DATA AUGMENTATION FOR DEEP NEURAL NETWORK ACOUSTIC MODELING

Xiaodong Cui, Vaibhava Goel, Brian Kingsbury

IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA

ABSTRACT

Data augmentation using label preserving transformations has been shown to be effective for neural network training to make invariant predictions. In this paper we focus on data augmentation approaches to acoustic modeling using deep neural networks (DNNs) for automatic speech recognition (ASR). We first investigate a modified version of a previously studied approach using vocal tract length perturbation (VTLP) and then propose a novel data augmentation approach based on stochastic feature mapping (SFM) in a speaker adaptive feature space. Experiments were conducted on Bengali and Assamese limited language packs (LLPs) from the IARPA Babel program. Improved recognition performance has been observed after both cross-entropy (CE) and state-level minimum Bayes risk (sMBR) training of DNN models.

Index Terms— deep neural networks, data augmentation, vocal tract length perturbation, stochastic feature mapping, automatic speech recognition.

1. INTRODUCTION

A good neural network model for pattern recognition should make predictions that are invariant to variations of the same class of patterns. This is usually done by traing the neural network using a large number of samples with abundant variations of the patterns to be recognized. However, this will pose a problem when there is only limited training data available. Under this condition, the neural network training is at the risk of over-fitting and hurting the classification robustness. One way to deal with this problem is data augmentation where the training set is artificially augmented by adding replicas of the training samples under certain types of transformations that preserve the class labels. Data generated under such label preserving transformations will improve the prediction invariance and generalization ability of the neural networks.

Data augmentation has been widely used in neural network based pattern recognition tasks [1][2][3][4], especially in image recognition where transformations such as translation, deformation and reflection [1][4] have led to significant improvements in recognition accuracy. In the past few years, deep neural networks (DNNs) have made dramatic impact in acoustic modeling and have delivered the state-of-the-art performance in automatic speech recognition (ASR) [5][6][7][8]. However, work related to data augmentation for ASR based on DNNs has been rarely reported. Most recently, a data augmentation scheme based on vocal tract length perturbation (VTLP) was proposed in [3] and experiments on the TIMIT database using deep convolutional neural networks (CNNs) showed decent improvements in phone error rate (PER).

In this paper, we exploit data augmentation approaches to deal with limited training data in deep neural network (DNN) acoustic modeling for large vocabulary continuous speech recognition (LVCSR). We first investigate a modified version of VTLP proposed in [3] and then propose a novel label preserving transformation scheme based on stochastic feature mapping (SFM) in a speaker-adaptive feature space to augment the training data. Experiments were carried out on limited language packs (LLPs) of two Indian languages, Bengali and Assamese, under the IARPA Babel program [9].

The remainder of the paper is organized as follows. Section 2 gives the details of our implementation of VTLP and also the SFM approach to generate transformed input features for deep neural network training. Experimental results on Bengali and Assamese LLPs under both cross-entropy (CE) training and Hessian-free (HF) sequence training of the hybrid DNN acoustic models are presented in Section 3 followed by a discussion and future work in Section 5.

2. DATA AUGMENTATION

2.1. Feature Space

The data augmentation schemes to be investigated in this paper are based on a speaker adaptive feature space whose extraction pipeline is shown in Fig. 1. This is also the feature pipeline used by the IBM speaker adaptive ASR systems in the Babel evaluation [10][11].

In this pipeline 13-dimensional mean-normalized perceptual linear prediction (PLP) features with vocal tract length normalization (VTLN) [12] are used as the fundamental acoustic features. After taking into the context (CTX) information by splicing adjacent 9 frames, linear discriminant analysis (LDA) is used to project the feature dimensionality down to 40. The components of LDA features are further decorrelated by a global semi-tied covariance (STC) matrix [13]. For implementation convenience, the two transformation matrices from LDA and STC are combined together to create only one transformation matrix and we still refer to this feature space as the LDA space. In this LDA space, speaker adaptive training (SAT) using feature space maximum likelihood linear regression (FMLLR) (i.e. constrained maximum likelihood linear regression (cMLLR) [14]) is applied to reduce the speaker variability.

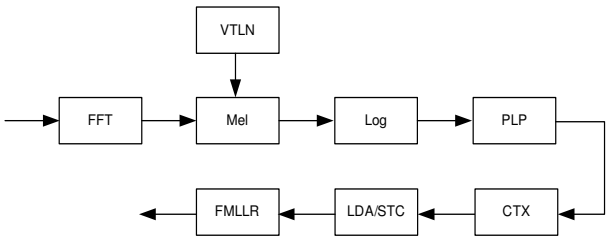

Fig. 1. Speaker adaptive feature extraction pipeline.

