Background Model Based Posterior Probability for Measuring Confidence

Peng Liu, Ye Tian, Jian-Lai Zhou and Frank K. Soong

Microsoft Research Asia, Beijing, China
{t-peliu,t-yetian,jlzhou,frankkps}@microsoft.com

Abstract

Word posterior probability (WPP) computed over LVCSR word graphs has been used successfully in measuring confidence of speech recognition output. However, for certain applications the word graph is too sparse to warrant reliable WPP estimation. In this paper, we incorporate subword units as background models to generate a subword graph for estimating posterior probability. Experiments on both English and Chinese databases show that syllable background models can repopulate the dynamic hypothesis space for effective computation of confidence measure. The resultant posterior probability confidence measure achieves 94.3% and 95.2% Out-Of-Vocabulary (OOV) word detection / rejection in English and Chinese, respectively. Correspondingly, confidence error rates are at 6.0% and 6.4%, respectively.

1. Introduction

Automatic Speech Recognition (ASR) has made substantial progress in the past few decades. But in real applications, it still encounters various hurdles in delivering robust performance across operating conditions like varying background noise, speaking styles, accents, etc. To deal with this robustness problem, many Confidence Measures (CM) [1] have been proposed to measure the recognition reliability.

Generally, the methods proposed for computing CM can be grouped into three categories [2]. In the first category, CM computing methods are based on so-called predictor features [3]. But until now, the predictor features are not good enough to separate well the correctly recognized words from misrecognized ones. In the second category, the CM is treated as a hypothesis testing problem [4]. However, it can be rather difficult for some cases to model complex alternative hypotheses. The last category, the posterior probability based CM, is estimated in the framework of standard Maximum A Posterior (MAP) [5]. Posterior probability is a good candidate for CM for its good dynamic range, i.e., between 0 and 1, and its statistically sounding nature to measure the quality of decision. Superior performance has been demonstrated by using the posterior probability [5][6].

In general, it is not easy to estimate the posterior probability precisely. Practically, certain assumptions need to be made and some approximations need to be established. The word graph based methods [5], which estimates the posterior probability by the forward-backward algorithm [5], does the job fairly well.

However, for some Context Free Grammar (CFG) based applications, e.g., command control, the word graph generated by the decoder can be too sparse to warrant reliable computation of the posterior probability. Also, for such applications, it is very important for the ASR system to detect / reject the Out-Of-Vocabulary (OOV) words besides to reject misrecognized in-vocabulary words reliably. In a sparse word graph, the best path commonly becomes a dominating one and ends up with a high posterior probability regardless of its correctness.

In this paper, we propose to use subword units as background models in conjunction with the decoding CFG to generate a reliable background model graph for measuring confidence. Architecture of background model based confidence measure is presented, and two subword model sets, phonemes and syllables, are investigated. The performance is measured by rejection and OOV detection on both Chinese and English databases.

The rest of the paper is organized as follows: In section 2, we give a discussion of word graph sparseness in a CFG application and introduce our background model based posterior probability. In section 3, we describe the databases and CM testing schemes. The experiments are given in section 4. In section 5, we give our conclusions.

2. Background model based posterior probability

2.1. Word graph sparseness in CFG based ASR

Word Posterior Probability (WPP) has been successfully applied to measuring confidence in Large Vocabulary Continuous Speech Recognition (LVCSR) [5][6]. Given a feature stream $\boldsymbol{o}_1^T$ of observation from frame 1 to *T*, the posterior probability of a word $[w;s,t]$ with initial frame *s* and final frame *t* is:

$$p\left([w;s,t]\middle|\boldsymbol{o}_1^T\right)=\sum_{\substack{\forall I,[w';s',t']_1^I,\\ \exists i,1\le i\le I,[w_i';s_i',t_i']=[w;s,t]}}\frac{\prod_{n=1}^{I}p\left(\boldsymbol{o}_{s_n'}^{t_n'}\middle|w_n'\right)}{p\left(\boldsymbol{o}_1^T\right)} \quad (1)$$

where $W=[w';s',t']_1^I$ denote any legal word sequence with length *I*. The denominator representing the probability of observations can be calculated as follows:

$$p\left(\boldsymbol{o}_1^T\right)=\sum_{W}p\left(\boldsymbol{o}_1^T\middle|W\right)p(W) \quad (2)$$

