

Verification of Speech Recognition Results Incorporating In-domain Confidence and Discourse Coherence Measures

Ian R. LANE^{†,††} and Tatsuya KAWAHARA^{†,††}, *Members*

SUMMARY Conventional confidence measures for assessing the reliability of ASR (automatic speech recognition) output are typically derived from “*low-level*” information which is obtained during speech recognition decoding. In contrast to these approaches, we propose a novel utterance verification framework which incorporates “*high-level*” knowledge sources. Specifically, we investigate two measures: *in-domain confidence*, the degree of match between the input utterance and the application domain of the back-end system, and *discourse coherence*, the consistency between consecutive utterances in a dialogue session. A joint confidence score is generated by combining these two measures with an orthodox measure based on GPP (generalized posterior probability). The proposed framework was evaluated on an utterance verification task for spontaneous dialogue performed via a (English/Japanese) speech-to-speech translation system. Incorporating the two proposed measures significantly improved utterance verification accuracy compared to using GPP alone, realizing reductions in CER (confidence error-rate) of 11.4% and 8.1% for the English and Japanese sides, respectively. When negligible ASR errors (that do not significantly affect translation) were ignored, further improvement was achieved for the English side, realizing a reduction in CER of up to 14.6% compared to the GPP case.

key words: *speech recognition, confidence measure, utterance verification, in-domain confidence, discourse coherence*

1. Introduction

Current state-of-the-art speech recognition technologies are not robust against acoustic mismatch caused by noise, channel mismatch, or speaker variability, or linguistic inconsistencies such as ill-formed utterances, OOV (out-of-vocabulary) words or OOD (out-of-domain) input. In order to develop effective spoken language systems based on this technology, it is necessary to assess the confidence of the speech recognition hypothesis (or individual words within this hypothesis) and design an appropriate repair strategy based on this information.

Confidence measures have been investigated for a wide range of tasks within spoken language systems. For example, the systems in [1], [2] prompted users to re-speak or re-phrase, when recognition output had low confidence; in [3]–[5], dialogue strategy was switched to

overcome poor recognition; and in [6], [7], the robustness of semantic analysis was enhanced by weighting multiple plausible recognition hypotheses. To realize effective performance in these frameworks, it is vital to define an effective measure of recognition confidence.

Various approaches have been proposed for confidence scoring, and these can be roughly classified into three categories: feature-based approaches, explicit model-based approaches, and posterior probability-based approaches. Feature-based methods, such as [4], [8], [9], assess confidence according to a set of specific features (e.g., word duration, acoustic and language model back-off, and word graph density) by explicitly training classifiers to discriminate between correctly recognized and erroneously recognized words or utterances. Explicit model-based schemes [10]–[12] compare the candidate model to a competing model (an anti-model, background model or set of cohort models) via a likelihood ratio test. Posterior probability-based approaches, including [13]–[15], estimate the posterior probability of a recognized entity (word or utterance) considering all competing hypotheses (typically in an N-best list or word graph).

All these approaches, however, estimate recognition confidence considering only the “*low-level*” information available during ASR decoding (for example, normalized acoustic and linguistic likelihoods, and confusability with competing hypotheses). On the other hand, there are apparently knowledge sources outside the ASR framework, such as information about the application domain and discourse flow, which have not been well exploited for assessing recognition confidence.

There are several works [6], [7], [16]–[19] that consider such “high-level” knowledge by incorporating measures based on parse quality [6], [7], [16], [17], number and order of semantic slots filled [7], [16], and dialogue state of the system [17]–[19]. These approaches, however, are limited to simple, small-vocabulary spoken dialogue tasks. During development, such approaches require structured knowledge about the application domain, for example; hand-crafted grammars, definite semantic slots and task keywords, and definite dialogue states. Thus, they are not scalable to complex tasks or applications where such knowledge cannot be manually created. For many spoken language systems, including, question-answering systems, spoken document retrieval, and speech-to-speech transla-

[†]The authors are with the School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan

^{††}The authors are with Spoken Language Communication Research Laboratories, Advanced Telecommunications Research Institute International, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan

