Performance Improvement of Dialog Speech Translation by Rejecting Unreliable Utterances

Toshiyuki Takezawa †‡ and Tohru Shimizu †‡

† National Institute of Information and Communications Technology
‡ ATR Spoken Language Communication Research Laboratories
2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan
{toshiyuki.takezawa, tohru.shimizu}@{nict.go.jp, atr.jp}

Abstract

We discuss how to measure the reliability of recognized utterances based on a confidence measure, and applied it to a dialog speech translation system. In this study, we employ generalized word posterior probability (GWPP), a confidence measure for verifying recognized words, and expand it to measure the reliability of recognized utterances. We confirmed the performance improvement by applying the rejection technique to a dialog speech translation system from Japanese to English. We conducted two kinds of performance evaluation. One is a ranking evaluation of translation output by human evaluators. The other is to measure the machine output against human results by a paired-comparison method. Both of them yield significant improvements.

Index Terms: confidence measure, dialog speech translation.

1. Introduction

Today's state-of-the-art speech recognition technology is still not sufficiently robust for use in various conditions. These include speaker variability, a mismatch between the training and testing channels, interference from environmental noise, and so on. For these reasons, output from large-vocabulary continuous speech recognition (LVCSR) systems usually contains errors. For speech-to-speech translation applications, it is desirable to locate this erroneous recognized string output from an LVCSR. These erroneous words can then be rejected by tagging them with a low confidence measure.

There have been many approaches proposed for measuring the confidence of speech recognition output. They can be roughly classified into three categories:

- Feature-based
- Explicit model based
- Posterior probability based

Feature based approaches try to assess the confidence according to selected features (e.g., acoustic stability, acoustic and language model back-off, hypothesis density, part-of-speech, and word duration) [1]. Explicit model based approaches require models for both the candidate class as well as the competing classes (e.g., an anti-model or a filler model). Hypothesis testing (e.g., a likelihood ratio test or Bayesian factor) between the candidate and competing class models is then applied to assess a confidence measure for deciding acceptance or rejection [2]. In the posterior probability based approach, a statistical confidence measure is obtained by estimating the posterior probabilities of a recognized unit (e.g., word) from all the acoustic observations [3].

To measure the reliability of recognized words in automatic speech recognition (ASR), Soong et al. propose a generalized word posterior probability (GWPP) [4, 5]. We used GWPP as a confidence measure and applied it to dialog speech translation.

2. Confidence measure based on the generalized word posterior probability

2.1. Generalized word posterior probability

In maximum a posteriori (MAP) based speech recognition, the best recognized word string w_1^{*M} (M: string length) is obtained by maximizing the corresponding string posterior probability (SPP) as

$$\begin{aligned} w_1^{*M} &= \max_{w_1^M} \arg p(w_1^M|x_1^T) \\ &= \max_{w_1^M} \arg \frac{p(x_1^T|w_1^M)p(w_1^M)}{p(x_1^T)} \\ &= \max_{w_1^M} \arg p(x_1^T|w_1^M)p(w_1^M), \end{aligned} \tag{1}$$

where $p(x_1^T|w_1^M)$ is the acoustic model probability; $p(w_1^M)$, the language model probability; and $p(x_1^T)$, the acoustic observation probability, respectively. The denominator $p(x_1^T)$ can be ignored in maximization since it is independent of the choice of the recognized word sequence.

The word posterior probability (WPP) is defined in the same way as the SPP. In continuous speech recognition, 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

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

where a word hypothesis is defined by the corresponding triple, $[w;s,t]$; x_s^t is the sequence of acoustic observations; M, the number of words in a string hypothesis; $p(x_1^T)$, the probability of the acoustic observations; and T, the length of the complete acoustic

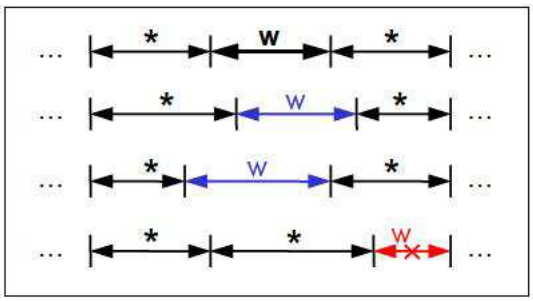

Figure 1: Illustration of the time registration relaxation

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.

GWPP is a generalization of WPP to take into account the following three issues in computing WPP:

Reduced search space The 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, $p(x_1^T)$.

