states) [25], the theoretical underpinning of the proposed approach is applicable to any type of speech units.

2.2.2. Dictionary Learning and Sparse Coding Algorithms

Building on our previous work on dictionary learning for sparse modeling of posterior features [23], we use the online dictionary learning [18] algorithm for solving l_1 sparse coding problem expressed as

$$\arg \min_{\mathbf{D}, \mathbf{A}} \sum_{t=1}^T \|\mathbf{z}_t - \mathbf{D} \alpha_t\|_2^2 + \lambda \|\alpha_t\|_1, \text{ s.t. } \|d_j\|_2^2 \leq 1 \forall j \quad (1)$$

where $\mathbf{A} = [\alpha_1 \dots \alpha_T]$ and d_j denotes each atom of the dictionary.

Class-specific data of senone posterior features is obtained through GMM-HMM based forced alignment on training data, which is then used to learn individual over-complete basis set \mathbf{D}_ℓ for each senone subspace \mathcal{S}_ℓ using dictionary learning algorithm. These class-specific dictionaries are concatenated into a larger dictionary $\mathbf{D} = [\mathbf{D}_1 \dots \mathbf{D}_\ell \dots \mathbf{D}_L]$ for subspace-sparse acoustic modeling. Since any posterior feature obtained from DNN lies in a union of subspaces $\cup_{\ell=1}^L \mathcal{S}_\ell$, a test posterior feature z can be reconstructed using the atoms of dictionary \mathbf{D} . According to SSR property, only the atoms associated to the correct class (underlying subspace) of z will be used for sparse representation.

It may be noted that dictionary learning approach is fundamentally different from dictionary construction using a random subset [3, 10] of training features since we use all of the training data to compute an over-complete basis set for sparse representation which is far smaller (less than 3% in case of Numbers'95 database) than the actual collection size yet more effective in sparse representation [23].

2.3. Enhanced Acoustic Modeling

We use group sparsity based hierarchical Lasso algorithm [19] for sparse coding to enforce group sparsity in α based on the internal partitioning of dictionary \mathbf{D} into senone-specific sub-dictionaries \mathbf{D}_ℓ . The high dimensional group sparse representation α is computed for each DNN output posterior feature z by sparse recovery over \mathbf{D} . Projection of a test posterior feature z on training data space is given by computing $\mathbf{D}\alpha$.

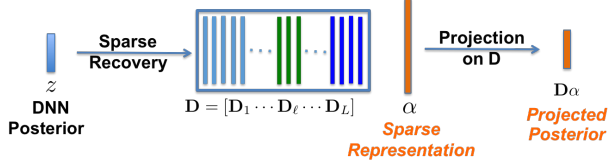


Fig. 1. DNN output senone posteriors z are projected to the space of training posteriors using $D\alpha$. Resulting projected posteriors are used for typical decoding in DNN-HMM framework.

Note that $D\alpha$ is an approximation of posterior feature z based on ℓ_1 -norm sparse reconstruction using atoms of D . Consequently, it has the same dimension as z and it is forced to lie in a probability simplex by normalization. Figure 1 summarizes this procedure.

3. EXPERIMENTAL ANALYSIS

In this section, we provide empirical analysis of the theoretical results established in Section 2. These experiments confirm that the information bearing components of DNN class-conditional probabilities indeed live in a very low-dimensional space. Exploiting this structure enables enhancement of DNN based acoustic models and removes the effect of high-dimensional noise leading to improvement in DNN-HMM speech recognition performance.

3.1. Database and Speech Features

We use Numbers’95 database for this study where only the utterances consisting of digits are considered (more details in [23]). The phoneset includes 27 phones and accordingly 557 context dependent tied states referred to as senones are learned by forced alignment of the training data using Kaldi speech recognition toolkit [26]. A DNN is trained using sequence discriminative training [27] with 3 hidden layers each having 1024 nodes. For every 10 ms speech frame, the DNN input is a vector of MFCC+ Δ + $\Delta\Delta$ features with a context of 9 frames ($39 \times 9 = 351$ dimension). The DNN output is a vector of posterior probabilities corresponding to 557 senone classes. We use DNN posteriors as features z for dictionary learning and sparse coding (??).

3.2. Low-rank Posterior Reconstruction

As explained in Sections 2.2–2.3, DNN posteriors are used to learn senone-specific dictionaries D_ℓ from the training data. Number of atoms n_ℓ in each senone dictionary D_ℓ is approximately 100. A value of $\lambda = 0.2$, optimized on development data, was used for sparse coding to get sparse representations α . Subsequently $D\alpha$ projected posterior probabilities are computed for the test data. Sparsity leads to selection of a few subspaces of the training data resulting in new test posteriors which (1) live in low-dimensions, (2) are projected onto the subspace of the training posteriors, and (3) separated from the subspaces of other senone classes. We investigate these properties below through further analysis.