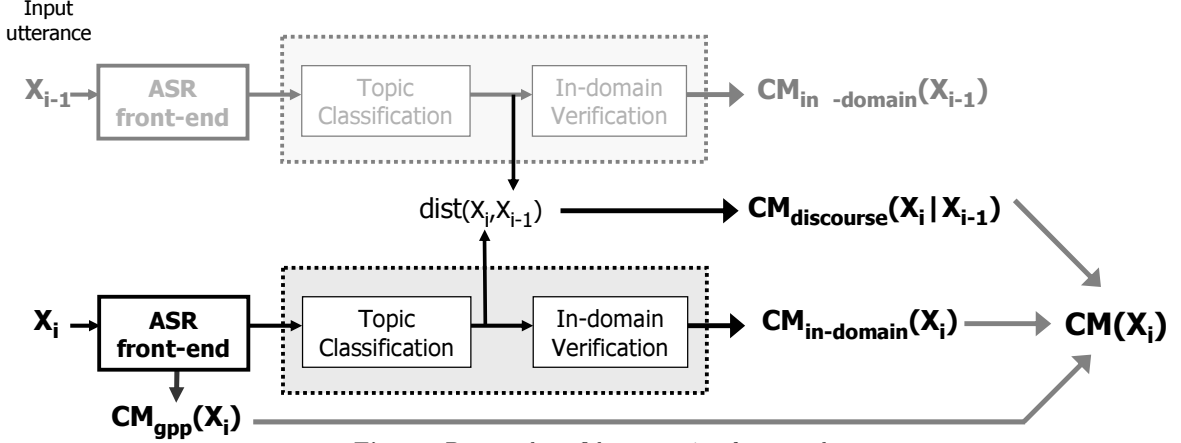


Fig. 1 Proposed confidence scoring framework

- $CM_{in-domain}(X_i)$: *In-domain confidence* of input utterance X_i
 $CM_{discourse}(X_i)$: *Discourse coherence* between input utterance X_i and preceding utterance in dialogue X_{i-1}
 $CM(X_i)$: *Joint confidence score* of utterance X_i . Overall recognition confidence score, incorporating above two measures with GPP (generalized posterior probability) of ASR result $CM_{gpp}(X_i)$.

tion systems, such knowledge is not available, and thus, these schemes cannot be simply adopted.

To overcome the limitations of these techniques we propose a more general confidence scoring framework. Specifically we introduce two novel measures that are related to knowledge sources external to the ASR framework: *in-domain confidence* and *dialogue coherence*. The first, *in-domain confidence*, is a measure of match between the input utterance and the application domain of the back-end system. The second, *discourse coherence*, is a measure of the consistency between consecutive utterances in a dialogue session. A joint confidence score is generated by combining these two measures with a conventional measure based on the GPP (generalized posterior probability) of the ASR output.

2. Proposed Confidence Scoring Framework

Typical spoken language systems consist of two main sub-systems: an ASR (automatic speech recognition) front-end, which generates a recognition hypothesis (or N-best list of recognition hypotheses) for each input utterance, and an NLP (natural language processing) back-end, which performs semantic understanding, dialogue management, and response generation. While conventional approaches [8]–[15] generate confidence measures based on the information obtained during decoding in the ASR front-end, this paper focuses on the incorporation of “high-level” knowledge sources from the back-end system.

The specific approach proposed in this paper is depicted in Figure 1. The knowledge sources exploited here relate to two dissimilar aspects of spoken lan-

guage and both are expected to be useful for identifying recognition errors that are difficult to detect using only acoustic and linguistic likelihoods from the ASR front-end. The first measure, *in-domain confidence*, $CM_{in-domain}(X_i)$, is a measure of topic consistency between the input utterance and the application domain of the back-end system. This measure is intended to detect recognition errors that are caused by mismatch of domain (Figure 2, Example A) and erroneous hypotheses that are not semantically coherent (Figure 2, Example B). The second measure, *discourse coherence*, $CM_{discourse}(X_i|X_{i-1})$, verifies the consistency between consecutive utterances in a dialogue session. This measure is designed to detect erroneous hypotheses that are not consistent, in terms of topic, with the previous utterance in the dialogue (Figure 2, Example C). A joint confidence score $CM(X_i)$ is defined by combining these two “high-level” measures with an orthodox measure based on the GPP (generalized posterior probability) of the ASR output [12], $CM_{gpp}(X_i)$. The combined measure will be effective not only in detecting acoustic and linguistic mismatch but also semantic inconsistency, at either the utterance or dialogue level. In the following sections, we describe in detail the *in-domain confidence* and *discourse coherence* measures, followed by the joint confidence score.

3. In-domain Confidence

In-domain confidence, $CM_{in-domain}(X)$, is a measure of the semantic relevance of an utterance with respect to the application domain of the back-end system. This measure was originally designed to detect OOD (out-of-domain) utterances in our previous work [20], and

