ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Discriminative Training Using Non-uniform Criteria for Keyword Spotting on Spontaneous Speech

Chao Weng[1], *Biing-Hwang (Fred) Juang*[1], *Daniel Povey*[2]

[1]Center for Signal and Image Processing, Georgia Institute of Technology, Atlanta, GA, USA
[2]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA

chao.weng@ece.gatech.edu, juang@ece.gatech.edu, dpovey1@jhu.edu

10.21437/Interspeech.2012-172

Abstract

In this work, we investigate the feasibility of applying our prior works on discriminative training (DT) using non-uniform criteria to a keyword spotting task on spontaneous conversational speech. One of DT methods, minimum classification error (MCE), is recast and efficiently implemented in the weighted finite state transducer (WFST) framework to fit a keyword spotting task. To validate our approach, we evaluate it on a conversational speech task, the credit card use subset of Switchboard, in both kinds of keyword spotting scenarios: one is when a large vocabulary continuous speech recognition (LVCSR) decoder is available, the other is when a simple word-loop grammar of limited vocabulary is used. The results show our approach performs well in both cases, achieving 2.77% and 3.15% figure of merits (FOMs) absolute improvements over the baseline respectively.

Index Terms: LVCSR, keyword spotting, DT, non-uniform criteria, WFST

1. Introduction

The complexity of a general automatic speech recognition (ASR) task is characterized along two dimensions: the size of the vocabulary and the speaking style[1]. One usually can expect higher than 90% word accuracy, which is calculated using the Levenshtein distance between the label and the fully recognized transcriptions, on a large vocabulary task with the dictation speaking style, e.g., Wall Street Journal (WSJ) task. This accuracy, however, would dramatically decrease for a spontaneous conversational task (e.g., Switchboard) where the speaking style is much less constrained dialogue rather than the transcribed monologue. Thus, recognition of spontaneous speech requires a paradigm shift from speech recognition to understanding where underlying messages of the speaker are extracted from some significant *keywords*, instead of transcribing all the spoken words[2]. One good example is AT&T's How-May-I-Help-You (HMIHY)[3] system, in which the concept of *salient* words is proposed to evaluate the various word significance using the mutual information between each word and the call-type for the call-routing services.

Therefore, keyword spotting based techniques for the ASR problem become crucial in spontaneous speech scenarios. Keywords spotting, called "gisting" in the 1970s, was primarily based on template matching using dynamic time warping (DTW)[4]; recent systems are mostly based on hidden Markov models (HMMs) which constitute a layer of foundation on the concept of statistical modeling. In spite of several engineering successes, such as AT&T's VRCP (voice recognition call processing) [5], a formal formulation of the keyword spotting problem in the spirit of hypothesis testing remains elusive. In [6], the system employs N whole-word HMMs to represent N keywords, and an additional model, i.e. the filler or garbage model, to represent those non-keyword speech signals, and then the spotting is carried out by performing the task of (N+1)-word recognition, possibly followed by a hypothesis testing step using likelihood ratios. To better capture the characteristics of non-keywords, more filler models and the related structures are introduced into the grammar network, which treats the keyword spotting problem as an (N+M)-word recognition problem. Obviously, the performance of these whole-word model systems are often hampered by the issue of insufficient training data, thus giving rise to the modern use of sub-word based models. More reliable model estimation may be achieved by constructing keyword models as concatenations of phonetic HMMs. More recently, benefited from large vocabulary continuous speech recognition (LVCSR) techniques, a two-stage approach [7] is often shown to deliver good word-spotting results. In the first stage, the approach uses an LVCSR decoder to produce a set of hypothesized transcriptions, from which the presence of keywords are detected and verified in the second stage. The key issue in this approach is that the two stages are isolated and most likely designed under different criteria. LVCSR systems are trained to minimize the word error rate (WER) in general, without placing any emphasis on those keywords.

