Robust Verification of Recognized Words in Noise

Wai-Kit Lo  Frank K. Soong  Satoshi Nakamura

Spoken Language Translation Research Labs, ATR, Kyoto, Japan

{waikit.lo, frank.soong, satoshi.nakamura}@atr.jp

Abstract

In this paper we investigate robust word verification in noise using the generalized word posterior probability (GWPP). In computing GWPP, reduced search space, relaxed time registrations of hypothesized words in the word graph, and optimal acoustic and language model weights are employed. The sensitivity of word verification errors with respect to the parameters of GWPP was tested under different SNR conditions. We found that around the optimal parameter settings, there exists a relatively stable region where the total number of word verification errors is fairly insensitive (robust) to the exact choice of the optimal values.

1. Introduction

Proper verification of large vocabulary, continuous speech recognition (LVCSR) output can provide additional reliability check of recognized words. Since the current state-of-the-art speech recognition technology is still susceptible to various environmental factors, performance of speech recognition can degrade significantly in a noisy environment. It will be beneficial to improving the performance of a speech input system if the reliability of recognized words can be assessed. This reliability information can help the system to decide whether to confirm only a portion of uncertain spoken input (low reliability) or to accept the recognized results (high reliability), instead of reprompting users frequently. As a result, a user-friendlier speech interface can be implemented. On the other hand, reliabilities of recognized words can also be useful for a speech translation system. When automatic speech translation is performed based on LVCSR output, the word reliability information can be employed to down-weight those unreliable recognized words for improving the overall translation performance.

Word reliability information is usually obtained by measuring the confidence of each recognized word in LVCSR output. In this paper, we concentrate on using one single metric, the generalized word posterior probability (GWPP) [1], for its well-behaved dynamic range, i.e., between 0 and 1, and a statistical rigor in the Bayesian sense and simplicity to compute it algorithmically. GWPP is an extension or generalization of the original word posterior probability [2,3]. It addresses three issues in the word posterior probability computation for improving the performance of word verification of LVCSR output. They are (1) reduced search space, (2) relaxed time registration of words; and (3) reweighted contributions from acoustic and language models.

Specifically, performance of GWPP in a noisy environment at different SNR’s will be investigated in details. We will show that GWPP is a robust confidence measure for word verification under different SNR conditions.

2. Generalized Word Posterior Probability under Noisy Conditions

Generalized word posterior probability (GWPP), used here to measure confidence of words recognized in LVCSR, is based on the posterior probability of word in an LVCSR. The word posterior probability is the probability of a focused word, given the acoustic observations of a sentence and a statistical speech recognizer. The larger the word posterior probability, the more likely the focused word is correctly recognized.

GWPP generalizes this word posterior probability by addressing three practical issues mentioned in the last section.

The resulting formula is given in Eqn. 1.

\[
p([w; s, t], x^T) = \sum_{w \in \Lambda_{\text{LVCSR}}} \prod_{m=1}^M p^a(w_m x_m^T) \cdot p^l(w_m x_m^T) / p(x^T)
\]

2.1. Reduced search space

The search space for LVCSR is usually very large. It is therefore unavoidable to prune partial hypotheses during the decoding process to a smaller and tractable subset. A reduced search space can be either a word graph [1,2] or an N-best list of string hypotheses, usually generated as a by-product of the recognizer. This reduced and compact search space is also handy for computing GWPP. When computing GWPP, the acoustic observation probability (denominator in Eqn. 1) is approximated by the summation of the product of reweighted acoustic and language model probabilities for all hypotheses in the reduced search space, \(\sum_{w \in \Lambda_{\text{LVCSR}}} \prod_{m=1}^M p^a(w_m x_m^T) \cdot p^l(w_m x_m^T)\).
noise contaminations. This reduced dynamic range of the hypothesis likelihoods results in a larger word graph. Albeit the increased search space, the recognition performance is still degraded because the correct hypotheses along with many other competing hypotheses all have lower likelihoods. As shown in Table 1, word graph densities (\# of links / \# of words) of word graphs obtained using the same decoder increases with decreased SNR’s.

$$\begin{array}{c|c|c|c|c|c}
\text{SNR (clean)} & \text{30dB} & \text{20dB} & \text{10dB} & \text{5dB} \\
\hline
\text{WGD} & 86.4 & 91.0 & 121.0 & 232.9 & 407.7 \\
\end{array}$$

2.2. Relaxed time registration
For a good word verification performance, the time registrations (starting and ending time) of the focused word need to be relaxed in GWPP computation. As a result, a candidate word is considered as the occurrence of the same word as long as it bears the same word identity and overlaps in time with the focused word [3,4]. This relaxation is necessary because the time registration is only a by-product of the decoding process in LVCSR and it can be affected by many factors, e.g., pruning threshold, acoustic and language model resolutions, etc. In practice, it is well known that the same word token in different decoded string hypotheses can have slight deviations at either starting or ending time. In noisy speech recognition, deviations in time registrations of the same word can be even more prominent than the clean data case. For clean speech data, we have observed that there is a relatively smaller difference in performance with or without applying time registration relaxation in computing GWPP. However, the difference is much larger for noisy speech. It is because the noise has made the determination of starting and ending time frames more uncertain in the decoding process.