Obviously, we cannot take all possible word sequence into consideration. In LVCSR application, a word graph *G* generated by the decoder with a beam-width is rich enough to contain many likely hypothesis. Based on the word graph, the acoustic probability can be approximated. In practice, WPP is calculated by considering all the hypotheses with the same word *w* identity and time overlapped with interval (*s*, *t*), because they can be regarded as reappearances [6]:

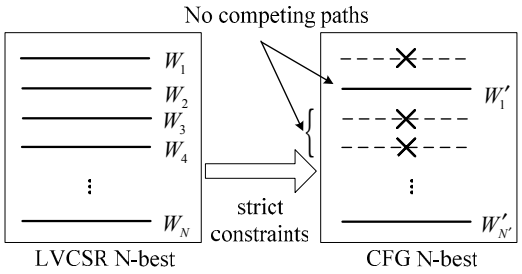

Figure 1. A schematic diagram of word graph sparseness

$$p\left(\left[w;s,t\right]\middle|\boldsymbol{o}_1^T\right)\approx \sum_{\substack{\forall I,[w';s',t']_1^I\subseteq G,\\ \exists i,1\le i\le I,w'_i=w,(s'_i,t'_i)\cap(s,t)\neq\varnothing}} \frac{\prod_{n=1}^{I} p\left(\boldsymbol{o}_{s'_n}^{t'_n}\middle|w'_n\right)}{p\left(\boldsymbol{o}_1^T\right)} \quad (3)$$

However, in some CFG constrained ASR applications, the lexical and language model constraints can limit the number of hypotheses. As a result, WPP cannot be estimated reliably. The concept can be clarified from the viewpoint of N-best hypotheses. Given an utterance, let W_1, W_2, ..., W_N represents the corresponding top N-best string hypotheses extracted from a word graph generated in a LVCSR decoder. If N-grams in LVCSR are replaced by the stricter CFG grammar, some candidates in N-best list will be eliminated, as illustrated in figure 1. In many cases, we have $p\left(\boldsymbol{o}_1^T\right)\approx p\left(\boldsymbol{o}_1^T\middle|W_1'\right)p\left(W_1'\right)$ and $p\left(W_1'\middle|\boldsymbol{o}_1^T\right)\approx 1$, or equivalently, the best path becomes the dominant one, even if it is incorrect. Hence, the word posterior probability is no longer reliable for measuring confidence due to the *sparse* graph.

2.2. Background model graph for measuring confidence

To alleviate this graph sparseness problem, we need to recover the string candidates eliminated by the stricter decoding constraints in CFG or to refill the search space by appropriate hypotheses. One approach is to use another LVCSR decoder to generate a separate word graphs, but it may be too time-consuming and impractical.

However, for posterior probability calculation, we don't need the exact word sequences. It is enough if we can obtain those pronunciation sequences that can approximate the input utterance in the dynamic HMM space. That is to say, it is necessary to introduce qualified competitors for the dominating path. So we propose to use some generalized background models in the decoder. In this decoder, N-gram at background model level can be considered. In this paper, since CFG is used as the decoding constraints, we use a background model looping in the background decoder. The graph generated by the background decoder is named as *background model graph*. Based on this graph, Model based Posterior Probability (MPP) can be calculated for measuring confidence.

The system architecture for MPP is depicted in figure 2. Besides a conventional decoder, another background decoder directed by a free background model loop is established.

2.3. Word sequence normalization and background model graph penalization

The best word sequence of I words $W=\left[w;\tau,t\right]\big|_1^I$ is need to be

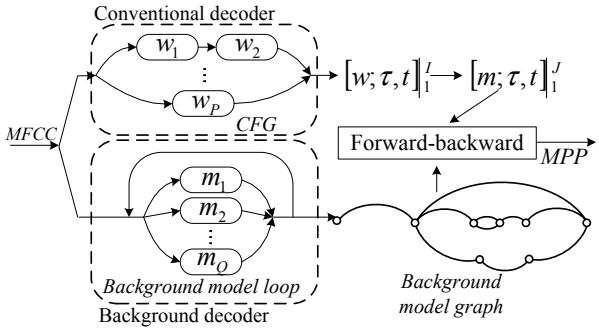

Figure 2. System architecture for MPP

first decomposed into a corresponding model sequence before its posterior probability can be calculated via the background model graph. The word sequence is alternatively represented by the model sequence $M=\left[m;\tau,t\right]\big|_1^J$, where J is the corresponding model, sequence length, as illustrated in figure 2.

