A Keyword-Boosted sMBR Criterion to Enhance Keyword Search Performance in Deep Neural Network Based Acoustic Modeling

I-Fan Chen\(^1\), Nancy F. Chen\(^2\), and Chin-Hui Lee\(^1\)

\(^1\) School of Electrical and Computer Engineering, Georgia Institute of Technology
\(^2\) Institute for Infocomm Research, Singapore

ichen8@gatech.edu, nfychen@i2r.a-star.edu.sg, chl@ece.gatech.edu

Abstract

We propose a keyword-boosted state-level minimum Bayes risk (sMBR) criterion for training DNN-HMM hybrid keyword search systems by enhancing acoustic detail of a given list of target keyword terms. The rationale behind the proposed discriminative training strategy is to place more acoustic modeling emphasis on states appearing in the given keywords. We observed a relative gain of 1.7~6.1\% in actual term weighted value (ATWV) performance with the proposed keyword-boosted SMBR training over the conventional SMBR systems when tested on the IARPA Babel program’s Vietnamese limited-language-pack task. A detailed result analysis suggests that the proposed sMBR objective function effectively improves the ATWV scores by boosting the probability of detecting keywords appearing in the system output with an increased correct and insertion rates in the decoded lattices.

Index Terms: keyword spotting, keyword search, spoken term detection, deep neural network, discriminative training

1. Introduction

Keyword search (KWS) is a task of detecting a set of preselected keywords and key-phrases in continuous speech [1-3]. A common realization of today’s KWS systems is to use a large vocabulary continuous speech recognition (LVCSR) system as a detector to establish keyword hypotheses [3-5]. By converting input speech into text documents using the underlying LVCSR systems, KWS systems can use techniques developed for text search to finalize the decisions. Since an LVCSR system is utilized for keyword detection, many discriminative training techniques including MCE (Minimum Classification Error) [6], MMI (Maximum Mutual Information) [7, 8], bMMI (boosted Maximum Mutual Information) [9], MPE (Minimum Phone Error) [8, 10], MWE (Minimum Word Error) [8], sMBR (state-level Minimum Bayes Risk) [11], etc., can also be used to improve KWS system performances. Because such discriminative training methods were originally designed for LVCSR systems geared at reducing the system word error rate (WER), all words are being considered equally putting no emphasis on the set of special keywords.

In KWS tasks, the systems only care about whether the preselected keywords are detected or not. For words unrelated to the target keywords, their detection results are not as critical. In [12], a non-uniform MCE training method is proposed for conventional GMM-HMM systems to address this objective difference in KWS tasks by putting more weights on the error cost function of keyword data. Since sMBR training has been shown to outperform other discriminative training criteria for deep neural network (DNN) models [13], we propose, in this paper, a modified objective function of state-level minimum Bayes risk (sMBR) training for deep neural network models in KWS systems by putting more emphasis on keyword-related states in acoustic modeling.

In Section 2, conventional sMBR training for DNN is presented. The proposed keyword-boosted sMBR (KW-boosted sMBR) objective function is also formulated in the section. A series of experiments, investigating the characteristics of the proposed KW-boosted sMBR objective function, is presented in Section 3. Finally conclusions are drawn and future work is discussed in Section 4.

2. State-level Minimum Bayes risk (sMBR) Training for DNNs

The sMBR training has been shown to be one of the most effective discriminative training for DNN-HMM hybrid systems in speech recognition tasks [13]. In this section, we review the sMBR training for a DNN-HMM system and propose the modified sMBR criterion for KWS systems.

2.1. sMBR Training Criterion

The objective function of the sMBR training criterion is designed to minimize the expected state-level errors over the whole training utterances [13, 14]:

\[ F_{SMBR}(\lambda) = \sum_{u \in O} \sum_{s} p(O_u | s) P(W) A(W | W_u) / \sum_{s} p(O_u | s) P(W) \]

(1)

where \( \lambda \) represents the acoustic model parameters, \( \kappa \) is the acoustic scaling factor, \( O_u = \{ o_{u1}, \ldots, o_{ux} \} \) is the observation sequence of the training utterance \( u \), \( S_p \) and \( S_w \) are the HMM state sequences corresponding to the word sequence \( W \) and \( W_u \), \( W \) is the reference word sequence for the sentence \( u \); and \( A(W, W_u) \) is the raw state accuracy which could be expressed as

\[ A(W, W_u) = \sum_{s} \text{StateAcc}(s(t), s_u(t)) \]

