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

International Symposium on Chinese Spoken
Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

GENERALIZED POSTERIOR PROBABILITY FOR MINIMIZING VERIFICATION ERRORS AT SUBWORD, WORD AND SENTENCE LEVELS

Wai Kit LO *Frank K. SOONG* *Satoshi NAKAMURA*
{waikit.lo, frank.soong, satoshi.nakamura}@atr.jp

Spoken Language Translation Research Labs, ATR, Kyoto

ABSTRACT

Generalized posterior probability, a statistical confidence measure, is tested in this study for verifying optimally the recognized units at the subword, word and sentence levels. We developed the generalized posterior probability by analyzing the exponential weights of the acoustic and language model scores to minimize the total verification errors at different unit levels. Experimental results have demonstrated the effectiveness of this generalized confidence measure for verifying Chinese LVCSR output. The Chinese Basic Travel Expression Corpus (BTEC) is used for evaluation and the relative improvement of confidence error rate (CER) over the baseline performance is 47.76% for sentences, 27.31% for words and 4.64% for subwords.

1. INTRODUCTION

The current state-of-the-art speech recognition technology is not robust to changes such as noise, channel mismatch, speaker variability, etc. Verification, selective acceptance or rejection, of the recognition output of a large vocabulary continuous speech recognition (LVCSR) system is then necessary. By assessing the confidence of speech recognition results properly, appropriate actions can then be taken. This will improve the overall performance 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 both subjectively and objectively. For example, only recognized words with low reliabilities need to be confirmed by a machine prompt via a dialogue. On the other hand recognized words with high reliabilities can be accepted without confirmation to reduce the number of dialogue turns. In an automatic speech translation system, we can use the confidence measures to weight corresponding reliabilities of recognized words 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., word) given all the acoustic observations [3, 4].

In this study the generalized posterior probability is extended from word to subword and sentence levels for verification of recognized subwords and sentences in an LVCSR. The approach is tested on a Chinese database.

2. GENERALIZED POSTERIOR PROBABILITY

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

In continuous speech recognition, the conventional word posterior probability (WPP) is 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

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

where a word hypothesis is defined by the corresponding triple, *[w; s, t]*; x_s^t is 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 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

0-7803-8678-7/04/$20.00 ©2004 IEEE

reduced search space (e.g., word graph or N-best list), rather then the original full search space, is used when computing the GWPP, including the acoustic observation probability, $p(x_1^T)$ (see Eqn. 1).

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, a by-product of the search, 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 with the same identity and overlapping in time registrations are considered as reappearances.

c) Reweighted acoustic and language model likelihood: In continuous speech recognition, assumptions are made to facilitate efficient parametric modeling and decoding process. Also incompatibilities among the components in the recognition process exist. They include:
   - Difference in dynamic range: In theory, acoustic likelihoods computed by using continuous Gaussian mixture probability density functions have an unbounded dynamic range. The language model likelihoods, if based on the statistical n-grams, lie between 0 and 1.
   - Difference in the frequency of computation: Acoustic likelihoods are computed every frame but language model likelihoods are computed only 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 almost always pruned. A word graph or an N-best list of string hypotheses is used.

In order to compensate the above discrepancies, the acoustic and language model weights are jointly adjusted to optimize the word verification performance and a generalized word posterior probability (GWPP) is thus obtained. The exponential weights of the acoustic and language models are labeled as α and β, respectively. The corresponding GWPP is defined as

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

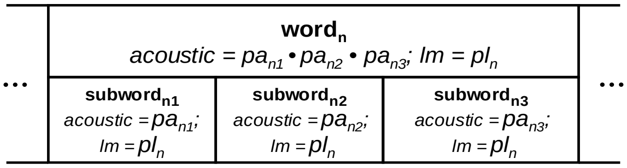
It has been demonstrated that GWPP achieves robust word verification performance at different search beam widths [5], signal-to-noise ratios [6], etc., a clear evidence to demonstrate that it is a good confidence measure for verifying recognized words.

3. GENERALIZED POSTERIOR PROBABILITIES FOR SUBWORDS AND SENTENCES

GWPP can be extended to other recognition units, shorter or longer than word, like subword or sentence. The former one is essentially useful for a language like Chinese where subword plays an important role in speech communication. Subword units investigated in this study are monosyllabic characters. The longer units of sentences are universally useful for LVCSR of all languages.

3.1. Subword level

In order to obtain subword level acoustic scores, likelihood scores from all frames fall between the subword boundaries are multiplied. Subword boundaries are derived from phoneme boundaries obtained in the decoding process. Since the word level acoustic scores are also obtained based on frame likelihoods, the products of subword level and word level acoustic scores are preserved. An example break-down of a word level acoustic score into corresponding subword scores is given in Figure 1.