Moreover, background model graphs need to be appropriately penalized for making the correct outputs distinctive and rejecting misrecognized output. For each arc in the background model graph, penalty is imposed as:

$$\log p'\left(\boldsymbol{o}_\tau^t\middle|m\right)=\log p\left(\boldsymbol{o}_\tau^t\middle|m\right)-penalty\cdot\left(t-\tau+1\right) \quad (4)$$

Finally, MPP is normalized by the total number of J, to facilitate a universal threshold for rejection:

$$p\left(M\middle|\boldsymbol{o}_1^T\right)=\left\{\prod_{j=1}^{J} p\left(\left[m_j;s_j,t_j\right]\middle|\boldsymbol{o}_1^T\right)\right\}^{\frac{1}{J}} \quad (5)$$

2.4. Background model selection

There are two criteria to select appropriate background model units: 1) They should characterize the utterance in HMM space well; 2) A word sequences can be decomposed into a sequence of model units unambiguously. Accordingly, we have adopted some subword units as exact models rather than ordinary general filler models. Two candidate model sets: phonemes and syllables in both English and Chinese are investigated

2.4.1. Phoneme based background models

There are only about 40 phonemes in English and around 70 toneless syllable initials and finals in Chinese, so they can be a natural choice for subword based background model sets.

2.4.2. Syllable based background models

It is natural to use syllable background models in syllable based languages like Chinese where there are only slightly over 400 syllables in the whole inventory. But for English where number of syllables exceeds 15,000, it is rather cumbersome to use all in the background model loop. To reduce its size, we can cluster syllables into a smaller set or prune those syllables with lower frequencies.

First, we cluster similar syllables together by considering the following rules: /b/→/p/, /g/→/k/, /d/→/t/, /ŋ/→/n/, /z/→/s/, /ʒ/→/ʃ/, /v/→/f/, /dʒ/→/tʃ/, /ð/→/θ/, where we cluster consonants with similar place of articulations together, but level vowels intact.

Also we calculate the syllable frequency in a large dictionary of 221,268 entries, and use the frequency count to prune out syllables of low occurrence.

3. Experimental setup

Our experiments were performed on two test databases: an English database of 5838 utterances recorded by 92 speakers and a Chinese database of 548 utterances recorded by 548 speakers [7]. Both databases were recorded with close-talking headset microphones.

39-dimensional MFCCs features are used to train HMM triphone models. All experiments were performed with HTK package.

3.1. CFG recognition grammars

The CFGs are built with all legal phrases arranged in parallel, as illustrated in the left part of figure 4. In English, the CFG grammar consists of 220 computer command words; In Chinese, the CFG grammar consists of 154 Chinese names of 2 to 3 Chinese characters. All phrases are short and the average phrase lengths in terms of syllables are 2.47 and 2.57, respectively in English and Chinese databases. Since every phrase can be regarded as either a word or a compound word integrally. We shall call it a word from now on.

The proposed CM approach was tested in the vocabulary word rejection where we use only half of the vocabulary to make a partial CFG, as illustrated in figure 3. In the partial CFG, half of the test utterances are OOV words and they should be rejected.

