Improving Deep Neural Network Acoustic Modeling For Audio Corpus Indexing Under The IARPA Babel Program

Xiaodong Cui[1], *Brian Kingsbury*[1], *Jia Cui*[1], *Bhuvana Ramabhadran*[1], *Andrew Rosenberg*[2]
Mohammad Sadegh Rasooli[3], *Owen Rambow*[3], *Nizar Habash*[4], *Vaibhava Goel*[1]

[1]IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA
[2]Department of Computer Science, Queens College, City University of New York, NY 11367, USA
[3]Department of Computer Science, Columbia University, NY 10027, USA
[4]Center for Computational Learning Systems, Columbia University, NY 10115, USA
{cuix,bedk,jiacui,bhuvana,vgoel}@us.ibm.edu, andrew@cs.qc.cuny.edu,
{rasooli,rambow}@cs.columbia.edu, habash@ccls.columbia.edu

Abstract

This paper is focused on several techniques that improve deep neural network (DNN) acoustic modeling for audio corpus indexing in the context of the IARPA Babel program. Specifically, fundamental frequency variation (FFV) and channel-aware (CA) features and data augmentation based on stochastic feature mapping (SFM) are investigated not only for improved automatic speech recognition (ASR) performance but also for their impact to the final spoken term detection on the pre-indexed audio corpus. Experimental results on development languages of Babel option period one show that the improved DNN acoustic models can reduce word error rates in ASR and also help the keyword search performance compared to already competitive DNN baseline systems.

Index Terms: deep neural networks, data augmentation, stochastic feature mapping, fundamental frequency variation, channel-aware

1. Introduction

Deep neural networks (DNNs) have achieved the state-of-the-art performance in acoustic modeling for large vocabulary continuous speech recognition (LVCSR) in recent years [1][2][3][4]. In the IARPA Babel program, which seeks high-performance spoken term detection systems [5], conversational telephony audio data is usually pre-indexed by LVCSR systems for keyword search in a later stage. In the IBM systems for the Babel base period evaluation, DNN-based acoustic models were found to outperform discriminatively trained GMM acoustic models [6][7].

This paper aims at improving DNN acoustic modeling for audio corpus indexing for keyword search with a focus on the limited language packs (LLPs) of development languages which only have limited amount of training data (10 hours of speech). Although there are multiple DNN architectures for acoustic modeling, such as hybrid DNN models and bottleneck/tandem models, this paper concentrates on enhancing the performance of the IBM hybrid DNN models only. Specifically, several techniques that have shown to be helpful will be presented. First, the fundamental frequency variation (FFV) spectrum is included when constructing the feature space. The FFV features describe instantaneous change of fundamental frequency and provide prosody information of speech signals. It will be shown that FFV features can help improve the ASR performance as complementary features and are especially helpful for tonal languages. Second, to deal with channel variation, channel-aware input features are employed for DNN acoustic model learning by augmenting the original speech features with utterance-based cepstral means. Third, data augmentation based on stochastic feature mapping (SFM) [21] is used to deal with data sparsity for DNN training of LLPs. The SFM-based data augmentation scheme artificially augments the training set using replicas of training samples under certain label-preserving transformations. The new data is created by performing voice conversion between a source speaker and a target speaker in some designated feature space. It can be shown that the SFM-based data augmentation consistently improves both ASR and KWS performance. The effectiveness of the aforementioned techniques is demonstrated through experiments on development languages of Babel option period one. Their impact on both ASR and KWS will be investigated.

The remainder of the paper is organized as follows. Section 2 describes the baseline hybrid DNN systems that are used in the Babel base period evaluation. Improved ASR configurations from the baseline DNN acoustic models are provided in Section 3 with details of FFV and channel-aware features and SFM-based data augmentation given in the subsections, respectively. Experimental results on five development languages of Babel option period one are presented in Section 4.

2. Baseline DNN Systems

The IBM baseline hybrid DNN systems have been introduced in [6][7] in detail, and will only be briefly described here. The baseline hybrid DNN systems use speaker adaptive features as input. The construction of the feature space is illustrated in Fig.1. In this feature space 13-dimensional mean-normalized perceptual linear prediction (PLP) features with vocal tract length normalization (VTLN)[8] are used as the fundamental acoustic features. Acoustic context (CTX) information is taken into account by splicing adjacent 9 frames of PLP features. Linear discriminant analysis (LDA) is then used to project the feature dimensionality down to 40. The LDA features are further decorrelated by a global semi-tied covariance (STC) [9] matrix on top of which speaker-adapted training (SAT) using feature space maximum likelihood linear regression (FMLLR) (i.e. constrained MLLR (cMLLR))[10] is applied to reduce the speaker variability. Finally, 9 frames of FMLLR features

are concatenated together (CTX2) as the input features to the DNNs.