Discriminative training (DT) is a general technique to boost the recognition accuracy of an LVCSR system; it optimizes the model parameters to minimize recognition errors. If we envision word-spotting as a recognition task in which only the recognition accuracy of some words (i.e., those keywords) out of all possible words in the vocabulary needs to be maximized, an adaptation of the fundamental principle of error minimization in DT may present a new paradigm for performance enhancement in word-spotting. Such an adaptation would call for the introduction of non-uniform error cost embedded in discriminative training and our prior works[8][9][10] on DT using non-uniform criteria seem a perfect candidate. In this work, we will investigate the feasibility of applying the DT using non-uniform criteria to a keyword spotting task on spontaneous conversational speech. One of DT methods, minimum classification error (MCE)[11], will be recast and implemented in the weighted finite state transducer (WFST) framework to fit a keyword spotting task. Then we will test trained models on a conversational speech task with two kinds of scenarios: one is when a LVCSR decoder is available, and the other is a simple decoder with a word loop of limited vocabulary. The remainder of this paper is organized as follows: the MCE DT method is recast and augmented for a keyword spotting task in Section 2. Then the implementation of the augmented MCE in the WFST

framework is discussed in Section 3. Experiments and results on credit card use subset of Switchboard corpus are reported in Section 4.

2. Non-uniform MCE for Keyword Spotting

2.1. General MCE DT framework

General MCE training is a DT method for pattern recognition with the aim of direct minimization of the *empirical error rate*. In the speech recognition scenario, let $X_r, r = 1, \cdots, R$, be the utterances in the training set, W_r be the label word transcription for X_r and W be the certain selected hypothesis events. The discriminant function for a hypothesis W is defined as,

$$g_\Lambda(X_r, W) = \log P_\Lambda^\alpha(X_r|W)P_\Lambda^\beta(W). \quad (1)$$

Thus the misclassification measure takes the following form,

$$d_\Lambda(X_r) = -g_\Lambda(X_r, W_r)+\log\left[\frac{1}{|W|}\sum_{W\neq W_r}\exp[g_\Lambda(X_r, W)]\eta\right]^{\frac{1}{\eta}}. \quad (2)$$

$P_\Lambda(X_r|W)$,$P_\Lambda(W)$ denote the acoustic and language models, and α and β are scaling factors respectively. Finally, with proper smoothing using the sigmoid function, the objective function is formulated as,

$$\mathcal{L}_\Lambda = \sum_{r=1}^{R}\ell(d_\Lambda(X_r)), \quad (3)$$

where

$$\ell(d) = \frac{1}{1+\exp(-\gamma d+\theta)}. \quad (4)$$

As can be seen from Eq.(3), the objective function forms a smoothed approximation of *empirical errors* in the training set. The model parameters can be optimized iteratively to minimize the objective function via generalized probabilistic descent (GPD) as in original MCE work or gradient descent (GD) and extended Baum-Welch (EBW) after proper rewriting the objective function as in [12].

2.2. Non-uniform MCE Extension for Keyword Spotting

In [10], we already explored one way to extend the traditional MCE to embed the error cost, which converts the minimization of *error rate* to *error cost*. In that extension, to assign the different error cost on the model (phoneme) level, the error cost is selected from a *cross-layer* confusion matrix according to the label and most probable hypothesis phonetic transcription in certain time interval, i.e., only the phoneme with highest posterior (calculated on decoded phoneme lattice) among the hypothesis space is considered. In a keyword spotting scenario, however, we need to broaden the hypothesis space to capture more possible false alarm occurrences. We generalize our previous extension to accommodate all hypothesis in the decoded lattice, with proper simplification for efficient implementation, the objective function of non-uniform MCE can be written as,

$$\mathcal{L}_\Lambda = \sum_{r=1}^{R}\epsilon_r(t)\ell(d_\Lambda(X_r)), \quad (5)$$

here the error cost $\epsilon_r(t)$ is a function of time (frame) rather than a fixed value through the rth utterance , and the smoothed misclassification measure remain the same as in Eq.(3).