(2)

In Eq. (2), \( T_u \) is the number of frames of the utterance \( u \), \( s(t) \) and \( s_u(t) \) are corresponding HMM states of word sequences \( W \) and \( W_u \) at time \( t \). The StateAcc\((s, s_u)\) is defined as follows:

\[ \text{StateAcc}(s, s_u) = \begin{cases} 1 & \text{if correct state, namely } s = s_u \\ 0 & \text{if incorrect state, namely } s \neq s_u \end{cases} \]

(3)

2.2. Training DNN-HMM hybrid systems with sMBR criterion

DNN-HMM hybrid systems use DNNs in replace of GMMs to provide log-likelihood for the HMM states in an LVCSR system. For an observation \( o_{u0} \) corresponding to time \( t \) in utterance \( u \), a pseudo log-likelihood of state \( s \) provided by DNN is
\[
\log p(o_s | s) = \log y_s(s) + \log p(s)
\]
where \( p(s) \) is the prior probability of state \( s \) calculated from the training data, and \( y_s(s) \) is the posterior probability for state \( s \) given observation \( a_s \) estimated by DNN and is defined as
\[
y_s(s) = p(s | a_s) = \frac{\exp[a_s(s)]}{\sum_{s'} \exp[a_s(s')]} \tag{4}
\]
where \( a_s(s) \) is the activation at the output layer corresponding to state \( s \).

The DNNs can be trained with sMBR criterion using the gradient decent approach \([11]\). By first differentiating Eq. (1) with respect to \( \log p(o_s/r) \), we get
\[
\frac{\partial F_{\text{MBR}}}{\partial \log p(o_s | r)} = \kappa \cdot \gamma_{\text{MBR}}(s) = \kappa \cdot \gamma_{\text{MBR}}^\text{state acc}(s), \tag{5}
\]
where \( \gamma_{\text{MBR}}(s) \) is the average state accuracy of all paths in the lattice of sentence \( u \) passing through state \( r \) at time \( t \), \( \gamma_{\text{MBR}}^\text{state acc}(s) \) is the average state accuracy of all paths in the lattice, and \( \gamma_{\text{MBR}}^\text{state acc}(s) \) is a factor as defined in the MPE training method \([8]\).

Using Eq (5), the gradient for activation function \( a_s(s) \) can be derived as \([11, 13]\)
\[
\frac{\partial F_{\text{MBR}}}{\partial a_s(s)} = \sum_t \frac{\partial F_{\text{MBR}}}{\partial \log p(o_s | r)} \cdot \frac{\partial \log p(o_s | r)}{\partial a_s(s)} = \frac{\partial F_{\text{MBR}}}{\partial a_s(s)} = \kappa \cdot \gamma_{\text{MBR}}(s).
\tag{6}
\]
The gradient then can be used to update the parameters in the whole network using back-propagation procedure.

### 2.3. Keyword-boosted sMBR

In Eq (2) and (3), the raw state accuracy computation considers all states as equally important disregarding their source words. However, in a KWS system, accuracies of keyword terms are more important than the rest of the terms in the vocabulary. To adapt the conventional sMBR objective function to KWS tasks, we modify Eq (3) with a boosting weight \( \alpha \) for keyword terms in the system:
\[
\text{StateAcc}(s_k, s_{\text{ref}}) = \begin{cases} 
\alpha & \text{if } s_k = s_{\text{ref}} \text{ and } t \in \text{[KW Seg]} \\
1 & \text{if } s_k = s_{\text{ref}} \text{ and } t \not\in \text{[KW Seg]} \\
0 & \text{if } s_k \neq s_{\text{ref}}
\end{cases}
\tag{7}
\]
where ‘[KW Seg]’ is the set of time segments for keyword words in the reference sentences. Note the resulting keyword-boosted sMBR has the same computational complexity as the conventional sMBR training since the only difference between them is the weight distribution in the objective function.