2.3. Reweighted acoustic and language model
The third issue in computing GWPP is on the reweighting of contributions from acoustic and language models. Reweighting the acoustic and language model likelihoods is to compensate for the incompatibilities between these two likelihoods as well as some convenient but inaccurate assumptions in the modeling process. They include:

a) Difference in the dynamic range: Acoustic likelihood is measured against a probability density function, as parameterized in the commonly used Gaussian mixtures of HMM’s. It has a much larger dynamic range than language model probability, with a probability value between 0 and 1.

b) Difference in frequency of computation: While acoustic likelihood is computed every frame, language model probability is evaluated only at the end of each word.

c) Independency assumption: When computing the acoustic likelihood, adjacent frames are assumed to be statistically independent.

d) Reduced search space: The computation is performed based on the reduced search space of a word graph or N-best list, not the original full search space.

In practice, the introduction of these weights offers further control on the relative importance of the ranked hypotheses. For larger weights, more emphasis is put on the higher rank hypotheses. Smaller weights, on the other hand, take more hypotheses into consideration in computing the GWPP. In the extreme case, when the weights are set to infinity, only the best hypothesis is considered. Assigning zero weights will treat all hypotheses in the search space equally and the values of acoustic and language model likelihoods are ignored.

In our previous work [4], we have shown that the optimal acoustic and language model weights should be set to a point \((\alpha, \beta)\) close to the origin. With these small values, more hypotheses are then considered in computing GWPP. Under noisy conditions, the incompatibilities and modeling assumptions mentioned above still exist. Among them, the largest change is the dynamic range of the acoustic likelihoods. When speech data become noisier, acoustic likelihoods are smaller than those obtained from clean data. As shown in the experimental results later, the importance of the acoustic likelihoods has also become smaller and more emphasis is put on the language model. However, the optimal weights locate in a region close to the origin. Change of the relative importance between acoustic and language models does not affect the word verification performance significantly. In the following sections, we will look into the effect of additive noise at different SNR’s on the optimal parameter values used in computing GWPP.

3. Experiments
3.1. Setups
3.1.1. Data preparation
The speech corpus used in our experiments is the Basic Travel Expression Corpus (BTEC), a large vocabulary, continuous, read Japanese corpus [5]. It is compiled for a travel domain speech-to-speech translation task. In particular, the development and test sets used in this work have 508 and 510 utterances, consisting 3442 and 3545 words, respectively. Each test set has 10 speakers (gender balanced) reading different sentences related to traveling.

Noise data used in our experiments is real noise collected from a busy public train station concourse in Japan. Data was sampled at 48kHz using a DAT recorder and then down-sampled to 16kHz. Portions of the down-sampled noise are then randomly selected and digitally added to the clean speech utterances at different SNR’s (5, 10, 20 and 30dB’s) to make the noisy data.

3.1.2. LVCSR
Speech recognition is carried out by using the ATRASR LVCSR from ATR [6]. 25 acoustic features (12 MFCC + 12 ΔMFCC + Δpower) are used and the acoustic HMM’s are trained using clean data. The LVCSR runs in multi-pass with a 47k word lexicon and a word bigram language model to generate word graphs. After generation of the initial word graphs, language model rescoring is performed using a word trigram language model. The final recognition results are obtained from the rescored word graphs. Specifically for the experiments presented in this paper, a wide beam width is used in the decoding process.
3.1.3. Evaluation metrics

Evaluation of the confidence measure is carried out in a word verification task. Two metrics are used in our evaluations: confidence error rate (CER) \([3]\) and modified word accuracy (mAcc). CER is based on the ratio between the total number of verification errors and the total number of recognized words in the LVCSR output. It is defined as

\[
\text{CER} = \frac{\#\text{false acceptance} + \#\text{false rejection}}{\#\text{recognized words}} \times 100\%. \quad (2)
\]

Modified word accuracy, a measure of word accuracy after word verification, is defined as

\[
\text{mAcc} = \frac{\#\text{recog} - \#\text{delete} - \#\text{subs} - \#\text{insert} - \#\text{FR} + \#\text{CR}}{\#\text{ref} - \#\text{FR} - \#\text{CRsub}} \times 100\%
\]

where \#recog: the no. of recognized words; \#delete, \#subs, \#insert: deletion, substitution and insertion errors; \#ref: words in reference transcriptions; \#FR and \#CR: false and correct rejections; \#CRsub: correct rejections of substituted words.

