REFINEMENTS OF REGRESSION-BASED CONTEXT-DEPENDENT MODELLING OF DEEP NEURAL NETWORKS FOR AUTOMATIC SPEECH RECOGNITION

Guangsen Wang, Khe Chai Sim

School of Computing, National University of Singapore, Republic of Singapore, 117417

ABSTRACT

The data sparsity problem of context-dependent (CD) acoustic modelling of deep neural networks (DNNs) in speech recognition is addressed by using the decision tree state clusters as the training targets. The CD states within a cluster cannot be distinguished during decoding. This problem, referred to as the *clustering* problem, is not explicitly addressed in the current literature. In our previous work, a regression-based CD-DNN framework was proposed to address both the data sparsity and the clustering problems. This paper investigates several refinements for the regression-based CD-DNN including two more representative state approximation schemes and the incorporation of sequential learning. The two approximations are obtained based on the statistics learned from the training data. Sequential learning is applied to both broad phone DNN detectors and the regression NN. The proposed refinements are evaluated on a broadcast news transcription task. For the cross-entropy systems, the two approximations perform consistently better than our previous work. Consistent performance gain over the corresponding cross-entropy trained systems is also observed for both the baseline CD-DNN and the regression model with sequential learning.

Index Terms— Context Dependent Modelling, Deep Neural Network, Logistic Regression, Canonical State Modelling, Articulatory Features, Sequential Learning

1. INTRODUCTION

Context-dependent (CD) acoustic modelling in automatic speech recognition (ASR) raises an important issue of how to reliably handle the large number of CD phones that grows exponentially with the width of the context. In addition, a considerable number of them have limited number of occurrences or even unseen in the training corpus. To address this data sparsity problem, parameter sharing [1, 2, 3] is used. However, this leads to the *clustering* problem where the CD states that share the same parameters will yield the same acoustic scores given the observations.

Over the past few years, the development of machine learning algorithms [4] and General-purpose computing on graphics processing units have made possible the training of Deep Neural Networks (DNNs). To handle the data sparsity problem, decision tree state clusters [1, 3] are often used as CD-DNN training targets [5, 6, 7, 8]. However, the clustering problem is not explicitly addressed.

In our previous work [9], a regression-based CD-DNN modelling approach was proposed. Multiple sets of state clusters are used to represent the canonical states. Each set divides all the CD states into simpler disjoint clusters, which are easier to model, circumventing the data sparsity problem. These clusters are obtained based on the broad phone contexts defined according to the articulatory features. DNNs are used to obtain the posterior probabilities of the broad phone state clusters. The concatenated log posteriors of the DNN detectors form the canonical state space. Logistic regression is then used to transform the canonical states into the final state output probabilities. However, directly training the logistic regression model is difficult due to the large number of distinct CD states, many of which have very limited training data. To address this data sparsity problem, regression parameter tying is performed. Based on some approximations, the regression model can be viewed as a sparse two-layer neural network with dynamically connected weights and its parameters can be trained with the cross-entropy criterion. More interestingly, by carefully designing the broad phone state clusters such that each CD state can be uniquely identified using the canonical state representation, the resulting regression-based CD-DNN is able to model each CD state distinctly, yielding a better context resolution compared to the conventional state clustering approach.

The frame-based cross-entropy criterion is not optimum for sequential classification tasks like speech recognition. Therefore, lattice-based sequential learning of NN/HMM systems was proposed in [10]. It has been shown that (D)NNs with sequential learning perform significantly better than the corresponding cross-entropy systems [11, 12, 13].

In this paper, two refinements are investigated for the regression-based CD-DNN framework [9]. Firstly, the intuition-based approximation in [9] does not aim at optimising the objective function directly, which is not desirable. Therefore, two more representative state approximations are proposed based on the statistics learned from the training data. Secondly, sequential learning is applied to estimate the regression-based CD-DNN parameters.