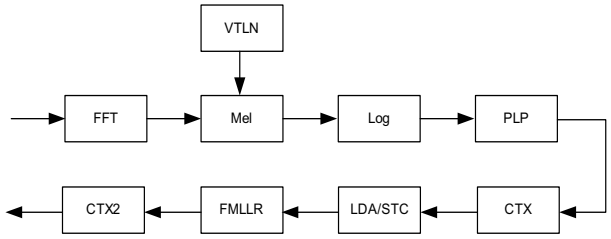

Figure 1: Speaker-adapted feature extraction pipeline of the baseline hybrid DNN systems.

The DNN acoustic models are composed of 5 hidden layers with sigmoid activation functions. Each hidden layer consists of 1024 hidden units. There are 1000 units in the softmax output layer. The DNN models are initialized by layer-wise discriminative pre-training. After initialization, they are optimized by 15 iterations of stochastic gradient training based on cross-entropy (CE) followed by 30 iterations of Hessian-free (HF) sequence training based on state-level minimum Bayes risk (sMBR)[4].

Compared to the training recipe described in [6] and [7], the baseline DNN model training in this paper is modified to a two-pass training scheme. After the sMBR training in the first pass, the DNN models are used to re-align the training data to re-generate the frame-wise labels. In the second pass, the refined targets are used for another 15 iterations of CE training followed by 30 iterations of sMBR-based sequence training.

3. Improved ASR Configurations

This section will present the improved ASR configurations for DNN acoustic modeling which include FFV and channel-aware features and SFM-based data augmentation.

3.1. Fundamental Frequency Variation (FFV) Features

Fundamental frequency variation (FFV) features are a frame-level vector representation describing instantaneous change in fundamental frequency $F0$ [11][12]. They have a variety of desired properties. For instance, FFV features are instantaneous: their estimation does not rely on adjacent frames. In addition, they are continuous and well-defined even in unvoiced frames. It was reported in previous work that FFV features when used in conjunction with other acoustic features can improve performance in applications such as ASR and speaker identification[13][14]. FFV features have also been applied in other ASR systems for Babel keyword search [26][27].

Similar to PLP or MFCC features, FFV features are computed from the speech samples within a window. Following the procedure in [12] with slight modification, the FFT magnitude spectra are first computed over the left and right halves of the analysis window and the FFV spectrum is then obtained by evaluating the vanishing point products between the dilated or compressed magnitude spectra. The FFV spectrum is then compressed using a filterbank with 7 filters resulting 7 dimensions indicating the magnitude of rise/fall of the fundamental frequency. Details of implementation of FFV and discussion with respect to its impact on bottleneck DNN configurations are given in [28].

The FFV features provide auxiliary prosody information of speech signals which will be especially helpful for tonal languages such as Cantonese, Vietnamese or Lao. The 7-dimensional FFV features are used as complementary features to augment the original 13-dimensional PLP features as shown in Eq.1.

$$\boldsymbol{o}_t = [\boldsymbol{x}_t^{(\mathsf{PLP})},\ \boldsymbol{x}_t^{(\mathsf{FFV})}] \tag{1}$$

The augmented 20-dimensional features are not only used as DNN input features but also to generate the decision tree whose quinphone states are output units of the hybrid DNN acoustic models. In this way, the decision tree can "see" the prosody information. The discriminative speaker adaptive GMM models are trained using the prosody-aware decision tree to create the initial frame-wise labeling for CE training of the DNN acoustic models.

3.2. Channel-aware (CA) Features

Most of the speech data in the Babel program is collected through wireless telephone channels. Therefore, dealing with channel variation in the DNN framework is an important issue. To that end, channel-aware features are investigated.