One may not easily grasp how the non-uniform MCE works only from the objective function in Eq.(5). But if we write down the gradient of this objective function (for simplification, we let $\eta = 1$ and ignore the factor $1/|W|$ in Eq.(2)),

$$\nabla\mathcal{L}_\Lambda = \sum_{r=1}^{R}\sum_{t=1}^{T_r}\gamma\ell(d_\Lambda(X_r))[1-\ell(d_\Lambda(X_r))] \epsilon_r(t)(-\gamma_{jm}^{W_r}(t)+\gamma_{jm}^{W\neq W_r}(t))\frac{\partial\log\mathcal{N}_{jm}(x_t^r,\Lambda)}{\partial\Lambda}, \quad (6)$$

where $\mathcal{N}_{jm}(x_t^r, \Lambda)$ is the corresponding Gaussian of certain model and mixture. $\gamma_{jm}^{W_r}(t)$ and $\gamma_{jm}^{W\neq W_r}(t)$ are *occupancy probabilities* of the Gaussian at certain frame t among the label and hypothesized transcriptions respectively. The value of the non-uniform error cost function $\epsilon_r(t)$ at tth frame can be absorbed into corresponding occupancy probabilities which implies we can repeat a training procedure by the scale of $\epsilon_r(t)$. From the other perspective, assuming the training sample $X_r(t)$ is drawn from certain distribution $\mathcal{D}(X_r(t))$, employing non-uniform MCE with the error cost function embedded on the original training set is equivalent to employing the regular MCE on an artificial *resampled* training set which observes the distribution,

$$\hat{\mathcal{D}}(X_r(t)) = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\epsilon_r(t)\mathcal{D}(X_r(t))}{E_{\epsilon_r(t),X_r(t)\sim\mathcal{D}(X_r(t))}[\epsilon_r(t)]}. \quad (7)$$

It can be shown in [13] that any *error rate* minimizing classifier on this *resampled* distribution $\hat{\mathcal{D}}$ will accomplish expected *error cost* minimization on the original distribution $\mathcal{D}$ if the training samples are drawn independently from the respective distribution. Although there exist correlations between speech frames, we can still expect the expected cost to be minimized under the non-uniform MCE training.

Another issue is the design of error cost function for keyword spotting task. Since $\epsilon_r(t)$ is a function of time, it is straightforward to design it in a way that all frames labeled as keywords should be assigned a higher value. In the other hand, as mentioned earlier, those frames with high possibility recognized as keyword hypothesis should also be emphasized to prevent the false alarms, this can be efficiently done via searching keywords in the corresponding decoded lattice and recording the start and end frames. So the error cost function can be designed in the following form,

$$\epsilon_r(t) = \begin{cases} 2 & t\in\{t|W_r(t)\in keywords\}\cup\{t|W(t)\in keywords\} \\ 1 & otherwise \end{cases}. \quad (8)$$

The parameters update when assigned a higher error cost corresponding to correct keyword hypothesis will be canceled out while focuses on the detection miss or false alarms cases. One also can design the error cost function in a asymmetrical way to achieve desirable compromise between the detection miss and false alarm rate. Considering that the keyword's context frames may also need to be put additional emphasis, we can introduce roll-offs at the boundaries between keyword frames and non-keyword frames.

3. Non-uniform MCE Implementation in the WFST framework

In this section we will discuss how non-uniform MCE is implemented in the WFST framework. The WFST interpretation of

the decoding graph for an ASR problem can be written as,

$$HCLG = \min(\det(H \circ C \circ L \circ G)), \quad (9)$$

where H, C, L and G are the HMM structure, phonetic context-dependency, lexicon and grammar FST respectively, and $\circ$ is the WFST composition operations[14]. If we let an acceptor U (an acceptor is just a FST with the identical input and output labels) encode the acoustic scores of an utterance, then the full search graph for this utterance is expressed as [15],

$$S = U \circ HCLG. \quad (10)$$

A decoded lattice, which forms a compact representation of the hypothesis space for this utterance, is a beam pruned subgraph of S.