In what follows, we will investigate two label preserving transformations in this feature space to augment the training data.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

2.2. Vocal Tract Length Perturbation

VTLP was first proposed in [3] with experiments on TIMIT database where for each utterance in the training set a warping factor α is randomly chosen from $[0.9, 1.1]$ to warp the frequency axis. Therefore, the vocal tract length of the speaker is slightly perturbed to distort the original speech spectrum of the utterance to create a new replica of it.

In the IBM Attila toolkit [15], the vocal tract length warping factor is quantized between $[0.8, 1.25]$. As a result, the estimated warping factor α is an integer between $[0, 20]$ with 10 equivalent to the neutral warping factor 1.0. In this paper, we exploit a modified version of VTLP used in [3]. Instead of randomly selecting a warping factor, we use a deterministic perturbation:

$$\alpha \mapsto \{\alpha-4,\ \alpha-2, \alpha+2,\ \alpha+4\} \tag{1}$$

where the VTLN warping factor α for a speaker is first estimated and then perturbed in both positive and negative directions by small shifts (± 2 and ± 4) to give 4 more warping factors. The perturbed warping factors, if they are beyond $[0.8, 1.25]$, are clipped to 0.8 or 1.25 which corresponds to integer 0 or 20, respectively, in the Attila implementation. The 4 warping factors after perturbation are applied to the original speech signals to create 4 replicas of all the utterances under the same speaker.

The reason we choose deterministic perturbation rather than random perturbation is that speech features are not sensitive to a small distortion of the VTL warping factor. To guarantee an effective perturbation, we force a relatively large step away from the original warping factor. In addition, the speaker adaptive feature space investigated in this paper is different from that of [3]. In this feature space, VTL is explicitly estimated and used for speaker normalization. Based on observations of our pilot experiments, this method appears to be more helpful than that used in [3] for our Babel tasks.

2.3. Stochastic Feature Mapping

While VTLP augments data by perturbing a speaker, we also want to explore the possibility of augmenting data by converting a speaker's speech to another speaker. To that end, we attempt to answer the following question:

Suppose there is a speaker S who speaks an utterance u with label $\mathbf{W}$ which generates a sequence of features with N frames

$$\mathbf{O}^{(\mathsf{S})} = \{\mathbf{o}_1^{(\mathsf{S})}, \cdots, \mathbf{o}_N^{(\mathsf{S})}\} \tag{2}$$

Then for another speaker B what would the sequence of features

$$\mathbf{O}^{(\mathsf{B})} = \{\mathbf{o}_1^{(\mathsf{B})}, \cdots, \mathbf{o}_N^{(\mathsf{B})}\} \tag{3}$$

be if he/she were to speak the same utterance u under the same label $\mathbf{W}$?

Ideally, stereo data is needed for such a conversion, but unfortunately it is not available in most training scenarios. Therefore in this paper we investigate a stochastic feature mapping approach that estimates a mapping function for the feature conversion between the two speakers statistically.

Specializing to the speaker adaptive feature space we are using in Fig.1, to generate the target feature sequence in the FMLLR space, we first build a speaker dependent model for the target speaker B in the LDA space $\boldsymbol{\lambda}_{\mathsf{LDA}}^{(B)}$. This is done by model space maximum likelihood linear regression (MLLR) [16] based on a regression tree which dynamically determines the transformation granularity.

Given $\boldsymbol{\lambda}_{\mathsf{LDA}}^{(B)}$, we want to estimate a linear transformation of LDA feature sequence $\mathbf{O}_{\mathsf{LDA}}^{(\mathsf{S})}$ from speaker S such that the transformed feature sequence maximizes the likelihood against model $\boldsymbol{\lambda}_{\mathsf{LDA}}^{(B)}$:

$$\{\tilde{\mathbf{A}}, \tilde{\mathbf{b}}\} = \underset{\{\mathbf{A},\mathbf{b}\}}{\operatorname{argmax}} \log P\big(\tilde{\mathbf{A}}\mathbf{O}_{\mathsf{LDA}}^{(\mathsf{S})} + \tilde{\mathbf{b}} | \boldsymbol{\lambda}_{\mathsf{LDA}}^{(B)}\big) \tag{4}$$

Eq.4 is simply a cMLLR problem in the LDA feature space [14].