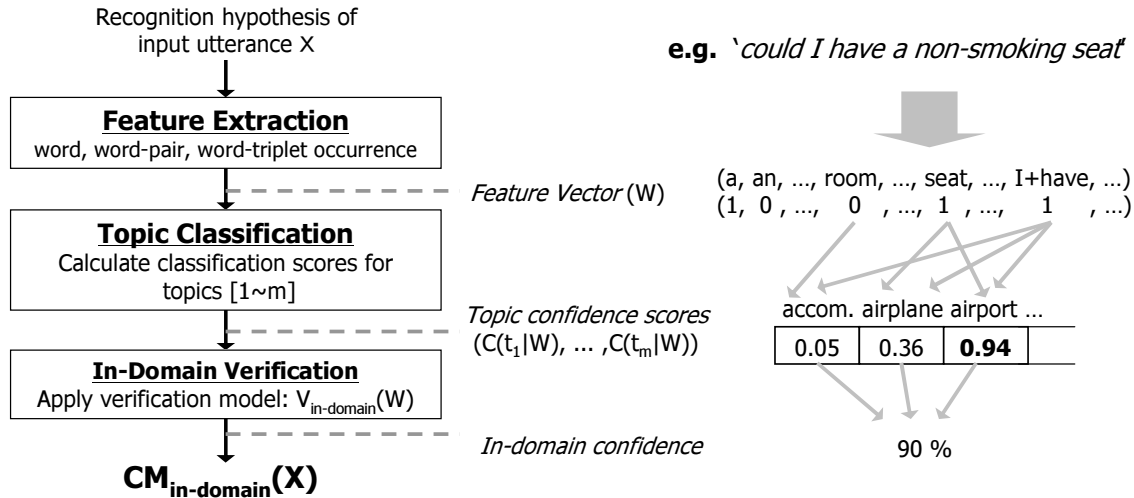


Fig. 3 In-domain confidence based on topic classification

Example A: Out-of-domain utterance, which is not correctly recognized
REF: How can I print this WORD file double-sided.
ASR: How can I open this word on the poolside
hypothesis not consistent by topic → <i>in-domain confidence low</i>
Example B: Erroneous recognition hypothesis
REF: I want to go to Kyoto, can I take a bus.
ASR: I want to go to Kyoto, can I take a bath.
hypothesis not consistent by topic → <i>in-domain confidence low</i>
Example C: Erroneous recognition hypothesis
Speaker A: Previous utterance $[X_{i-1}]$
REF: What type of shirt are you looking for?
ASR: What type of shirt are you looking for?
Speaker B: Current utterance $[X_i]$
REF: I'm looking for a white T-shirt.
ASR: I'm looking for a white teacher.
hypothesis of utterance X_i not consistent with utterance X_{i-1} → <i>discourse coherence low</i>
REF: correct transcription of utterance
ASR: speech recognition hypothesis

Fig. 2 Examples of errors handled by the proposed framework

we observed that most OOD utterances detected by this scheme contained speech recognition errors. Moreover, in-domain utterances with critical recognition errors tended to be rejected as OOD because the recognition hypothesis was not consistent in terms of topic class. Therefore, *in-domain confidence* is also expected to be useful for utterance verification, which detects recognition errors themselves.

To apply this scheme, we assume that the application domain consists of multiple topic classes. In this work, topic classes were pre-defined and the training set was hand-labeled appropriately. This data was then used to train the topic classification and in-domain verification models.

In-domain confidence is computed using the procedure shown in Figure 3. First, a feature vector (W) is generated based on lexical features in the recognition hypothesis. Next, a topic classification confidence vector $(C(t_1|W), \dots, C(t_m|W))$ is computed by applying support vector machines (SVM) trained for each in-domain topic class, to the feature vector W . Finally, an in-domain verification model $V_{in-domain}(W)$ is applied, realizing the final in-domain verification score. These three steps are briefly described in the following subsections.

3.1 Feature Extraction

A set of lexical features are defined by selecting word, word pair and word triplet tokens (consisting of word and POS (part-of-speech) information) in the training set that occur more frequently than a certain cutoff threshold. The feature vector W is composed of their occurrence counts, where each component is the count of a particular lexical feature.

Typically word baseform tokens are applied for topic classification, however, by incorporating word-

tense and POS information the system is expected to detect erroneous recognition hypotheses which contain mismatch of tense or POS. Similarly, by incorporating n-gram features (word pairs and word triplets), the ability to detect topics and topic mismatched is enhanced. We observed the effectiveness of both techniques for OOD utterance detection, thus we adopt them here.

3.2 Topic Classification

For topic classification, we adopt support vector machine (SVM) [21]. In earlier work [20] we compared various topic classification methods and concluded that SVM provides the most accurate and robust performance. During training, a discriminative SVM hyperplane H_j is trained for each topic class t_j using a one-vs.-all scheme, where sentences labeled with the current topic (t_j) are used as positive training examples and the remainder of the training set is used as negative examples. Since this space is very high-dimensional (up to 70,000 features), a linear kernel is adequate for classification.

Topic classification is performed by comparing the feature vector W to each SVM hyperplane. A score for topic t_j , $dist_{\perp}(W, H_j)$, is calculated as the perpendicular distance between W and topic t_j 's hyperplane (H_j). This value is positive if W is in-class, and negative otherwise. A confidence score $C(t_j|W)$ is generated by applying a sigmoid transformation to this distance.