The remainder of the paper is organised as follows. Section 2 reviews the regression-based CD-DNN framework. Section 3 introduces two approximations to the objective function so that the training procedure is computationally tractable. The incorporation of sequential learning for the regression NN is given in Section 4. Experimental results are presented in Section 5. Section 6 summarises the findings and concludes the paper.

2. REGRESSION-BASED CD-DNN OVERVIEW

To address the clustering problem of the CD-DNNs, an initial investigation was proposed in our previous work [9] by introducing a regression-based CD model for DNNs. Given some regression bases, each CD state can be *uniquely* defined. There are three main components of the regression-based CD-DNN: 1) canonical state vector generation 2) context dependent state vector mapping 3) multi-class logistic regression. We will give a brief review of these components here. The detailed description can be found in [9].

The canonical state vector $\bar{\boldsymbol{b}}_t$ is the concatenated log posteriors of all the DNN detectors given an input feature vector, $\boldsymbol{o}_t$. DNN detectors are used here to predict biphone clusters using different categories of broad phone contexts in table 1.

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

Place of articulation		Production manner		Voicedness		Miscellaneous	
Front Vowel	iy ih eh ae aw ey y	High Vowel	ih iy uh uw	Voiced	iy ih eh ey ae aa aw ay ah ao oy ow uh uw er b d dh g jh l m n ng r v w y z zh	Short Vowel	eh ih uh ae ah y oy
Central Vowel	ah er hh	Mid Vowel	ah eh ey ow er			Long Vowel	iy uw aa
Back Vowel	aa ao uh uw ay ow oy	Low Vowel	aa ae aw ay oy ao			Diphthong	ey aw ow ao
Coronal	d l n s t z r th dh	Fricative	jh ch s sh z f zh th v dh hh			ay	ay
Palatal	sh zh jh ch	Nasal	m n ng			Retroflex	er r
Labial	b f m p v w	Stop Cons	b p t d k g	Unvoiced	p f th t s sh ch k hh	Affricate	ch jh
Velar	g k ng	Approximant	w y l r			Alveolar	s z t d n l
Silence	sil	Silence	sil	Silence	sil	NonContinuent	p b g k
						Continuent	sh th dh hh m f ng v w zh
						Silence	sil

Table 1. Broad phone classes based on place of articulatory (A), production manner (M), voicedness (V) and miscellaneous (O)

For each CD state, s, a state descriptor $\boldsymbol{D}_s$ (see [9]) is used to map the high dimensional canonical state vector $\bar{\boldsymbol{b}}_t$ to a low dimensional CD state vector, $\boldsymbol{V}(s,t)[i] = \log \boldsymbol{b}_{t,i}[\boldsymbol{D}_s[i]]$, where $[i]$ indicates the ith vector element. $\boldsymbol{D}_s$ is an N-dimensional vector whose elements are the state cluster indices for each DNN detector. The design principle of the broad phone classes is to assign each CD state s to a *unique* descriptor $\boldsymbol{D}_s$ that is composed of simpler biphones clusters. Given a triphone state s, the N biphone clusters are obtained by mapping its left and right contexts to the broad phone groups defined in Table 1. The corresponding indices for the biphone clusters are held in its descriptor $\boldsymbol{D}_s$.

Finally, $\boldsymbol{V}(s,t)$ is transformed into the state output probabilities, $P(s|\boldsymbol{o}_t)$, by means of logistic regression:

$$P(s|\boldsymbol{o}_t) = \frac{\exp\left(\boldsymbol{w}_{c(s)}^T \cdot \boldsymbol{V}(s,t)\right)}{\sum_{s' \in S} \exp\left(\boldsymbol{w}_{c(s')}^T \cdot \boldsymbol{V}(s',t)\right)} \tag{1}$$

To address the data sparsity issue, all the triphone states within the same state cluster share the same regression weight $\boldsymbol{w}_c$, where $c(s) \in C$ is the triphone state cluster for state s. Although $\boldsymbol{w}_c$ are shared within a cluster, the $\boldsymbol{V}(s,t)$ term will result in a different state output probability since $\boldsymbol{D}_s$ is unique for each state. Since the denominator of equation 1 is independent of s, it is just a constant bias which can be ignored during decoding. Therefore, the log probability of each state can be easily computed:

$$\log P(s|\boldsymbol{o}_t) \propto \boldsymbol{w}_{c(s)}^T \cdot \boldsymbol{V}(s,t) \tag{2}$$