Relaxed time registration A word is defined as a triple by the word identity, its starting time and its ending time. The starting time and ending time of a word are affected by various factors, like the pruning threshold, model resolution, noise, and so on. 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.

Re-weighted acoustic and language model likelihoods In continuous speech recognition, assumptions are made to facilitate an efficient parametric modeling and decoding process. Incompatibilities also exist among components in the models. They include:

Difference in the dynamic range The acoustic likelihoods computed by using a continuous Gaussian mixture HMM are based on the probability density functions which, in theory, have an unbounded dynamic range while the language model likelihoods based on N-gram probabilities have values between 0 and 1.

Difference in the frequency of computation Acoustic likelihoods are computed every frame and language model probabilities are computed only once per word.

Independence assumption Neighboring acoustic observations are assumed to be statistically independent in computing the acoustic likelihoods, a convenient but obviously wrong assumption.

Reduced search space In practice, the search space is reduced by pruning a word graph, or an N-best list of hypotheses is generated.

Table 1: Experimental system

	High performance	High speed
J speech recognition	$RTF = 5$	$RTF = 1$
J-E translation	$RTF = 5$	$RTF = 1$

Table 2: Test sets

Name	BTEC	MAD	FED
Characteristics	Read speech	Dialog speech (Office)	Dialog speech (Airport)
# of speakers	20	12	6
# of utterances	510	502	155
# of word tokens	4,035	5,682	1,108
Average length	7.91	11.32	7.15
Perplexity	18.9	23.2	36.2

As shown in Figure 1, the word w in the top hypothesis is being spotted. Other strings with w appearing with intersecting time interval (the second and third string) will be included. A string with word w but no intersection (the last string) is excluded.

Obviously, the denominator term, $p(x_1^T)$, when summed over all string hypotheses in the reduced search space of a word graph or an N-best list, needs to be scaled by α and β accordingly. The optimal values of α and β are learned from given training or development data. The final generalized word posterior probability (GWPP) is given as

$$p([w;s,t]|x_1^T) = \sum_{\substack{\forall M,[w;s,t]_1^M \\ \exists n, 1\le n\le M \\ (s_n,t_n)\cap(s,t)\neq\phi}} \frac{\prod_{m=1}^{M} p^{\alpha}(x_{s_m}^{t_m}|w_m)\cdot p^{\beta}(w_m|w_1^M)}{p(x_1^T)}. \quad (3)$$

2.2. Confidence measure for an utterance

We newly define the confidence measure for a recognized utterance as the joint confidence of all component words in the recognized string. The GWPP of a word is a measure of its correctness, or a probability of a binomial distributed "correct word" event. The probability of a "correct utterance" event is then the product of all probabilities of component "correct word" events, assuming that all word events are statistically independent.

The product of the GWPPs of all recognized words in a recognized utterance is therefore proposed as an utterance level confidence ($CF_{utterance}$) as given below

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

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

In general, there are two kinds of decision errors: false rejection when a correctly recognized utterance is rejected and false acceptance when a mis-recognized utterance is accepted. The confidence error rate (CER) is used here as a performance measure for word acceptance/rejection decision. CER is defined as the ratio of all errors to the total number of recognized utterances. We

Table 3: Experimental results of rejection for Japanese speech recognition (%)

	BTEC		MAD		FED	
	$RTF = 5$	$RTF = 1$	$RTF = 5$	$RTF = 1$	$RTF = 5$	$RTF = 1$
Word accuracy (Original)	94.9	94.8	92.9	91.4	91.0	89.4
Utterance accuracy (Original)	82.4	82.4	62.2	60.2	69.0	65.8
CER	12.8	11.4	18.5	18.5	16.8	14.8
Utterance output rate (Rejection)	94.1	86.9	69.3	62.4	67.1	63.2
Utterance accuracy (Rejection)	87.1	91.0	83.9	85.9	91.4	91.8