3.1.4. Performance references

Two performance references are used for comparison in this work: the baseline and the reappearance rate \((rprate)\). The baseline is obtained by accepting all recognized words without rejection. All verification errors are made up from false acceptance. The baseline performance can be easily computed from the recognition output of any LVCSR.

The second reference is the “reappearance” rate. This reference reflects the confidence of a word by computing just the ratio between the number of string hypotheses containing the focused words to the total number of hypotheses in the reduced search space. This value can be efficiently computed with a word graph by setting both \(\alpha\) and \(\beta\) in Eqn. 1 to 0.

3.2. Word verification training and testing

To perform the word verification task, word graphs are first generated by using our LVCSR to recognize noise corrupted utterances at different SNR’s. A forward-backward algorithm is then applied to compute GWPP with the word graph for each word in the best string output. Word verification is then made by comparing these GWPP’s against a preset threshold. Words with GWPP’s below the threshold are rejected and the remaining words are accepted.

When computing the GWPP and making the acceptance / rejection decision, the acoustic model weight \((\alpha)\), language model weight \((\beta)\) and the threshold are determined from the development set. Full grid search in a coarse scale is performed first. The small region around the coarse optimal point (always situated near the origin) is then searched again in a finer scale for the optimal parameters. For each of the SNR’s, optimal parameters are obtained from the development set and applied to the test set. An example pair of error surfaces for the development set obtained from the coarse and fine grid search is shown in Fig 1.

4. Performance at Matched SNR’s

Albeit the recognition performance decreases with noisier condition, we found that the optimal parameter values \((\alpha, \beta)\) and threshold) for word verification using GWPP do not show significant fluctuations. In addition to the consistency of these optimal parameters, experiments in this section also show that the optimal ratio (yielding lower verification errors) between acoustic and language model weights, which equalized the contributions from the two knowledge sources, changes with noise levels.

![Figure 2: Word verification performance in CER and mAcc for noisy speech with SNR’s ranging from 5dB to 30dB. The optimal \((\alpha, \beta)\) determined from the test set are \((0.07,0.7), (0.06,0.6), (0.04,0.4)\) and \((0.06,0.5)\) for SNR’s at 5dB to 30dB, respectively.](image)

![Figure 1: Total error surfaces for the search of optimal parameters from development set in a coarse and a fine grid scale.](image)
The slope of this low-error valley increases as the SNR decreases.

Figure 3: Contour plots of word verification errors for four different SNR’s. The optimal ratios are indicated by the dotted lines overlaid on the contour plots.

The slope change of the low-error valley is due to the change in the reliability of acoustic likelihoods, relative to the language likelihoods, at different SNR’s. At low SNR’s, the acoustic information becomes less reliable than the corresponding language model likelihood. The acoustic likelihood, due to its intrinsic lower value in noise, also contributes less to the final GWPP, comparing to the language likelihood. Therefore, language model is more critical to make correct word verification decision. Consequently, the sensitivity or the change of acoustic weights in lower SNR conditions has less effect on the word verification errors than high SNR.

5. Performance at Cross-SNR’s

Parameters ($\alpha$, $\beta$ and threshold) determined from the development set at SNR’s different from that of the test set were also tested. The word verification performance is shown in Fig. 4. CER’s and mAcc’s are plotted for test data at different SNR’s (the horizontal axis) and the SNR’s of the development set are shown inside the legend box. Although the CER performances degrade slightly in mismatched SNR conditions, they are still significantly better than the baseline performance for all conditions. The mAcc performance at various cross-SNR conditions shows similar robustness as the CER.

These results demonstrate further the robustness of the GWPP parameters as a confidence measure of recognized words in mismatched SNR conditions. As explained in the previous section, this low sensitivity to SNR’s, and hence noise robustness can be attributed to the low dependency of the parameters on noise level. Another reason is that these optimal parameters are rather close to the origin and as a result many string hypotheses are used in computing GWPP. The weighted average nature of GWPP has made it rather insensitive to changes of SNR’s in verifying recognized words reliably, even with parameters obtained from a mismatched SNR.

6. Conclusions

In this paper, the robustness of GWPP for word verification in noise has been investigated. It is shown that GWPP works well for measuring reliability of recognized words in LVCSR output under noisy conditions. The cross-SNR tests further show that optimal parameters (acoustic weight, language model weight and threshold) determined under mismatched SNR conditions are relatively robust and they yield significant performance improvement over the baseline results.

7. Acknowledgements

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

8. References