GENERALIZED POSTERIOR PROBABILITY FOR MINIMUM ERROR VERIFICATION OF RECOGNIZED SENTENCES

Wai Kit LO *Frank K. SOONG*[1]
{waikit.lo, frank.soong}@atr.jp

Spoken Language Translation Research Labs, ATR, Kyoto, Japan

ABSTRACT

Generalized posterior probability (GPP) is investigated in this paper as a statistical confidence measure for verifying recognized sentences of a large vocabulary continuous speech recognition system (LVCSR). We optimize the GPP by training the exponential weights of the acoustic and language models and decision threshold to minimize total verification errors. Two utterance level confidence measures: generalized utterance posterior probability (GUPP) and product of generalized word posterior probabilities (GWPP) of component words in a string hypothesis are tested. When evaluated on the Chinese Basic Travel Expression Corpus (BTEC), 47.9% and 53.9% relative improvement of utterance confidence error rate (CER) have been obtained for the GUPP and product of GWPPs confidence measures, respectively.

1. INTRODUCTION

Current state-of-the-art speech recognition technology is not robust to changes such as noise, channel mismatch, speaker variability, etc. Verifying recognition output of an LVCSR is then necessary. By assessing the confidence of speech recognition results, appropriate actions can then be taken. This will improve the overall performance, subjectively and objectively, of a spoken language system (e.g., a spoken dialogue system or an automatic speech translation system).

Confidence measures are useful for improving performance of spoken language systems by assessing reliability of recognition output. For instance, a spoken dialogue system only needs to confirm recognized words of low confidence. Recognized words of high confidence are accepted without further confirmation to avoid unnecessary dialogue turns. Another application is automatic speech translation. Confidence measures can be used to weight recognized words, as well as utterances, to facilitate appropriate translations.

There have been various approaches proposed for measuring confidence of speech recognition output. They can be roughly classified into three categories: i) feature based; ii) explicit model based; and iii) posterior probability based. Feature based approaches [1] try to assess the confidence according to selected features (e.g., word duration, part-of-speech, acoustic and language model back-off, word graph density, etc.) using some trained classifiers. Explicit model based approaches employ a candidate class model with competing models [2] (e.g., an anti-model or a filler model) and a likelihood ratio test is applied. The posterior probability based approach tries to estimate the posterior probabilities of a recognized entity (e.g., a word) given all acoustic observations [3, 4].

In this study we generalize the posterior probability approach to word and utterance levels. Verification experiments were performed on the Chinese Basic Travel Expression Corpus (BTEC) [5]. Since every sentence is contained in a recorded utterance in our corpus, verification of recognized sentences is then performed using the utterance level posterior probability and product of word level posterior probabilities.

2. UTTERANCE VERIFICATION USING GENERALIZED POSTERIOR PROBABILITIES

Generalized posterior probability (GPP) is a probabilistic confidence measure for verifying optimally the recognized entities at different levels, e.g., subword, word and utterance [6]. It was first applied to verification at the word level under various conditions [4, 7, 8].

In continuous speech recognition, the word posterior probability (WPP) can be computed by summing the posterior probabilities of all string hypotheses in the search space bearing the focused word, *w*, starting at time *s* and ending at time *t*, given as

$$p\left([w;s,t\ ]\,|\,x_1^T\right)=\sum_{\substack{\forall M,[w;s,t]_1^M\\ \exists n,1\le n\le M\\ w=w_n,\,s=s_n,\,t=t_n}}\frac{\prod_{m=1}^{M}p\left(x_{s_m}^{t_m}\mid w_m\right)\cdot p\left(w_m\mid w_1^{M}\right)}{p\left(x_1^T\right)} \quad (1)$$

where a word hypothesis is defined by the corresponding triple, *[w; s, t]*; $p(x_s^t|w_m)$ is the acoustic likelihood; $p(w_m|w_1^M)$, the language model likelihood; x_s^t, the sequence of acoustic observations; *M*, the no. of words in a string hypothesis; $p(x_1^T)$, the probability of the acoustic observations; *T*, the length of the complete acoustic observations. WPP can be computed for each recognized

[1] This author is currently with Microsoft Research Asia, Beijing, China.

word, without using any additional models (e.g., anti-models) from a word graph or N-best list generated during the decoding process.