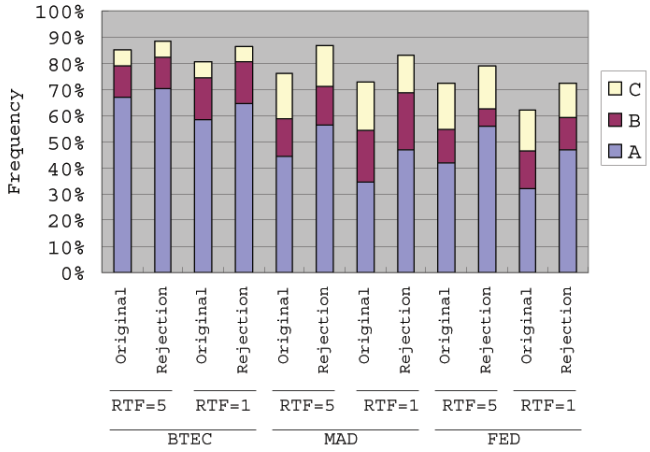

Figure 2: Speech translation experiment

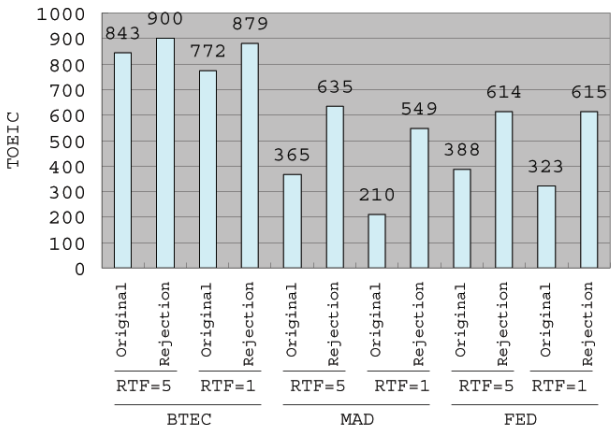

Figure 3: Estimated TOEIC score of the systems

then optimize the decision threshold to minimize total verification errors using the following equation:

$$CER = \frac{\#FalseAcceptance + \#FalseRejection}{\#Utterance}. \quad (5)$$

3. Evaluating the utterance rejection for dialog speech translation

If a user uses a small device, he or she may confirm the recognition result by himself or herself before translation. If the recognition result contains some fatal errors, the user can then avoid the translation. In order to reduce such operational costs by users, we employ automatic rejections of recognition candidates based on confidence measures. In this section, we conduct speech translation experiments from Japanese to English, and show the effectiveness of the rejection of recognition candidates for dialog speech translation.

3.1. Experimental system

We conducted speech translation experiments from Japanese to English. For Japanese speech recognition, we used ATRASR [6], which was built at ATR. For Japanese-to-English translation, we used corpus-based multiple engines such as SAT [7] and HPATR2 [8], which were built at ATR, in which the selector [9] module selects and outputs the best result.

Table 1 shows the experimental systems and conditions. The high-performance version had a real-time factor (RTF) of five, and the high-speed version had an RTF of one.

3.2. Test sets

Table 2 shows the test sets. BTEC contains read speech from basic travel expressions [10]. MAD contains dialog speech collected in an office using human typists and our translation system [11]. FED contains dialog speech collected at an airport using our speech-to-speech translation system [12].

The average utterance length of MAD is greater than those of BTEC and FED. The perplexity of BTEC is small, that of MAD is intermediate, and that of FED is large.

3.3. Speech recognition experiment

We used three kinds of test sets: BTEC, MAD, and FED, with two kinds of experimental conditions: a high-performance version ($RTF = 5$) and a high-speed version ($RTF = 1$). Table 3 shows the results.

The word accuracy (original) and the utterance accuracy (original) for the high-performance version ($RTF = 5$) are superior to those for the high-speed version ($RTF = 1$) under all conditions and in all test sets. The utterance output rate (rejection) for the high-performance version ($RTF = 5$) is also superior to that for

the high-speed version ($RTF = 1$).

3.4. Speech translation experiment

For the speech translation experiment from Japanese to English, we used two kinds of experimental conditions: a high-performance version ($RTF = 5$) and a high-speed version ($RTF = 1$). The input to the translation system from Japanese to English for the high-performance version ($RTF = 5$) was the output from the Japanese speech recognition system for the high-performance version ($RTF = 5$). The input to the translation system from Japanese to English for the high-speed version ($RTF = 1$) was the output from the Japanese speech recognition system for the high-speed version ($RTF = 1$).

We had the speech translation output evaluated and ranked by English native evaluators who can understand Japanese sufficiently. The output was evaluated into the following A, B, C, and D ranks.

A: Perfect.
B: Good.
C: Fair.
D: Nonsense.