To provide an insight into the dimension of the senone subspaces, we construct matrices of 1000 class-specific senone posteriors and compute the number of singular values required to preserve 95% variability of the data. Due to skewed distribution of the posteriors, we take their log prior to singular value decomposition. We refer to the number of required singular values as roughly the “Rank” of senone matrices. An ideal posterior feature should have its maximum component at the support indicating its associated class. Hence, we group the posteriors as “correct” if the max-

imum component corresponds to the correct class and “incorrect” if the maximum component corresponds to the incorrect class. Table 1 shows the average number of required singular values over all senones for DNN and projected posteriors. Another approach referred to as robust PCA based posteriors will be discussed in the subsequent section.

We can see that the “correct” posteriors live in a space which has far lower dimension than the space of “incorrect” posteriors. In other words, the information bearing components in “correct” senone posteriors are fewer resulting in matrices which have lower rank compared to “incorrect” posteriors. Given that the ranks are nevertheless very low (compared to the dimension of the senone posteriors which is 557), the “incorrect” posterior are exposed to a high-dimensional spurious noise. Therefore, to enhance the posterior probabilities,

the low-dimensional subspace has to be modeled/identified and the posterior has to be projected onto that space.

To further investigate the subspaces selected for sparse recovery, the values in sparse representation α for each class are summed to form α -sum vectors and the “Rank” of senone-specific α -sum matrices are computed. According to SSR property, it is expected that sparse recovery should select the subspaces from the underlying classes so the “Rank” of α -sum matrices has to be 1. In fact, we found that the empirical results averaged over the whole test set conformed to this theoretical insight indicating that

subspace sparse recovery leads to selection of the subspaces belonging to the underlying senone classes.

The class-specific dictionary learning for sparse coding enables us to model the non-linear manifold of the training data as a union of low-dimensional subspaces. A DNN posterior z from the test data may not lie on this manifold due to presence of high-dimensional noise embedded in its components. It is important to extract the low-dimensional structure in z while separating the effect of noise. Sparse coding does exactly this by finding the true underlying subspaces in sparse representation α and enables projecting z on the class-specific subspace of the training data manifold via $D\alpha$ reconstruction.

3.3. Low-rank and Sparse Decomposition

To further study the *true* underlying dimension of the senone-specific subspaces, we consider robust principle component analysis (RPCA) based decomposition of the senone posteriors [28]. The idea of RPCA is to decompose a data matrix M as

$$M = L + N \quad (2)$$

where matrix L has low-rank and matrix N is sparse (see Figure 2). Building upon the observations in Section 3.2, the low-rank component L corresponds to the enhanced posteriors while the high dimensional erroneous estimates are separated out in the sparse matrix N .

We collect posterior features for each senone from training data using *ground truth* based GMM-HMM forced alignment. RPCA decomposition is applied to data of each senone-class to reveal the *true* underlying dimension of the class-specific senone subspaces. The

	DNN	Projected	Robust PCA
Rank-Correct	36.6	11.9	7.6
Rank-Incorrect	45.5	21.7	11.7

Table 1. Comparison of “Rank” of DNN posterior matrix, projected posterior matrix and RPCA senone posterior matrix.

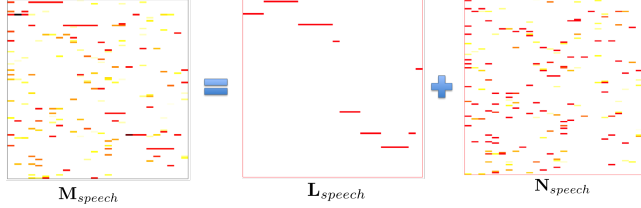


Fig. 2. Decomposing a DNN estimated senone posterior matrix $\mathbf{M}_{\text{speech}}$ into a low-rank matrix $\mathbf{L}_{\text{speech}}$ of enhanced posteriors and a sparse matrix $\mathbf{N}_{\text{speech}}$ of spurious noise.

rank of senone posteriors (i.e. rank of \mathbf{L}) obtained after RPCA decomposition for both “Correct” and “Incorrect” classes are listed in Table 1. We can see that the *true* dimension (7.6) of the class-specific subspaces of senone posteriors is indeed far lower than the DNN posteriors (36.6) and yet lower than the projected posteriors (11.9). Exploiting this multi low-rank structure of speech can lead to posterior enhancement via low-rank representation at utterance level [29].

The low-rank bottleneck layer based DNN is studied in [30] which shows that low-dimensional structuring of DNN architecture yields smaller footprint and faster training. In contrast, our proposed method suggests an added layer of sparse coding for structuring DNN outputs relying on the generic sparse and low-rank structures. Since, these generic structures are characterized from the training data, this approach enables us to handle mismatches in DNN train and test conditions.

