Recent advances in deep learning for speech research at Microsoft

Conference Paper *in* Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on · October 2013

DOI: 10.1109/ICASSP.2013.6639345

12 authors, including:

li Deng
Zhejiang Normal University
401 PUBLICATIONS **29,330** CITATIONS
SEE PROFILE

Kaisheng Yao
Ant Financial
62 PUBLICATIONS **2,779** CITATIONS
SEE PROFILE

Dong Yu
Tohoku University
171 PUBLICATIONS **23,384** CITATIONS
SEE PROFILE

Michael Seltzer
Microsoft
121 PUBLICATIONS **4,869** CITATIONS
SEE PROFILE

RECENT ADVANCES IN DEEP LEARNING FOR SPEECH RESEARCH AT MICROSOFT

Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael L. Seltzer, Geoff Zweig, Xiaodong He, Jason Williams, Yifan Gong, and Alex Acero
Microsoft Corporation, One Microsoft Way, Redmond, WA 98052, USA

ABSTRACT

Deep learning is becoming a mainstream technology for speech recognition at industrial scale. In this paper, we provide an overview of the work by Microsoft speech researchers since 2009 in this area, focusing on more recent advances which shed light to the basic capabilities and limitations of the current deep learning technology. We organize this overview along the feature-domain and model-domain dimensions according to the conventional approach to analyzing speech systems. Selected experimental results, including speech recognition and related applications such as spoken dialogue and language modeling, are presented to demonstrate and analyze the strengths and weaknesses of the techniques described in the paper. Potential improvement of these techniques and future research directions are discussed.

Index Terms— deep learning, neural network, multilingual, speech recognition, spectral features, convolution, dialogue

1. INTRODUCTION

For many years, speech recognition technology has been dominated by a "shallow" architecture using many Gaussians in the mixtures associated with HMM states to represent acoustic variability in the speech signal. Since 2009, in collaboration with researchers at University of Toronto and other organizations, we at Microsoft have developed deep learning technology that has successfully replaced Gaussian mixtures for speech recognition and feature coding at an increasingly larger scale (e.g., [24][19][53][39][7][8][44][54][13][56][30][48]). In this paper, we provide an overview of this body of work, with emphasis on more recent experiments which shed light onto the understanding of the basic capabilities and limitations of the current deep learning technology for speech recognition and related applications.

The organization of this paper is as follows. In Sections 2-5, we focus on several aspects of deep learning in the feature-domain with the theme of how deep models can enable the effective use of primitive, information-rich spectral features. The remaining sections are focused on the model-domain implementation of deep learning and on two application areas beyond acoustic modeling for speech recognition. Representative experimental results are shown to facilitate the analysis on the strengths and weaknesses of the techniques we have developed and illustrated in this paper.

2. BACK TO PRIMITIVE SPECTRAL FEATURES

Deep learning, sometimes referred as representation learning or (unsupervised) feature learning [3] sets an important goal of automatic discovery of powerful features from raw input data independent of application domains. For speech feature learning and for speech recognition, this goal is condensed to the use of primitive spectral [26]or possibly waveform [46] features.

Over the past 30 years or so, largely "hand-crafted" transformations of speech spectrogram have led to significant accuracy improvements in the Gaussian mixture model (GMM) based HMM systems, despite the known loss of information from the raw speech data. The most successful transformation is the non-adaptive cosine transform, which gave rise to Mel-frequency cepstral coefficients (MFCC) and the related PLP features. The cosine transform approximately de-correlates feature components, which is important for the use of diagonal GMMs. However, when GMMs are replaced by deep learning models such as deep neural nets (DNN), deep belief nets (DBN), or deep autoencoders (DAE), such de-correlation become irrelevant due to the very strength of the deep learning methods in modeling data correlation. Our early work [19] demonstrated such strength and in particular the benefit of spectrograms over MFCCs in effective coding of bottleneck speech features using DAE in an unsupervised manner. Subsequent work carried out at Stanford [40] generalized the use of DAE from single modality of speech to bimodal speech and visual features. This success partly inspired the mixed-band and multilingual DNNs to be described in Section 4.