### 3. Experiments and Discussions

A series of experiments was designed to investigate the characteristics of the proposed KW-boosted sMBR trained KWS systems. Key factors affecting the system performance, such as system lexicons and language models, boosting weight \( \alpha \) for keyword states, the number of keywords in the system, are studied in the following sections.

#### 3.1. Experimental Setup

Experiments were conducted on the Vietnamese limited language pack (LLP) provided by the IARPA Babel program used in the NIST OpenKWS13 Evaluation \([15]\). The training set consists of 10 hours of transcribed audio and roughly 70 hours of data without transcription. The audio data is conversational speech between two parties over a telephone channel, which can be landlines, cell phones, or phones embedded in vehicles, with the sampling rate set at 8000 Hz. The development set consists of 10 hours of conversational telephone speech. In this study, we built our systems using only the 10-hour transcribed data and no unsupervised training was used on the un-transcribed data. Furthermore, instead of using the whole 10-hour development data for system tuning, a 2-hour subset of the development set (denoted as dev2h in this paper) was used to speed up the tuning process.

The 15-hour evaluation part 1 data (denoted as evalpart1) was used for testing. Two keyword lists released by NIST for the OpenKWS13 Evaluation were used in this paper. An evaluation keyword list containing 4065 phrases, including out-of-vocabulary words not appearing in the training set, was used as the major keyword list for system development and performance evaluation; while the other development keyword list, consisting of 200 phrases, was used for investigating the effect of keyword list sizes to the proposed approach. The performance of keyword search was measured by Actual Term Weighted Value (ATWV) \([16]\), which was computed by
\[
ATWV = \frac{1}{K} \sum_{k} \left[ \frac{N_{\text{true}}(k)}{N_{\text{ref}}(k)} \right] \beta \frac{N_{\text{false}}(k)}{N_{\text{ref}}(k)} \tag{8}
\]
where \( K \) is the number of keywords, \( N_{\text{true}}(k) \) is the number of true keyword tokens that are not detected, \( N_{\text{false}}(k) \) is the number of false alarms, \( N_{\text{ref}}(k) \) is the number of keywords in reference, \( T \) is the number seconds of the evaluation audio, and \( \beta \) is a constant set at 0.999. In addition to ATWV, word error rate (WER) of the systems was compared as well.

All keyword search systems evaluated were LVCSR-based with hybrid DNN-HMM acoustic models built with the Kaldi toolkit \([17]\). The acoustic vectors were bottleneck features with an fMLLR transformation, and the input for the bottleneck DNNs is a concatenation of PLP, fundamental frequency (F0), and fundamental frequency variation (FFV) features used in \([3]\). Keyword-specific threshold normalization \([18]\) was used in all the KWS systems, and the common threshold after normalization is chosen to be 0.5.

For the pronunciation dictionaries, in addition to the original LLP lexicon provided by IARPA, we also used a grapheme-to-phoneme (G2P) method \([19]\) to estimate the pronunciation for out-of-vocabulary (OOV) words not present in the training text. They were then merged into the original LLP lexicon to form a G2P lexicon (called G2P Lex).

Two language models were used in the experiments. The first one is a basic trigram language model, trained by the 10-hour training text and denoted as original LM (orig LM) here. The second one is a context-simulated keyword LM interpolated language model (CS-KWLM Int) \([20]\), which strengthens the probabilities of \( n \)-grams used by key phrases in the system to increase the keyword detection rate while keeping the word error rate performance of the underlying LVCSR system nearly intact.

#### 3.2. Experimental Results

We first compared the proposed KW-boosted sMBR system with the cross-entropy trained DNN baseline and sMBR trained DNN systems in three different configurations, namely:
KW-boosted sMBR introduces a boosting-weight parameter $\alpha$ to control the weight ratio between the accuracies of keywords and non-keywords as presented in Eq. (7). In the following experiments, we investigated the effect of the KW-boosting weights on the system performance.