3.4. Enhanced DNN-HMM Speech Recognition

Continuous speech recognition is performed using DNN posteriors as well as projected posteriors in the framework of conventional hybrid DNN-HMM. HMM topology learned during training of the hybrid DNN-HMM is used for decoding the word transcription in all cases. Hence, all parameters of different ASR systems shown here are the same and the only difference is in terms of senone posterior probabilities at each frame which results in different best paths being decoded by the Viterbi algorithm.

To demonstrate the increased robustness in projected posteriors as compared to the DNN posteriors, we also compared their performance in noisy conditions where artificial white Gaussian noise was added at signal level to the test utterances at signal-to-noise (SNR) ratios of 10 dB, 15 dB and 20 dB. DNN trained on clean speech is used for computing posteriors from noisy test spectral features so that the artificially added noise acts as an unseen variation in the data for DNN. Comparison of ASR performance is shown in Table 2 in terms of Word Error Rate (WER) percentage.

We can see that the projected posteriors outperform DNN posteriors in all cases suggesting that projection based on $\mathbf{D}\alpha$ provides enhanced acoustic models for DNN-HMM decoding. We note that in all experiments, a consistent decrease in insertion and substitution errors is observed when using projected posteriors in place of DNN posteriors. This implies fewer wrong hypotheses being made in case of projected posteriors at word level as compared to DNN posteriors. A similar insight comes by comparing the GMM-HMM based forced senone alignment (ground truth) with senone alignments achieved by best Viterbi paths in projected posterior and DNN posterior systems. Senone classification error of 24.1% in case of DNN posteriors is reduced to 19.8% in case of projected posteriors. Improvement in senone alignments and subsequent reduction in WER proves superior quality of projected posteriors over DNN posteriors and supports the hypothesis that projection moves the test features closer to the subspace of the correct classes.

SNR	Posteriors	WER (%)	Ins	Del	Subs
Clean	RPCA	0.4	36	18	4
Clean	DNN	2.6	111	96	152
	Projected	2.2	72	100	137
20db	DNN	4.0	160	121	293
	Projected	3.5	90	162	233
15db	DNN	6.8	205	249	498
	Projected	6.2	130	298	442
10db	DNN	14.0	199	950	801
	Projected	13.9	117	1064	763

Table 2. Comparison of ASR performance using DNN posteriors and projected posteriors in clean and noisy conditions on Numbers’95 database. RPCA posteriors indicate an ideal enhancement through low-dimensional posterior reconstruction. Breakdown of WER in terms of insertions (Ins), deletions (Del), and substitutions (Subs) has also been shown out of a total of 13967 words in all test utterances.

Finally, RPCA posteriors (matrix \mathbf{L} obtained from low-rank and sparse decomposition as explained in Section 3.3) which have ranks close to the true underlying dimensions of senone subspaces perform extremely well in ASR (c.f. Table 2). WER of 2.6% using DNN posteriors (“Rank” 36.6) reduces to a WER of 2.2% using projected posteriors (“Rank” 11.9) i.e. a relative improvement of 15.4%, and when RPCA posteriors (“Rank” 7.6) are used, it is reduced to a mere 0.4%. Since RPCA based low-rank reconstruction of posteriors has been done using ground truth senone alignment, ASR performance in this case is the best case scenario and demonstrates the scope of improvement possible even after DNN based acoustic modeling.

4. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we demonstrated explicit modeling of low-dimensional structures in speech using dictionary learning and sparse coding over the DNN class conditional probabilities. We showed that albeit their power in representation learning, DNN based acoustic modeling still has room for improvement in 1) exploiting the union of low-dimensional subspaces structure underlying speech data and 2) acoustic modeling in noisy conditions. Using dictionary learning and sparse coding, DNN posteriors were transformed to projected posteriors which were shown to be more suitable acoustic models. Sparse reconstruction moves the test posteriors closer to the correct underlying class of the data by exploiting the fact that the true information is embedded in a low-dimensional subspace thus separating out the high dimensional erroneous estimates. Improvements in ASR performance were shown for both clean and noisy conditions paving the way towards an effective robust ASR framework using DNN in unseen conditions. The importance of low-dimensional structures was further confirmed through RPCA analysis.

The proposed method can be improved through discriminative dictionary learning for better class-specific subspace modeling. Furthermore, we will study the low-rank clustering techniques to enhance posterior probabilities exploiting their low-dimensional multi-subspace structure. Moreover, we will consider further analysis on challenging databases and in particular the case of accented non-native speech recognition. Projection of accented speech posteriors on dictionaries trained with native language speech can result in transformation of accented phonetic space to native phonetic space and lead to improvements in accented speech recognition task.

