DYNAMIC TUNING OF LANGUAGE MODEL SCORE IN SPEECH RECOGNITION USING A CONFIDENCE MEASURE

Sherif Abdou, Michael Scordilis
Department of Electrical and Computer Engineering, University of Miami
Coral Gables, Florida 33124, U.S.A.

ABSTRACT
Speech recognition errors limit the capability of language models to predict subsequent words correctly. An effective way to enhance the functions of the language model is by using confidence measures. Most of current efforts for developing confidence measures for speech recognition focus on applying these measures to the final recognition result. However, using these measures early in the search process may guide the search to more promising paths. In this work we propose to use a word-based acoustic confidence metric estimated from word posterior probability to dynamically tune the contribution of the language model score. The performance of this approach was tested on a conversational telephone speech corpus and results show significant reductions in recognition error rates.

1. INTRODUCTION
Confidence measures literature is dominated by descriptions of methods for converting the likelihoods and joint probabilities output by traditional HMM based decoders into useful confidence measures [1]. These methods can be classified according to two main approaches. The first is the normalization of these quantities in some way, usually using alternative models, so that they can be comparable across utterances [2]. The second is the training of some application specific post-classifier that takes as input a number of features, including likelihoods, language model probabilities and information from n-best decoding lists [3]. Such usage of confidence measures resembles a final stage filter that provides us with a recognition result with acceptable degree of confidence. A major shortcoming in such utilization of confidence measures is that they are applied to the final output of the decoder but they don't share in the improvement of the search process that leads to that output.

Recognition error analysis for the Switchboard data shows that the language model is more likely to predict a word incorrectly when the previous word is incorrect than if the previous word was correct [4]. About 87% of words proceeded by a correct word were correctly decoded but only 47% of words proceeded by incorrect word was correctly decoded.

To minimize the effect of recognition errors on the decoding of the subsequent parts of the input Rose et al [5] proposed the usage of confidence measures to build a joint acoustic-language model. In that model word level measures of acoustic confidence are extracted and the linguistic context is augmented to include encoded values of that acoustic confidence. To deal with the limited acoustic training corpus, which is generally over an order of magnitude smaller than the text training corpus, a back-off mechanism was used for the N-gram context that occurs infrequently or not at all with a given acoustic confidence.

In this work we take a different approach for using confidence measures in the decoding process. Instead of coding the confidence information into the language model parameters, we use confidence measures to dynamically modify the weight of the language model score during the search process. Based on the confidence of the current sequence of hypothesized words during search, the weight of its prediction is changed as a function of confidence. Incorporating a confidence measure in the decoding process enforce some constraints on the type of measure that can be used. The measure has to be computationally inexpensive so it doesn’t affect the efficiency of the search process. Also it should be extracted synchronously from the on line information that are available during the search process. In previous work we proposed the usage of a confidence measure estimated from rank statistics of the likelihood scores at the frame level [6]. In this work we use a confidence measure estimated from word posterior probability. This measure has the advantage of being derived only from the acoustic observations therefore it can be used as a tuning parameter for the language model score and it doesn’t require the training of alternative models such as the likelihood ratio based measures [2].

In the following parts, section 2 describes how to use confidence measures to guide the search process. Section 3 introduces the used confidence metric. Section 4 includes
experimental results for the proposed approach. The final conclusion and future work prospects are included in section 5.

2. USING CONFIDENCE TO GUIDE THE SEARCH

Given acoustic data $X$ corresponding to an utterance, the maximum a posteriori word sequence $W = w_1 \ldots w_n$ is that which maximizes the conditional probability $P(W|X)$. By Bayes rule

$$P(W / X) = \frac{P(X / W)P(W)}{P(X)} (1)$$

Since $P(X)$ is independent of the hypothesized word sequence, the maximization is simply done over the numerator of (1). In practice, the two components $P(X|W)$ (acoustic model score) and the prior probability of the word sequence $P(W)$ (language model score) have vastly different dynamic ranges due to the incorrect assumption imposed by most HMM approaches of independence between successive frames of data. The language model weight $LW$ was proposed to balance the two quantities [7].