3. REGRESSION PARAMETER ESTIMATION

Cross-entropy criterion was used in [9] as the objective function to optimise the regression parameters. The cross-entropy between the target state label vector, $y_t(s)$ and the state output probabilities $P(s|\boldsymbol{o}_t)$ predicted by the model is:

$$\mathcal{F}_{\text{XENT}} = -\sum_t^T \sum_{s \in S} y_t(s) \log P(s|\boldsymbol{o}_t) = -\sum_t^T \log P(s_t|\boldsymbol{o}_t)$$

where in the case of hard target labels $y_t(s) = 1$ if $s = s_t$ and $y_t(s) = 0$ otherwise. s_t is the correct state label at time t. Substituting equation 1 into the above objective function yields:

$$\mathcal{F}_{\text{XENT}} = -\sum_t^T \left\{ \boldsymbol{w}_{c(s)}^T \cdot \boldsymbol{V}(s,t) - \log \boldsymbol{Q}_{s_t} \right\} \tag{3}$$

$$\boldsymbol{Q}_{s_t} = \sum_{s' \in S} \exp\left(\boldsymbol{w}_{c(s')}^T \cdot \boldsymbol{V}(s',t)\right) \tag{4}$$

S denotes a set of all the triphone states. It is not feasible to directly optimise $\mathcal{F}_{\text{XENT}}$ because it will be computationally intractable to compute the summation over all the states for every time frame $\boldsymbol{o}_t$ in $\boldsymbol{Q}_{s_t}$. Therefore, instead of computing $\boldsymbol{V}(s,t)$ for all the states, we compute only one state, s_c, for each state cluster c. The rest of the states in that cluster will use the CD state vector of s_c when computing the objective function. s_c can be viewed as a representative state for cluster c. Therefore, the new objective function, $\mathcal{F}'_{\text{XENT}}$ can be obtained by replacing $\boldsymbol{Q}_{\boldsymbol{s_t}}$ with $\boldsymbol{Q}'_{\boldsymbol{s_t}}$:

$$\mathcal{F}'_{\text{XENT}} = -\sum_t^T \left\{ \boldsymbol{w}_{c(s)}^T \cdot \boldsymbol{V}(s,t) - \log \boldsymbol{Q}'_{s_t} \right\} \tag{5}$$

$$\boldsymbol{Q}'_{s_t} = \sum_{c \in C} N_c \exp\left(\boldsymbol{w}_c^T \cdot \boldsymbol{V}(s_c,t)\right) \tag{6}$$

where s_c is the representative state of cluster c. C is the set of state clusters and N_c is the number of states in cluster c. We further constrain that $\boldsymbol{V}(s_t,t)$ for the reference state s_t is computed directly and will not use the representative state approximation. In [9], we proposed to choose the triphone state with the largest number of training frames to represent the state cluster. However, the approximation may not necessarily yield the representative states that are optimum for $\mathcal{F}_{\text{XENT}}$. In the following, two approximation methods that determine the representative states based on the statistics from the training data will be described.

3.1. Frame-varying approximation

In order to achieve a better approximation, the frame-varying approximation method aims at finding the representative states so that minimising $\mathcal{F}'_{\text{XENT}}$ will result in a decrease in $\mathcal{F}_{\text{XENT}}$. This can be achieved by finding the frame-dependent cluster state representatives, $s_c(t)$, such that $\mathcal{F}'_{\text{XENT}} \geq \mathcal{F}_{\text{XENT}}$ or $\boldsymbol{Q}'_{s_t} \geq \boldsymbol{Q}_{s_t}$. This requires $s_c(t)$ satisfy the following constraints:

$$\boldsymbol{w}_c^T \cdot \boldsymbol{V}(s_c(t),t) \geq \boldsymbol{w}_c^T \cdot \boldsymbol{V}(s',t), \qquad \forall t, s' \in c \tag{7}$$

