INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Non-Uniform Boosted MCE Training of Deep Neural Networks for Keyword Spotting

Zhong Meng, Biing-Hwang (Fred) Juang

School of Electrical and Computer Engnineering, Georgia Institute of Technology
75 5th Street NW, Atlanta, GA 30308, USA
zhongmeng@gatech.edu, juang@ece.gatech.edu

Abstract

Keyword spotting can be formulated as a non-uniform error automatic speech recognition (ASR) problem. It has been demonstrated [1] that this new formulation with the non-uniform MCE training technique can lead to improved system performance in keyword spotting applications. In this paper, we demonstrate that deep neural networks (DNNs) can be successfully trained on the non-uniform minimum classification error (MCE) criterion which weighs the errors on keywords much more significantly than those on non-keywords in an ASR task. The integration with a DNN-HMM system enables modeling of multi-frame distributions, which conventional systems find difficult to accomplish. To further improve the performance, more confusable data is generated by boosting the likelihood of the sentences that have more errors. The keyword spotting system is implemented within a weighted finite state transducer (WFST) framework and the DNN is optimized using standard backpropagation and stochastic gradient decent. We evaluate the performance of the proposed framework on a large vocabulary spontaneous conversational telephone speech dataset (Switchboard-1 Release 2). The proposed approach achieves an absolute figure of merit improvement of 3.65% over the baseline system.

Index Terms: discriminative training, deep neural networks, minimum classification error, non-uniform criterion, keyword spotting

1. Introduction

Large vocabulary continuous speech recognition (LVCSR) has achieved extraordinary performance when the speech is read or dictated. For instance, a word accuracy higher than 90% can be expected on the Wall Street Journal task. However, this performance decreases tremendously on a spontaneous conversational speech recognition task [2] as it consists of a stream of words with no overt lexical marking of punctuations and disfluencies (i.e, filled pauses, repetitions, repairs and false starts) may occur frequently in a natural conversation [3]. However, in real applications, it is more important to semantically understand a spontaneous speech rather than to recognize its word transcription. Moreover, the semantic meaning generally resides in a set of keywords in the spoken utterances. Therefore, keyword spotting techniques become crucial for spontaneous conversational speech recognition tasks.

Many techniques have been proposed for the keyword spotting task. In [4], an optimum dynamic programing (DP) based time-normalization algorithm is proposed for spoken word recognition. In 1990s, a hidden Markov model (HMM) based keyword spotting system is proposed within the framework of hypothesis testing [5]. In [6], a set of hypothesized word transcriptions are first generated by the LVCSR decoder and the keywords are then detected and verified. Although good performance is achieved, the two stages in this approach are isolated and optimized based on different criteria. To circumvent this problem, the keyword spotting is formulated as a non-uniform error LVCSR task and the method of *non-uniform minimum classification error (MCE)* is proposed in [1]. In conventional LVCSR, discriminative training (DT) is applied to refine the models with the objective of minimizing the recognition errors without any emphasis on the keywords. However, with non-uniform error LVCSR, the non-uniform error cost is embedded in the DT process to minimize the errors of some words (i.e., keywords) out of all possible words in the vocabulary. This idea is implemented efficiently in the weighted finite state transducer (WFST) framework and has shown some improvement over the baseline system. Moreover, this work is built upon a GMM-HMM system where GMMs are used to model the probability distribution of input features that are associated with each state of HMM. With an adequate number of mixture components, GMMs are able to accurately model any kind of distribution. The parameters of GMMs can be fine-tuned discriminatively to minimize the non-uniform MCE objective specially designed for keyword spotting.

However, GMMs with diagonal covariance matrices are not good at handling highly correlated frames and the concatenation of neighboring frames will inevitably bring about the curse of dimensionality issue during model training. Recently, deep neural networks (DNN) with multiple hidden layers are trained to model the multi-frame distributions over *senones* (tied triphone states) as its output and have achieved remarkable performance improvement on almost all challenging LVCSR tasks [2, 7]. The resulting DNNs learn a hierarchy of nonlinear intermediate representations at the middle layers that capture the complex statistical characteristics in data and the multilayer perceptron (MLP) at the upper layers classifies the intermediate representation to different senones. These intermediate representations are first generated through the generative pre-training of a stack of restricted Boltzmann machine (RBM) and are then discriminatively fine-tuned to predict the senones with a certain objective through backpropagation. By using densely connected DNN for acoustic modeling, the high correlations between frames can be well extracted and reflected in the intermediate representation and the distribution of a concatenation of several consecutive speech frames within a long context window can be robustly modeled [8].