When the linear transformation $\{\tilde{\mathbf{A}}, \tilde{\mathbf{b}}\}$ is in place, the LDA feature sequence for the target speaker B can be obtained by

$$\mathbf{O}_{\mathsf{LDA}}^{(\mathsf{B})} = \tilde{\mathbf{A}}\mathbf{O}_{\mathsf{LDA}}^{(\mathsf{S})} + \tilde{\mathbf{b}} \tag{5}$$

Assume $\{\mathbf{A}^{(\mathsf{B})}, \mathbf{b}^{(\mathsf{B})}\}$ is the SAT FMLLR transformation for the speaker B in the speaker adaptive feature space in Fig.1, then we have

$$\begin{aligned} \mathbf{O}_{\mathsf{FMLLR}}^{(\mathsf{B})} &= \mathbf{A}^{(\mathsf{B})}\mathbf{O}_{\mathsf{LDA}}^{(\mathsf{B})} + \mathbf{b}^{(\mathsf{B})} \\ &= \mathbf{A}^{(\mathsf{B})}(\tilde{\mathbf{A}}\mathbf{O}_{\mathsf{LDA}}^{(\mathsf{S})} + \tilde{\mathbf{b}}) + \mathbf{b}^{(\mathsf{B})} \end{aligned} \tag{6}$$

From Eq.6 we can see that $\mathbf{O}_{\mathsf{FMLLR}}^{(\mathsf{B})}$, which is the converted feature sequence for the same utterance with the same label as $\mathbf{O}_{\mathsf{FMLLR}}^{(\mathsf{S})}$, is obtained by a composition of two linear transforms: One maps the LDA features from speaker S to speaker B, and the other transforms the mapped features from the LDA space to the FMLLR space for speaker B. Therefore, as a label preserving transformation, it essentially performs the "voice conversion" between two speakers in the designated feature space.

The implementation details of the above SFM approach is illustrated in Algorithm 1.

Algorithm 1 Data augmentation by Stochastic Feature Mapping

M $\leftarrow$ number of replicas ;

S $\leftarrow$ number of speakers ;

for $i \leftarrow 1, \cdots, S$ **do**

  get speaker dependent model λ_i in LDA space by MLLR using all utterances from speaker i ;

end for

for $i \leftarrow 1, \cdots, S$ **do**

  for $j \leftarrow 1, \cdots, M$ **do**

    randomly select a new speaker k as the target speaker ;

    estimate cMLLR transformation matrix $\{\tilde{\mathbf{A}}, \tilde{\mathbf{b}}\}$ based on model λ_k of the target speaker and all utterances from speaker i in LDA space according to Eq.4;

    map all utterances from speaker i to the target speaker k using $\{\tilde{\mathbf{A}}, \tilde{\mathbf{b}}\}$ in the LDA feature space ;

    transform mapped utterances using SAT FMLLR transformation $\{\mathbf{A}^{(\mathsf{k})}, \mathbf{b}^{(\mathsf{k})}\}$ of speaker k;

  end for

end for

After data augmentation, multiple replicas of the original training data are created. The augmented training data (both original and replicas) will be used for the DNN training.

3. EXPERIMENTAL RESULTS

Experiments were carried out using the limited language packs (LLPs) of Bengali and Assamese, two development languages from option period 1 of the Babel program. The Bengali LLP comprises 23.8 hours of telephony data for the training data set and 20.1 hours of telephony data for the development set. The Assamese LLP comprises 24.3 hours of telephony data for the training data set and 20.0 hours of telephony data for the development set. Both training data sets consist of scripted and conversational speech while the development sets consist of conversational speech only. Specifically, the Bengali training set is composed of 19.9 hours of conversational data and 3.9 hours of scripted data. The Assamese training set is composed of 20.0 hours of conversational data and 4.3 hours of scripted data. All the data is sampled at 8 KHz. Approximately 40%-50% of the audio is speech which indicates quite limited data for training.

Hybrid DNNs and the aforementioned speaker adaptive features are used for acoustic models. The input to the DNNs consists of nine frames of 40-dimensional FMLLR features generated by the feature extraction pipeline illustrated in Fig.1. The DNNs have five hidden layers each of which is composed of 1,024 hidden units with a sigmoid activation function and a softmax output layer with 1,000 or 2,000 quinphone context dependent states depending on the GMM/HMM models they are associated with. There are two types of GMM/HMM models used in the IBM systems [10]. One is the baseline SAT GMM/HMM with 1,000 quinphone states and 6,000 Gaussians. The other is a bootstrap and aggregation based GMM/HMM [17][10] with 2,000 quinphone states and 120,000 Gaussians. Since the GMM/HMM models only affect the DNN training via the alignments, there is no restructuring in this case.