The search problem becomes for the sequence of words that maximizes

$$\text{Score} (W / X) = P(X / W)P(W)^{LW} (2)$$

Given a hypothesized word sequence, $W = w_1 \ldots w_i$, and the corresponding acoustic segment $X^i_1$, the score of a hypothesized extension $w_i \ldots w_j$ is

$$\text{Score} (W^i_{i+1} / X^i_1) = \text{Score} (W^i_j / X^i_1) P(X^i_{j+1} / W^i_{i+1}) P(W^i_{i+1})^{LW} (3)$$

$LW$ is usually used as a static parameter typically determined empirically to optimize the performance. We propose to enhance the functionality of $LW$. Besides being used just for scaling, it can be used as a dynamic parameter that tunes the usage of the language model. Making $LW$ a function of $C(W_{i+1})$, the confidence of the current and history word hypothesis. So (3) will be modified to:

$$\text{Score}(W^i_{i+1} / X^i_1) = \text{Score}(W^i_{i+1} / X^i_1) P(X^i_{j+1} / W^i_{i+1}) P(W^i_{i+1})^{LW(C(W_{i+1}))} (4)$$

If $LW_0$ is the determined static value for the language weight. And since $P(W_{i+1})$ is less than 1, $LW(C(W_{i+1}))$ should be less than $LW_0$ when $C(W_{i+1})$ is high so that predictions by more confident hypotheses will have higher score and inversely $LW(C(W_{i+1}))$ should be greater than $LW_0$ when $C(W_{i+1})$ is low. In this work we used the following functional form:

$$LW(C(W_{i+1})) = LW_0 \frac{2}{1 + \exp(r \ast (C(W_{i+1}) - C_0))} (5)$$

$$C(W_{i+1}) = \frac{\sum_{j<i} C(w_j)}{i + 1} (6)$$

This functional form was chosen so that $LW$ changes smoothly around $LW_0$ with gradient concentrated around the operation point $C_0$ which represents the threshold value associated with the confidence measure used. The constant $r$ controls the rate of change around $LW_0$. If the used language model is a bigram model the history is reduced to just the previous word and expression (6) becomes:

$$C(W_{i+1}) = \frac{C(w_i) + C(w_{i+1})}{2} (7)$$

Figure 1 shows an example for of the variation of $LW$ as a function of $C(W_{i+1})$ for some values of $r$.

![Figure 1: LW as function of C(W_{i+1}), LW_0=6.5, C_0=0.65](image)

3. CONFIDENCE MEASURES BASED ON WORD POSTERIOR PROBABILITIES

Posterior probabilities were first exploited as confidence measures by Renals et al [8] for hybrid recognition systems. Contrary to standard HMM systems that are based on the estimation of probabilistic distribution functions for the observations’ likelihood, hybrid systems utilize neural networks to estimate the HMM states’ probability given an acoustic observation as input. These posterior probabilities are discriminative by nature as they sum up to unity over all HMM states and were one of the major advantages for the hybrid systems. In order to
transform the HMM-based likelihoods into posterior probabilities, Bays’ rule can be applied. Consider we have the likelihood observation function \( p(x/q) \) for a HMM state \( q \) and acoustic observation \( x \), the local posterior state probability \( p(q/x) \) can be calculated from:

\[
p(q/x) = \frac{p(x/q) p(q)}{p(x)} \tag{8}
\]

Ordinary HMM systems don’t have a dedicated model for \( p(q) \), the state priors, or for \( p(x) \), the unconditional acoustic likelihood. \( P(q) \) can easily be estimated by forced alignment of some training data with their reference utterances to the state level.

For the estimation of \( p(x) \) two methods are possible:

On the one hand, a separate pdf, usually called catch all model, can be trained on all the training data to model the general distribution of the acoustic observations. In order to achieve a good resolution, usually large number of mixture components is needed, so that the estimation of \( p(x) \) with this independent modeling is computationally expensive. Kamppari et al[9] proposed a bottom-up clustering method of all the Gaussian components in the \( p(x) \) model that can reduce the computation effort up to 95% with very minor degradation in performance.

On the other hand, \( p(x) \) can be estimated as the sum over all the states’ likelihood functions weighted by the states’ priors according to

\[
p(x) = \sum_q p(q) p(x/q) \tag{9}
\]

Semi-continuous HMM with tied states, such as the one that we used in this work, offer the ideal structure for estimating \( p(x) \) according to (9). Since in this type of systems all pdfs share a common codebook of Gaussian distributions, the computation of \( p(x) \) turns out to be a weighted sum of the Gaussian codebook. Its evaluation is as simple as the evaluation of a single observation input.