The authors would like to thank Chao Weng at Georgia Institute of Technology for his help on MCE implementation.

http://dx.doi.org/10.21437/Interspeech.2016-642

Therefore, we propose the *non-uniform MCE training of DNN* for keyword spotting, in which a DNN is discriminatively trained to minimize the empirical error cost. The backpropagation error based on non-uniform MCE is derived for updating the parameters in DNNs. When applying this to LVCSR, a sequence of decoded words will be produced similar to the usual word error rate (WER) based LVCSR, excepted that the keywords will have fewer recognition errors. To further improve the performance, we generate more data from the more confusable hypothesized word sequences by boosting the likelihood of hypothesized word sequences proportional to their difference from the label transcription. Therefore, *non-uniform boosted MCE (BMCE) training of DNN* is proposed by integrating this data augmentation strategy. Experiments are conducted a large-scale spontaneous conversational telephone speech (CTS) dataset. The proposed method has achieved 3.65% absolute figure of merit (FOM) gain over the baseline system using cross entropy as the objective on "Credit Card Use" topic of Switchboard-1 Release 2.

In Section 2, we discuss how the non-uniform BMCE criterion is embedded into the DNN training for keyword spotting. In Section 3, we show how the non-uniform BMCE is implemented in the WFST framework. In Section 4, experimental results on Switchboard dataset are shown and discussed. We draw our conclusion in Section 5.

2. Non-Uniform BMCE Training of DNN for Keyword Spotting

Conventionally, DNNs are trained to model the distribution of the senones based on a cross-entropy criterion in LVCSR tasks. A senone-level alignment on the training set is used as the labels for training the DNN. However, the DNNs trained through distribution estimation do not necessarily lead to the minimization of the recognition error rate. In [9], maximum mutual information (MMI) [10, 11], minimum phone error (MPE) [12, 13], state-level minimum Bayes risk (sMBR) [14, 15, 16] and boosted MMI [17] are used as the objective for DNN training. Although these discriminative traing methods are able to improve the performance over the traditional cross-entropy based methods, they do not directly minimize an objective function which is related to the recognition error rate. To circumvent this problem, MCE was proposed to directly minimize the empirical error rate and is widely used in GMM-based LVCSR systems.

Keyword spotting can be formulated as a LVCSR task in which some recognition units (i.e., keywords) are more significant than others. More specifically, the LVCSR designed for keyword spotting should be able to generate a decoded word sequence in which keywords have fewer recognition errors than the normal LVCSR system. To satisfy this requirement, we introduce the non-uniform MCE objective for the training of DNNs in the LVCSR task. Instead of minimizing the empirical error rate for conventional MCE, the non-uniform MCE training of DNN is aimed at minimizing the empirical error cost. This can be realized by embedding the non-uniform error cost function into the MCE objective on the frame level to emphasize both the miss detection errors and the false alarm errors on the keywords. Strictly speaking, the error cost function should be individually assigned to each pair of words in the vocabulary to take care of all kinds of recognition errors [18]. Our formulation is a simplified version of the general non-uniform MCE for fast and easy implementation.

A two-stage training approach based on the standard error backpropagation procedure is applied to optimize the non-uniform MCE objective. In the first stage, the gradients of the non-uniform MCE objective with respect to the activations at the output layer are calculated and then backpropagated to derive the gradients for all the parameters of the DNN in the second stage. We will derive this important gradient below.

Assume that the training data is given by training utterances $r = \{1, \ldots, R\}$. $X_r = \{x_{r1}, \ldots, x_{rT_r}\}$ is the sequence of observations for utterance r, W_r is the word sequence in the reference (label transcription) for utterance r. W is one of all the word sequences in the decoded speech lattice for utterance r. $S_W = \{s_{W1}, \ldots, s_{WT}\}$ is the senone sequence corresponding to W, where s_{Wt} is the senone which frame x_{rt} is aligned with.

The output of the DNN for senone s is the posterior probability $p(s|x_{rt})$ obtained by a softmax function.

$$p(s|x_{rt}) = \frac{\exp[a_{rt}(s)]}{\Sigma_{s'} \exp[a_{rt}(s')]} \tag{1}$$

where $a_{rt}(s)$ is the activation for senone s at the output layer. The pseudo log-likelihood of observation x_{rt} given senone s is

$$\log p(x_{rt}|s) = \log p(s|x_{rt}) - \log p(s) + \log p(x_{rt}) \tag{2}$$

where $p(s)$ is the prior probability of senone s estimated from the training set and $p(x_{rt})$ is the probability of observation x_{rt} which is independent of the word sequence and can be ignored.