Six KW-boosted sMBR systems with $\alpha = 2, 4, 5, 6, 8,$ and $10$ were trained, and the ATWV and WER performances of the systems on dev2h with the $\text{orig} \text{ Lex} + \text{LM}$ test configuration are shown, respectively, in Figure 1 and Figure 2. Notice that since the conventional sMBR system can also be considered as a KW-boosted sMBR system with $\alpha = 1$, we also included its performance in Figure 1 and Figure 2 for comparison. From the plots, it is clear that $\alpha$ plays an important role to the performances. ATWV at $\alpha = 1$ (sMBR system) is about 0.22, and it is improved as $\alpha$ increased till $\alpha = 0.5$, in which the KW-boosted sMBR system achieved an ATWV of 0.255, which is about a 16% relative improvement over the sMBR system. The ATWV then decreased to 0.235 when $\alpha = 8$ and slightly increased to 0.24 when $\alpha = 10$. As shown in Figure 1, the relation between the system $\alpha$ and the ATWV performance is not simple, and thus a development set is required for tuning this parameter. On the other hand, from Figure 2 it is clear that the WER of KW-boosted sMBR system increases as $\alpha$ increasing, as expected.

In Table 2 the WER was reduced from row 2 for baseline to row 3 for sMBR. However it was increased slightly from row 3 to row 4 since the KW-boosted sMBR training criterion put much less focus on words not used by the key-phrases. Notice that all the KWS systems were trained with the original lexicon and LM in the training phase and then were tested with the three test configurations. System parameters and training iterations are selected with the dev2h set.

### Table 1. ATWV (4065 keywords) of systems on the evalpart1 test set with three test configurations.

<table>
<thead>
<tr>
<th>ATWV (evalpart1)</th>
<th>(i) $\text{orig} \text{ Lex} + \text{LM}$</th>
<th>(ii) G2P lex + LM</th>
<th>(iii) G2P lex + CS-KWLM Int</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.1895</td>
<td>0.1883</td>
<td>0.3044</td>
</tr>
<tr>
<td>sMBR</td>
<td>0.2069</td>
<td>0.2093</td>
<td>0.3287</td>
</tr>
<tr>
<td>KW-boosted sMBR</td>
<td>0.2104</td>
<td>0.2220</td>
<td>0.3396</td>
</tr>
</tbody>
</table>

In Table 2 the WER was reduced from row 2 for baseline to row 3 for sMBR. However it was increased slightly from row 3 to row 4 since the KW-boosted sMBR training criterion put much less focus on words not used by the key-phrases.

### Table 2. WER (4065 keywords) of systems on the evalpart1 test set with three test configurations.

<table>
<thead>
<tr>
<th>WER (evalpart1)</th>
<th>(i) $\text{orig} \text{ Lex} + \text{LM}$</th>
<th>(ii) G2P lex + LM</th>
<th>(iii) G2P lex + CS-KWLM Int</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>66.6</td>
<td>66.6</td>
<td>67.0</td>
</tr>
<tr>
<td>sMBR</td>
<td>65.0</td>
<td>65.0</td>
<td>66.0</td>
</tr>
<tr>
<td>KW-boosted sMBR</td>
<td>66.1</td>
<td>65.9</td>
<td>66.3</td>
</tr>
</tbody>
</table>

3.2.2. Effect of keyword set size on KW-boosted sMBR training

In the previous experiments we built our KWS systems using the 4065-keyword evaluation list. However, since the idea of KW-boosted sMBR is to put more weights on the accuracies of words being considered, it is interesting to know whether the percentage of the interested words in the system vocabulary affects the performance of the KW-boosted sMBR systems. In this following, we built our KW-boosted systems with the 200-keyword development list. Like in Section 4.2.1, the relations of ATWV and WER vs. $\alpha$ are illustrated in Figure 3 and Figure 4, respectively.

The 4065-keyword list used 2155 words in the system's 3210-word vocabulary (namely about two thirds of words in the vocabulary); while the 200-keyword list covers 408 words (about an eighth) in the system vocabulary. Comparing Figure 1 and 3, the first thing we noticed is that the $\alpha$-ATWV curves are very different for these two keyword sets. For the 200-keyword set, unlike in the 4065-keyword systems, ATWV of the systems dropped when $\alpha$ increased from 1 to 2. In Figure 3, the ATWV of KW-boosted sMBR systems oscillated between 0.22 and 0.19 and is worse than the sMBR system till $\alpha = 6$, where the KW-boosted sMBR system start getting better than the sMBR system. When $\alpha = 8$ the proposed system achieved the best ATWV at 0.2124, however the number dropped again when $\alpha = 10$. This observation shows that when the keyword set size is smaller, the system performance is less stable as $\alpha$ changes. It also shows that, the optimal parameter for KW-
boosted sMBR is different for different keyword sets, and the best way for finding the optimal \( \alpha \) is tuning it with a small development data. The ATWV of the dev2h tuned 200-keyword KW-boosted sMBR system (with \( \alpha = 8 \)) on evalpert1 is 0.1453, which is again better than the 200-keyword sMBR system whose ATWV is at 0.1349.

