INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Non-Uniform MCE Training of Deep Long Short-Term Memory Recurrent 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

It has been shown in [1, 2] that improved performance can be achieved by formulating the keyword spotting as a non-uniform error automatic speech recognition problem. In this work, we discriminatively train a deep bidirectional long short-term memory (BLSTM) - hidden Markov model (HMM) based acoustic model with non-uniform boosted minimum classification error (BMCE) criterion which imposes more significant error cost on the keywords than those on the non-keywords. By introducing the BLSTM, the context information in both the past and the future are stored and updated to predict the desired output and the long-term dependencies within the speech signal are well captured. With non-uniform BMCE objective, the BLSTM is trained so that the recognition errors related to the keywords are remarkably reduced. The BLSTM is optimized using back-propagation through time and stochastic gradient descent. The keyword spotting system is implemented within weighted finite state transducer framework. The proposed method achieves 5.49% and 7.37% absolute figure-of-merit improvements respectively over the BLSTM and the feedforward deep neural network baseline systems trained with cross-entropy criterion for the keyword spotting task on Switchboard-1 Release 2 dataset.

Index Terms: automatic speech recognition, keyword spotting, long short-term memory, recurrent neural networks, acoustic modeling, discriminative training

1. Introduction

Automatic speech recognition (ASR), with the purpose of generating accurate word-level transcription, is a crucial step towards robust speech understanding. However, ASR is faced with great challenges when the size of the vocabulary is large and the speaking style is flexible. The frequent occurrences of words streams with no overt lexical marking of punctuations and disfluencies (i.e, filled pauses, repetitions, repairs and false starts) in a natural conversation drastically degrade the performance of ASR system on spontaneous conversational speech. Fortunately, the main idea of the spontaneous conversational speech can be well understood by accurately recognizing a set of keywords that are *semantically important*. Therefore, keyword spotting is an important technique for the accurate understanding of spontaneous conversational speech.

One conventional approach for keyword spotting is the keyword/filler hidden Markov model (HMM) within the framework of hypothesis testing [3, 4]. Recently, a query-by-example approach is successfully used to detect user-specified keywords by comparing the neural networks learned representation of keyword with that of the test audio segment [5, 6]. It is well suited for on-device application with small memory footprint and low computational cost solutions. In [7, 8, 9], a set of hypothesized word transcriptions are first generated by the large vocabulary speech recognition (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. In [2, 1], keyword spotting is formulated as a 1-step *non-uniform error* LVCSR instead of a 2-step traditional approach of uniform error speech-to-text conversion plus text-based search. With the non-uniform error cost function as the discriminative training objective, the acoustic model is optimized so that the errors of some words (i.e., keywords) out of all possible words in the vocabulary are minimized. The system in [1] is built upon the deep neural networks (DNN)-HMM, where DNN are used to model the multi-frame input feature distribution over *senones* (tied triphone states) as its output.

However, DNNs can only make use of limited context information by taking a fixed-size window of speech frames as the input to make the prediction. Although they successfully model the high correlations between frames within a fixed and short time interval, they fail to capture the long-term dependencies within the entire speech signal and are not able to handle dynamic speaking rates. By using recurrent nueral networks (RNN), the network activations of the previous time step are fed as the input of the network to assist in making predictions at the current time step. The cycles in a RNN allows it to store and update the context information about the past inputs in its internal state for an amount of time that is not fixed a priori, but rather depends on its weights and on the input data [10]. Therefore, RNNs are able to exploit a dynamically changing contextual window over the input sequence rather than a static one as in the fixed-sized window used with DNN. The long short-term memory (LSTM) network [11] is a kind of RNN specially designed for capturing *long-term* temporal context information. It overcomes the diminishing gradient problem that comes along the RNN training with a special gating mechanism to control the information to be added or removed to the internal cell state. To also exploit future context information to assist in making current prediction, bidirectional LSTM (BLSTM) networks [12] are introduced to process the input sequence in both directions with two separate hidden layers which are then fed forward together to the same output layer.