The frame-level discriminative function for W and misclassification measure are given by

$$g(x_{rt}, s_{Wt}; \Lambda) = \log[p(x_{rt}|s_{Wt})^{\kappa} p(s_{Wt})] \tag{3}$$

where $p(x_{rt}|s_{Wt})$ and $p(s_{Wt})$ denote the acoustic and language models respectively, κ is the acoustic model scaling factor and Λ is a set of model parameters.

$$d(x_{rt}; \Lambda) = -g(x_{rt}, s_{W_r t}; \Lambda) + \log \left\{ \frac{1}{N-1} \sum_{W \neq W_r} \exp[g(x_{rt}, s_{Wt}; \Lambda)\eta] \right\}^{\frac{1}{\eta}} \tag{4}$$

where N is the total number of hypothesized word sequences. By varying the positive number η, the significance of the competing classes can be adjusted.

By embedding the misclassification measure Eq. (4) into a sigmoid function for smoothing, the objective function of the non-uniform MCE training of DNN is given by

$$\mathcal{L}_{NUMCE}(\Lambda) = \sum_{r=1}^{R} \sum_{t=1}^{T_r} \epsilon_r(t) l(d(x_{rt}; \Lambda)) \tag{5}$$

where $\epsilon_r(t)$ is the error cost function at the frame level, $l(\cdot)$ is the sigmoid which takes the form

$$l(d) = \frac{1}{1 + \exp(-\alpha d + \beta)} \tag{6}$$

The slope of the sigmoid curve can be adjusted by α and β is normally set to 0. The objective function in Eq. (5) is essentially a smoothed approximation of the *empirical error cost*. Note that when the error cost function is fixed to 1 for all t (i.e., $\epsilon_r(t) = 1$), Eq. (5) degrades to the objective function of MCE, which

is a smoothed approximation of the *empirical error rate* on the training set.

The derivative of Eq. (5) with respect to the activation $a_{rt}(s)$ at the output layer is

$$\begin{aligned}\frac{\partial \mathcal{L}_{NUMCE}(\Lambda)}{\partial a_{rt}(s)} &= \sum_q \frac{\partial \mathcal{L}_{NUMCE}(\Lambda)}{\partial \log p(x_{rt}|q)} \frac{\partial \log p(x_{rt}|q)}{\partial a_{rt}(s)} \\ &= \alpha \epsilon_r(t) l\left(d(x_{rt};\Lambda)\right)\left[1 - l\left(d(x_{rt};\Lambda)\right)\right] \\ &\quad \kappa\left[\delta_{s_{W_rt}:s} - \gamma_{rt}^{W\neq W_r}(s)\right] \end{aligned} \tag{7}$$

where $\gamma_{rt}^{W\neq W_r}(s)$ is the posterior of being in senone s at time t, computed over the denominator lattice of the utterance r, and the lattice of utterance r excluding the path corresponding to the word sequence W_r, $\log p(x_{rt}|q)$ is the log-likelihood of x_{rt} given senone q, and $\delta_{s_{Wt}:s}$ is the Kronecker delta function defined as

$$\delta_{s_{Wt}:s} = \begin{cases} 1, & s_{Wt} = s \\ 0, & s_{Wt} \neq s \end{cases} \tag{8}$$

For easy implementation, $d(X_{rt};\Lambda)$ is used as an approximation of $d(x_{rt};\Lambda)$. Eq. (7) is the error to be backpropagated to derive the gradients for all the parameters of DNN.

To minimize the recognition errors on the keywords, the error cost function $\epsilon_r(t)$ should be designed in such a way that all the recognition error cost associated with the keywords are emphasized. More specifically, the initial $\epsilon_r(t)$ for the frames labeled as keywords in the label transcription (denoted by K_1) should be greater than 1 to reduce the miss detection errors. Also the initial $\epsilon_r(t)$ for the frames aligned with keywords on the hypothesized word sequences (denoted by K_2) other than the label transcription should be greater than 1 to prevent the false alarm errors. The $\epsilon_r(t)$ for the frames aligned with non-keywords in all the word sequences in the decoded speech lattice for utterance r should be 1. The error cost function can be adjusted adaptively through iterations using a AdaBoost-like scheme as is proposed in [19]. We multiply $\epsilon_r(t)$ with a decay factor β if the frame x_{rt} is correctly classified at the current training iteration.