So a confidence score for a phone \( q_k \) with hypothesized start time \( n_s \) and end time \( n_e \) can be estimated using the local posterior probability of (8) for the Viterbi state sequence and time duration normalization by:

\[
CM_{\text{post}}(q_k) = \frac{1}{n_e - n_s} \sum_{n_s}^{n_e} \log \left( \frac{p(x^n/q_k^n)}{p(x^n)} \right) \tag{10}
\]

Another confidence measure was proposed in [8] by excluding the state priors and just using the conditional likelihood \( p(x/q) \) scaled by the unconditional acoustic likelihood \( p(x) \).

\[
CM_{\text{SL}}(q_k) = \frac{1}{n_e - n_s} \sum_{n_s}^{n_e} \log \left( \frac{p(x^n/q_k^n)}{p(x^n)} \right) \tag{11}
\]

Word level measures can be derived from the phone level measures either by taking the arithmetic mean or the geometric mean over all the phone constituents of that word. The value of this confidence measure will be in the log domain, which is a zero-centered score where positive scores are good and negative scores are bad. To have a probability interpretation of our scores they need to be mapped to the \([0,1]\) interval. A decision tree was used to do that mapping. Based on the step function defined by the decision tree a piecewise linear mapping function was chosen and applied to the log posterior values.

4. EXPERIMENTS AND RESULTS

Two experiments were conducted. The target of the first one was to evaluate the performance of the different confidence measures that proposed in the previous section. We evaluated \( CM_{\text{post}} \) versus \( CM_{\text{SL}} \) using different alternations in the method of estimating \( p(x) \), using catch all model or using the weighted sum of the states’ likelihood according to (9), and the method of deriving the word level measures, using the arithmetic mean or the geometric mean.

For this experiment we used a test set of 2400 utterances from the DARPA/CUMMMNICATOR domain, telephony conversations for the task of airline, car rental and hotel reservations [10]. For training the catch all model we used a development set of 3500 utterances. We used Sphinx II speech decoder [11]. This decoder uses a semi-continuous acoustic model that consists of 10,000 tied state senones and a class-based trigram language model. For a vocabulary size of 1760 words the baseline system achieved 80.7% recognition accuracy.

Figure (2) shows the receiver operating characteristic (ROC) curve for the 8 different confidence measures and their relative performance.

From the curves of Figure (2) it is clear that \( CM_{SL} \) do better than \( CM_{post} \) leading to the conclusion that the priors do not improve performance. Also the geometric mean is significantly better than the arithmetic mean. This result may be due to the characteristic of the geometric mean, which allows a single low score to pull down the score of the whole word, where as an arithmetic mean can be immune to that single low score. There was no significant difference in performance of the catch all model and the weighted sum method with the advantage of the latter one for not requiring the training of a new model. The best...
performing measure, $G$-$CM_{SL}$–$SUM$, was elected to be used in the second experiment.

The target of the second experiment was to evaluate the performance of the proposed confidence guided search approach. The proposed modification was applied to the first pass of SPHINX II, which is a lexical-tree Viterbi search. Because of the inherited constraint of lexical tree structure for considering language model score at word end frame not at word entry frame, at that time there would be enough information to calculate our confidence metric and dynamically adapt the language weight value.

For this experiment we used a test set of 5000 utterances. Table (1) shows the error rate of the baseline system and the modified one for different values of $r$. The value $r=0$ simulates the baseline decoder with $LW=LW_0$. A relative 4.5% reduction in word error rate was achieved using the guided search modification. The error rate was sensitive to the change in the value of $r$, when $r$ gets large the performance starts to degrade. This can be justified by the increase of the dynamic range of $LW$, which makes it highly sensitive to any small changes in $C(W)$.

![Fig 2: The ROC curves](image)

**Table 1. Error rate for different values of $r$**

<table>
<thead>
<tr>
<th>Decoder Type</th>
<th>$r$</th>
<th>Word Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0</td>
<td>19.30%</td>
</tr>
<tr>
<td>Modified</td>
<td>1</td>
<td>18.43%</td>
</tr>
<tr>
<td>Modified</td>
<td>2</td>
<td>18.42%</td>
</tr>
<tr>
<td>Modified</td>
<td>3</td>
<td>18.47%</td>
</tr>
</tbody>
</table>

5. CONCLUSION AND FUTURE WORK

In this work we used a confidence metric to guide the search in speech recognition. The dynamic tuning of the language model weight parameter proved to be effective for performance improvement. At the end frame of every word hypothesis we make a decision. When we have high confidence, we increase the contribution of the language model score and decrease it for low confidence. We used a non-expensive confidence metric that is calculated synchronously from the available information during the search. This metric proved to be a good predictor for word classification as correct or incorrect hypothesis. Right now, the decision is made depending on the joint confidence of the current and previous word hypotheses. We plan to extend this work for the cases when we have high confidence only for one of these words, we should back off to the unigram language model score not completely reduce the language model score.

6. REFERENCES

[10] [http://communicator.colorado.edu/#downloads](http://communicator.colorado.edu/#downloads)