Therefore, we propose the *non-uniform boosted minimum classification error (BMCE) training of deep BLSTM* acoustic model for keyword spotting in spontaneous conversational speech. We define the empirical error cost for non-uniform MCE and derive the backpropagation error for the BLSTM. The BLSTM is optimized using backpropagation through time and stochastic gradient descent. With the non-uniform BMCE trained BLSTM acoustic model, the LVCSR decoder is able to generate word transcription with significantly reduced recognition errors on the keywords. To further improve the performance, we boost the likelihood of hypothesized word se-

http://dx.doi.org/10.21437/Interspeech.2017-583

quences proportional to their difference from the label transcription. The experiments are performed on Switchboard-1 Release 2 dataset, which is large-scale spontaneous conversational telephone speech (CTS) dataset. The proposed method achieves 5.49% and 7.37% absolute figure-of-merit (FOM) improvements respectively over the BLSTM and DNN baseline systems trained with cross-entropy criterion for the keyword spotting task on "Credit Card Use" topic of Switchboard-1 Release 2 dataset.

2. Deep BLSTM acoustic model

The LSTM network, a special kind of RNN with purpose-built memory cells to store information, have been successfully applied to many sequence modeling tasks. Recently, LSTMs based acoustic modeling has achieved improved performance over DNNs [13, 14] and conventional RNNs [15, 16] for LVCSR as they are able to model temporal sequences and long-range dependencies more accurately than the others especially when the amount of training data is large. LSTM has been successfully applied in both the LSTM-HMM hybrid systems [17, 18, 19, 20] and the end-to-end system [21, 22, 23].

For acoustic modeling, the LSTM takes in a sequence of input speech frames $X = \{x_1, \ldots, x_T\}$ and computes the hidden vector sequence $H = \{h_1, \ldots, h_T\}$ by iterating the equation below

$$h_t = \text{LSTM}(x_t, h_{t-1}) \tag{1}$$

where $\text{LSTM}(\cdot)$ denote the hidden layer function of the LSTM. In this paper, we implement Eq. (1) with the LSTM introduced in [24] as follows

$$\begin{align}
i_t &= \sigma(W_{x,i}x_t + W_{h,i}h_{t-1} + W_{c,i}c_{t-1} + b_i) \tag{2}\\
f_t &= \sigma(W_{x,f}x_t + W_{h,f}h_{t-1} + W_{c,f}c_{t-1} + b_f) \tag{3}\\
c_t &= f_t \odot c_{t-1} + i_t \odot \tanh(W_{x,c}x_t + W_{h,c}h_{t-1} + b_c) \tag{4}\\
o_t &= \sigma(W_{x,o}x_t + W_{h,o}h_{t-1} + W_{c,o}c_t + b_o) \tag{5}\\
h_t &= o_t \odot \tanh(c_t) \tag{6}
\end{align}$$

where i, f, o, c are the input gate, forget gate, output gate and cell state respectively, all of which are the same dimension as the hidden units vector h_t, σ is the logistic sigmoid function and $\odot$ stands for point-wise product. The weight matrix subscripts indicates the input and the gate (e.g., $W_{h,i}$ is the hidden-input gate matrix, etc.) The weight matrices from the cell to gate vectors (e.g. $W_{c,i}$, etc.) are diagonal.

A deep LSTM stacks multiple LSTM hidden layers on top of each other, with the output sequence of one layer forming the input sequence for the next hidden layer. With deep architecture, we are able to progressively learn higher level representations of the acoustic data and capture the high correlations between speech frames within a dynamic size of context window. The deep BLSTM processes the sequence of speech frames from both directions. It computes the forward hidden vector sequence $\overrightarrow{H} = \{\overrightarrow{h}_1, \ldots, \overrightarrow{h}_T\}$, the backward hidden vector sequence $\overleftarrow{H} = \{\overleftarrow{h}_1, \ldots, \overleftarrow{h}_T\}$ and the output sequence of senone posterior $Y = \{y_1, \ldots, y_T\}$ by iterating the backward layer from $t = T$ to 1, the forward layer from $t = 1$ to T and then updating the output layer. For each time, the output from both the forward and backward hidden layers are concatenated and then fed as the input of the next forward and backward hidden layers or the output layer.