We assume that the translation accuracy is an accumulation from A to C. Figure 2 shows the results. The remaining parts indicate the rank D, which means nonsense output. The bar graphs for rejection indicates the rate in the remaining utterances after rejection.

We also conducted experiments to calculate the estimated TOEIC score of the system by the paired comparison method [13]. This method gives an objective evaluation result, namely a score for the Test of English for International Communication (TOEIC) [14], which is a widely used measure of English communication capability for non-native English speakers, by measuring machine output against human translation results. Figure 3 shows the results.

According to [14], the range of TOEIC scores is from 10 to 990 and the skill levels are in five grades, such as the following:

860-990 can usually communicate adequately as a non-native speaker.
730-860 is capable of communicating appropriately in most situations.
470-730 has sufficient knowledge for daily activities and conducting business within certain limits.
220-470 is capable of minimal communication in ordinary conversation.
10-220 is not able to communicate adequately.

According to Figure 2, we find that the rejection reduces the utterances of rank D, which indicates nonsense output. This means that the translation output after the rejection contains much more meaningful results, such as those of ranks A, B, and C. According to Figure 3, the estimated TOEIC scores of the systems for the rejection are superior to those for the original ones.

4. Conclusions

In this paper, we discussed how to measure the reliability of recognized utterances based on a confidence measure, and applied it to a dialog speech translation. We employed GWPP, a confidence measure for verifying recognized words, and expanded it to measure the reliability of recognized utterances. We confirmed the performance improvement by applying the rejection technique to a dialog speech translation system from Japanese to English. According to the experimental results, the estimated TOEIC score of our system for BTEC read speech is 900, which means that the user of our system is expected to be able to communicate adequately as a non-native speaker, and that for dialog speech such as MAD and FED is 600, which means that the user of our system is expected to have sufficient knowledge for daily activities and conducting business within certain limits.

5. References

[1] Kemp, T., and Schaaf, T., "Estimating confidence using word lattices," *Proc. of EUROSPEECH*, pp. 827–830, 1997.

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

[3] Wessel, F., Schluter, R. Macherey, K., and Ney, H., "Confidence measures for large vocabulary continuous speech recognition," *IEEE Trans. SAP*, Vol. 9, pp. 288–298, 2001.

[4] Soong, F. K., Lo, W. K. and Nakamura, S., "Generalized word posterior probability (GWPP) for measuring reliability of recognized words," *Proc. of SWIM*, 2004.

[5] Soong, F. K., Lo, W. K., and Nakamura, S., "Optimal acoustic and language model weights for minimizing word verification errors," *Proc. of ICSLP*, 2004.

[6] Itoh, G., Ashikari, Y., Jitsuhiro, T., and Nakamura, S., "Summary and evaluation of speech recognition integrated environment ATRASR," *Autumn Meeting of the Acoustical Society of Japan*, 1-P-30, 2004.

[7] Watanabe, T. and Sumita, E, "Example-based decoding for statistical machine translation," *Proc. of MT Summit IX*, pp. 410-417, 2003.

[8] Imamura, K., Okuma, H., and Sumita, E., "Practical approach to syntax-based statistical machine translation," *Proc. of MT Summit X*, pp. 267-274, 2005.

[9] Akiba, Y., Watanabe, T., and Sumita, E., "Using language and translation models to select the best among outputs from multiple MT systems," *Proc. of COLING*, pp. 8-14, 2002.

[10] Kikui, G., Sumita, E., Takezawa, T., and Yamamoto, S., "Creating corpora for speech-to-speech translation," *Proc. of EUROSPEECH*, pp. 381–384, 2003.

[11] Takezawa, T. and Kikui, G., "A comparative study on human communication behaviors and linguistic characteristics for speech-to-speech translation," *Proc. of LREC*, pp. 1589–1592, 2004.

[12] Kikui, G., Takezawa, T., Mizushima, M., Yamamoto, S., Sasaki, Y., Kawai, H., and Nakamura, S., "Monitor experiments of ATR speech-to-speech translation system," *Autumn Meeting of the Acoustical Society of Japan*, 1-7-10, 2005.

[13] Sugaya, F., Takezawa, T., Yokoo, A., and Yamamoto, S., "Proposal of an evaluation method for speech translation capability by comparing a speech translation system with humans and experiments using the method," *IEICE Trans. Inf. & Syst.*, Vol. J84-D-II, No. 11, pp. 2362–2370, 2001.

[14] TOEIC: Test of English for International Communication, http://www.toeic.or.jp/.