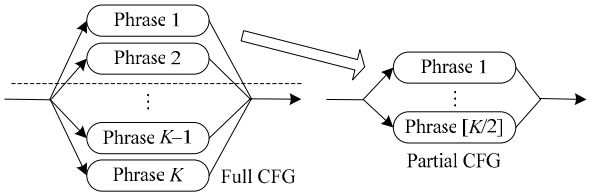

Figure 3. CFGs used in experiments

3.2. Rejection performance measure

To evaluate the rejection performance, we use the Confidence Error Rate (CER) defined as:

$$\mathrm{CER} = \frac{\#\,\text{false acceptances} + \#\,\text{false rejections}}{\#\,\text{recognized words}}$$

The definition can be clarified as follows: Given the measured confidence and a rejection threshold, the output will be accepted as correct or rejected as false. If an incorrectly recognized word is accepted, it is a false acceptance. On the other hand, if a correctly recognized word is rejected, it is a false rejection. CER is the number of the incorrect tags divided by the total number of recognized words, which is equivalent to:

$$\mathrm{CER} = \mathrm{FAR} + \mathrm{FRR} \qquad (6)$$

where FRR the false rejection rate; and FAR denotes the false acceptance rate.

CER can be further decomposed into:

$$\mathrm{CER} = P_{OOV}\,(1\text{-}\mathrm{ODR}) + (1 - P_{OOV})(\mathrm{FRR}_{IV} + \mathrm{FAR}_{IV}) \qquad (7)$$

where P_{OOV} is the OOV word rate of all the test utterances; ODR, the OOV word detection rate; and the subscript IV denotes In-Vocabulary. The overall measure of FRR and FAR can easily be calculated given the variables in (7).

4. Experimental results

4.1. Recognition tests

The recognition test results on the two databases are listed in table 1. In the experiments for Word Error Rate (WER), the full CFG grammars were used, i.e., no OOV words, and in those for the last two columns, only partial CFG grammars were used and about half of the input utterances are OOV words, which are shown as P_{OOV}.

Table 1. Recognition test results

#Utter	WER(%)	P_{OOV}(%)	WER_{IV}(%)
5838	3.85	48.53	1.83
548	5.47	49.64	3.62

4.2. Syllable set selection in English

To determine the syllable background model set in English, we tried to reach a trade-off between CER and the set size experimentally. In this part, partial CFG grammars were used.

We used a degenerated background model graph, where only the best path in the graph was saved. The rejection threshold was fixed at 0.5, and the penalty was experimentally tuned to its best value.

The results shown in figure 4 suggested that: 1) More syllables don't always lead to a better performance, and a set of 1,000 syllables set is adequate for measuring confidence effectively. 2) Rule based clustering does not seem to be helpful in clustering syllables. The result of the clustered syllable set performs worse than that of the unclustered set.

Syllable based posterior probability performs better than phoneme based one. It indicates that a syllable loop structure can describe the dynamic evolution of input utterance dynamically more succinctly than a phoneme loop one. The vowel nucleus in each syllable serves as a decoding anchor to make a good guardian "cohort" around the correct hypothesis for rejection.

4.3. Rejection tests

We compare the performances of MPP and WPP based CMs by rejection tests using partial CFGs.

In the proposed rejection approach, MPP of correct word is usually close to 1, while that of incorrect word is close to 0, as mentioned in 2.3, so the CER performance is quite insensitive to the rejection threshold. Observing that for incorrect words, MPP distributes more widely because parts of the output model sequence may be correct, the threshold is set to 0.9 in our experiments.

The value of WPP is commonly close to 1 due to the word graph sparseness, so the threshold of WPP based rejection is also set to 0.9. The acoustic scaling factor is set to an empirical value of 0.075.

In MPP tests, the background model graphs were generated by a decoder with 4 tokens. In WPP tests, graphs were generated by using a decoder with 20 tokens. For syllable

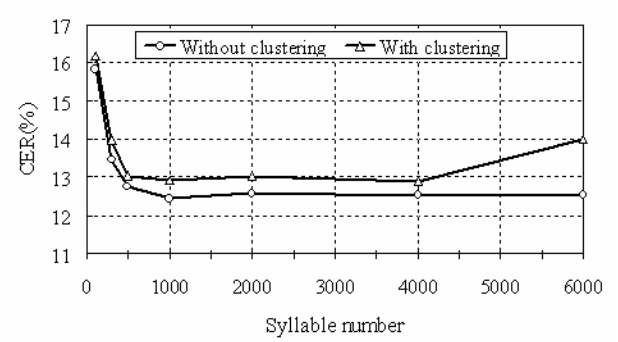

Figure 4. Rejection performances versus syllable number

based MPP in English, 1,000 most frequent syllables were used. The penalties were tuned to their best values.