Assume the channel condition is stationary for each utterance u. In the cepstral domain, one has

$$\boldsymbol{x}_t^u = \boldsymbol{s}_t^u + \boldsymbol{h}^u \tag{2}$$

where $\boldsymbol{x}_t^u$ and $\boldsymbol{s}_t^u$ are the observed and speech cepstra of frame t of the utterance u, respectively, and $\boldsymbol{h}^u$ is the channel distortion which can be treated as constant convolutional noise. Assume the average of the speech cepstra in the utterance is approximately zero. It follows that

$$\boldsymbol{h}^u \approx \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t^u \tag{3}$$

where T is the total number of frames in utterance u. Hence, the cepstral mean on the RHS of Eq.3 can be used as an indicator of the channel condition. Using cepstral mean to mitigate the channel distortion has been well-known and been widely used in the feature extraction for partial removal of the convolutional noise. Cepstral mean normalization (CMN) is also applied in the feature space of the baseline hybrid DNN systems as described in Section 2. However, the channel distortion, especially for wireless channels, can be highly nonlinear even in the cepstral domain. Therefore, providing an indicator of the channel condition together with the input feature in the training may help the DNN acoustic models, which have multiple layers of nonlinearity, cope with the nonlinearity imposed by the channel distortion. In this work, the utterance-based PLP mean is appended to the input speaker-adapted acoustic features across all frames of the utterance as shown in Eq.4

$$\boldsymbol{o}_t = [\boldsymbol{x}_{t-4}^{(\mathsf{FMLLR})}, \cdots, \boldsymbol{x}_t^{(\mathsf{FMLLR})}, \cdots, \boldsymbol{x}_{t+4}^{(\mathsf{FMLLR})}, \boldsymbol{h}^u] \tag{4}$$

Using indicative constant vectors together with input features to DNNs has been seen to help the learning of DNNs. For instance, in [15], noise-aware features are used for noise robust ASR where the average of the first few frames of features is used to augment the original speech features. In [16], I-vectors with coefficients describing the speaker position in a certain subspace are used to augment the original speech features. The employment of channel-aware features investigated in this paper is similar in spirit.

3.3. Data Augmentation

Data augmentation based on label-preserving transformations for neural network training can help the networks make robust predictions that are invariant to those transformations. Hence, it has been a commonly used method in pattern recognition [17][18][19][20]. Augmented data can also alleviate the data sparsity issue for the neural network training, which is crucial for the performance of the Babel LLPs consisting of only about 10 hours of speech data.

In acoustic modeling, label-preserving transformation based data augmentation is not a novel practice. For example, multi-style training (MST) where clean speech signals are artificially corrupted by adding background noise to improve the noise robustness of acoustic models also belongs to this family. However, DNN-oriented data augmentation in terms of speaker variability has not been extensively investigated except for [20]. In this paper, a data augmentation approach based on stochastic feature mapping (SFM) proposed in [21] will be used to augment the training data for the DNN acoustic modeling of the Babel LLPs.

SFM augments the training data by converting the feature sequence of one speaker (source speaker S) in each utterance to the feature sequence of another speaker (target speaker B) under the same transcripts. A speaker-dependent acoustic model is first trained for the target speaker for his/her speaker-specific acoustic characteristics, which is accomplished via model space MLLR [10] in this work. Given the feature extraction pipeline in Fig.1, SFM is performed by a composition of two linear transformations as shown in Eq.5

$$\boldsymbol{O}_{\mathrm{B}}^{(\mathrm{FMLLR})} = \boldsymbol{A}_{\mathrm{B}}\boldsymbol{O}_{\mathrm{B}}^{(\mathrm{LDA})} + \boldsymbol{b}_{\mathrm{B}} = \boldsymbol{A}_{\mathrm{B}}(\tilde{\boldsymbol{A}}\boldsymbol{O}_{\mathrm{S}}^{(\mathrm{LDA})} + \tilde{\boldsymbol{b}}) + \boldsymbol{b}_{\mathrm{B}} \quad (5)$$

where $\boldsymbol{O} = \{\boldsymbol{o}_t\}_{t=1}^{\mathrm{T}}$ is the sequence of T frames of features for an utterance. The first linear transformation denoted by $\{\tilde{\boldsymbol{A}}, \tilde{\boldsymbol{b}}\}$ transforms the feature sequence from the LDA space of the source speaker to the LDA space of the target speaker. The second linear transformation denoted by $\{\boldsymbol{A}_{\mathrm{B}}, \boldsymbol{b}_{\mathrm{B}}\}$ is the speaker adaptive transformation that transforms the feature sequence from the LDA space of the target speaker to the FMLLR space of the target speaker. For each speaker in the training set, all his/her utterances are mapped to multiple randomly selected speakers in the training set under the same transcripts. Obviously, this label-preserving transformation can enrich the acoustic information for the selected speakers, thus augmenting the training data in an informative way.

