Speed or Accuracy?
A Study in Evaluation of Simultaneous Speech Translation

Takashi Mieno, Graham Neubig, Sakriani Sakti, Tomoki Toda, Satoshi Nakamura

Graduate School of Information Science, Nara Institute of Science and Technology, Japan
{mieno.takashi.mh1,neubig,ssakti,tomoki,s-nakamura}@is.naist.jp

10.21437/Interspeech.2015-498

Abstract

Simultaneous speech translation is a technology that attempts to reduce the delay inherent in speech translation by beginning translation before the end of explicit sentence boundaries. Despite best efforts, there is still often a trade-off between speed and accuracy in these systems, with systems with less delay also achieving lower accuracy. However, somewhat surprisingly, there is no previous work examining the relative importance of speed and accuracy, and thus given two systems with various speeds and accuracies, it is difficult to say with certainty which is better. In this paper, we make the first steps towards evaluation of simultaneous speech translation systems in consideration of both speed and accuracy. We collect user evaluations of speech translation results with different levels of accuracy and delay, and using this data to learn the parameters of an evaluation measure that can judge the trade-off between these two factors. Based on these results, we find that considering both accuracy and delay in the evaluation of speech translation results helps improve correlations with human judgements, and that users placed higher relative importance on reducing delay when results were presented through text, rather than speech.
Index Terms: simultaneous speech translation, evaluation

1. Introduction

In traditional speech translation systems, it is standard to first segment speech recognition results into full sentences, then perform translation sentence-by-sentence [1]. However, as sentences can be relatively long, particularly in the case of formal speech such as lectures or presentations, this method can cause a significant delay between the speaker's original utterance and the presentation of translation results. Due to this fact, there has been a recent surge in interest in *simultaneous speech translation*, in which translation starts before explicit sentence boundaries. Within this framework, the main question is how to reduce the delay without causing a decrease in translation accuracy, and a wide variety of methods have been proposed to tackle this problem [2, 3, 4, 5, 6].

However, the task of translating before a full sentence has been observed is inherently difficult, even for humans, and as a result most previous work has noted that there is a trade-off between speedy presentation of translation results and production of high-quality translations, as shown in Figure 1.[1] This fact raises several central questions affecting the usefulness of these systems: How important is it to reduce the delay? When there is a trade off between accuracy and speed, which should we choose? Knowing the answer to these questions is essential if we hope to perform effective evaluation or optimization of simultaneous speech translation systems.

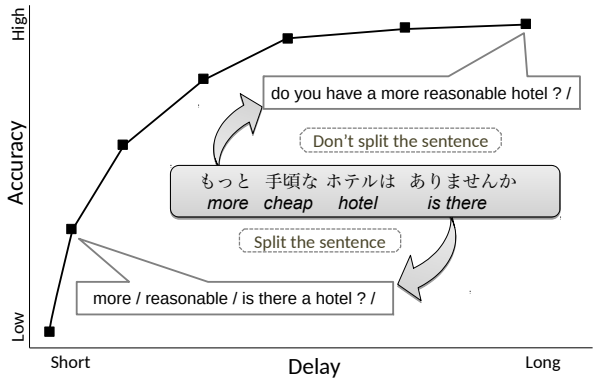

Figure 1: An example of the trade-off between speed and accuracy in simultaneous speech translation

In this paper, we make a first step towards answering these questions by devising an evaluation measure for simultaneous speech translation that simultaneously considers delay and accuracy (Section 2). Specifically, we first present annotators with multiple translations of various accuracies and delays along with the original video, and have the evaluators rank the results according to their preference (Section 3). Using these ranked translation results, we learn a classifier that takes delay and accuracy as input, and automatically learns weights of delay and accuracy that allow us to correctly reflect these human evaluation results (Section 4).

In our experiments, we use this method to create an evaluation measure on data from English-Japanese translation of TED Talks (Section 5).[2] We present results to the annotators in two presentation modalities: *text subtitles* and *read speech*, which simulate speech-to-text (S2T) translation and speech-to-speech translation (S2S) respectively. As a result of experiments, we find that the proposed evaluation measure considering both accuracy and delay achieves better correlation with human results than evaluation measures that consider each of the elements individually. We also found significant differences between presentation modalities, with users placing more emphasis on delay when results were presented by text than when results were presented by speech.

[1] This example is from Japanese-English translation for clarity, but the remainder of our examples and experiments target English-Japanese translation.