Generalized word posterior probability (GWPP) is a generalization of WPP to take into account of three issues in computing WPP:

a) Reduced search space: Search space in recognition is almost always pruned to make the search tractable. A reduced search space (e.g., word graph or N-best list) is used when computing GWPP, including the acoustic observation probability, $p(x_1^T)$.
b) Relaxed time registration: A word is defined as a triple by the *word identity*, its *starting* and *ending time*. The starting and ending time of a word is affected by various factors like the pruning threshold, model resolution, noise, etc. It is therefore desirable to relax the time registrations for deciding whether the same word reappears in a different string hypothesis. In GWPP, words in the search space with the same identity and overlapping in time are considered as reappearances.
c) Reweighted acoustic and language model likelihoods: In continuous speech recognition, assumptions are made to facilitate efficient parametric modeling and decoding process. Incompatibilities also exist among components in the models. They include:
   - Difference in dynamic range: In Gaussian mixture models, acoustic likelihoods obtained from pdf have an unbounded dynamic range. Language model likelihoods, based on the statistical n-grams, however, lie between 0 and 1.
   - Difference in frequency of computation: Acoustic likelihoods are computed every frame, while language model likelihoods are computed once per word.
   - Independence assumption: Neighbouring acoustic observations are assumed to be statistically independent in computing the acoustic likelihoods.
   - Reduced search space: The full search space is always pruned to a word graph (or an N-best list).

The acoustic and language models weights are labeled as α and β, respectively. When reweighting the acoustic and language model likelihoods, these weights are jointly trained to optimize word verification performance. The corresponding GWPP is:

$$p([w;s,t]\mid x_1^T)=\sum_{\substack{\forall M,[w;s,t]_1^M \\ \exists n,1\le n\le M \\ w=w_n \\ (s_n,t_n)\cap(s,t)\neq\phi}} \frac{\prod_{m=1}^{M} p^{\alpha}\left(x_{s_m}^{t_m}\mid w_m\right)\cdot p^{\beta}\left(w_m\mid w_1^M\right)}{p\left(x_1^T\right)} \quad (2)$$

GWPP has been demonstrated to achieve robust word verification performance at different search beam widths [7], signal-to-noise ratios [8], etc. These are clear evidences to show the appropriateness and effectiveness of this confidence measure in verifying recognized words.

2.1. Generalized Utterance Posterior Probability

For utterance verification, a generalized utterance posterior probability (GUPP) is defined similarly. Deciding whether the utterance is correctly recognized does not require a sharp focus on the specific misrecognized word components when compared to word verification. It only measures the statistical confidence of the whole hypothesized utterance.

Definition of the GUPP is similar to that of its word counterpart (GWPP), where the reduced search space, reweighted acoustic and language model likelihoods are similarly applied. However, the time registration relaxation of beginning and ending of a utterance is no longer necessary since all string hypotheses share the same utterance boundaries. As a result, the GUPP is defined as

$$GUPP = \frac{pa^{\alpha}\cdot pl^{\beta}}{\sum_{\forall\ hypotheses} pa^{\alpha}\cdot pl^{\beta}} \quad (3)$$

where *pa* is the acoustic model score and *pl*, the language model score of the hypothesized utterance; α and β, the acoustic and language model weights, respectively. The resultant GUPP is between 0 and 1 where a value close to 1 implies higher confidence on the correctness of an utterance.

Although application of the language model scaling factor is commonly used in LVCSR, optimal language model scaling can only change the ranking of hypotheses. In order to optimally reject incorrectly recognized utterances, both acoustic and language model weights are jointly trained to minimize rejection errors.

2.2. Product of GWPPs

A new way to measure the confidence of a recognized utterance is based on the joint confidence of all component words in the recognized string. GWPP of a word is a measure of its correctness, or a probability of a binomial distributed "word correct" event. The probability of an "utterance correct" event is then the product of all probabilities of component "word correct" events, assuming that all word events are statistically independent. The product of GWPPs of all recognized words in a recognized utterance is therefore proposed as a utterance level confidence as given below

$$CF_{sentence} = \prod_{i=1}^{M} GWPP(w_i) \quad (4)$$

where M is the total number of words in the string hypothesis.

3. EXPERIMENTAL SETUP

3.1. Speech recognition

The LVCSR used in this study is the ATR speech recognition system [9], running in multi-pass with a word bigram language model and a 16k word lexicon. The

feature parameters included 12 MFCC, 12 ΔMFCC and Δpower. Word graphs were generated and then rescored using a word trigram language model to obtain the final recognition output. The word recognition accuracy is 91%.

3.2. Corpus

The speech corpus used for evaluation was a large vocabulary, continuous, read Chinese speech database in the Chinese Basic Travel Expression Corpus (BTEC) [5]. It was compiled and collected for a travel domain speech-to-speech translation project. We used two subsets of utterances as the development and test sets. Both speakers and utterances in these sets are mutually exclusive. We summarize the information in Table 1.