4. Experimental Results

Experiments are carried out on the LLP development sets of the five development languages of Babel option period one including four non-tonal languages (*Bengali, Assamese, Haitian Creole and Zulu*) and one tonal language (*Lao*). Each development set has 20 hours of audio containing about 10 hours of conversational speech.

4.1. Baseline DNN Models

Table 1 shows the word error rates (WERs) of baseline DNN models and discriminatively trained GMM models on LLPs of four non-tonal development languages. The first column presents the performance of the DNN models under the two-pass training scheme. With refined class labeling using the DNN alignments in the second pass, 0.5% to 1.3% absolute improvements are observed on four languages after sMBR-based sequence training over the first pass using the alignments generated by FMPE+MPE GMM models. As a comparison, the WER (71.4%) of using a maximum-likelihood-based speaker adaptive GMM model for generating alignments is also presented for Bengali, which is worse than that (68.9%) of using alignments generated by a discriminatively trained GMM model (FMPE+MPE) and significantly worse than that (67.6%) of using alignments generated by a DNN model.

The second column of the table presents the performance of the best IBM discriminative GMM models (FMPE+MPE+MLLR) on the LLPs of the four languages. For *Haitian Creole*, the DNN and GMM models are about the same. For *Bengali*, *Assamese* and *Zulu*, DNN models are 1.6%, 2.9% and 2.5% absolute better than the GMM models, respectively.

language	alignment	DNN	GMM
Bengali	GMM	71.4	**69.2**
	FMPE+MPE	68.9	
	DNN	**67.6**	
Assamese	FMPE+MPE	67.2	**69.6**
	DNN	**66.7**	
Haitian Creole	FMPE+MPE	64.4	**63.4**
	DNN	**63.5**	
Zulu	FMPE+MPE	74.7	**76.4**
	DNN	**73.9**	

Table 1: WERs(%) of DNN models and discriminatively trained GMM models on LLPs of four non-tonal development languages.

4.2. FFV and Channel-Aware Features

Table 2 shows the impact of the FFV and channel-aware features on the performance of the DNN models on a tonal language (*Lao*) and a non-tonal language (*Zulu*).

language	Zulu		Lao	
model	CE	sMBR	CE	sMBR
DNN	78.9	**73.9**	67.1	**63.3**
DNN + FFV	77.6	**73.5**	64.3	**60.6**
DNN + FFV + CA	76.0	**72.8**	63.0	**60.2**

Table 2: WERs(%) of DNN models with FFV and channel-aware features on Zulu and Lao LLPs.

From the table, including FFV helps reduce the WERs of the DNN models for both languages, specifically, 0.4% absolute for *Zulu* and 2.6% absolute for *Lao*. Note that FFV is particularly helpful for *Lao*, the tonal language. On top of FFV, channel-aware features can further improve the performance of the DNN models. The improvements are relatively larger after CE training (1.6% absolute for *Zulu* and 1.3% absolute for *Lao*). The gains decrease after the sequence training (0.7% absolute for *Zulu* and 0.4% absolute for *Lao*). The speculation is that the utterance-based PLP mean integrated in the channel-aware features provides a certain degree of sequence information to the frame-wise CE training that each individual frame does not have. However, in the sMBR training, that type of sequence information can be partially learned by the sequence training.

4.3. SFM-Based Data Augmentation

Table 3 shows the performance of the SFM-based data augmentation on the four non-tonal languages using the conventional

language		WER	
		CE	sMBR
Bengali	Baseline DNN	71.1	**67.6**
	BS DNN	70.2	66.6
	BS DNN + SFMx4	68.4	**65.2**
Assamese	Baseline DNN	72.2	**66.7**
	BS DNN	72.8	66.5
	BS DNN + SFMx4	69.9	**63.4**
Haitian Creole	Baseline DNN	65.8	**63.5**
	BS DNN	66.9	62.8
	BS DNN + SFMx4	62.3	**59.1**
Zulu	Baseline DNN	78.9	**73.9**
	BS DNN	79.1	73.6
	BS DNN + SFMx4	77.0	**71.4**

Table 3: WERs(%) of DNN models using SFM-based data augmentation on the LLPs of the four non-tonal development languages.