More recent experiments at Microsoft demonstrate noticeably lower speech recognition errors using large-scale DNNs when moving from MFCCs back to more primitive filter-bank features (i.e., a Mel-scaled spectrogram with no cosine transforms). Table 1 is a summary of these experiments, where the DNN-HMM speech recognizer in a voice search task makes use of 72 hours of audio training data with over 26 million frames. The relative error rate reduction going from MFCC to filter-banks shown in Table 1 is comparable to that which we also observed for the TIMIT phone recognition task. Note the use of raw FFT features has not resulted in even lower errors, suggesting that current DNN training cannot automatically learn Mel-like filter weights. The same difficulty is also found for learning or improving delta-like features as shown in the bottom two rows of Table 1.

Table 1*: Comparing MFCC with filter-bank features*

Systems (Features: static+Δ+ΔΔ)	Word error rate
Best GMM-HMM (MFCCs; fMPE+BMMI)	34.7%
DNN (MFCCs)	31.6%
DNN (256 log FFT bins)	32.3%
DNN (29 log filter-banks)	30.1%
DNN (40 log filter-banks)	**29.9%**
-Static 40-log-filter-banks only (11-frames)	31.1%
-Static 40-log-filter-banks only (19-frames)	30.5%

One advantage of MFCC is its automatic normalization (after removing C_0) of power variation arising from different microphone gains associated with different data sets. When spectral (or time-domain) features are used, each feature component is subject to such variation [46]. However, when the data sets are obtained from the same source in training the DNN system (as is the case for the

task of Table 1), similar error rates are obtained with and without applying a sentence-level spectral feature normalization procedure (30.0% vs. 30.1%). But when the data sets are from diverse sources, we observed that the application of feature normalization procedures has reduced the error rate from 24.5% to 23.7%. Effective online normalization of features, however, is a cumbersome process in practical speech recognition scenarios. This raises a need for improving the current DNN method in handling amplitude or power variation across all spectral feature components. One potential solution is the use of rectified linear instead of sigmoid hidden units.

3. CONVOLUTION ON SPECTRAL FEATURES

Compared with MFCCs, "raw" spectral features not only retain more information (including possibly redundant or irrelevant one), but also enable the use of convolution and pooling operations to represent and handle some typical speech invariance and variability --- e.g., vocal tract length differences across speakers, distinct speaking styles causing formant undershoot or overshoot, etc. --- expressed explicitly in the frequency domain.

As a baseline, we explored a primitive convolutional neural net (CNN) [1] where the pooling configuration is fixed. The larger pooling size enforces a greater degree of invariance to frequency shifts while also running a greater risk of confusion among different speech sounds with similar formant frequencies. Based on detailed error analysis, we have developed a strategy for trading between invariance and discrimination. This strategy reduces the TIMIT core test set's phone recognition error rate to 19.7% from 20.4%. After regularizing the CNN using a variant of the "dropout" technique [25], the error rate drops further to 18.7%. Note all the above error analysis and the interpretation of the convolution and pooling operations in the CNN have been made possible after the change from the use of MFCC to spectral features. Details of this new deep CNN and error analysis are provided in [17].

4. LEARNING MULTI-TASK FEATURES

From its very original motivation, deep learning or representation learning algorithms are designed to make them especially powerful in multi-task scenarios that would benefit from universal or shared feature representations in the intermediate layer(s) of the deep architecture; e.g., [40]. In this section, we present and analyze two sets of speech recognition experiments to demonstrate that the DNN is a universal learner that effectively handles heterogeneous data from different acoustic sources and languages.

4.1 Mixed-Band DNN

In this set of experiments, we design the filter-banks in such a way that the narrowband (8-kHz) data are treated as wideband (16-kHz) data with half of the feature dimensions missing. We use the same filter-bank design that is described and used in [20]. For the 8-kHz data, the upper filter banks are padded with 0's in the multitask architecture shown in Figure 1a. The common layers extract features that correlate with both the narrowband and wideband data.

Experiments have been carried out on a large-scale speech recognition task, with the results summarized in Table 2; see details in [34]. The use of additional narrowband data, which is very different but highly correlated with wideband data (of primary