…	**word$_n$** *acoustic* = $pa_{n1}\bullet pa_{n2}\bullet pa_{n3}$; *lm* = pl_n			…
	subword$_{n1}$ *acoustic* = pa_{n1}; *lm* = pl_n	**subword$_{n2}$** *acoustic* = pa_{n2}; *lm* = pl_n	**subword$_{n3}$** *acoustic* = pa_{n3}; *lm* = pl_n	

Figure 1. Break-down of acoustic and language model scores from word level to subword level. *pa* **is the acoustic score for the subword segment and it is based on the boundaries obtained in the decoding process.** *pl* **is the language model score and is made to be the same at both word and composite subword levels.**

When deriving the language model scores from the word recognition output for computing the generalized subword posterior probability, we adopted an approach to take advantage of the higher (word) level language model. All composite subwords inherit the language model score of the corresponding word without modification. There are two reasons for assigning subword language model scores this way. First, it enables us to derive confidence measure at different levels using the same recognition output from a *single* LVCSR. If we changed the word language model to a character language model, the recognizer is then altered. Second, since we derived the subwords components from the corresponding words, probabilities of existence of these components are the same as those of the corresponding words.

With the acoustic and language model scores for the subwords, generalized subword posterior probabilities can then be computed in the same way as GWPP by using Eqn. 2. The only change is that the word, *[w; s, t]*, now represents a subword with identity *w*, and the starting and ending time, *s* and *t*, respectively. With these modifications, the recognition output can then be verified at a lower level using the generalized subword posterior probabilities.

3.2. Sentence level

At the sentence level, a generalized sentence posterior probability can be defined similarly. Deciding whether the sentence is correctly recognized does not pinpoint misrecognized parts more precisely when compared to words or subwords. But the main purpose of verifying a sentence is to statistically measure the confidence that the sentence is correctly recognized.

Definition of the generalized sentence posterior probability is similar to those of the word and subword counterparts. The reduced search space, reweighted acoustic and language model likelihoods are similarly applied. The major difference is that the time registration relaxation is no longer necessary, since all string hypotheses share the same sentence boundaries. As a result, the general sentence posterior probability is defined as

$$\frac{pa^{\alpha}\cdot pl^{\beta}}{\sum\limits_{\forall\, hypotheses} pa^{\alpha}\cdot pl^{\beta}} \qquad (3)$$

where pa is the acoustic score; pl, the language model score of a hypothesis; α and β, the acoustic and language model weights, respectively.

4. EXPERIMENTAL SETUP

4.1. Speech recognition

The LVCSR used in this study is the speech recognition system developed at ATR [7], 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 another word trigram language model to obtain the final recognition output. The word recognition accuracy is about 91%.

4.2. Corpus

The corpus used for evaluation is a large vocabulary, continuous, Chinese read speech database — the Chinese Basic Travel Expression Corpus (BTEC) [8, 9]. It was compiled and collected for a travel domain speech-to-speech translation project. We extracted two subsets of utterances as the development and test sets. 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
# sentences	841	3,437
# words	4,030	16,781
# characters	6,327	25,939

Table 1. Summary of the development and test sets extracted from the Chinese BTEC corpus

4.3. Verification

Generalized posterior probabilities at subword, word and sentence levels were computed separately. Optimal values for the acoustic and language model weights (α, β) and decision threshold were determined from the development set by a full grid search of the total error contour. Other efficient search algorithms (e.g., steepest descent, Downhill Simplex search) for parameter optimization have been proposed in [5]. These optimized parameters were then used in the test set for evaluation.

4.4. Evaluation Measure

Evaluation of the verification task is based on a normalized total decision error, or the confidence error rate (CER) [3]. Total errors include false acceptance (FA) of incorrectly recognized units and false rejection (FR) of correctly recognized units. The total is then normalized by the number of recognized units in the LVCSR output.

$$\text{CER} = \frac{\#\,\text{false acceptance} + \#\,\text{false rejection}}{\#\,\text{recognized words}} \times 100\% \qquad (4)$$

The 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.

A baseline was used for performance comparison in this work. It was obtained by accepting all recognition output without any rejection. All errors in the baseline were made up from false acceptance of incorrectly recognized units.

5. RESULTS AND DISCUSSIONS

The total verification error (#FA + #FR) contours at various acoustic and language model weights at word, subword and sentence levels are shown in Figure 2, 3 and 4, 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.

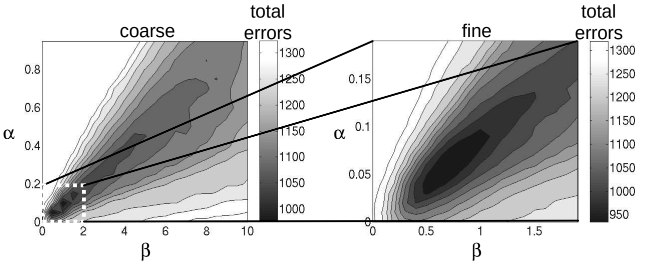


Figure 2. Total errors (test set) for word verification by using GWPP. The coarse scale plot shows equal error contours at different α and β values. Optimal parameters are determined using the fine scale plot.