In order to satisfy all the above constraints, equation 7 has to be evaluated for every $\boldsymbol{o}_t$ and triphone states of cluster c or upon every update of the regression weights, $\boldsymbol{w}_c$, rendering it computationally intractable. To reduce the complexity, instead of evaluating 7 for each member triphone state of cluster c, we can choose a subset of c as the candidates to get the representative state:

$$\boldsymbol{w}_c^T \cdot \boldsymbol{V}(s_c(t),t) \geq \boldsymbol{w}_c^T \cdot \boldsymbol{V}(s',t), \qquad \forall t, s' \in \psi_c \tag{8}$$

where ψ_c is a subset of triphone states of cluster c. In addition, ψ_c is assumed to be static (frame independent), which can be obtained before training the regression model. The representative state for cluster c at time t can then be obtained as:

$$s_c(t) = \arg\max_{s'} \boldsymbol{w}_c^T \cdot \boldsymbol{V}(s', t) \qquad \forall s' \in \psi_c \tag{9}$$

It is interesting to note that the state representative is changing per observation $\boldsymbol{o}_t$. Therefore, we referred to the approximation in equation 9 as "frame-varying" approximation. However, even with the approximation, the computational complexity is still quite high for a large number of regression targets since we have to compute equation 8 and 9 for each regression target given each $\boldsymbol{o}_t$. Consequently, we investigate the frame-varying approximation using only the monophone state regression targets.

Now how to get the subset/candidate triphone states ψ_c for each regression target (monophone state)? In this paper, we propose to use decision tree state clusters to define the candidate sets. Assume we can find a representative triphone state for each decision tree state cluster. Given a CI regression target, all the representative states of the corresponding decision tree state clusters are used as the candidate states for the frame-varying approximation. For example, to get the representative candidate set for target "/iy/[2]", $\psi_{/iy/[2]}$, we collect all the decision tree state clusters of "/iy/[2]" and get the corresponding representative states of these clusters. Given each training frame, the representative state for "/iy/[2]" is assumed to be one of these representative states (equation 9).

3.2. Frame-independent approximation

The only remaining problem is how to get the representative state for each decision tree state cluster, which requires another approximation. To this end, we constrain the representative states of the state clusters to be static (frame independent) so that it can be obtained once and reused in subsequent optimisation iterations. Therefore, the approximation is referred to as "frame-independent". Furthermore, by constraining $\boldsymbol{w}_c$ to be positive, a new set of constraints independent of $\boldsymbol{w}_c$ can be obtained as follows:

$$\boldsymbol{V}(s_c, t)[i] \geq \boldsymbol{V}(s', t)[i], \qquad \forall t, i \in [1..N], s' \in c \tag{10}$$

However, it may not be possible to find a static s_c, that can satisfy all the constraints in equation 10. Therefore, we propose finding the representative states that will satisfy *most* of the constraints:

$$s_c = \arg\max_{s' \in c} \mathcal{T}_{s'} \tag{11}$$

where $\mathcal{T}_{s'}$ denotes the number of times s' has the largest value of $\boldsymbol{V}(s', t)[i]$ among all states in $c(s')$, which is calculated as follows:

$$\mathcal{T}_{s'} = \sum_{i=1}^{N} \phi_{c(s'),i}[\boldsymbol{D}_{s'}[i]] \tag{12}$$