As in Eq.(3), the competing hypothesis for MCE training has to exclude the reference word transcription W_r. Naively removing the corresponding arcs of the label words will hurt the topological structures of the decoded lattice. Although this issue can be circumvented via subtracting the corresponding reference's statistics from the nominator and denominator of posterior in [12], we can come up with a clever solution that takes advantage of FST's difference operation to exclude the reference in the WFST framework. Since only strings that are in the first FST but not in second are retained in the result after the FST's difference, we can first compile a linear FST which only accepts the label word string then subtracted from the decoded lattice (FST).

However, as in Eq.(10), the input symbols of a direct decoded FST for an utterance are HMM states, which apparently is not a legal operand for the FST difference. In [15], a compact representation of the decoded lattice is proposed whereby the decoded lattice is a FST (FSA) with identical input and output symbols being words, while the acoustic, language score and the state alignment strings are all encoded in to the weight using a special semiring as follows, let (c, s) be a pair of the cost c (including both acoustic and language cost) and a state symbol sequence s,

$$(c, s) \otimes (c', s') = (c + c', (s, s')), \quad (11)$$

$$(c, s) \oplus (c', s') = \begin{cases} (c, s) & if\ c < c' \\ (c', s') & if\ c > c' \\ (c, s) & if\ len(s) < len(s') \\ (c', s') & if\ len(s) > len(s') \end{cases}, \quad (12)$$

where (s, s') is the concatenation of s and s', if both the costs and the length of state strings are identical, the $\oplus$ will return the pair whose string appears first in dictionary order. With this FST representation of the decoded lattice, we can conduct the FST difference and generate the lattice for MCE training as,

$$FST_r^{MCE} = FST_{compact}(W) - FST(W_r). \quad (13)$$

$FST_{compact}(W)$ is the compact encoding as described above of pruned search graph of r utterance; $FST(W_r)$ is the compiled FST from label transcription.

With FST_r^{MCE} available, $\gamma_{jm}^{W \neq W_r}(t)$ can be directly evaluated using forward-backwards on it. Then for the error cost function embedding of the non-uniform MCE implementation, we only need a vector with its length equal to the frame number of the utterance, and search both label alignments and FST_r^{MCE}, recording the union of time intervals where the keywords occur and setting the corresponding value of the vector, then use this vector to scale the posteriors when updating the model parameters. Therefore, in the WFST framework, non-uniform MCE can be implemented efficiently without significant additional overheads compared to the regular MCE.

Table 1: Non-uniform MCE over Baseline for LVCSR case

Method	FOM	Absolute Improvement
Baseline	83.59%	N/A
MCE	85.34%	1.75%
Non-uniform MCE	86.36%	2.77%

4. Experiments and Results

We evaluate the non-uniform MCE training for keyword spotting on a spontaneous conversational task, Switchboard-1 Release 2, which is a collection of 2438 two-sided telephone conversations among 543 speakers (302 male, 241 female). Each pair of callers is introduced a topic for discussion and there are about 70 topics were provided. The training utterances are selected through all Switchboard corpus. We construct the test set for keyword spotting as follows: we first extract the conversations which is on the topic of "CREDIT CARD USE" (this information is not included in the release, but can be downloaded from Switchboard LDC official website) as the test utterances and 18 keywords are selected for the spotting evaluation based on their relevance to the topic and occurrence, which are "BANK", "CARD", "CASH", "CHARGE", "CHECK", "MONTH", "ACCOUNT", "BALANCE", "CREDIT", "DOLLAR", "HUNDRED", "LIMIT", "MONEY", "PERCENT", "TWENTY", "VISA", "DISCOVER", "INTEREST". The baseline system is built using Kaldi Speech Recognition Toolkit[16], cross-word triphone models represented by 3-state left-to-right HMMs (5-state HMMs for silence) are trained using MLE on about half the data of whole Switchboard Corpus. A tri-gram language model is trained on the whole transcriptions of the dataset for decoding. The input features are MFCCs coupled with their linear discriminant analysis (LDA) and maximum likelihood linear transform (MLLT) and feature-space maximum likelihood linear regression (fMLLR) for speaker adaptation during later iterations. The WER of the baseline system on HUB5 English evaluation set is 33.4%.