To reduce the number of trainable parameters and alleviate the computational complexity, we introduce a separate linear projection layer after each BLSTM layer as in [18]. We connect each hidden layer to a recurrent projection layer with reduced number of units before recurrently feeding the projection layer back to the BLSTM input. The deep BLSTM acoustic model in this work is formulated as follow.

$$\begin{align}
\overrightarrow{h}_t^1 &= \text{LSTM}_1^{\text{forward}}(x_t, \overrightarrow{h}_{t-1}^1) \tag{7}\\
\overleftarrow{h}_t^1 &= \text{LSTM}_1^{\text{backward}}(x_t, \overleftarrow{h}_{t+1}^1) \tag{8}\\
\overrightarrow{h}_t^n &= \text{LSTM}_n^{\text{forward}}(p_t^{n-1}, \overrightarrow{h}_{t-1}^n), \quad n = 2, \ldots, N \tag{9}\\
\overrightarrow{p}_t^n &= W_{\overrightarrow{h}^n, \overrightarrow{p}^n} \overrightarrow{h}_t^n, \quad n = 1, \ldots, N \tag{10}\\
\overleftarrow{h}_t^n &= \text{LSTM}_n^{\text{backward}}(p_t^{n-1}, \overleftarrow{h}_{t+1}^n), \quad n = 2, \ldots, N \tag{11}\\
\overleftarrow{p}_t^n &= W_{\overleftarrow{h}^n, \overleftarrow{p}^n} \overleftarrow{h}_t^n, \quad n = 1, \ldots, N \tag{12}\\
p_t^n &= (\overleftarrow{p}_t^n, \overrightarrow{p}_t^n), \quad n = 1, \ldots, N \tag{13}\\
y_t &= \text{softmax}(W_{p^N y} \tanh(p_t^N) + b_y) \tag{14}
\end{align}$$

where $\text{LSTM}_n^{\text{forward}}(\cdot)$ and $\text{LSTM}_n^{\text{backward}}(\cdot)$ denote the forward and backward n^{th} hidden layer functions of the LSTM respectively. $\overrightarrow{p}_t^n$ and $\overleftarrow{p}_t^n$ are the projection vectors of forward and backward hidden vectors $\overrightarrow{h}_t^n$ and $\overleftarrow{h}_t^n$ respectively at the n^{th} layer. $W_{\overrightarrow{h}^n, \overrightarrow{p}^n}$ and $W_{\overleftarrow{h}^n, \overleftarrow{p}^n}$ are the projection matrices. p_t^n is the concatenation of forward and backward projection vectors $\overrightarrow{p}_t^n$ and $\overleftarrow{p}_t^n$. y_t is the senone posterior output vector given input speech frame x_t.

3. Non-uniform BMCE training of deep BLSTM acoustic model for keyword spotting

With cross-entropy or connectionist temporal classification criterion, the BLSTM acoustic models are trained to model either the senone distribution given an input speech frame or the distribution over all possible phone/character sequences conditioned on a given input speech frame sequence. They do not necessarily lead to minimized recognition error rate in LVCSR tasks. Therefore, many discriminatvie training methods such as MCE [25], maximum mutual information (MMI) [26, 27], minimum phone error (MPE) [28], state-level minimum Bayes risk (sMBR) [29, 30] and boosted MMI [31] are proposed to further refine the DNN [32] and LSTM [33] acoustic model.

For the keyword spotting task based on LVCSR, our goal is to minimize the recognition error on the set of keywords, while the aforementioned methods focus on the minimization of recognition error rate on all possible words which are not suitable for the keyword spotting task. Therefore, we perform non-uniform BMCE training of the BLSTM acoustic model , in which the BLSTM is train to minimize the *empirical error cost* instead of the empirical error rate.

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

$$g(x_{rt}, s_{Wt}; \Lambda) = \log[p(x_{rt}|s_{Wt})^{\kappa} p(s_{Wt})] \tag{15}$$

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.

$$d(x_{rt};\Lambda) = \\ -g(x_{rt}, s_{W_r t};\Lambda) + \log\left\{\frac{1}{N-1}\sum_{W\neq W_r}\exp[g(x_{rt}, s_{Wt};\Lambda)\eta]\right\}^{\frac{1}{\eta}} \tag{16}$$