$$\phi_{c,i}[j] = \sum_{t=1}^{T} \delta\big(j - I_{i,t}^c\big) \tag{13}$$

$$\approx \sum_{t=1, c=c(s_t)}^{T} \delta\big(j - I_{i,t}^c\big) \tag{14}$$

$$I_{i,t}^c = \boldsymbol{D}_{\hat{s}_{i,t}^c}[i] \tag{15}$$

$$\hat{s}_{i,t}^c = \arg\max_{s' \in c} \boldsymbol{V}(s', t)[i] \tag{16}$$

where, among all the state members of c, $\phi_{c,i}[j]$ is the number of times the jth output of the ith detector has the largest value; $I_{i,t}^c$ denotes the index of the ith detector's output with the largest value at time t; the CD state which has the largest i-th element of $\boldsymbol{V}(s, t)$ is $\hat{s}_{i,t}^c$. $\delta(\cdot)$ is a Kronecker delta function. To reduce the computational complexity, each cluster, c, will only consider the training frames whose reference states (s_t) belong to c , which is the approximation in equation 14. Interestingly, the approximation in equation 11 can be viewed as a special case of the frame-varying approximation if ψ_c contains only one representative state.

4. SEQUENTIAL LEARNING OF REGRESSION NN

In [9], $\mathcal{F}'_{\text{XENT}}$ is trained with a 2-layer sparsely-connected regression NN by viewing $\boldsymbol{w}_s$ as weights and using the representative states as regression targets. As another refinement, we can apply sequential training [10] to the regression NN. In this paper, Minimum Phone Error (MPE) [14] criterion is used. Optimising the MPE criterion directly is difficult for the regression NN. Therefore, we aim at optimising the weak sense auxiliary function $\mathcal{G}_{MPE}$ [14]:

$$\mathcal{G}_{\text{MPE}} = \sum_{s=1}^{S} \sum_{t=1}^{T} \gamma_t^{\text{MPE}}(s) \log P(s|\boldsymbol{o}_t) \tag{17}$$

where $\gamma_t^{\text{MPE}}(s)$ is a "posterior" term defined in [14], which is computed as the differences between the average accuracies of all lattices passing state s at time t and the average accuracy of all the states at time t. Substituting equation 1 into the auxiliary function and applying weight sharing, we have

$$\mathcal{G}_{\text{MPE}} = \sum_{c \in C} \sum_{t=1}^{T} \gamma_t^{\text{MPE}}(s) \Big(\boldsymbol{w}_s^T \cdot \boldsymbol{V}(s, t) - \log \boldsymbol{Q}'_{s_t} \Big)$$

where $\boldsymbol{Q}'_{s_t}$ is defined in equation 6. The gradient of $\mathcal{G}_{\text{MPE}}$ with respect to $a_t(s)$ is also $\gamma_t^{\text{MPE}}(s)$. Recall we do not use state cluster representative for the reference triphone state where $\gamma_t^{\text{MPE}}(s)$ is (mostly) positive. On the other hand, the occupancy is negative for the competing states. To maximise $\mathcal{G}_{\text{MPE}}$, we need to maximise $\boldsymbol{Q}'_{s_t}$ which is the same as the cross-entropy training criterion. Therefore, all the approximations proposed in the last section can be used here naturally to optimise the weak sense auxiliary function. The sequential learning is incorporated by using the lattice-based gradient $\gamma_t^{\text{MPE}}(s)$ in the EBP of the training of the regression NN.

5. EXPERIMENTAL RESULTS

5.1. Experimental Setup and Baseline Systems

We evaluate the proposed refinements for the regression-based CD-DNN on a broadcasting news transcription task using the Topic Detection and Tracking - Phase 3 (TDT3) corpus [15] with 100 hours of English speech. The phone set contains 40 phones including silence. Each phone HMM is modelled with 5 states including 3 emitting states. The features are the standard 39-dimensional PLPs. Each triphone state in the baseline GMM/HMM is modelled with 20 components. The testing set is the F0 portion of the Hub4-97 evaluation set. The language model is trained using the Gigaword English corpus and the TDT3 transcriptions with a 58K vocabulary list. The word error rate is obtained from a bigram full decoding and a trigram lattice rescoring.