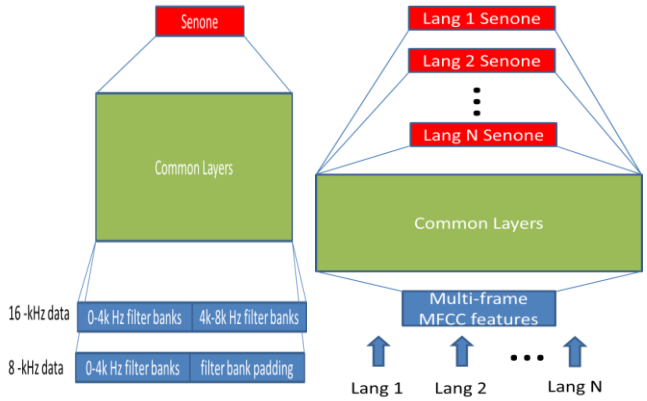

***Figure 1**: a) left: DNN training/testing with mixed-band acoustic data with16-kHz and 8-kHz sampling rates; b) right: Illustrative architecture for multilingual DNN*

business interest to us), has reduced the error rate from 30.0% to 28.3%, amounting to 5.7% relative error reduction with the number of test words being 26,757. In our group's previous work, we made several attempts to exploit narrowband data (plentiful from earlier telephone-based applications) to benefit training the wideband speech models in the GMM-HMM framework without success. Switching to the DNN created a quick success.

***Table 2**: DNN performance on wideband and narrowband test sets (a multitask-learning setup) using mixed-bandwidth training data.*

Training Data	Test WER (Wideband)	Test WER (Narrowband)
Wideband only	30.0%	71.2%
Narrowband only	-	29.0%
Wideband+Narrowband	**28.3%**	29.3%

4.2 Multi-Lingual DNN

Multilingual speech recognition is of high practical value. It has a long history of research, making use of many sources of prior knowledge including phonology and speech production [43][16] and of model adaptation [35] or neural net initialization [49][51]. However, given the very nature of multitask machine learning (as reviewed in [11]), multilingual speech recognition is best suited for the DNN where the intermediate hidden layer(s) is expected to provide universal representations across multiple languages' acoustic data that are highly correlated.

We developed and experimentally evaluated the multilingual DNN architecture shown in Figure 1b. It has the input and hidden layers shared by all languages, but separate output layers are made specific to each language. In the training phase, the multilingual DNN is exposed to the training acoustic data from all languages. Given a training data point, regardless of the language, all shared DNN parameters are updated, but we learn only the top-layer weights corresponding to the correct language. After the training, the entire DNN except the top layer can be considered as the feature extractor shared across all languages.

Using this language-universal feature extractor, we readily construct a powerful monolingual DNN for any target language as follows. First, the top layer and its connection to the hidden layer trained previously are discarded. Then, a new top layer corresponding to the target language's senone set is built and the

weights to the language-universal hidden layer are trained using the limited training data from the target language.

The multilingual DNN has been evaluated on a Microsoft-internal speech recognition task. In the training set, we used French (FRA), German (DEU), Spanish (ESP), and Italian (ITA) as the resource-rich languages, with 138 hours, 195 hours, 63 hours, and 93 hours of speech data, respectively. In Table 3, the final WERs are compared on a FRA test set for two monolingual FRA DNNs: one is trained using only FRA data and the other extracted from the multilingual (FRA+DEU+ESP+ITA) DNN. The latter DNN gives 3.5% fewer errors than the former DNN.

***Table 3**: Comparing DNN word error rates on a resource-rich task (FRA training data=138 hrs) w. & wo other languages*

Speech Recognizers	WER on FRA
DNN trained with only FRA data	28.1%
DNN trained with FRA + DEU + ESP+ ITA	27.1%

For cross-lingual evaluation of the multilingual DNN, we used 9 hours of training data from U.S. English (ENU) as the resource-limited target language, with typical results presented in Table 4. Retraining only the top layer gives lower errors than retraining all layers due to the data sparsity in ENU. Adding three more source languages in training further reduces recognition errors. We see that the multilingual DNN provides an effective structure for transferring information learnt from multiple languages to the DNN for a resource-limited target language due to phonetic information sharing.

***Table 4**: Comparing DNN word error rates on a resource-limited task (ENU training data=9 hrs) w. & wo other languages*.

Speech Recognizers	WER on ENU
DNN trained with only ENU data	30.9%
+FRA, retrain all layers with ENU	30.6%
or +FRA, retrain the top layer with ENU	27.3%
or +FRA+ DEU+ ESP+ITA, retrain top layer	25.3%

5. NOISE-ROBUST INTERNAL FEATURES

A main benefit of the DNN as the acoustic model is its ability to discover representations that are stable with respect to variations in the training data. One significant source of such variations is environmental noise. In order to evaluate the noise-robustness of DNN-based acoustic models, we performed a series of experiments using Aurora 4, a medium-vocabulary corpus based on WSJ0.