[2] http://www.ted.com

Copyright © 2015 ISCA

2. The Evaluation Function

In order to achieve our goal of creating an evaluation measure that jointly considers both accuracy and delay, we define a scoring function

$$s(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}), \quad (1)$$

where $\boldsymbol{x}$ is a displayed speech translation result, and $\boldsymbol{\phi}$ is a function that calculates a feature vector from $\boldsymbol{x}$. This is a general formulation, but in this paper we assume that $\boldsymbol{\phi}(\boldsymbol{x})$ calculates exactly two features: some measure of translation accuracy, and delay. $\boldsymbol{w}$ is a vector that specifies the relative importance of the features in the feature vector. Our goal in this paper is to learn this feature vector based on human evaluation data, which will allow us to: 1) learn a scoring function that can evaluate existing speech translation systems, 2) examine the learned weights in $\boldsymbol{w}$, telling us something about the subjective relative importance of speed and accuracy in simultaneous speech translation systems.

3. Data Collection

In this section, we describe how to collect human evaluations used as training data for the function described in the previous section. Note that we focus on the general data collection approach, and discuss the actual data used in our experiments in Section 5.

3.1. Creation of Data to Evaluate

The first step in obtaining human evaluations is creating the data to evaluate. In this work, we use video data as input, as the visual stimulus of the video can help the user judge the extent to which the translation results are delayed. We set the length of each video to be 4-5 sentences, which we judged (through trial and error) to be enough to evaluate the translation accuracy, but not too much to be a burden on the evaluators. In addition, we were careful to select segments that did not strongly rely on the previous context, and that had a clear start of the utterance.

3.2. Presentation of Translation Results

The next step is creation and display of translation results. The creation of translation results, like in standard speech translation evaluation, can be done by running a translation system on the input sentences or input speech. The presentation of results, on the other hand, poses unique challenges for this specific task, and thus we discuss it in some detail.

3.2.1. Modality of Presentation

When presenting speech translation results, we can think of two modalities of presentation: *speech* (for S2S translation) and *text* (for S2T translation).

In order to evaluate output in the speech modality, it is necessary to create speech data from translation results. In this work, we consider two methods to do so: the use of a text-to-speech (TTS) system, or having a human read the results and record natural speech. Preliminary experiments showed that listening to TTS results over a long recording session was tiring for annotators, and thus to prevent loss of concentration from affecting our results, we decided to use recorded speech in this paper. Recording is performed sentence-by-sentence. These recorded sentences are then added to the original video at the appropriate timing (discussed in Section 3.2.2). When doing so, we follow the common protocol in voice-over translation [7] of reducing the volume of the original speech to a low, but audible level, and overlaying the translated speech.

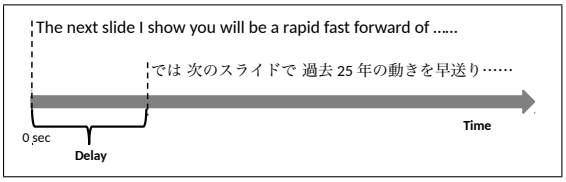

Figure 2: An example of delay

Evaluating in the text modality is relatively simple. The videos we use in evaluation already have English subtitles, so for each English subtitle segment, we manually align the translated text that corresponds to this segment. Then, the translated segments can be displayed at the same timing as the English segments, possibly with the addition of delay described in the following section.

3.2.2. Delay of Results

The next step in the process is to consider the *delay* that occurs in simultaneous speech translation. In order to do so, we treat a translation that begins at exactly the same time as the source utterance as a translation with zero delay. Utterances with delay are created by beginning presentation of the translation results later than the source utterance, as shown in Figure 2.

In the case of text input, this delay is performed for each subtitle segment by simply delaying the subtitle segment by the appropriate number of seconds. In the case of speech input, this delay is performed for each sentence, with the translated speech for each sentence starting the appropriate number of seconds after the start of the sentence in the original speech. One thing to note is that in the speech data, due to the length of the translated text, in some cases the length of the read speech can be longer than that of the sentence in the original utterance. When this occurs, to prevent two translated speech sentences from overlapping, the latter segment is delayed just enough to prevent overlap.

3.3. Evaluation of Translation Results

The scoring function in Section 2 takes an input $\boldsymbol{x}$ and returns a score indicating the quality of the presented result. Perhaps the most obvious way to create training data for this function is to have a human annotator watch the video, and assign a score, for example on a scale of 1-5. However, in contrast to traditional MT evaluation for adequacy or fluency [8], when considering both delay and translation accuracy at the same time it is not trivial to come up with a standard that specifies in which cases a particular score should be assigned.

As a way to overcome this problem, we opt to have human evaluators assign not an absolute score, but make relative comparisons between the outputs of multiple systems. Specifically, we have evaluators watch videos with multiple translation results of varying accuracies and delays, and ask the evaluators to rank them based on how "easy to understand" they are. Evaluators are allowed to re-play the videos as many times as they wish, and asked to base their decisions solely on the content and timing of the speech, and ignore other factors such as speech speed, voice quality, or intonation.