3.3 In-domain Verification

In the final stage of OOD detection, an in-domain verification model $V_{\text{in-domain}}(W)$ is applied to the topic confidence vector $(C(t_1|W), \dots, C(t_m|W))$. We adopt a linear discriminant model, as shown in Equation 1.

$$V_{\text{in-domain}}(W) = \sum_{j=1}^m \lambda_j C(t_j|W) \quad (1)$$

$C(t_j|W)$: classification score of topic t_j for feature vector W
 m : number of topic classes

The linear discriminant weights $\{\lambda_1, \dots, \lambda_m\}$ are trained using only in-domain examples by applying deleted interpolation of topics and the gradient probabilistic descent (GPD) algorithm [22], as presented in [20]. During each training iteration, a single topic class t_i is set to be temporarily OOD, and the corresponding vector component $C(t_i|W)$ is removed from the verification model. The discriminant weights of the remaining topic classifiers $\{\lambda_j, 1 \leq j \leq m, j \neq i\}$ are then estimated using GPD. Upon completion of all iterations,

the final model weights $\{\lambda_1, \dots, \lambda_m\}$ are calculated by averaging over all interpolation steps.

A confidence measure $CM_{\text{in-domain}}(X)$ is generated by applying a sigmoid transformation to the resulting verification score.

$$CM_{\text{in-domain}}(X) = \text{sigmoid}\left[V_{\text{in-domain}}(W)\right] \quad (2)$$

4. Discourse Coherence

In spoken language systems, a user's response is typically related to the preceding utterance in the dialogue, either a system prompt in a spoken dialogue system, or the counterpart's input in a speech-to-speech translation system. If a series of utterances are not coherent in terms of discourse flow, it is likely that a recognition error occurred in one of these utterances. To incorporate this information, we introduce a novel confidence measure, *discourse coherence*, which verifies topic consistency across consecutive utterances.

Discourse coherence is defined as the distance between the current utterance (X_i) and the preceding utterance in the dialogue (X_{i-1}) in the topic-confidence space derived by topic classification. We investigate three distance measures for this purpose: cosine distance, Euclidean distance, and weighted-Euclidean distance.

For a topic confidence vector of the current utterance $(C(t_1|W_i), \dots, C(t_m|W_i))$ and that of the preceding utterance, $(C(t_1|W_{i-1}), \dots, C(t_m|W_{i-1}))$, cosine distance is defined as:

$$dist_{\text{cosine}}(W_i, W_{i-1}) = \frac{\sum_{j=1}^m C(t_j|W_i) \cdot C(t_j|W_{i-1})}{\sqrt{\sum_{j=1}^m C(t_j|W_i)^2 \sum_{j=1}^m C(t_j|W_{i-1})^2}} \quad (3)$$

Euclidean distance as:

$$dist_{\text{Euclidean}}(W_i, W_{i-1}) = \sqrt{\sum_{j=1}^m \left(C(t_j|W_i) - C(t_j|W_{i-1})\right)^2} \quad (4)$$

and weighted-Euclidean distance as:

$$dist_{\text{weighted}}(W_i, W_{i-1}) = \sqrt{\sum_{j=1}^m \lambda_j \cdot \left(C(t_j|W_i) - C(t_j|W_{i-1})\right)^2} \quad (5)$$

For the weighted-Euclidean case, the discriminant weights $(\lambda_1, \dots, \lambda_m)$ used during in-domain verification (sub-section 3.3) are applied.

The *discourse coherence* confidence score, $CM_{\text{discourse}}(X_i|X_{i-1})$, is generated by applying a sigmoid transformation to the resulting distance. This score is high when the topic-distance between the two utterances is close.

$$CM_{\text{discourse}}(X_i|X_{i-1}) = \text{sigmoid}\left[\text{dist}(W_i, W_{i-1})\right] \quad (6)$$

5. Joint Confidence by Combining Multiple Measures

A joint measure of recognition confidence $CM(X)$ is realized by combining the two proposed confidence measures described in the previous sections with an orthodox measure based on GPP [12], $CM_{\text{gpp}}(X_i)$.

5.1 Generalized Posterior Probability

Generalized posterior probability (GPP) assesses the confidence of a recognition hypothesis in terms of confusability with competing hypotheses during ASR decoding. It is formulated as described in [12]. At the word-level a generalized word posterior probability $GWPP(x_j)$ is calculated as the posterior against all competing hypotheses within the word graph. At the utterance-level the generalized posterior probability is defined as the geometric mean of the individual GWPPs of component words in the utterance (Equation 7). We adopt this measure in the proposed framework.

$$CM_{\text{gpp}}(X) = \left(\prod_{j=1}^l GWPP(x_j)\right)^{\frac{1}{l}} \quad (7)$$

x_j : j -th word in recognition hypothesis of X
 l : number of words in X

5.2 Joint Confidence Score

A joint confidence score, $CM(X_i)$, is defined by combining all three measures via a linear weighted model.