To achieve better performance for keyword spotting, we boost the likelihood of the hypothesized word sequences that have a higher phone error relative to the label transcription, which is equivalent to generating more data from the more confusable hypothesized word sequences. For non-uniform boosted MCE, the misclassification measure can be re-written as

$$\begin{aligned} d(x_{rt};\Lambda) &= -g(x_{rt}, s_{W_rt};\Lambda) \\ &+ \log\left\{\frac{1}{N-1}\sum_{W\neq W_r}\exp\{g(x_{rt}, s_{Wt};\Lambda) - bA(p_{Wt}, p_{W_rt})]\eta\}\right\}^{\frac{1}{\eta}} \end{aligned} \tag{9}$$

where b is the boosting factor and $A(p_{Wt}, p_{W_rt})$ is the frame-level raw phone accuracy of a sentence W given the label transcription W_r, i.e.,

$$A(p_{Wt}, p_{W_rt}) = \begin{cases} 1, & p_{Wt} = p_{W_rt} \\ 0, & p_{Wt} \neq p_{W_rt} \end{cases} \tag{10}$$

where p_{Wt} is the raw phone which frame x_{rt} is aligned with and $P_W = \{p_{W1}, \ldots, p_{WT}\}$ is the phone sequence corresponding to word sequence W.

The objective function and backpropagation error of non-uniform BMCE can be derived correspondingly.

3. Implementation of Non-Uniform BMCE in the WFST Framework

The non-uniform MCE is implemented within the WFST framework. As is mentioned in [20], a decoded lattice of an utterance is generated by a beam pruning on the full searching graph which is a composition of the WFST U and the HCLG graph. U, H, C, L, G encode the acoustic score of the utterance, HMM structure, phonetic context-dependency, lexicon and grammar respectively. The decoded lattice is a compact representation of the hypothesis space for the utterance. The lattice is converted to a compact version for higher efficiency.

For an utterance r, the competing hypothesis for non-uniform MCE training has to exclude the label transcription W_r as is shown in Eq. (4). This is accomplished by taking the *difference* operation of WFST. Assuming that $L_r(W)$ is the compact lattice for utterance r and $WFST(W_r)$ is the compiled WFST for W_r, the lattice representing the competing hypothesis in non-uniform MCE training is given by

$$L_r^{NUMCE} = L_r(W) - WFST(W_r) \tag{11}$$

The posterior $\gamma_{rt}^{W\neq W_r}(s)$ in Eq. (7) can be obtained by performing forward-backward on L_r^{NUMCE}.

In WFST framework, non-uniform BMCE training of DNN can be easily implemented based on non-uniform MCE. The extra computation involved is to subtract b times the frame-level raw phone accuracy $A(p_{Wt}, p_{W_r})$ from the scaled acoustic log-likelihood on each arc at time t in the lattice while performing forward-backward on L_r^{NUMCE}. This can be viewed as a modification of the contribution from language model on each arc.

4. Experiments

4.1. Dataset Description

We evaluate the performance of the proposed framework on a large-scale CTS task, i.e., the 110 hours Switchboard-1 Release 2 (LDC97S62). It consists of 2348 two-sided telephone conversations from 543 speakers (302 males and 241 females) in the United States. One topic is assigned to each of the conversation between two callers and about 70 topics in total are provided in the corpus.

For the keyword spotting task, the conversations on the topic of "Credit Card Use" (including 5649 utterances) are used as the test set and around 100k utterances selected from the rest of the Switchboard corpus form a training set with about 110 hours of speech. 18 keywords are selected for the spotting evaluation, which are BANK, CARD, CASH, CHARGE, CHECK, MONTH, ACCOUNT, BALANCE, CREDIT, DOLLAR, HUNDRED, LIMIT, MONEY, PERCENT, TWENTY, VISA, DISCOVER, INTEREST. For both tasks, the Mississippi State transcripts and the 30K-word lexicon released with those transcripts are used. The lexicon contains pronunciations for all words and word fragments in the training data.