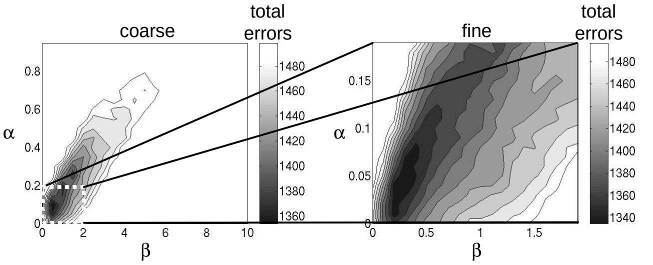


Figure 3. Contour plots of total errors for subword (character) verification using the generalized subword posterior probability.

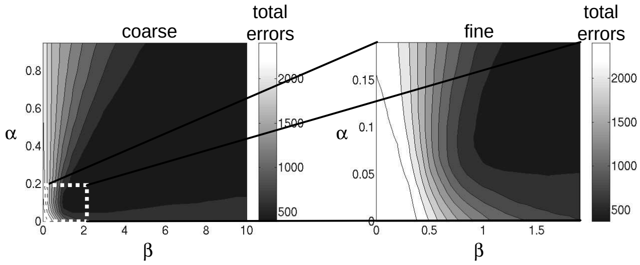


Figure 4. Contour plots of total errors for sentence verification using the generalized sentence posterior probability.

In general, better verification performance (darker region) is found near the lower left corner. As mentioned in [4, 5], 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. In the extreme case when both α and β are set to zero, all hypotheses in the reduced search space are taken into account, regardless of their acoustic and language model likelihoods, by counting the occurrences of the focused unit.

Figure 2 and 3 show that the error characteristics of verification at word and subword levels are similar. However, the subword level coarse scale error contour shows a smaller optimal region than that of the word level. This means that verification at the subword level is more sensitive to the proper choice of acoustic and language model weights.

The total error contours for sentence level verification are depicted in Figure 4. 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. These imply that neither the acoustic nor the language model score can be ignored when assessing the confidence of a recognized sentence. The best verification performance is obtained when α=0.16 and β=1.8. Contrary to the case of subword and word verification, the number of verification errors at the origin, (0, 0), is very large. This is because recognized sentences do not reappear in the search space. As a result, verification by counting just the reappearance is not reliable at the sentence level.

Figure 5 shows the verification performance in CER. It is observed that at higher level (e.g., sentence), the baseline CER is much higher. It is because the sentence recognition accuracy is much lower than those at word and subword levels. The relative improvement of verification at sentence level is also the highest (47.76%), compared to subword (4.64%) and word (27.31%) verifications. More importantly, results in Figure 5 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 where parameters are determined by using the test set.

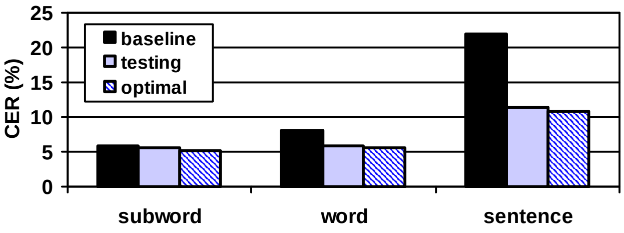


Figure 5. Verification performance (in CER, a normalized total errors) at various levels. Consistent verification performance improvement over baseline is achieved by using generalized posterior probabilities as the confidence measure.

6. SUMMARY

Optimal verification of recognition output at various levels (subword, word and sentence) is investigated by using the generalized posterior probability. This statistical approach takes into account of the three issues in the computation of posterior probabilities. Results showed that when parameters obtained from the development set are used in the test set for evaluation, very small degradation in performance, with respect to the upper bound optimal verification performance, is observed. Relative improvements of verification performance over the baseline are 4.64%, 27.31%, and 47.76% for subwords, words and sentences, respectively.

7. ACKNOWLEDGEMENTS

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

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

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

[7] T. Shimizu, H. Yamamoto, H. Masataki, S. Matsunaga, and T. Sagisaka, "Spontaneous dialogue speech recognition using cross-word context constrained word graph," *Proc. ICASSP1996*, pp.145-148.

[8] J. S. Zhang, M. Mizumachi, F. K. Soong, and S. Nakamura, "An introduction to ATRPTH: a phonetically rich sentence set based Chinese Putonghua speech database developed by ATR," *Proc. ASJ Fall Meeting 2003*, pp.167-168.

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