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. (16) into a sigmoid function for smoothing, the objective function of the non-uniform MCE training of BLSTM is given by

$$\mathcal{L}_{NUMCE}(\Lambda) = \sum_{r=1}^{R}\sum_{t=1}^{T_r}\epsilon_r(t)l(d(x_{rt};\Lambda)) \tag{17}$$

where $\epsilon_r(t)$ is the error cost function at the frame level, $l(\cdot)$ is the sigmoid function with adjustable slope α. The objective function in Eq. (17) 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. (17) 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. (17) with respect to $a_{rt}(s)$, the activation for senone s at the output layer is

$$\begin{aligned}\frac{\partial\mathcal{L}_{NUMCE}(\Lambda)}{\partial a_{rt}(s)} &= \sum_q \frac{\partial\mathcal{L}_{NUMCE}(\Lambda)}{\partial\log p(x_{rt}|q)}\frac{\partial\log p(x_{rt}|q)}{\partial a_{rt}(s)} \\ &= \alpha\epsilon_r(t)l\left(d(x_{rt};\Lambda)\right)\left[1 - l\left(d(x_{rt};\Lambda)\right)\right] \\ &\quad \kappa\left[\delta_{s_{W_r t}:s} - \gamma_{rt}^{W\neq W_r}(s)\right]\end{aligned} \tag{18}$$

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 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 obtained by subtracting the log senone prior $\log p(q)$ from the log senone posterior $\log(y_t)$ in Eq. (14), and $\delta_{s_{Wt}:s}$ is the Kronecker delta function defined as

$$\delta_{s_{Wt}:s} = \begin{cases} 1, & s_{Wt} = s \\ 0, & s_{Wt} \neq s \end{cases} \tag{19}$$

For easy implementation, $d(X_{rt};\Lambda)$ is used as an approximation of $d(x_{rt};\Lambda)$. Eq. (18) is the error to be backpropagated through time to derive the gradients for all the parameters of BLSTM.

To minimize the recognition errors on the keywords, the error cost function $\epsilon_r(t)$ should be designed as follows so that all the recognition error cost associated with the keywords are emphasized.

$$\epsilon_r(t) = \begin{cases} K_1, & t \in \{t|W_r(t) \text{ is a keyword}\} \\ K_2, & t \in \{t|W(t) \text{ is a keyword}, W \neq W_r\} \\ 1, & \text{otherwise} \end{cases} \tag{20}$$

where $W_r(t)$ is the word which x_{rt} is aligned with in the label transcription and $W(t)$ is the word which x_{rt} is aligned with in the hypothesis word sequences and $K_1 > 1, K_2 > 1$.

The error cost function can be adjusted adaptively through iterations using a AdaBoost-like scheme as is proposed in [34]. We multiply $\epsilon_r(t)$ with a decay factor β if the frame x_{rt} is correctly classified at the current training iteration.

The non-uniform BMCE is further realized by boosting 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 as in [1].

The non-uniform BMCE training of BLSTM is implemented within the WFST framework. For an utterance r, we subtract the label transcription W_r from the decoding lattice to generate the competing hypothesis in Eq. (16) for non-uniform BMCE training. This is realized by taking the *difference* operation of WFST as follows.

$$L_r^{NUBMCE} = L_r(W) - WFST(W_r) \tag{21}$$

where $L_r(W)$ is the compact lattice for utterance r and $WFST(W_r)$ is the compiled WFST for W_r. We perform forward-backward on L_r^{NUBMCE} to obtain the posterior $\gamma_{rt}^{W\neq W_r}(s)$ in Eq. (18).

4. Experiments

4.1. Dataset Description

We evaluate the performance of the proposed framework on a large-scale CTS task, i.e., the 300 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 300 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 [35]. The GMM-HMMs are trained with the 300 hour training data using maximum-likelihood (ML) criterion. Each cross-word triphone is modeled by a 3-state left-to-right GMM-HMM (a 5-state HMM for silence). The 40 dimensional input features are Mel-frequency cepstral coefficient coupled with their linear discriminant analysis and maximum likelihood linear transform and feature-space maximum likelihood linear regression. The trigram language model is trained on 3M words of the training transcripts.

For training the BLSTM and DNN, the 36 dimensional log Mel filterbank features are extracted and then concatenated with 3 dimensional pitch features (consisting of probability of voicing, log pitch and delta log pitch) [36] to form a 39 dimensional "log Mel filterbank + pitch" feature.

To build the BLSTM-HMM baseline system, we stack 4 BLSTM hidden layers together and add a softmax output layer on the top to represent the 8861 senones posteriors. Each forward or backward hidden layer has 512 hidden units and is connected to a 256 dimensional recurrent projection layer. The forward and backward projection layers are concatenated together (to form a 512 dimensional vector) and fed as the input of the next BLSTM hidden layer. After appending delta and delta-delta coefficients to the 39 dimensional "log Mel filterbank + pitch" features, we use the 117 dimensional features with globally normalized zero mean and unit variance as the input to the BLSTM. The BLSTM is randomly initialized and then trained (initial learning rate 0.00002) to minimize the cross-entropy (CE) criterion using senone-level forced alignment generated by the GMM-HMM system as the target.