4.2. Baseline System

The baseline ASR system is built with Kaldi Speech Recognition Toolkit [21]. Initially, a GMM-HMM system is built. Each cross-word triphone is modeled by a 3-state left-to-right GMM-HMM (a 5-state HMM for silence). The GMM-HMMs are ini-

tially trained with the 110 hour training data using maximum-likelihood (ML) criterion and then refined with MMI and MCE criteria. The input features are obtained as follows. First, 9 frames (4 on each side of the current frame) of 13-dimensional Mel-frequency cepstral coefficient (MFCCs) are spliced together and projected down to 40 dimensions using linear discriminant analysis (LDA). Then a single semi-tied covariance (STC) transform is performed on the features obtained by LDA. Then speaker adaptive training is performed using a single feature-space maximum likelihood linear regression (FMLLR) transform estimated for each speaker. The resulting feature after FMLLR is called LDA+STC+FMLLR feature and is used for training the GMM-HMMs. The trigram language model (LM) is trained on 3M words of the training transcripts.

Then we bulid a DNN-HMM system based on the GMM-HMM system. Initially, we pre-train a deep belief network (DBN) containing stacked restricted Boltzmann machines (RBMs) that are trained generatively in a layerwise fashion. Then we create a DBN-DNN by adding a softmax output layer that contains one unit for each possible senone of each HMM. The DBN-DNN is fine-tuned to train a DNN with cross-entropy objective using stochastic gradient decent (SGD) (initial learning rate 0.008). The input to the DNN is an 11 frame (5 frames on each side of the current frame) context window of the 40 dimensional LDA+STC+FMLLR features globally normalized to have zero mean and unit variance. The resulting baseline DNNs has 7 layers (including 6 hidden layers), where each hidden layer has 2048 neurons, and 4280 output units.

4.3. Results for Keyword Spotting

The baseline DNN is trained with non-uniform BMCE objective for keyword spotting. A set of alignments and lattice generated by decoding the training data using the DNN previously trained with cross-entropy are used to produce the error for backpropagation. We compare the WER performance of the non-uniform BMCE based keyword spotting system with other state-of-arts DT methods (e.g., MMI, sMBR) on Switchbaord 1 data set in Table 1. The non-uniform BMCE trained DNN system achieves the best performance with an FOM of 79.61% at $K_1 = 5.0$, $K_2 = 5.0$, $\beta = 0.1$ and provides 3.65%, 1.73%, 1.04% and 0.74% absolute FOM improvements over baseline system, DNN MMI system, DNN sMBR system and DNN BMCE system, respectively. The best FOM is achieved when the learning rate is 0.0001, the slope of sigmoid α is 0.002 and the boosting factor b is set at 0.07.

To investigate the impact of the initial values of the error cost function on the keyword spotting performance, we show the FOMs for different set of initial values K_1, K_2 and decaying factors β (Here, we only focus on the balanced case where K_1 and K_2 are equal.). Note that when $K_1 = K_2 = 1$, the system is equivalent to a BMCE system which minimizes the recognition error with no emphasis on the keywords. As a general trend, the FOM first increases rapidly and then fluctuates around a certain level as K_1 and K_2 continues to grow. We also plot the FOM values with respect to K_1 and K_2 for different values of β in Fig. 1. The figure shows that the FOM curve for $\beta = 0.1$ is the highest and the FOM curve for $\beta = 0.5$ is the lowest when $K_1 = K_2 \leq 5$. However, when $K_1 = K_2 \geq 6$, the FOM curve for $\beta = 0.1$ becomes the lowest and FOM curve for $\beta = 0.5$ becomes the highest. This implies that a larger (smaller) value of initial error cost function K_1, K_2 requires higher (lower) decay factor β to achieve high FOM. We also notice that the fluctuation of FOM is the highest when $\beta = 0.1$ and is mitigated when decaying factor β is set to 0.3 and 0.5.

System	K_1	K_2	Decay	FOM (%)
GMM ML	-	-	-	68.91
GMM MCE	1	1	-	70.78
GMM MMI	-	-	-	70.60
DNN (baseline)	-	-	-	75.96
DNN MMI	-	-	-	77.88
DNN sMBR	-	-	-	78.57
DNN MCE	1	1	-	78.39
DNN BMCE	1	1	-	78.87
DNN Non-Uniform BMCE	2.0	2.0	0.1	79.36
	2.0	2.0	0.3	79.36
	2.0	2.0	0.5	79.36
	3.0	3.0	0.1	79.56
	3.0	3.0	0.3	79.56
	3.0	3.0	0.5	79.47
	4.0	4.0	0.1	79.46
	4.0	4.0	0.3	79.46
	4.0	4.0	0.5	79.29
	5.0	5.0	0.1	**79.61**
	5.0	5.0	0.3	79.36
	5.0	5.0	0.5	79.30
	6.0	6.0	0.1	79.07
	6.0	6.0	0.3	79.37
	6.0	6.0	0.5	79.46
	7.0	7.0	0.1	79.44
	7.0	7.0	0.3	79.47
	7.0	7.0	0.5	79.43
	8.0	8.0	0.1	79.19
	8.0	8.0	0.3	79.28
	8.0	8.0	0.5	79.39

Table 1: The FOM results of the GMM-HMM and DNN-HMM systems trained with different objectives for keyword spotting on Credit Card Use subset of Switchboard-1 Release 2.