We conduct both MCE and non-uniform MCE (with the embedding of the error cost function defined in Eq.(8)) in 4 iterations. The lattices used for both DT methods is exactly as described in Section 3, the γ, θ in Eq.(4) is set to 0.08 and 0 respectively. All trained models are evaluated on Switchboard subset of credit card use in terms of figure of merit (FOM) and receiver operating characteristic (ROC) curve. To completely validate our approach, we test both MCE and non-uniform MCE trained models in two scenarios: one is when a LVCSR decoder is available for keyword spotting, the other is only a simple word-loop decoding grammar is allowed and the vocabulary size is limited. The keyword spotting is conducted using the decoded word alignments and scores since we want to observe the impacts of the boosted acoustic models. We report FOMs and generate ROC curves for both scenarios.

4.1. Keyword Spotting with LVCSR decoder

For the LVCSR decoder case, the trained tri-gram grammar network is used for the decoding, so there are no out-of-vocabulary terms (OOVs). We list overall FOMs of baseline, MCE and non-uniform MCE in Table 1 and draw the ROC curves for these three case in Fig.1. We can see non-uniform MCE achieves the highest FOM which is 2.77% and 1.02% absolute improvement over baseline and regular MCE respectively. As shown in Fig.1, non-uniform MCE also obtains higher accuracy through all FA rate.

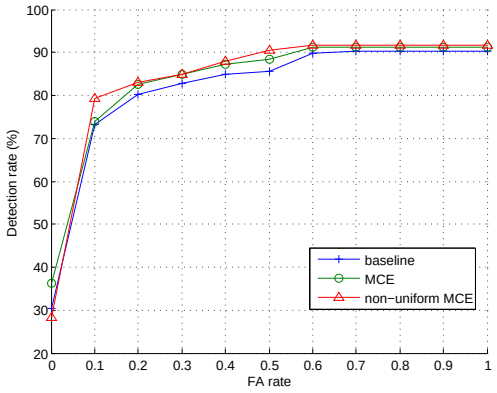

Figure 1: ROC curves for the LVCSR case

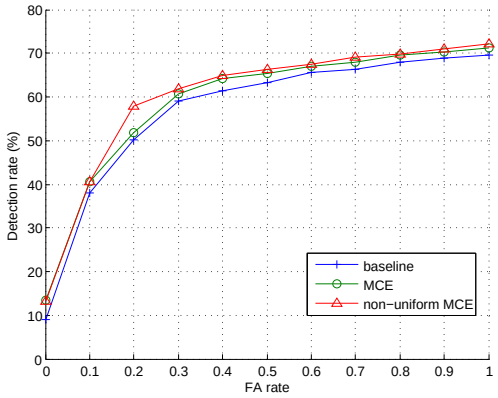

Figure 2: ROC curves for 1000 Vocabulary word-loop grammar

Table 2: FOMs with Different Vocabulary Size

Vocabulary Size	100	500	1000	1500	2000
FOMs (%)	36.46	55.47	58.26	59.62	59.93

4.2. Keyword Spotting with word-loop grammar

For the limited vocabulary case, we just use word-loop grammar (zero-gram LM) for the decoding, which is similar to the $N + M$ word recognition. To determine the appropriate vocabulary size, we extract most 100, 500, 1000, 1500 and 2000 frequent words among the training transcription unioned with the 18 selected keywords to form the vocabulary, and use them to construct the word-loop decoding grammar respectively. The corresponding FOMs is listed in Table2. As can be seen, the significant impact of OOVs begins to disappear from 1000. With the compromise of FOM performance and vocabulary size, we use 1000+18 case to report the results. We listed the FOMs and generate the ROC curves in Table3 and Fig.2. It is shown that in this case non-uniform case still works best.