For the system WER, similar trends were observed for both the 4065-keyword and 200-keyword sets. When comparing Figures 2 and 4, it is easy to see that WER of the 200-keyword systems increased faster with increasing \( \alpha \) than the 4065-keyword system because non-keyword words, whose accuracies are less cared in KW-boosted sMBR training, occupy a much higher percentage in the 200-keyword systems.

### 3.2.3. System 1-best output analysis

To realize why the KW-boosted sMBR criterion worked for keyword search tasks, we further analyzed the performance of the 1-best sentence outputs of both conventional sMBR and KW-boosted sMBR systems. By categorizing the system vocabulary into keyword-word (KWW) and non-keyword word (non-KWW) groups, we evaluated the correction, substitution, deletion, and insertion rate regarding words in these two groups for the two systems. The 4065-keyword evaluation list was used in the experiments.

Table 3 shows the performance of KWW and non-KWW for sMBR and KW-boosted sMBR systems under the orig Lex+LM test configuration on the dev2h data, where the ATWV of KW-boosted sMBR system (0.2564) outperformed the ATWV of the sMBR system (0.2101) by 20% relative. In Table 3, comparing the KWW performance of both systems, the KW-boosted sMBR system had higher correction and insertion rates, and is with much lower rate for deletion. While for the non-KWWs, the KW-boosted sMBR system has high deletion rate and was with lower rate for correction and insertion rate. In other words, the KW-boosted sMBR system tended to generate keyword words in the system output while reducing the output of non-keyword words. In Eq. (7), the training criterion putting more weight on the reference KWW segments implies the criterion penalizes KWW misses more than insertions of KWWs. As a result, the KW-boosted sMBR trained systems are prone to generating keyword words with higher scores in the decoding lattices. Since ATWV is a metric with higher emphasis on misses than on false alarms [21], this tendency provides KWS systems a better chance to get good performance result on ATWV. On the other hand, this also explains the non-trivial relations between the boosting factor \( \alpha \) and ATWV for the KW-boosted sMBR system since with inappropriately estimated \( \alpha \), a KW-boosted sMBR system may generate too many keyword words and result in high false alarm rate which greatly degrades the ATWV performance. This suggests that a possible improvement for the current KW-boosted sMBR criterion is to take the keyword words appearing in lattice hypotheses into consideration as well so that the false alarm of KWWs can be also penalized with higher weight.

Similar results were also observed in the G2P Lex + CS-KWLM Int test configuration, which are shown in Table 4. The KW-boosted sMBR system has higher correction and insertion rates for KWWs, while for non-KWWs the system has higher deletions.

### 4. Conclusion and Future Work

In this paper, a modified objective function of sMBR training for DNN-HMM hybrid systems is proposed to enhance the performance of keyword search tasks. By putting more weight on acoustic modeling of keyword states, the KW-boosted sMBR system is capable of detecting more keywords while reducing false alarms. Experimental results show that the proposed KW-boosted sMBR trained systems outperform the baseline sMBR systems by a relative increase of 1.7 ~ 6.1% in ATWV in all test configurations. An analysis of experimental results also show that tuning the keyword-boosting weight \( \alpha \) with different keyword sets is a key to KW-boosted sMBR training. The current KW-boosted criterion only focuses on keyword segments in the reference sentences and therefore the resulting systems tend to have reduced keyword miss rates. In the future we plan to introduce the boosting weight to keyword words in hypothesized sentences as well so that both miss and false alarm rates of keywords can be minimized at the same time. Keyword-boosted LM is also a critical research topic.

### 5. Acknowledgements

This collaboration effort uses the IARPA Babel Program Vietnamese language collection release babel107b-v0.7 with the LimitedLP training set.
6. References