DNN is trained using TNet [16]. Up to 5 hidden layers with 2048 hidden units per layer are trained. The training labels are obtained

Table 2. Output dimensions and trigram WERs of DNN Detectors

Detector Type	Output Dimension	XENT	MPE
A(L)-S	939	15.5	11.7
S+A(R)		15.9	11.0
M(L)-S	939	15.7	11.6
S+M(R)		16.4	12.0
V(L)-S	354	15.9	11.6
S+V(R)		16.6	11.8
O(L)-S	1173	15.0	11.5
S+O(R)		15.9	11.2

from the forced alignments using the corresponding GMM/HMMs. The input window size of the DNN input layer is 15 frames, rendering 585 input units. The hybrid DNN/HMM system is set up using Kaldi [17]. The trigram WER for CI DNN is 16.0% while the best CD-DNN with 2303 state clusters has a trigram WER of 14.8%.

The input of the 2-layer regression NN $\bar{\boldsymbol{b}}_t$ is the concatenation of all the log posteriors from the 8 broad phone DNNs with a dimension of 6810. The WER of the 8 broad phone DNNs under both cross-entropy and MPE training is given in table 2. All 8 cross-entropy trained broad phone DNN detectors perform worse than the baseline CD-DNN with 2303 state clusters. This is expected since they have a significantly smaller number of clusters and consider only one side of the triphone context. The incorporation of MPE training provides significant performance gain over the cross-entropy systems. However, these MPE DNN detectors still perform significantly worse than the MPE CD-DNN baseline with a trigram WER of 10.6%.

5.2. Refinements for regression-based CD-DNN

The cluster representative is approximated with the triphone state with the largest number of training examples in [9]. This approximation is referred to state frequency (SF) based approximation . The frame-independent (FI) and frame-varying (FV) approximations are the proposed two refinements. Recall a candidate representative state set is required for the frame-varying approximation. Therefore, we choose a decision tree with 2303 state clusters to define the candidate set. To do this, we obtain a representative for each of the 2303 clusters according to the frame-independent approximation in equation 11. The representative states are then used as the state candidates for the frame-varying approximation with CI state regression targets. We refer to the canonical state space produced by the cross-entropy trained DNN detectors as "XENT canonical space" and the one produced by MPE trained DNN detectors as "MPE canonical space". Based on the XENT canonical space, the regression NN is only trained with cross-entropy criterion for all three approximations. For the MPE canonical space, both cross-entropy and MPE criteria are used. Note for the MPE regression model, 4 iterations of MPE training are performed based on the best cross-entropy system. The trigram WERs of these configurations are given in table 3, where all the regression NNs have only CI states as regression targets.

For the XENT canonical state space, both the frame-independent and frame-varying approximations outperform the state frequency approximation in [9]. The frame-independent approximation performs better than the frame-varying approximation. For the MPE canonical space, even with the cross-entropy trained regression NNs, significant improvements are obtained for all the approximations. This shows the advantages of the MPE canonical space over the XENT space. In addition, consistent performance gain has also been observed for all the approximations with MPE trained regression NN compared to the corresponding cross-entropy systems. The state frequency based approximation has the comparable performance to the frame-varying approximation. Interestingly, the best performance of the regression model on the MPE canonical space is obtained with frame-varying approximation for both the XENT and MPE trained regression NN instead of the frame-independent approximation.

Table 3. WER comparison of different representative state approximation methods

Training criterion		WER		
Detectors	Regression NN	SF	FI	FV
XENT	XENT	12.7	11.7	12.0
MPE	XENT	10.5	10.6	10.3
	MPE	10.4	10.5	9.8

To explain this, we need to investigate how the approximations are used for both cross-entropy and MPE criteria. Cross-entropy is a frame-based learning criterion, where the objective function is optimised per frame and each frame is independent of each other. Recall the frame-independent approximation is obtained to optimise the cross-entropy criterion directly in equation 11. On the other hand, the frame-varying approximation relies on the frame-independent approximations to *indirectly* optimise the cross-entropy objective function. Therefore, the frame-independent approximation may be more suitable for the cross-entropy criterion and the XENT canonical state space. For the MPE canonical state space, the frame-independent approximation deviates more with the MPE criterion since MPE is a sequential criterion which explores the relations among frames. However, for frame-varying approximation, as it has the flexibility of approximating a representative state for each frame, the mismatch between the cross-entropy criterion which is used to obtain the approximations and the MPE criterion can be reduced by dynamically choosing the representative state for each frame. Therefore, it may be more consistent with the MPE canonical space thus has consistently better performance than frame-independent approximation. In addition, the best performance (9.8%) is significantly better than the baseline MPE trained CD-DNN (10.6%) with a p-value of 0.001 reported using SCTK [18].