For DNN-HMM baseline system, we first pre-train a deep belief network (DBN) containing stacked restricted Boltzmann machines that are trained generatively in a layerwise fashion. The DBN is then fine-tuned to train a DNN with cross-entropy objective using stochastic gradient descent (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 39 dimensional "log Mel filterbank + pitch" 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 the output layer has 8861 units.

4.3. Results for Keyword Spotting

The BLSTM in baseline system is then trained with the non-uniform BMCE criterion for keyword spotting. We generate the forced alignment and denominator lattice of the training data using the baseline BLSTM, compute posterior $\gamma_{rt}^{W\neq W_r}(s)$ from the difference lattice L_r^{NUBMCE}, impose error cost function $\epsilon_r(t)$ on the frames aligned with keywords and compute errors in Eq. (18) for backpropagation through time. For comparison, we also discriminatively train baseline BLSTM with MMI, sMBR and BMCE criteria.

In Table 1, we show the FOM results with respect to different initial error costs K_1, K_2 and decay factors β. The system achieves the highest FOM 85.42% when $K_1 = K_2 = 10$ and $\beta = 0.3$, which is 4.49% and 1.23% absolute improvements over the baseline BLSTM and sMBR trained BLSTM. The best FOM is achieved when the learning rate is 0.00001, the slope of sigmoid α is 0.002 and the boosting factor is set at 0.07. We also observe that the FOM first increases as K_1 and K_2 grows and then gradually decreases when K_l and K_2 are larger than 10. The FOM increases or decreases more rapidly when the decay factor is smaller.

As a comparison, the DNN in the baseline system is also discriminatively trained with different criteria for the keyword spotting task. In Table 2, we observe the same trend of FOM variation with respect to the initial error cost function and decay factor as that of the BLSTM. By comparing Table 2 and Table 1, we see that the non-uniform BMCE trained BLSTM achieves 7.37% and 4.88% absolute FOM gains over cross-entropy trained DNN and non-uniform BMCE trained DNN. Under other uniform error discriminative training criteria, the BLSTM in general leads to about 4.0%-4.5% absolute FOM improvements over the DNN, which are much larger than the 2.3% absolute FOM gain BLSTM achieves under cross-entropy criterion. The large FOM improvement of BLSTM over DNN verifies its strong capability of modeling long-term dependencies and high correlations between speech frames that spans over long dynamic time intervals.

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

System	K_1	K_2	β	FOM (%)
BLSTM CE (baseline)	1	1	-	80.93
BLSTM MMI	1	1	-	83.25
BLSTM sMBR	1	1	-	84.19
BLSTM BMCE	1	1	-	84.21
BLSTM Non-Uniform BMCE	7	7	0.3	84.69
	7	7	0.5	85.01
	8	8	0.3	85.02
	8	8	0.5	85.15
	9	9	0.3	84.98
	9	9	0.5	84.63
	10	10	0.3	**85.42**
	10	10	0.5	85.08
	11	11	0.3	85.27
	11	11	0.5	84.96
	12	12	0.3	85.01
	12	12	0.5	85.05
	13	13	0.3	84.99
	13	13	0.5	85.01

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

System	K_1	K_2	β	FOM (%)
DNN CE	1	1	-	78.06
DNN MMI	1	1	-	79.05
DNN sMBR	1	1	-	79.37
DNN MCE	1	1	-	79.24
DNN BMCE	1	1	-	79.48
DNN Non-Uniform BMCE	6.0	6.0	0.5	80.24
	8.0	8.0	0.5	80.44
	10.0	10.0	0.5	80.5
	14.0	14.0	0.5	80.54
	16.0	16.0	0.5	80.45

5. Conclusions

In this paper, a BLSTM-HMM acoustic model is successfully trained using non-uniform BMCE criterion for the keyword spotting task on spontaneous conversational speech. We show that non-uniform BMCE criterion achieves higher performance than other discriminative training criteria on the keyword spotting task. We also show that a significant FOM gain can be achieved by using BLSTM acoustic model instead of DNN due to its strong capability of capturing long-term dependencies within speech signal.