	Development	Test
# speakers	4 M + 4 F	16 M + 16 F
# utterances	841	3,437
# words	4,030	16,781
# characters	6,327	25,939

Table 1. Summary of the development and test sets selected from the Chinese BTEC corpus and used in our experiments.

3.3. Verification

Generalized posterior probabilities at word and utterance levels were computed separately and corresponding optimal values for the acoustic and language model weights (α, β) and rejection thresholds were determined from the development set by a full grid search of the total error surface. Other efficient search algorithms (e.g., steepest descent, Downhill Simplex search) for parameter optimization have also been proposed in [7]. These optimized parameters thus trained in the development set were then used in the test set.

3.4. Evaluation Measure

Evaluation of verification performance was based on a normalized total verification errors — confidence error rate (CER) [3]. Total errors include false acceptance (FA) of incorrectly recognized units and false rejection (FR) of correctly recognized units. This sum is then normalized by the number of recognized units in the LVCSR output.

$$\mathrm{CER} = \frac{\#\text{false acceptance} + \#\text{false rejection}}{\#\text{recognized units}} \times 100\% \quad (5)$$

CER is 1 when all correctly recognized units are rejected and all incorrectly recognized units (insertions and substitutions) are accepted. A CER of 0 means that all units are correctly verified.

In our experiments, a baseline was also used for performance comparison. It was obtained by accepting all recognition output without any rejection.

4. RESULTS AND DISCUSSIONS

The total verification error contours at various acoustic and language model weights are shown in Figure 1 and 2 for word and utterance, respectively. The coarse scale plots show the contours of total errors over the full range of parameters. Fine scale contours of lower error regions are shown in a smaller range.

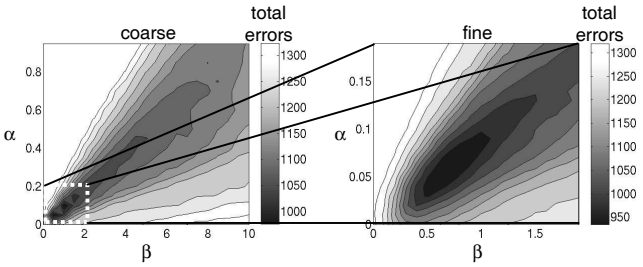

Figure 1. Total error surfaces (test set) for word verification using GWPP. The coarse scale plot shows equal error contours at different α and β values. Optimal parameters are determined using the fine scale plot.

Figure 1 shows the total error contours when word verification is carried out using the GWPP. In general, better verification performance (darker region) is found near the lower left corner. As mentioned in [4, 7], when larger values of α and β are used, more emphasis is put on higher ranked hypotheses. The smaller α and β are, the more hypotheses are taken into account. Therefore, small values of α and β imply more hypotheses are taken into consideration when computing the GPP making it more reliable. In the extreme case, when both α and β are set to zero, all hypotheses in the reduced search space are taken into account equally by simply counting reappearances of the focused word.

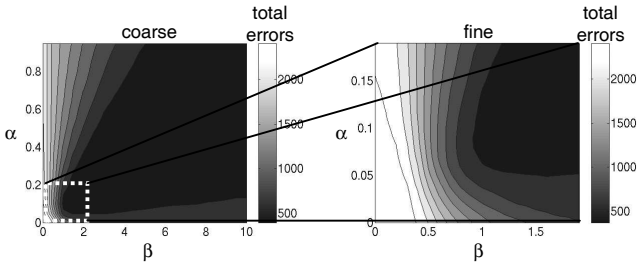

Figure 2. Contour plots of total errors (test set) for utterance verification using the generalized utterance posterior probability.

The total error contours for utterance level verification are depicted in Figure 2. It is observed that the number of errors is very large along the y-axis where the language model weight is zero. Similar phenomenon is observed when the acoustic model weight is zero, or along the x-axis. These imply that neither the acoustic nor the language model score can be ignored when assessing the confidence of a recognized utterance using GUPP. The best verification performance is obtained when α=0.16 and β=1.8. Contrary to the case of word verification using GWPP, the number of verification errors at the origin, (0, 0), is very large. This is because recognized utterances do not reappear in the search space (i.e., single occurrence).

As a result, verification by just counting the reappearance is not reliable at the utterance level.