features (Fig.1). The data augmentation is conducted in conjunction with bootstrapped DNN models which are denoted by BS DNN in the table. The only difference between the baseline DNN models and the bootstrapped DNN models is that bootstrapped DNN models use quinphone states from the decision tree of bootstrapped GMM models. Those models are aggregated speaker adaptive models trained using multiple sets of resampled training data [22], which are designed to deal with data sparsity. As a result, the BS DNNs have 2000 units in the softmax output layer while the baseline DNNs have 1000 units. With data augmentation, the training data set is significantly increased. Therefore, increasing the number of parameters by using the BS DNNs is a reasonable choice. In the experiments in Table 3, SFM generates four replicas of the training data which are combined with the original training set to train the DNN models. As observed from the table, SFM-based data augmentation has consistently improved the performance of the DNN models. Specifically, the WERs are reduced by 2.4% absolute for *Bengali*, 3.3% absolute for *Assamese*, 4.4% absolute for *Haitian Creole* and 2.5% absolute for *Zulu*.

4.4. Performance on KWS

Since the automatic speech recognizers are used to pre-index an audio corpus for keyword search, it is important and necessary to investigate not only the performance of the speech recognizers but also the eventual performance of the KWS systems.

The KWS performance is measured by the term-weighted value (TWV) [23] defined as a function of the probability of missed detections and the probability of false alarms. In this paper, maximum term-weighted value (MTWV) which is the best TWV achievable by varying the threshold that defines the YES/NO decisions in the postings list is used for reporting the KWS performance.

The IBM KWS system is based on a two-pass implementation of weighted finite state transducer (WFST) indexing and search [6][7]. The query list is split into in-vocabulary (IV) and out-of-vocabulary (OOV) queries according to the ASR lexicon. For IV queries, each query is converted into a lexical finite state acceptor which is then composed with the lexical index. For OOV queries, each query is converted by a grapheme-to-phoneme converter to a phonetic finite-state acceptor which is then composed with the phonetic index. A phoneme-to-phoneme confusion model is used in the composition. A cascaded search strategy is used [24]. In this strategy, the word index is first searched for IV queries. If no results are returned or the query term is OOV then the phonetic index is searched. Sum-to-one normalization [25] is applied to each term in the postings list.

Table 4 and Table 5 demonstrate the WERs and MTWVs of the DNN acoustic models for *Lao* and *Zulu* LLPs, respectively. The search is conducted on a 50-hour audio corpus. The search uses the 2000 keywords in the development keyword list from each language. Note that automatic segmentation of the audio corpus is used in the experiments so the WERs are slightly different from those reported in other ASR experiments in this Section. In addition, lattices are not fully optimized in terms of the KWS performance.

In Table 4, it can be seen that the improved ASR configurations have not only improved the WERs but also the MTWVs. With FFV features, the MTWV improves from 0.3774 to 0.3876 which is further improved slightly by including channel-aware features to 0.3894. The SFM-based data augmentation with 4 replicas of training data significantly boosts the MTWV from 0.3876 to 0.4312 when FFV features are used. But when combining FFV, CA and SFM, the overall KWS performance (MTWV=0.4267) slightly drops compared to only using FFV and SFM (MTWV=0.4312).

In Table 5, in addition to word and phonetic indices, morphological index [29] is also used for Zulu LLP. From the table, it can be seen that SFM-based data augmentation improves MTWV from 0.2202 to 0.2442. When adding FFV and CA, the MTWV is further improved to 0.2601. In the meantime, the WERs are also improved along with the MTWVs.

language	Lao	
model	WER	MTWV
DNN	62.8	0.3774
DNN + FFV	60.4	0.3876
DNN + FFV + CA	60.0	0.3894
BS DNN + FFV + SFMx4	**58.3**	**0.4312**
BS DNN + FFV + CA + SFMx4	58.3	0.4276

Table 4: WERs(%) and MTWVs of DNN models on Lao LLP.

language	Zulu	
model	WER	MTWV
DNN	73.1	0.2202
BS DNN + SFMx4	72.1	0.2442
BS DNN + FFV + CA + SFMx4	**70.5**	**0.2601**

Table 5: WERs(%) and MTWVs of DNN models on Zulu LLP.

5. Acknowledgements

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD/ARL) contract number W911NF-12-C-0012. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government. This effort uses the IARPA Babel Program language collection releases IARPA-babel102b-v0.5a, IARPA-babel103b-v0.4b, IARPA-babel201b-v0.2b, IARPA-babel203b-v3.1a and IARPA-babel206b-v0.1e limited language packs.