$$CM(X_i) = \lambda_{\text{gpp}} * CM_{\text{gpp}}(X_i) + \lambda_{\text{in-domain}} * CM_{\text{in-domain}}(X_i) + \lambda_{\text{discourse}} * CM_{\text{discourse}}(X_i|X_{i-1}) \quad (8)$$

where $\lambda_{\text{gpp}} + \lambda_{\text{in-domain}} + \lambda_{\text{discourse}} = 1$

Utterance verification is performed by comparing this score with a predefined threshold (φ). If the score is greater than this threshold, the input hypothesis is verified to be correct; otherwise it is assumed to contain recognition errors and rejected. The model weights $(\lambda_{\text{gpp}}, \lambda_{\text{in-domain}}, \lambda_{\text{discourse}})$ and decision threshold (φ) are trained to minimize verification errors on a development set.

6. Experimental Evaluation

The performance of the proposed confidence scoring framework was evaluated on utterance verification for spontaneous dialogue via the ATR speech-to-speech translation system [23]. This system operates on a travel-conversation domain and performs bi-directional translation between English and Japanese.

The ATR "basic travel expressions" corpus [24] was used for training of the language models applied during speech recognition and the topic classification and in-domain verification models. This corpus consists of 14 topic classes (e.g., accommodation, shopping, transit, etc.) and 400k training sentences for each language side. Evaluation data, which are different from the above corpus, consist of natural spoken dialogue between native English and native Japanese speakers via the ATR speech-to-speech translation system. Dialogue data were collected based on a set of pre-defined scenarios, relating to the travel domain. A summary of these data is shown in Table 1.

6.1 Baseline Speech Recognition Performance

First, the performance of the English and Japanese ASR systems were evaluated. ASR was performed using the ATR speech recognition system, ATRASR [25]. For acoustic analysis, 12-dimensional MFCC, energy and first order derivatives were computed. Acoustic models consisted of gender-dependent triphone HMMs, trained using the successive state-splitting algorithm [26]. English models consisted of 2800 shared states

Table 1 Summary of evaluation data

	set-1	set-2
# dialogues	184	185
English side		
# utterances	1808	1761
WER	14.2%	13.7%
SER	54.5%	52.2%
Japanese side		
# utterances	1857	1791
WER	10.8%	10.2%
SER	45.0%	42.7%

WER: Word error rate
SER: Sentence error rate

with 5 Gaussian mixture components per state, set up for 43 phones, and the Japanese models consisted of 2,100 shared states with 5 Gaussian mixture components per state, set up for 26 phones. Lexicons consisting of 16k and 20k words were applied for the English and Japanese sides, respectively. During recognition, word graphs were initially generated by applying a bi-gram language model and these were then rescored using a trigram language model to obtain the final recognition output. Speech recognition performance (WER, SER) for the Japanese and English dialogue sides is shown in Table 1.

6.2 Evaluation Measures

To recover from speech recognition errors, spoken language systems typically confirm those words which are critical to task success. However, in speech-to-speech translation, a recognition error of a single word can cause unintelligible translation results to be produced. Also, defining a set of "keywords" that affect translation performance is not trivial. For speech-to-speech translation tasks the simplest and most effective method to handle speech recognition errors is to prompt users to re-phrase the entire input (so long as it is in-domain). Thus, verification is formulated as rejecting entire utterances if they contain one or more recognition errors (in sub-section 6.7 we relax this constraint and reject only critical errors that affect translation). System performance was evaluated by CER (confidence error rate) [27] defined by Equation 9. Errors include false acceptance (FA) of incorrectly recognized utterances and false rejection (FR) of correctly recognized utterances.

$$CER = \frac{\#false\ acceptance + \#false\ rejection}{\#utterances} \quad (9)$$

Experiments were performed using a 2-fold evaluation procedure. Evaluation data were split into two sets. First, set-1 was used as development data and the weights $\{\lambda_{gpp}, \lambda_{in-domain}, \lambda_{discourse}\}$ and decision threshold (φ) in Equation 8 were trained. The system performance was then evaluated on set-2. This

Table 2 Baseline system performance

dev. set	test set	CER	
		Accept All	GPP
English side			
set-2	set-1	54.5%	17.8%
set-1	set-2	52.2%	18.7%
Average		53.5%	18.2%
Japanese side			
set-2	set-1	45.0%	21.0%
set-1	set-2	42.7%	20.4%
Average		43.8%	20.7%

Accept All: assume all utterances correctly recognized
GPP: Generalized Posterior Prob. based verification

Table 3 Performance of *in-domain confidence* measure

Side	IC	GPP	GPP+IC
English	31.7%	18.2%	16.6% (9.1%)
Japanese	33.1%	20.7%	19.4% (6.4%)

GPP: Generalized Posterior Prob.
IC: *in-domain confidence*

process was then repeated using set-2 as development data and set-1 as test data. The average over these two cases was used as the final evaluation measure.