The results in Table 5 compare the performance of four systems on the Aurora 4 task. The first is the baseline GMM-HMM system with no compensation. The second system [21] represents the state of the art in noise robustness for HMM-based speech recognition, combining MPE discriminative training and noise-adaptive training (e.g., [31][12]) to compensate for noise and channel mismatch. The third system uses a log-linear model with features derived from HMM likelihoods [41]. The final system is a DNN-HMM with 7 hidden layers and 2000 hidden units per layer. This system uses no explicit noise compensation algorithm. The DNN-HMM significantly outperforms the other systems. In addition, the DNN-HMM result was obtained in a single pass, while the previous two systems require multiple passes for adaptation. These results clearly demonstrate the inherent robustness of the hidden-layer features in the DNN to unwanted variability from noise and channel mismatch.

***Table 5:** Word error rate (%) for all four test sets (A, B, C, and D) of the Aurora 4 task. DNN outperforms GMM systems*

	A	B	C	D	AVG
GMM-HMM (Baseline)	12.5	18.3	20.5	31.9	23.9
GMM (MPE+VAT)	7.2	12.8	11.5	19.7	15.3
GMM + Deriv. Kernels	7.4	12.6	10.7	19.0	14.8
DNN (7x2000)	5.6	8.8	8.9	20.0	13.4

6. DNN ADAPTATION

Adapting DNN acoustic models is more difficult than adapting GMMs. We have recently investigated the affine transformation and several of its variants for adaptation of the top hidden layer [52]. The feature-space discriminative linear regression (fDLR) method [2] with an affine transformation on the input layer is also evaluated. We have implemented stochastic gradient descent (SGD) and batch update methods for the above adaptation techniques. Both implementations lead to significant reduction of word error rates on top of a baseline DNN system. Shown in Table 6, on a large vocabulary speech recognition task, a SGD implementation of the fDLR and the top softmax layer adaptation is shown to reduce word errors by 17% and 14%, respectively, compared to the baseline DNN performance. Using a batch update for adapting the softmax layer reduces recognition errors by 10%.

We have recently developed a KL-distance based regularization method [33] to improve robustness of the DNN system under the condition of a small number of adaptation utterances [55]. As shown in Table 7, on a large vocabulary system, the method shows 6% to 20% relative error reductions using 5 to 200 supervised adaptation utterances compared with the baseline DNN. (For the unsupervised case, the improvement is somewhat less.)

***Table 6:** DNN adaptation using SGD and batch implementations*

Speech Recognition Systems	WER	WERR (%)
GMM-HMM	43.6%	
DNN	34.1%	-
DNN + AdaptSoftMax (SGD)	29.4%	13.9
DNN + fDLR (SGD)	28.5%	16.8
DNN + AdaptSoftMax (batch)	30.9%	9.3

***Table 7:** Word error rates for varying number (200, 50, and 5) of adaptation utterances. DNN baseline error rate 34.1%.*

Adaptation Methods	200	50	5
fDLR	28.5%	30.4%	36.5%
KL-regularization	27.5%	28.4%	32.1%

7. RECURRENT NETWORKS FOR LANGAUGE MODELING

In the approach described here, we explore the capability of a neural net to combine and exploit information of diverse types, and apply it to the task of language modeling. This approach has been proposed in dialog systems with a feed-forward net [57] and more recently for recurrent nets [47][37]. In all these approaches the basic idea is to augment the input to the network with information above-and-beyond the immediate word history. In [37], we propose the architecture of Fig. 2, which adds a side-channel of information to the basic recurrent network of [38]. By providing a side-channel consisting of a slowly changing Latent Semantic Analysis (LSA) representation of the preceding text in the Penn Treebank data, we improved perplexity over a Kneser-Ney 5-gram

model with a cache from 126 to 110 – to our knowledge the best single-model perplexity reported for this dataset. Interestingly, these gains hold up even after interpolating a standard recurrent net with a cache model, indicating that the context-dependent recurrent net is indeed exploiting the topic information of the LSA vector, and not just implementing a cache. In subsequent experiments, we have found that this architecture is able to use other sources of information as well; for example, in initial experiments with a voice-search application, conditioning on a latent-semantic analysis representation of a user's history has reduced the perplexity of a language model from 135 to 122.