The training of DNNs is carried out in three stages. The DNNs are first initialized by layer-wise discriminative pre-training. The state-level target labels are generated by Viterbi alignment using some existing acoustic models. After the DNNs are initialized, cross-entropy (CE) training is conducted at the frame level. The optimization in both discriminative pre-training and CE training uses a mini-batch based stochastic gradient descent (SGD) algorithm with frame randomization. Finally, the DNNs are further optimized using the Hessian-free (HF) sequence training under the state-level minimum Bayes risk (sMBR) criterion [8]. Once the HF sequence training is over, the obtained DNNs are used to re-align the training data to refine the state-level target label and another round of three-stage training is conducted to get the final DNN models.

For data augmentation, VTLP is implemented as described in Sec. 2.2. Four additional replicas are created by perturbing the estimated VTL warping factor according to Eq.1. To make a comparison, four additional replicas are also produced using SFM by randomly selecting 4 new speakers from the training set. Both VTLP and SFM only augment conversational data in the training sets of the two languages. There are 124 speakers for the conversational data in the Bengali training set and 138 speakers for the conversational data in the Assamese training set.

The development sets of the two LLPs are used as test sets which are decoded using dynamic decoders [15]. Bi-gram language models (LMs) are used for both languages since they give the best held-out perplexity. There are 11K words in the Bengali test dictionary and 9.4K words in the Assamese test dictionary. Note that in [3] the test set is also augmented by VTLP based utterance transformations and noticeable gains on TIMIT database have been observed by various aggregation methods in the decoding process. However, we only found very marginal gains by adding utterance variations to the test set in some pilot experiments on Bengali LLP. In addition, the aggregation of such variations in decoding significantly increased the decoding time. So in this paper we don't employ data augmentation for the test sets and just run the decoding in the conventional way.

	Bengali		Assamese	
	CE	sMBR	CE	sMBR
Baseline SAT DNN	71.1	67.6	72.2	66.7
BS SAT DNN	70.2	66.6	72.8	66.3
BS SAT DNN + VTLP	68.5	65.3	70.1	64.7
BS SAT DNN + SFM	68.4	65.4	69.9	64.1

Table 1. Word error rates (WERs) of DNN acoustic models of Bengali and Assamese LLPs after cross-entropy (CE) and state-level minimum Bayes risk (sMBR) sequence training.

Table 1 shows the performance of the hybrid DNN acoustic models of Bengali and Assamese LLPs with and without data augmentation. Word error rates (WERs) of both frame-level CE training and sequence level sMBR training are presented. The first two rows of Table 1 are the DNN models without data augmentation. DNN models based on the bootstrapped SAT GMM/HMM (BS SAT DNN) improve WERs over the baseline SAT DNNs (1.0% absolute for Bengali and 0.4% absolute for Assamese after sMBR sequence training) mainly due to its larger number of output states via data resampling and aggregation and better alignments for the targets.

In the last two rows of the table, VTLP and SFM are applied on the BS SAT DNN models. By adding 4 replicas of the transformed conversational data to the original training sets, VTLP obtains 1.7% absolute improvement after CE and 1.3% absolute improvement after sMBR sequence training over BS SAT DNN models for Bengali; 2.7% absolute improvement after CE and 1.6% absolute improvement after sMBR sequence training over BS SAT DNN models for Assamese. Similarly, SFM obtains 1.8% absolute improvement after CE and 1.2% absolute improvement after sMBR sequence training over BS SAT DNN models for Bengali; 2.9% absolute improvement after CE and 2.2% absolute improvement after sMBR sequence training over BS SAT DNN models for Assamese. If compared to the baseline SAT DNN models, the best improvement by bootstrap/aggregation and data augmentation is 2.3% and 2.6% absolute after sMBR sequence training for Bengali and Assamese, respectively.