6.3 Performance of GPP-based Verification

For the baseline performance, utterance verification using GPP alone was evaluated. The CERs of this system ("GPP"), and a reference case where all hypotheses are assumed to be correct ("Accept All") are shown in Table 2. The performance of the "Accept All" case matches the SER of the respective ASR systems, and thus CERs are close to 50% for both language sides. The GPP-based baseline system significantly reduced CERs to 18.2% and 20.7% for the English and Japanese sides, respectively.

6.4 Effect of *In-domain Confidence* Measure

In the first experiment, the effectiveness of the proposed *in-domain confidence* measure was evaluated. A confidence score was generated for each utterance using either *in-domain confidence* ("IC"), generalized posterior probability ("GPP"), or a weighted combination of these measures ("GPP+IC"), and the resulting score was used for utterance verification. The average CERs for the three cases are shown in Table 3.

When only *in-domain confidence* was considered ("IC"), the performance was much lower than the GPP case. However, combining both measures via a linear weighted model ("GPP+IC") realized significant reductions in CER of 9.1% and 6.4% for the English and Japanese sides, respectively, compared to using the GPP measure alone.

The *in-domain confidence* measure is effective in detecting errors that result in a mismatch of domain

Table 4 Performance of *in-domain confidence* measure

Side	GPP	GPP+DC		
		cosine	Euclid.	w-Euclid
English	18.2%	18.1%	17.2%	17.0% (6.5%)
Japanese	20.7%	20.5%	20.1%	19.9% (4.2%)

GPP: Generalized Posterior Prob.
DC: *discourse coherence*
cosine: cosine distance
Euclid.: Euclidean distance
w-Euclid: weighted-Euclidean distance

or inconsistency of topic within an utterance. However, this measure is not sensitive to recognition errors of topic independent words, especially function words, which make up the majority of recognition errors. Combining this measure with GPP realizes a significant reduction in verification errors compared to using the GPP measure alone, demonstrating the effectiveness of incorporating “*high-level*” knowledge into the utterance verification process.

6.5 Effect of *Discourse Coherence* Measures

Next, *discourse coherence* was incorporated and the three distance measures described in Section 4 were compared. The performance of each distance measure when combined with GPP is shown in Table 4.

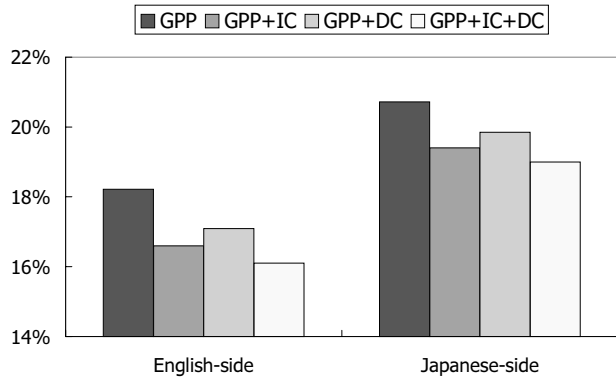
For the English side, the weighted-Euclidean distance provided the best performance, obtaining a reduction in CER of 6.5% compared to the GPP baseline. Similar performance was also gained for the Euclidean case. However, using the cosine distance obtained little improvement over the baseline system. A similar tendency was also observed for the Japanese side.

For the weighted-Euclidean case, we adopted the discriminant weights from in-domain verification to weight coordinates. These are not necessarily optimal for the *discourse coherence* measure. Some other method to estimate these weights may further improve the performance.

6.6 Performance of Joint Confidence Score

Next, both *in-domain confidence* and *discourse coherence* measures were incorporated into utterance verification. The performance of the GPP baseline, the individual measures combined with GPP, and a joint measure combining all three scores are shown in Figure 4.

For the English side, incorporating *in-domain confidence* (“*GPP+IC*”) and *discourse coherence* (“*GPP+DC*”) reduced the CER to 16.6% and 17.0%, respectively, compared to using the GPP measure alone (CER = 18.2%). These correspond to relative reductions in CER of 9.1% and 6.5%, respectively. Incorporating both measures jointly (“*GPP+IC+DC*”) provided reduction in CER of 11.4% (from 18.2%

**Fig. 4** Joint confidence score performance

GPP: Generalized Posterior Prob.
IC: *in-domain confidence*
DC: *discourse coherence*

to 16.1%). Similar performance was gained for the Japanese side with a relative reduction in CER of 8.1% (from 20.7% to 19.0%) when both “*high-level*” measures were incorporated.

The two proposed measures, *in-domain confidence* and *discourse coherence*, focus on dissimilar aspects of semantic consistency, and thus incorporating both measures improved utterance verification performance compared to using either measure individually. This demonstrates the combined effectiveness of the two measures.