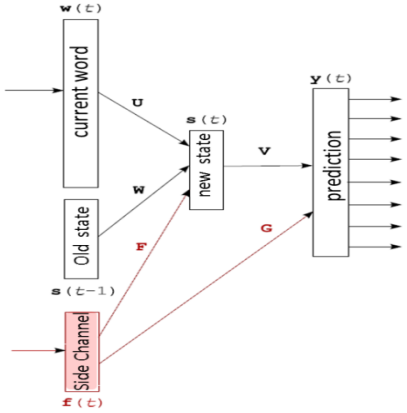

Figure 2: *Recurrent neural network with side-channel information*

8. STATE TRACKING FOR SPOKEN DIALOGUE

We have also begun applying deep learning to spoken dialogue systems, specifically the component of *dialog state tracking*. The objective of dialog state tracking is to assign a probability of correctness to user goals at time *t* given the history of the dialogue from $0 \ldots t-1$, and historical information about the user. For example, in a bus timetable application, a user goal is the user's location, intended destination, and desired arrival or departure date and time, the dialogue history includes everything the system has asked so far and all of the spoken language understanding results observed, and the historical information about the user includes which locations they have asked for in the past. In practice, probabilities are assigned to the subset of most promising goals, and also to a special class that indicates that none of the goals is correct. Deep networks are attractive here because there are many interactions among features that predict the correctness of a user goal.

One way of framing the dialogue state tracking problem is to construct a binary classifier that scores candidate user goals as correct or incorrect in isolation; normalizing the scores yields a distribution over all goals. Following this approach, we recently explored the application of the deep stacking network or DSN [13][14] to this task. Our initial experiments show its performance is on par with state-of-the-art classifiers. Table 8 summarizes the preliminary results using a slightly tuned DSN on a corpus of dialogs from the spoken dialog challenge 2010 [5] and 2011-2012, where the percent accuracy indicates how often the correct user goal was identified. Results are similar to our strongest baseline -- a tuned, highly optimized maximum entropy classifier. In future work we plan to conduct an evaluation in an end-to-end dialog system, and to tackle the technical challenge of instance-dependent sizes of the classes and feature dimensions by incorporating structure into the deep learning architectures.

Table 8: *Goal tracking accuracy for five slots using a baseline maximum entropy model and a DSN. Experiments were done on a fixed corpus of dialogs with real users.*

	Baseline	DSN
Bus route	58.0%	**58.1%**
Origin location	56.4%	**57.1%**
Destination location	**66.5%**	65.4%
Date	83.9%	**84.6%**
Time	**63.1%**	62.5%

The input to the dialog state tracking component of the full dialogue system comes from the speech understanding component. We have also explored the use of various versions of deep learning models for this task, with highly promising results reported in [15][50].

9. SUMMARY AND DISCUSSION

This paper provides selected samples of our recent experiments on applying deep learning methods to advancing speech technology and related applications, including feature extraction, acoustic modeling, language modeling, speech understanding, and dialogue state estimation.

A major theme we adopt in writing this overview goes to the very core of deep learning --- automatic learning of representations in place of hand-tuned feature engineering. To this end, we presented experimental evidence that spectrogram features of speech are superior to MFCC with DNN, in contrast to the earlier long-standing practice with GMM-HMMs. New improvements on DNN architectures and learning are needed to push the features even further back to the raw level of acoustic measurements.

Our and other's work over past few years has demonstrated that deep learning is a powerful technology; e.g. on the Switchboard ASR task the word error rate has reduced sharply from 23% in the GMM-HMM system as prior art to as low as 13% currently [32][48]. Our future work on deep learning research is directed towards three largely orthogonal directions: 1) more effective deep architectures and learning algorithms, including enhancing recently developed techniques (e.g., [4][9][17][42]); 2) scaling deep model training with increasingly larger data sets [6][10][48][29]; and 3) extending the applications of deep learning models to other areas of speech and language processing, and beyond (e.g., preliminary and promising applications to speech synthesis [36], end-to-end speech understanding and translation [22][58], recognition confidence measure [28], and information retrieval [18]).

ACKNOWLEDGEMENTS: Some experiments discussed in this paper were carried out with contributions from our summer interns: A. Metallinou, O. Abdelhamid, and T. Mikolov.

REFERENCES

[1] O. Abdel-Hamid and A. Mohamed, H. Jiang, and G. Penn. "Applying convolutional neural network concepts to hybrid NN-HMM model for speech recognition," ICASSP, 2012.