5. Conclusion

In this paper, we demonstrate DT using non-uniform criteria can be successfully applied to a keyword spotting task. Both MCE and non-uniform MCE are implemented efficiently via taking advantage of the FST's difference operations in the WFST framework. Experimental results show that the approach performs well even with quite simple error cost function. In the future work, we will explore more sophisticated form of the error cost function with which better performance is expected.

6. References

[1] B.-H. Juang and S. Furui, Automatic recognition and understanding of spoken language A first step towards natural human-machine communication, Proc. IEEE, 88, 8, pp. 1142-1165, 2000

[2] B.-H. Juang, From speech recognition to understanding: Shifting paradigm to achieve natural human-machine communication, Proc. 16th ICA and 135th Meeting ASA, pp. 617-618, 1998

[3] A. L. Gorin, B.A. Parker, R.M. Sachs and J.G. Wilpon, "How may I help you?," Proc. Third IEEE Workshop on Interactive Voice Technology for Telecommunications Applications, pp.57-60, 1996

[4] Sakoe, H. and Chiba, S., "Dynamic programming algorithm optimization for spoken word recognition," IEEE Transactions on Acoustics, Speech and Signal Processing, 26(1) pp. 43-49, 1978

[5] J. G. Wilpon and D. B. Roe, "AT&T Telephone Network Applications of Speech Recognition," Proc. COST232 Workshop, Rome, Italy, Nov. 1992.

[6] M. Rahim, C. Lee and B.-H. Juang, "Discriminative utterance verification for connected digits recognition," IEEE Transactions on Speech and Audio Processing, p. 266-277, 1997.

[7] R. C. Rose, Keyword detection in conversational speech utterances using hidden Markov model based continuous speech recognition, Computer Speech and Language, vol.9, pp. 309-333, 1995.

[8] Q. Fu, D. S. Mansjur, and B.-H. Juang, Non-uniform error criteria for automatic pattern and speech recognition, in Proc. ICASSP2008, 2008, pp. 1853-1856.

[9] C. Weng and B.-H. Juang, Recent development of discriminative training using non-uniform criteria for cross-level acoustic modeling, in Proc. ICASSP2011, 2011, pp. 5332-5335.

[10] C. Weng and B.-H. Juang, A comparative study of discriminative training using non-uniform criteria for cross-layer acoustic modeling, in Proc. ICASSP2012, 2012, p. 4089-4092.

[11] B.-H. Juang and S. Katagiri, Discriminative learning for minimum error classification, IEEE Trans. Signal Process., vol. 40, pp. 3043-3054, Dec. 1992.

[12] R. Schluter, W. Macherey, B. Muller, and H. Ney, Comparison of discriminative training criteria and optimization methods for speech recognition, Speech Communications, vol. 34, pp. 287-310, 2001.

[13] B. Zadrozny, J. Langford and N. Abe, "Cost-sensitive learning by cost-proportionate example weighting," in Proc. ICDM 2003, 2003, p 435-442.

[14] M. Mohri, F. Pereira, and M. Riley, Weighted finite-state transducers in speech recognition, Computer Speech and Language, vol. 20, no. 1, pp. 69-88, 2002.

[15] D. Povey, M. Hannemann et. al, "Generating exact lattices in the WFST framework", in Proc. ICASSP 2012, 2012, p. 4213-4216.

[16] D. Povey, A. Ghoshal, et al., The Kaldi Speech Recognition Toolkit, in Proc. ASRU, 2011.

Table 3: Non-uniform MCE over Baseline with 1000 vocabulary size

Method	FOM	Absolute Improvement
Baseline	58.26%	N/A
MCE	60.42%	2.16%
Non-uniform MCE	61.41%	3.15%