6.7 Content-based Utterance Verification

Finally, we explore “*high-level*” content-based utterance verification in which the effect of ASR errors on the back-end NLP system is considered. In this paper, a machine translation back-end system is assumed, thus, ASR errors that have little or no effect on the translation result (as identified in [28]) can be ignored. Specifically, erroneous recognition of function words, difference in noun plurality and POS were disregarded in this experiment. Applying this criterion, 164 English and 175 Japanese utterances that contained negligible recognition errors were treated as being correctly recognized, which equates to a reduction in CER of around 10% for the “*Accept All*” case (CERs of 48.3% and 39.0% for the English and Japanese sides, respectively). The utterance verification performance for the GPP and the proposed joint confidence score is shown in Figure 5.

For the proposed approach (“*joint*”), CER was reduced to 15.1% and 18.3% for the English and Japanese sides, respectively. This relates to relative reductions in CER of 14.6% and 7.8% compared to an equivalent system using GPP alone. The performance for the Japanese-side was similar to that gained in Section 6.6, however, for the English side, much larger improvement was gained. Difference in noun-plurality

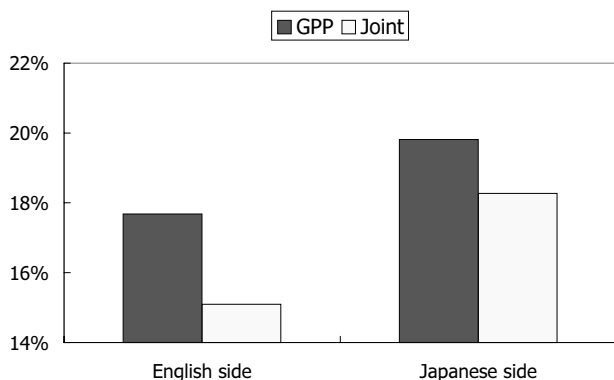


Fig. 5 Performance of content-based utterance verification

GPP: Generalized Posterior Prob.
 Joint: Proposed joint confidence score

accounted for a large number of trivial ASR errors in English and these errors were found not to significantly affect the proposed measures. For Japanese, however, noun-plurality is not applicable, and erroneous recognition of function words often reduced “joint” confidence and also degraded translation quality, thus, the improvement with content-based verification was small.

7. Conclusion

We have investigated a novel confidence measure framework that incorporates “*high-level*” knowledge. Specifically, two confidence measures were proposed: *in-domain confidence*, the degree of match between the input utterance and the application domain of the backend system, and *discourse coherence*, the consistency between consecutive utterances in a dialogue session. Experimental evaluations were performed on spontaneous dialogue via the ATR speech-to-speech translation system. The two proposed measures were effective in improving utterance verification accuracy, and the CER was reduced by 11.4% (for the English case) compared to using generalized posterior probability (GPP) alone. Furthermore, when minor ASR errors that do not affect translation were ignored, such as noun plurality, the proposed approach obtained a further reduction in CER for the English case.

In this paper, we evaluated the proposed confidence scoring with a speech-to-speech translation system. However, this framework is not limited to this task and can easily be incorporated into other spoken language systems, for example, call-routing and spoken dialogue systems. Before implementing the proposed framework, however, an adequate set of pre-defined topic classes is required. These can be defined by call destinations in a call-routing system, or sub-domains in a spoken dialogue system. Automatic generation of these classes via sentence clustering, as described in [29], should also be explored.

Acknowledgement: The research reported here was supported in part by a contract with the National Institute of Information and Communications Technology entitled, “A study of speech dialogue translation technology based on a large corpus”.