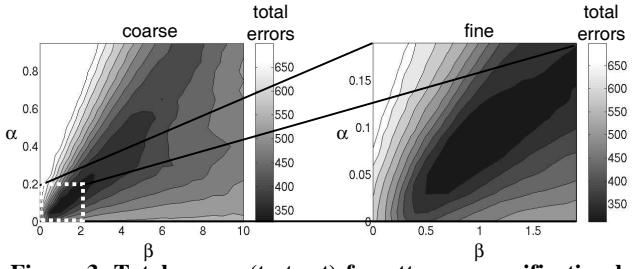

Figure 3. Total errors (test set) for utterance verification by using the product of GWPPs from component words.

When the product of GWPPs of component words is used as the confidence measure for utterance verification, the total verification error contours are shown in Figure 3. It is observed that for verification at utterance level using product of GWPPs, the error contours are more similar to that of GWPP-based word verification (c.f. Figure 1). The optimal region lies close to the origin, where both α and β are small. The best verification performance is obtained when α=0.07 and β=0.7, where small values of α and β are preferred as explained before.

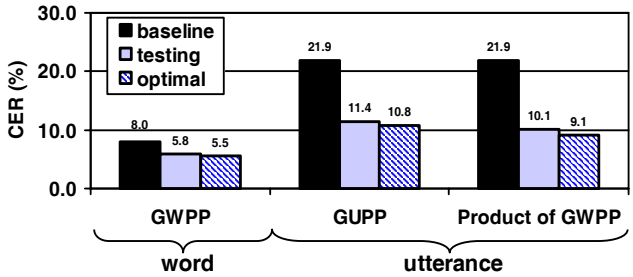

Figure 4. Verification performance in CER at word and utterance levels. Utterance verification using GUPP and product of GWPPs are shown together.

Figure 4 shows the word and utterance level verification performance using the GPP approach. It is observed that at the utterance level, absolute value of the CER is higher. This is because utterance recognition accuracy is lower than word accuracy, hence a higher baseline CER. By applying verification using GPP, relative improvement in CER at utterance level using GUPP is higher (47.9%) than that of word level using GWPP (27.5%). Furthermore, using product of GWPPs of component words in utterance verification can further reduce the already low CER. Comparing with the baseline, a 53.9% relative improvement in CER is obtained. More importantly, results shown in Figure 4 confirm that parameters (α, β and threshold) determined from the development set achieve a verification performance very close to the optimal performance, which is the upper bound obtained from closed test set tuning.

The performance improvement achieved by the GUPP over the baseline hinges on the fact that by proper weighting of the acoustic and language model likelihoods of the utterance hypotheses, utterance verification errors can be reduced. Furthermore, by considering the reappearances of component words with relaxed time registrations in the product of GWPPs, further improvement in utterance verification performance is achieved. Product of GWPPs of all component words also conforms well to the notion that an utterance is correct when all of its component words are correct.

5. SUMMARY

Optimal verification of recognition output at word and utterance levels is investigated by using the generalized posterior probability. Using the product of GWPPs of component words as a new confidence measure further enhances verification at utterance level. Experimental results showed consistent verification performance improvement when parameters obtained from the development set are used in the test set for evaluation. Relative improvement of verification performance obtained at utterance levels using GUPP is 47.9%. When the product of component GWPPs of all words in a string hypothesis is used, relative improvement in utterance verification performance is further increased to 53.9%

6. ACKNOWLEDGEMENTS

This research was supported in part by the National Institute of Information and Communications Technology.

7. REFERENCES

[1] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," *Proc. EuroSpeech1997*, pp.827-830.

[2] M. G. Rahim, C. H. Lee, and B. H. Juang, "Discriminative utterance verification for connected digits recognition," *IEEE Trans. SAP*, vol. 5, 1997, pp.266-277.

[3] F. Wessel, R. Schluter, K. Macherey, and N. Hermann, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Trans. SAP*, vol. 9, 2001, pp.288-298.

[4] F. K. Soong, W. K. Lo, and S. Nakamura, "Generalized word posterior probability (GWPP) for measuring reliability of recognized words," *Proc. SWIM2004*.

[5] H. Kashioka, "Grouping synonymous sentences from a parallel corpus," *Proc. LREC2004*, pp.391-394.

[6] W. K. Lo, F. K. Soong, and S. Nakamura, "Generalized posterior probability for minimizing verification errors at subword, word and sentence levels," *ISCSLP2004*, pp.13-16.

[7] F. K. Soong, W. K. Lo, and S. Nakamura, "Optimal acoustic and language model weights for minimizing word verification errors," *Proc. ICSLP2004*.

[8] W. K. Lo, F. K. Soong, and S. Nakamura, "Robust verification of recognized words in noise," *Proc. ICSLP2004*.

[9] T. Shimizu et al., "Spontaneous dialogue speech recognition using cross-word context constrained word graph," *Proc. ICASSP1996*, pp.145-148.