5. 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.

6. REFERENCES

- [1] Jeff A Bilmes, "What HMMs can do," *IEICE TRANSACTIONS on Information and Systems*, vol. 89, no. 3, pp. 869–891, 2006.
- [2] Yoshua Bengio, "Learning deep architectures for AI," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.
- [3] Tara N Sainath, Bhuvana Ramabhadran, Michael Picheny, David Nahamoo, and Dimitri Kanevsky, "Exemplar-based sparse representation features: From TIMIT to LVCSR," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2598–2613, 2011.
- [4] G. Saon and Jen-Tzung Chien, "Bayesian sensing hidden markov models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 43–54, Jan 2012.
- [5] Li Deng and Xiao Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 1060–1089, 2013.
- [6] Jort Gemmeke and Bert Cranen, "Noise robust digit recognition using sparse representations," *Proceedings of ISCA 2008 ITRW Speech Analysis and Processing for knowledge discovery*, 2008.
- [7] Li Deng, "Switching dynamic system models for speech articulation and acoustics," in *Mathematical Foundations of Speech and Language Processing*, pp. 115–133. Springer New York, 2004.
- [8] Simon King, Joe Frankel, Karen Livescu, Erik McDermott, Korin Richmond, and Mirjam Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.
- [9] Leo J Lee, Paul Fieguth, and Li Deng, "A functional articulatory dynamic model for speech production," in *ICASSP. IEEE*, 2001, vol. 2, pp. 797–800.
- [10] Jort F Gemmeke, Tuomas Virtanen, and Antti Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.
- [11] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdelrahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
- [12] Dong Yu, Michael L Seltzer, Jinyu Li, Jui-Ting Huang, and Frank Seide, "Feature learning in deep neural networks-studies on speech recognition tasks," *arXiv preprint arXiv:1301.3605*, 2013.
- [13] Jian Kang, Cheng Lu, Meng Cai, Wei-Qiang Zhang, and Jia Liu, "Neuron sparseness versus connection sparseness in deep neural network for large vocabulary speech recognition," in *ICASSP*, April 2015, pp. 4954–4958.
- [14] Dong Yu, Frank Seide, Gang Li, and Li Deng, "Exploiting sparseness in deep neural networks for large vocabulary speech recognition," in *ICASSP*, 2012, pp. 4409–4412.
- [15] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [16] Tasha Nagamine, Micheal L. Seltzer, and Nima Mesgarani, "Exploring how deep neural networks form phonemic categories," *Interspeech*, 2015.
- [17] Abdel rahman Mohamed, Geoffrey Hinton, and Gerald Penn, "Understanding how deep belief networks perform acoustic modelling," in *ICASSP. IEEE*, 2012, pp. 4273–4276.
- [18] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research (JMLR)*, vol. 11, pp. 19–60, 2010.
- [19] Pablo Sprechmann, Ignacio Ramirez, Guillermo Sapiro, and Yonina C Eldar, "C-HiLasso: A collaborative hierarchical sparse modeling framework," *Signal Processing, IEEE Transactions on*, vol. 59, no. 9, pp. 4183–4198, 2011.
- [20] Tara N Sainath, Bhuvana Ramabhadran, David Nahamoo, Dimitri Kanevsky, and Abhinav Sethy, "Sparse representation features for speech recognition," in *Interspeech*, 2010, pp. 2254–2257.
- [21] R. A. Cole, M. Noel, T. Lander, and T. Durham, "New telephone speech corpora at CSLU," 1995.
- [22] Ehsan Elhamifar and Rene Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2765–2781, 2013.
- [23] Pranay Dighe, Afsaneh Asaei, and Hervé Bourlard, "Sparse modeling of neural network posterior probabilities for exemplar-based speech recognition," *Speech Communication*, 2015.
- [24] Afsaneh Asaei, Benjamin Picart, and Hervé Bourlard, "Analysis of phone posterior feature space exploiting class-specific sparsity and MLP-based similarity measure," in *ICASSP*, 2010.
- [25] Dong Yu and Li Deng, *Automatic Speech Recognition - A Deep Learning Approach*, Springer, 2015.
- [26] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, et al., "The kald speech recognition toolkit," 2011.
- [27] Karel Veselý, Arnab Ghoshal, Lukáš Burget, and Daniel Povey, "Sequence-discriminative training of deep neural networks," 2013.
- [28] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright, "Robust principal component analysis?," *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 11, 2011.
- [29] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, , no. 99, pp. 1–1, 2013.
- [30] Tara N Sainath, Brian Kingsbury, Vikas Sindhwani, Ebru Arisoy, and Bhuvana Ramabhadran, "Low-rank matrix factorization for deep neural network training with high-dimensional output targets," in *ICASSP. IEEE*, 2013, pp. 6655–6659.