References

- [1] T. Takezawa, T. Morimoto, Y. Sagisaka, N. Campbell, H. Iida, F. Sugaya, A. Yokoo, and S. Yamamoto, “A Japanese-to-English speech translation system: ATR-MATRIX,” *Proc. ICSLP*, pp. 957–960, 1998.
- [2] W. K. Lo, and F. K. Soong, “Generalized posterior probability for minimum error verification of recognized sentences,” *In Proc. ICASSP*, pp. 85–89, 2005.
- [3] C. Ramond, Y. Esteve, F. Bechet, R. DeMori, and G. Damnati, “Belief confirmation in spoken dialogue systems using confidence measures,” *In Proc. ASRU*, 2003.
- [4] R. San-Segundo, B. Pellon, and W. Ward, “Confidence Measures for Dialogue Management in the CU communicator system,” *In Proc. ICASSP*, vol. 4, pp. 697–700, 2000.
- [5] G. Bouwman, J. Sturn, and L. Boves, “Incorporating confidence measures in the Dutch timetable information system developed in the ARISE project,” *In Proc. ICASSP*, vol. 1, pp. 493–496, 1999.
- [6] C. Pao, P. Schmid, and J. Glass, “Confidence scoring for speech understanding systems,” *In Proc. ICSLP*, pp. 815–818, 1998.
- [7] D. Guillevic, S. Gandrabur, and Y. Normandin, “Robust semantic confidence scoring,” *In Proc. ICSLP*, pp. 853–856, 2002.
- [8] S. Cox and R. Rose, “Confidence measures for the switchboard database,” *In Proc. ICASSP*, pp. 511–514, 1996.
- [9] T. Schaaf and T. Kemp, “Confidence measures for spontaneous speech recognition,” *In Proc. ICASSP*, vol. 2, pp. 887–890, 1997.
- [10] T. Kawahara, C.-H. Lee, and B.-H. Juang, “Flexible speech understanding based on combined key-phrase detection and verification,” *In Proc. IEEE transactions on speech and audio processing*, issue 6, vol. 6, pp. 558–568, Nov. 1998.
- [11] G. Bouwman, L. Boves, and J. Koolwaaaji, “Weighted phone confidence measures for automatic speech recognition,” *In Proc. COST249 workshop on voice operated telecom services*, pp. 59–62, 2000.
- [12] W. Lo, F. Soong, and S. Nakamura, “Generalized posterior probability for minimizing verification errors at sub-word, word and sentence levels,” *In Proc. International symposium on Chinese spoken language processing*, pp. 13–16, 2004.
- [13] A. Gunawardana, H.-W. Hon, and L. Jiang, “Word-based acoustic confidence measures for large vocabulary speech recognition,” *In Proc. ICSLP*, vol. 3 pp. 791–794, 1998.
- [14] G. Bernardis and H. Bourland, “Improving posterior based confidence measures in hybrid HMM/ANN speech recognition systems,” *In Proc. ICSLP*, pp. 775–778, 1998.
- [15] E. Mengusoglu and C. Ris, “Use of acoustic prior information for confidence measures in ASR applications,” *In Proc. EUROSPEECH*, vol. 4, pp. 2557–2561, 2001.
- [16] D. Bohus and A. Rudnick, “Integrating multiple knowledge sources for utterance-level confidence annotation in the CMU communicator spoken dialog system,” *Technical Report CMU-CS-02-190*, 2002.
- [17] C. Raymond, F. Bechet, N. Camlin, R. De-Mori, and G. Damnati, “Semantic interpretation with error correction,” *In Proc. ICASSP*, vol. 1, pp. 29–32, 2005.
- [18] M. Walker J. Wright and I. Langkilde, “Using natural language processing and discourse features to identify under-

- standing errors,” In Proc. *International Conference on Machine Learning*, pp. 1111–1118, 2000.
- [19] S. Pradhan and W. Ward, “Estimating semantic confidence for spoken dialogue systems,” In Proc. *ICASSP*, vol. 1, pp. 233–236, 2002.
- [20] I. Lane, T. Kawahara, T. Matsui, and S. Nakamura, “Out-of-domain detection based on confidence measures from multiple topic classification,” In Proc. *ICASSP*, vol. 1, pp. 757–760, 2004.
- [21] T. Joachims, “Text categorization with support vector machines,” *Proc. European Conference on Machine Learning*, 1998.
- [22] S. Katagiri, C.-H. Lee, and B.-H. Juang, “New discriminative training algorithm based on the generalized probabilistic descent method,” In Proc. *IEEE Workshop NNSP*, pp. 299–300, 1991.
- [23] T. Takezawa, A. Nishino, K. Takashima, T. Matsui, and G. Kikui, “An experimental system for collecting machine-translation aided dialogues,” In Proc. *Proc. FIT2003*, Vol. 2, pp. 161–162, 2003.
- [24] T. Takezawa, M. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, “Towards a broad-coverage bilingual corpus for speech translation of travel conversations in the real world,” In Proc. *LREC*, pp. 147–152, 2002.
- [25] T. Shimizu et al., “Spontaneous dialogue speech recognition using cross-word context constrained word graph,” In Proc. *ICASSP*, pp. 145–148, 1996.
- [26] T. Jitsuhiro, T. Matsui, and S. Nakamura, “Automatic Generation of Non-uniform HMM Topologies Based on the MDL Criterion,” *IEICE Trans. Inf. & Syst.*, vol. E87-D, no. 8, pp. 2121–2129, 2004.
- [27] M. G. Rahim, C. H. Lee, and B. H. Juang, “Discriminative utterance verification for connected digits recognition,” *IEEE transactions on speech and audio processing*, vol. 5, pp. 266–277, 1997.
- [28] Y. Sawai, G. Kikui, and H. Yamamoto, “The relationship between speech recognition quality and translation performance in speech-to-speech translation,” *ATR Technical Report SLT-0088*, 2005. (In Japanese)
- [29] B. Carlson, “Unsupervised topic clustering of switchboard speech messages,” *Proc. ICASSP*, pp. 315–318, 1996.