6. References

[1] G. E. Hinton, S. Osindero and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol.18, pp.1527-1554, 2006.

[2] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Interspeech*, 2011, pp.437-440.

[3] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription, in *Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011, pp. 24-29.

[4] B. Kingsbury, T. Sainath and H. Soltau, "Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization," *Interspeech*, 2012.

[5] http://www.iarpa.gov/Programs/ia/Babel/babel.html

[6] J. Cui, X. Cui, B. Ramabhadran, J. Kim, B. Kingsbury, J. Mamou, L. Mangu, M. Picheny, T. N. Sainath and A. Sethy "Developing speech recognition systems for corpus indexing under the IARPA Babel program," *ICASSP*, 2013.

[7] B. Kingsbury, J. Cui, X. Cui, M. J. F. Gales, K. Knill, J. Mamou, L. Mangu, D. Nolden, M. Picheny, B. Ramabhadran, R.Schluter, A. Sethy, P. C. Woodland, "A high-performance Cantonese keyword search system," *ICASSP*, 2013.

[8] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, 1998, pp. 49-60.

[9] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 272-281, 1999.

[10] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, 1998, pp. 75-98.

[11] K. Laskowski, J. Edlund, and M. Heldner, "Learning prosodic sequences using the fundamental frequency variation spectrum," *Proceedings of Speech Prosody*, 2008.

[12] K. Laskowski, M. Heldner and J. Edlund, "The fundamental frequency variation spectrum," *Proceedings of FONETIK*, 2008.

[13] K. Laskowski, J. Edlund and M. Heldner, "An instantaneous vector representation of delta pitch for speaker-change prediction in conversational dialogue systems," *ICASSP* 2008.

[14] K. Laskowski and Q. Jin, "Modeling instantaneous intonation for speaker identification using the fundamental frequency variation spectrum," *ICASSP*, 2009.

[15] M. Seltzer, D. Yu and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," *ICASSP*,2013.

[16] G. Saon, H. Soltau, D. Nahamoo and M. Picheny, "Speaker adaptation of neural network acoustic models using I-vectors," *ASRU*,2013.

[17] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient based learning applied to document recognition," *Proceedings of IEEE*, vol. 86, no. 11, 1998, pp. 2278-2324.

[18] P. Y. Simard, D. Steinkraus and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," *International Conference on Document Analysis and Recognition (ICDAR)*, 2003, pp. 958-963.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton,"ImageNet classification with deep convolutional neural networks," *Neural Information Processing Systems (NIPS)*, 2012, pp. 1106-1114.

[20] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition, in *ICML Workshop on Deep Learning for Audio, Speech, and Language Processing*, 2013.

[21] X. Cui, V. Goel and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *ICASSP*, 2014.

[22] X. Cui, J. Xue, X. Chen, P. A. Olsen, P. L. Dognin, U. V. Chaudhari, J. R. Hershey, and B. Zhou, "Hidden Markov acoustic modeling with bootstrap and restructuring for low-resourced languages," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 8, 2012, pp. 2252-2264.

[23] J. G. Fiscus, J. G. Ajot, J. Garofalo, and G. Doddington, "Results of the 2006 spoken term detection evaluation, in *Proc. SIGIR Workshop on Searching Spontaneous Conversational Speech*, 2007, pp. 51-57.

[24] M. Saraclar and R. Sproat, "Lattice-based search for spoken utterance retrieval," *HLT-NAACL*, pp. 129-136, 2004.

[25] M. Montague and J. A. Aslam, "Relevance score normalization for metasearch," in *Proc. ACM International Conference on Information and Knowledge Management*, 2001, pp. 427-433.

[26] F. Metze, Z. A. W. Sheikh, A. Waibel, J. Gehring, K. Kilgour, Q. B. Nguyen and V. H. Nguyen, "Models of tone for tonal and non-tonal languages," *ASRU*, 2013.

[27] Babel internal slides and personal communication.

[28] J. Cui, B. Ramabhadran, X. Cui, A. Rosenberg, B. Kingsbury1, A. Sethy, "Recent improvements in neural network acoustic modeling for LVCSR in low-resource languages," *Interspeech*, 2014, accepted.

[29] M. S. Rasooli, N. Habash, O. Rambow and T. Lippincott, "Unsupervised morphology-based vocabulary expansion," *The 52nd Annual Meeting of the Association for Computational Linguistics*, 2014.