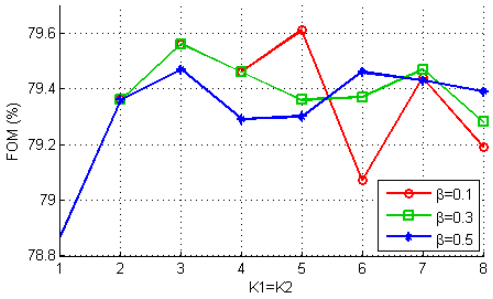

Figure 1: FOM of non-uniform BMCE system with respect to K_1 and K_2 for different decay factors.

5. Conclusions

In this paper, we show that DNNs can be dicriminatively trained using non-uniform BMCE criterion for keyword spotting. The proposed system is implemented in a WFST framework. Experiments are conducted on Switchboard-1 Release 2 dataset. DNN trained with non-uniform BMCE achieves 3.65% and 1.04% absolute FOM gain over the baseline system and DNN sMBR system. system,

6. References

[1] C. Weng and B.-H. Juang, "Discriminative training using non-uniform criteria for keyword spotting on spontaneous speech," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 2, pp. 300–312, Feb 2015.

[2] 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: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[3] E. Shriberg, "Spontaneous speech: How people really talk and why engineers should care," in *in Proc. European Conf. on Speech Communication and Technology (Eurospeech*, 2005, pp. 1781–1784.

[4] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 1, pp. 43–49, Feb 1978.

[5] M. Rahim, C.-H. Lee, and B.-H. Juang, "Discriminative utterance verification for connected digits recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 5, no. 3, pp. 266–277, May 1997.

[6] R. Rose, "Keyword detection in conversational speech utterances using hidden markov model based continuous speech recognition," *Computer Speech & Language*, vol. 9, no. 4, pp. 309 – 333, 1995. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0885230885700150

[7] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, Jan 2012.

[8] J. Pan, C. Liu, Z. Wang, Y. Hu, and H. Jiang, "Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: Why dnn surpasses gmms in acoustic modeling," in *Chinese Spoken Language Processing (ISCSLP), 2012 8th International Symposium on*, Dec 2012, pp. 301–305.

[9] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks." in *INTERSPEECH*, 2013, pp. 2345–2349.

[10] L. Bahl, P. Brown, P. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden markov model parameters for speech recognition," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86.*, vol. 11, Apr 1986, pp. 49–52.

[11] V. Valtchev, J. Odell, P. Woodland, and S. Young, "MMIE training of large vocabulary recognition systems," *Speech Communication*, vol. 22, no. 4, pp. 303 – 314, 1997. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639397000290

[12] D. Povey and P. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, May 2002, pp. I–105–I–108.

[13] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, University of Cambridge, 2005.

[14] M. Gibson and T. Hain, "Hypothesis spaces for minimum bayes risk training in large vocabulary speech recognition." in *INTERSPEECH*. Citeseer, 2006.

[15] J. Kaiser, B. Horvat, and Z. Kacic, "A novel loss function for the overall risk criterion based discriminative training of hmm models," in *Sixth International Conference on Spoken Language Processing*, 2000.

[16] D. Povey and B. Kingsbury, "Evaluation of proposed modifications to mpe for large scale discriminative training," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, April 2007, pp. IV–321–IV–324.

[17] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted mmi for model and feature-space discriminative training," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, March 2008, pp. 4057–4060.

[18] Q. Fu, Y. Zhao, and B.-H. Juang, "Automatic speech recognition based on non-uniform error criteria," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 3, pp. 780–793, March 2012.

[19] C. Weng and B.-H. Juang, "Adaptive boosted non-uniform mce for keyword spotting on spontaneous speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 6960–6964.

[20] D. Povey, M. Hannemann, G. Boulianne, L. Burget, A. Ghoshal, M. Janda, M. Karafiat, S. Kombrink, P. Motlicek, Y. Qian, K. Riedhammer, K. Vesely, and N. T. Vu, "Generating exact lattices in the wfst framework," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 4213–4216.

[21] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.