[2] V. Abrash, H. Franco, A. Sankar, and M. Cohen, "Connectionist speaker normalization and adaptation," Eurospeech, 1995.

[3] Y. Bengio. "Representation learning: A review and new perspectives," IEEE Trans. PAMI, special issue Learning Deep Architectures, 2013.

[4] Y. Bengio, N. Boulanger, and R. Pascanu. "Advances in optimizing recurrent networks," ICASSP, 2013.

[5] A. Black, et al, "Spoken dialog challenge 2010: Comparison of live and control test results," SIGdial Workshop, 2011.

[6] X. Chen, A. Eversole, G. Li, D. Yu, and F. Seide, "Pipelined back-propagation for context-dependent deep neural networks," Interspeech, 2012.

[7] G. Dahl, D. Yu, L. Deng, and A. Acero, “Large vocabulary continuous speech recognition with context-dependent DBN-HMMs,” ICASSP, 2011.

[8] G. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large vocabulary speech recognition,” IEEE Trans. Audio, Speech, Lang. Proc., vol. 20, pp. 30–42, 2012.

[9] G. Dahl, T. Sainath, and G. Hinton. “Improving DNNs for LVCSR using RELU and dropout,” ICASSP, 2013.

[10] J. Dean et al., “Large scale distributed deep networks,” NIPS, 2012.

[11] L. Deng and X. Li. “Machine learning paradigms for speech recognition: An overview,” IEEE Trans. Audio, Speech & Lang. Proc., Vol. 21, No. 5, May 2013.

[12] L. Deng, A. Acero, M. Plumpe, and X. Huang. “Large-vocabulary speech recognition under adverse acoustic environments,” Proc. ICSLP, 2000.

[13] L. Deng, D. Yu, and J. Platt. “Scalable stacking and learning for building deep architectures,” ICASSP, 2012.

[14] L. Deng and D. Yu. “Deep convex net: A scalable architecture for speech pattern classification,” Interspeech, 2011.

[15] L. Deng, G. Tur, X. He, and D. Hakkani-Tur, “Use of kernel deep convex networks and end-to-end learning for spoken language understanding,” IEEE SLT, 2012.

[16] L. Deng. “Integrated-multilingual speech recognition using universal phonological features in a functional speech production model,” ICASSP, 1997.

[17] L. Deng, O. Abdel-Hamid and D. Yu. “A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion,” ICASSP, 2013.

[18] L. Deng, X. He, and J. Gao. “Deep stacking networks for information retrieval,” ICASSP, 2013.

[19] L. Deng, M. Seltzer, D. Yu, A. Acero, A. Mohamed, and G. Hinton, “Binary coding of speech spectrograms using a deep auto-encoder,” Interspeech, 2010.

[20] X. Fan, M. Seltzer, J. Droppo, H. Malvar, and A. Acero, “Joint encoding of the waveform and speech recognition features using a transform codec,” ICASSP, 2011.

[21] F. Flego and M. Gales, “Factor analysis based VTS and JUD noise estimation and compensation,” Cambridge University, Tech. Rep. CUED/FINFENG/TR653, 2011.

[22] X. He, L. Deng, D. Hakkani-Tur, G. Tur, “Multi-style adaptive training for robust cross-lingual spoken language understanding,” ICASSP, 2013.

[23] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean. “Multilingual acoustic models using distributed deep neural networks,” ICASSP, 2013.

[24] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition,” IEEE Sig. Proc. Mag., vol. 29, 2012.

[25] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, & R. Salakhutdinov. “Improving neural networks by preventing co-adaptation of feature detectors,” arXiv: 1207.0580v1, 2012.

[26] H. Hermansky. “Speech recognition from spectral dynamics,” Sadhana (Indian Academy of Sciences), 2011, pp. 729-744.

[27] J. -T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong. “Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers” ICASSP, 2013.

[28] P. Huang, K. Kumar, C. Liu, Y. Gong, and L. Deng. “Predicting speech recognition confidence using deep learning with word identity and score features,” ICASSP, 2013.

[29] P. Huang, L. Deng, M. Hasegawa-Johnson, X. He. “Random features for kernel deep convex networks,” ICASSP, 2013.

[30] B. Hutchinson, L. Deng, and D. Yu, “Tensor deep stacking networks,” IEEE Trans. PAMI, 2013, to appear.