6. References

[1] Z. Meng and B.-H. Juang, "Non-uniform boosted mce training of deep neural networks for keyword spotting," in *Proceedings of INTERSPEECH*, 2016, pp. 770–774.

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

[3] J. R. Rohlicek, W. Russell, S. Roukos, and H. Gish, "Continuous hidden markov modeling for speaker-independent word spotting," in *International Conference on Acoustics, Speech, and Signal Processing,*, May 1989, pp. 627–630 vol.1.

[4] R. C. Rose and D. B. Paul, "A hidden markov model based keyword recognition system," in *International Conference on Acoustics, Speech, and Signal Processing*, Apr 1990, pp. 129–132 vol.1.

[5] G. Chen, C. Parada, and T. N. Sainath, "Query-by-example keyword spotting using long short-term memory networks," in *Proceedings of ICASSP*, April 2015, pp. 5236–5240.

[6] R. Prabhavalkar, R. Alvarez, C. Parada, P. Nakkiran, and T. N. Sainath, "Automatic gain control and multi-style training for robust small-footprint keyword spotting with deep neural networks," in *Proceedings of ICASSP*, April 2015, pp. 4704–4708.

[7] D. R. Miller, M. Kleber, C.-L. Kao, O. Kimball, T. Colthurst, S. A. Lowe, R. M. Schwartz, and H. Gish, "Rapid and accurate spoken term detection," in *INTERSPEECH*, 2007, pp. 314–317.

[8] G. Chen, S. Khudanpur, D. Povey, J. Trmal, D. Yarowsky, and O. Yilmaz, "Quantifying the value of pronunciation lexicons for keyword search in lowresource languages," in *2013 ICASSP*, May 2013, pp. 8560–8564.

[9] J. Mamou, B. Ramabhadran, and O. Siohan, "Vocabulary independent spoken term detection," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2007, pp. 615–622.

[10] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, Mar 1994.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[12] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, Nov 1997.

[13] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

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

[15] C. Weng, D. Yu, S. Watanabe, and B. H. F. Juang, in *Proceedings of ICASSP*.

[16] O. Vinyals, S. V. Ravuri, and D. Povey, in *Proceedings of ICASSP*.

[17] A. Graves, N. Jaitly, and A. r. Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, Dec 2013, pp. 273–278.

[18] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling." in *INTERSPEECH*, 2014, pp. 338–342.

[19] Z. Meng, S. Watanabe, J. Hershey, and H. Erdogan, "Deep long short-term memory adaptive beamforming networks for multichannel robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, March 2017.

[20] H. Erdogan, T. Hayashi, J. R. Hershey, T. Hori, C. Hori, W.-N. Hsu, S. Kim, J. Le Roux, Z. Meng, and S. Watanabe, "Multichannel speech recognition: Lstms all the way through," in *The 4th International Workshop on Speech Processing in Everyday Environments*, 2016.

[21] A. Graves, A. r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 6645–6649.

[22] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, 2015, pp. 577–585.

[23] D. Amodei, R. Anubhai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, J. Chen, M. Chrzanowski, A. Coates, G. Diamos *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," *arXiv preprint arXiv:1512.02595*, 2015.

[24] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with lstm recurrent networks," *Journal of machine learning research*, vol. 3, no. Aug, pp. 115–143, 2002.

[25] B.-H. Juang, W. Hou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 257–265, May 1997.

[26] L. Bahl, P. Brown, P. de Souza, and R. Mercer, "Maximum mutual information estimation of hidden markov model parameters for speech recognition," in *ICASSP*, vol. 11, Apr 1986, pp. 49–52.

[27] 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.

[28] D. Povey and P. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *Proceedings of ICASSP*, vol. 1, May 2002, pp. I–105–I–108.

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

[30] 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.

[31] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted mmi for model and feature-space discriminative training," in *Proceedings of ICASSP*, March 2008, pp. 4057–4060.

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

[33] Andrew, H. Sak, F. de Chaumont Quitry, T. Sainath, and K. Rao, "Acoustic modelling with cd-ctc-smbr lstm rnns," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015, pp. 604–609.

[34] C. Weng and B.-H. Juang, "Adaptive boosted non-uniform mce for keyword spotting on spontaneous speech," in *Proceedings of ICASSP*, May 2013, pp. 6960–6964.

[35] 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.

[36] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proceedings of ICASSP*, May 2014, pp. 2494–2498.