From the data augmentation perspective, as observed from Table 1, VTLP and SFM have comparable performance after sMBR sequence training on Bengali while SFM is 0.6% absolute better than VTLP after sMBR sequence training on Assamese. If we increase the number of replicas, VTLP tends to saturate or slightly degrade while SFM continues to improve. This is demonstrated in Table 2 which shows the CE WERs of Bengali DNN models using data augmentation by VTLP and SFM generating different numbers of replicas. In this table, 8 replicas of data are generated by VTLP using denser perturbation, namely $\{\pm 1, \pm 2, \pm 3, \pm 4\}$, of the current speaker's VTL warping factor. As a result, the performance does not improve after CE training with more augmented data. It actually degrades slightly from 68.5% to 68.8%. This is because acoustic features are not sensitive to small distortions to VTL warping factors. Given the range of warping factor (usually $[0.8, 1.2]$), we can imagine that VTLP will plateau if we keep adding new data by perturbing in this range for one speaker. Nevertheless, SFM does not appear to have this problem as long as we have a good diversity of speakers. This can explain why its performance can still keep improving when increasing the transformed data from 68.4% to 68.0%.

	VTLP	SFM
4 replicas	68.5	68.4
8 replicas	68.8	68.0

Table 2. Word error rates (WERs) of DNN acoustic models of Bengali LLP after cross-entropy (CE) training using data augmentation by VTLP and SFM with different numbers of replicas.

4. RELATION TO PRIOR WORK

Although following the same concept of improving the prediction invariance of neural networks, data augmentation in speech recognition for DNN based acoustic model training is quite different from that in other pattern recognition tasks, such as image recognition [4][1][2], since the ways the features are extracted are so different.

This paper is motivated by [3] which has shown promising data augmentation results on ASR using VTLP. We want to extend VTLP to LVCSR tasks such as ASR in the Babel program and have investigated a modified version of VTLP. That is, instead of randomly choosing a warping factor for each utterance of a speaker, we deterministically perturb the estimated VTL warping factor of a speaker. We also proposed a novel data augmentation approach based on SFM for utterance transformation. SFM estimates a maximum likelihood linear transformation in some feature space of the source speaker against the speaker dependent model of the target speaker. Different from VTLP which perturbs a speaker, SFM explicitly maps the features of a speaker to some target speaker based on a statistically estimated linear transformation.

5. DISCUSSION AND FUTURE WORK

Inspired by voice conversion, stochastic feature mapping is a label preserving transformation that augments the training data for neural network models by mapping speech features from a source speaker to a target speaker. It does not create new speaker information, but rather generates more acoustic context/variation information (from the utterance labels of the source speaker) for the target speaker that does not exist in the original training data.

SFM performs equivalent "voice conversion" in some designated feature space in the sense of stochastic mapping. The acoustic information of the target speaker is contained in his/her speaker dependent model. The mapping between the two speakers is statistically estimated so it does not rely on any particular spectral manipulation. Although in this paper we describe the mapping in the speaker adaptive feature space for the IBM systems as an example, SFM as a transformation approach in general can be applied to any feature space for the data augmentation purpose.

Given that VTLP attempts to perturb the VTL of a speaker himself/herself while SFM attempts to map the acoustics of the speaker to another speaker, the two approaches might be complementary. In fact, some of our pilot experiments at the CE stage have shown positive evidence to support this speculation. The combination of the two approaches for better data augmentation performance in under investigation.

Although data augmentation is most helpful for neural network training when the training data is limited, as indicated in the experimental results of the Babel limited language packs in this paper, we also would like to explore its effectiveness when the training data is relatively ample, for instance, the Babel full language packs which have around 200 hours of data and approximately 100 hours of speech.

6. ACKNOWLEDGEMENTS

This effort uses the IARPA Babel Program Assamese language collection release babel102b-v0.4 limited language pack and Bengali language collection release babel103b-v0.3 limited language pack. 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.

7. REFERENCES

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

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

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

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

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

[6] 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.

[7] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," in *IEEE Signal Processing Maganize*, November 2012, pp. 82–97.

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

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

[10] 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," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 6753–6757.

[11] B. Kingsbury, J. Cui, X. Cui, M. J. F. Gales, K. Knill, J. Mamou, L. Mangu, D. Nolden, M. Picheny, B. Ramabhadran, R. Schlüter, A. Sethy, and P. C. Woodland, "A high-performance Cantonese keyword search system," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 8277–8281.

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

[13] 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.

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

[15] H. Soltau, G. Saon, and B. Kingsbury, "The IBM Attila speech recognition toolkit," in *Spoken Language Technology Workshop (SLT)*, 2010, pp. 97–101.

[16] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[17] 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, pp. 2252–2264, 2012.