[31] O. Kalinli, M. L. Seltzer, J. Droppo, A. Acero, “Noise adaptive training for robust automatic speech recognition”, IEEE Trans. Audio, Speech & Lang. Proc., vol. 18, no. 8, pp. 1889-1901, 2010.

[32] B. Kingsbury, T. Sainath, and H. Soltau, “Scalable minimum Bayes risk training of DNN acoustic models using distributed Hessian-free optimization,” Interspeech, 2012.

[33] X. Li and J. Bilmes, “Regularized adaptation of discriminative classifiers,” ICASSP, 2006.

[34] J. Li, D. Yu, J. -T. Huang, and Y. Gong, “Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM,” IEEE SLT, 2012.

[35] H. Lin, L. Deng, D. Yu, Y. Gong, A. Acero, and C.-H. Lee, “A study on multilingual acoustic modeling for large vocabulary ASR,” ICASSP, 2009.

[36] Z. Ling, L. Deng, and D. Yu. “Modeling spectral envelopes using restricted Boltzmann machines for statistical Parametric speech synthesis,” ICASSP, 2013.

[37] T. Mikolov and G. Zweig, “Context Dependent Recurrent Neural Network Language Model,” Proc. SLT, 2012.

[38] T. Mikolov, M. Karafiat, J. Cernocky, and S.Khudanpur, “Recurrent neural network based language model,” Interspeech, 2010.

[39] A. Mohamed, D.Yu, L. Deng, “Investigation of full-sequence training of deep belief networks for speech recognition,” Interspeech,2010.

[40] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, “Multimodal deep learning,” ICML, 2011.

[41] A. Ragni and M. Gales, “Derivative kernels for noise robust ASR”, Proc. ASRU, 2011.

[42] T. Sainath, A. Mohamed, B. Kingsbury, B. Ramabhadran, “Deep convolutional neural networks for LVCSR,” ICASSP, 2013.

[43] T. Schultz and A. Waibel, “Multilingual and cross-lingual speech recognition,” DARPA Workshop on Broadcast News Transcription and Understanding, 1998.

[44] F. Seide, G. Li, and D. Yu, “Conversational speech transcription using context-dependent deep neural networks,” Interspeech 2011.

[45] M. L. Seltzer, D. Yu, and Y. Wang, “An investigation of deep neural networks for noise robust speech recognition,” ICASSP, 2013.

[46] H. Sheikhzadeh and L. Deng. “Waveform-based speech recognition using hidden filter models: Parameter Selection and sensitivity to power normalization,” IEEE Trans. Speech & Audio Proc., Vol.2, pp. 80-91, 1994.

[47] Y. Shi, P. Wiggers and C.M. Jonker, “Towards recurrent neural network language models with linguistic and contextual features,” Interspeech, 2012.

[48] H. Su, G. Li, D. Yu, and F. Seide, “Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription,” ICASSP, 2013.

[49] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," Proc. SLT, 2012.

[50] G. Tur, L. Deng, D. Hakkani-Tur, and X. He, “Towards deeper understanding: Deep convex networks for semantic utterance classification,” ICASSP, 2012.

[51] N. Vu, W. Breiter, F. Metze, and T. Schultz, “An investigation on initialization schemes for multilayer perceptron training using multilingual data and their effect on ASR performance,” Interspeech, 2012.

[52] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong, “Adaptation of context-dependent deep neural networks for automatic speech recognition,” IEEE SLT, 2012.

[53] D. Yu, L. Deng, and G. Dahl, “Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition,” NIPS Workshop on Deep Learning, 2010.

[54] D. Yu, F. Seide, G. Li, and L. Deng, “Exploiting sparseness in deep neural networks for large vocabulary speech recognition,” ICASSP, 2012, pp. 4409–4412.

[55] D. Yu, K. Yao, H. Su, G. Li and F. Seide, “KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition,” ICASSP 2013.

[56] D. Yu, L. Deng, and F. Seide. “The deep tensor neural network with applications to large vocabulary speech recognition,” IEEE Trans Audio, Speech, & Lang. Proc. vol. 21, no. 2, pp. 388-396, Feb, 2013.

[57] F. Zamora-Martinez, S. Espana-Boquera, M.J. Castro-Bleda, and R. De-Mori, “Cache neural network language models based on long-distance dependencies for a spoken dialog system,” ICASSP, 2012.

[58] Y. Zhang, L. Deng, X. He, and A. Acero. “A novel decision function and the associated decision-feedback learning for speech translation,” ICASSP, 2011.