Table 2. Rejection performance on the English database

	FAR_{IV}(%)	FRR_{IV}(%)	ODR(%)	CER(%)
WPP	1.40	0.37	12.43	43.41
Pho-MPP	1.06	3.73	84.82	9.83
Syl-MPP	0.60	5.62	**94.25**	**6.00**

Table 3. Rejection performance on the Chinese database

	FAR_{IV}(%)	FRR_{IV}(%)	ODR(%)	CER(%)
WPP	3.26	0.72	5.15	48.55
Pho-MPP	0.72	7.97	94.12	7.29
Syl-MPP	0.72	7.25	**95.22**	**6.39**

The rejection test results are listed in tables 2 and 3. As shown in the two tables, in CFG based ASR, WPP is incapable of measuring confidence effectively even 20 tokens are used, but MPP delivers a much more robust performance. As a result, the OOV detection and overall rejection performances are improved significantly. It is observed that the FRR_{IV} of WPP is lower than those of MPPs. This is due to the fact that the overall CERs have been minimized with respect to the rejection threshold, which implies that WPP is usually too indecisive to reject any utterances due to word graph sparseness.

4.4. Graph density

Additionally, the relationship between background model graph density and CER is studied. The corresponding results are shown in figure 5, where CER is plotted against the token number used in decoding, which in turn yields different graph density. When token number equals 1, the background model graph is a minimal degenerated one. From the figure we observe that the CER performance is not very sensitive to the graph density. The background model based posterior probability works as a robust confidence measure over a wide range of graph density.

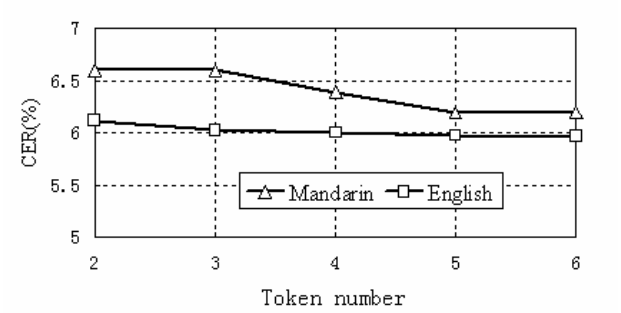

Figure 5. CER versus background model graph size

5. Conclusions

In this paper, posterior probability based confidence measure is used for verifying speech recognition output with a relatively sparse CFG. Specifically subword based background models are incorporated to construct subword graphs where confidence of recognized words can be accessed reliably. Both phones and syllables are investigated for making background models in Chinese and English recognition tasks. Using syllables as background models, 94.3% and 95.2% of the OOV words are successfully detected and rejected in Chinese and English recognition tests, respectively. The corresponding confidence error rates are at 6.0% and 6.4%.

6. References

[1] C.-H. Lee, "Statistical confidence measures and their applications", *Proc. ICSP 2001*, Daejeon, Korea, August, 2001.

[2] H. Jiang. "Confidence measures for speech recognition: a survey", *Speech Communication*, 45(4): 455-470, 2005.

[3] M. C. Benitez, A. Rubio and A. Torre, "Different confidence measures for word verification in speech recognition", *Speech Communication*, 32(1-2):79-94, 2000.

[4] R. C. Rose, B.-H. Juang and C.-H. Lee, "A training procedure for verifying string hypothesis in continuous speech recognition", *Proc. ICASSP 1995*: 281-284, 1995.

[5] F. Wessel, R. Schluter, K. Macherey and H. Ney, "Confidence measures for large vocabulary continuous speech recognition", *IEEE Trans. Speech and Audio Proc.*, 9(3):288-298, 2001.

[6] W. K. Lo, F. K. Soong and S. Nakamura, "Generalized posterior probability for minimizing verification errors at subword, word and sentence levels", *Proc. ISCSLP 2004*:13-16, Hong Kong, December, 2004.

[7] http://www.speecon.com.

[8] S. J. Young, N. H. Russell and J. H. S. Thornton, *Token passing: a simple conceptual model for connected speech recognition systems*, Technical Report, Dept. Engineering, Cambridge University, July 31, 1989.