6. CONCLUSION

In this paper, several refinements were applied to the previously proposed regression-based CD-DNN framework, including two representative state approximations as well as the incorporation of sequential learning. The proper approximation of the upper bound for the objective function is essential for the success of the regression model. Two approximations were investigated based on the statistics learned from the training data including a frame-independent approximation and a frame-varying approximation. The incorporation of the sequential learning of the regression NN was also studied under both approximations. The refinements are evaluated on a broadcast news transcription task. The two new approximations perform consistently better than the one in our previous work. Significant performance gain is observed using the MPE criterion for both baseline CD-DNN and the regression-based CD-DNN. The best MPE trained regression-based CD-DNN with the frame-varying approximation performs significantly better than the MPE trained baseline CD-DNN.

7. ACKNOWLEDGMENT

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

8. REFERENCES

[1] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *HLT*, 1994, pp. 307–312.

[2] Jerome R. Bellegarda and David Nahamoo, "Tied mixture continuous parameter modeling for speech recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 38, no. 12, pp. 2033–2045, 1990.

[3] Guangsen Wang and Khe Chai Sim, "An investigation of tied-mixture GMM-based triphone state clustering," in *ICASSP*, 2012, pp. 4717–4720.

[4] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 2006, 2006.

[5] George E. Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Trans. on ASLP*, vol. 20, pp. 30–42, 2012.

[6] Frank Seide, Gang Li, Xie Chen, and Dong Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *ASRU*, 2011, pp. 24–29.

[7] Tara N. Sainath, Brian Kingsbury, Bhuvana Ramabhadran, Petr Fousek, Petr Novák, and Abdel rahman Mohamed, "Making deep belief networks effective for large vocabulary continuous speech recognition," in *ASRU*, 2011, pp. 30–35.

[8] 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," *IEEE Signal Processing Magazine*, vol. 29, pp. 82–97.

[9] Guangsen Wang and Khe Chai Sim, "Context-dependent modelling of deep neural network using logistic regression," in *to appear IEEE 2013 Workshop on Automatic Speech Recognition and Understanding*, 2013.

[10] Brian Kingsbury, "Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling," in *IEEE., ICASSP*, 2009, pp. 3761–3764.

[11] Guangsen Wang and Khe Chai Sim, "Sequential classification criteria for NNs in automatic speech recognition," in *Interspeech*, 2011, pp. 441–444.

[12] Brian Kingsbury, Tara N. Sainath, and Hagen Soltau, "Scalable minimum bayes risk training of deep neural network acoustic models using distributed hessian-free optimization," in *INTERSPEECH*, 2012.

[13] Karel Vesely, Arnab Ghoshal, Lukas Burget, and Daniel Povey, "Sequential-discriminative training of deep neural networks," in *Proceedings of Interspeech 2013*, 2013.

[14] D. Povey, *Discriminative training for large vocabulary speech recognition*, PhD thesis, Cambridge University, 2004.

[15] David Graff, Chris Cieri, Stephanie Strassel, and Nii Martey, "The TDT-3 Text And Speech Corpus," in *in Proceedings of DARPA Broadcast News Workshop*. 1999, pp. 57–60, Morgan Kaufmann.

[16] "Neural Network Trainer TNet, online at http://speech.fit.vutbr.cz/software/neural-network-trainer-tnet," .

[17] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[18] "Speech Recognition Scoring Toolkit, http://itl.nist.gov/iad/mig/tools/," .