In order to ensure that the evaluators fully understand the content of the original utterance, we first show them a manually

translated reference. This is necessary when translated results are presented by speech, as the source sentence, while audible, is overlapped with the target and not possible to hear accurately. This is less necessary when presenting information by subtitles, but as evaluators are required to be native speakers of the target language, but not the source language, displaying a reference can help ensure that the original content was understood correctly, and thus we display a reference in this case as well.

4. Learning Parameters through Ranking

Next, we describe the process used to learn the parameters $\boldsymbol{x}$ of the evaluation function. Specifically, this data fits naturally in the framework of *learning to rank*, and in this research we used RankSVM [9], the most standard method in this framework. For a single input video, the training data for learning to rank takes the form $\{\langle \boldsymbol{\phi}(\boldsymbol{x}_i), y_i \rangle\}_{i=1}^{m}$, where m is the number of translation candidates for this video, and $y_i \in \{1, 2, ...\}$ is the rank of each candidate assigned by the human evaluator.

Learning to rank attempts to learn a function $f(\boldsymbol{x})$ that returns a higher score for inputs with better ranks (in other words, lower numbers). If we define this function as $f(\boldsymbol{\phi}(\boldsymbol{x})) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x})$, for any pair of feature vectors $\boldsymbol{\phi}(\boldsymbol{x}_i) \neq \boldsymbol{\phi}(\boldsymbol{x}_j)$ in the training data, this can be expressed as

$$\begin{aligned} y_i < y_j \quad &\Leftrightarrow \quad f(\boldsymbol{\phi}(\boldsymbol{x}_i)) > f(\boldsymbol{\phi}(\boldsymbol{x}_j)) \\ &\Leftrightarrow \quad \boldsymbol{w}^T(\boldsymbol{\phi}(\boldsymbol{x}_i) - \boldsymbol{\phi}(\boldsymbol{x}_j)) > 0. \end{aligned}$$

In order to find a weight vector $\boldsymbol{w}$ that satisfies this condition, the RankSVM considers each pair of training instances, and generates training data for a binary classifier for each pair of indices (i, j)

$$\langle \boldsymbol{\phi}(\boldsymbol{x}_i) - \boldsymbol{\phi}(\boldsymbol{x}_j), z_{i,j} \rangle \tag{2}$$

where the true label is defined as positive or negative based on the difference between the manually annotated ranks

$$z_i = \begin{cases} +1 & y_i < y_j \\ -1 & y_i > y_j \end{cases}. \tag{3}$$

This binary data can then be used to train a standard binary classifier such as SVMs, yielding a trained weight vector $\boldsymbol{w}$ that can distinguish between better and worse inputs.

5. Experiments

5.1. Experimental Setting

5.1.1. Data

In this work, we use data from TED Talks, using an English-Japanese test set from the Simultaneous Translation Corpus [10]. We have evaluators watch videos with three different translation results for a single input speech, and rank the three results from 1 to 3 based on how easy the content was to understand, disallowing ties. Each evaluator ranked 20 videos a piece, with 15 evaluators for the experiment using speech, and 10 evaluators for the experiment using text, resulting in a total of 900 and 600 pairwise comparisons between systems.

Based on the procedure described in Section 3.1, we chose 20 sections of videos from TED Talks ranging from 20-30 seconds and with an average of 4.45 sentences. The videos were chosen so that half contained slides, which we hypothesized may increase the importance of delay. The speech for the translation results was recorded by two speakers, a male speaker in the case that the speaker in the original TED talk was male, and a female speaker when the TED speaker was female.

Table 1: BLEU+1, RIBES, and Adequacy for TED subtitles, interpreters with 15 and 4 years of experience, Travatar, and Moses.

	TED	I-15	I-4	Trav	Mos
BLEU+1	0.38	0.14	0.11	0.20	0.16
RIBES	0.82	0.59	0.53	0.67	0.59
Adeq	0.89	0.57	0.45	0.48	0.38

5.1.2. Translation Data

In order to ensure that our findings are as widely applicable as possible, we generated 5 types of translations results for each video, using as wide a variety of methods as possible. Specifically, we used the original TED subtitles, 2 types of results from human simultaneous interpreters (with 15 and 4 years of experience respectively) from the simultaneous translation corpus, and 2 types of results generated by machine translation (using the phrase-based Moses [11] and tree-based Travatar [12] toolkits).

As measures of translation accuracy, we used 3 metrics, 2 automatic and 1 requiring manual human annotation. For manual evaluation, we used a 1-5 adequacy score [8], taking the average score of 3 annotators, and finally scaling the score to be between 0 and 1. As automatic metrics we used sentence-level BLEU+1 [13], and RIBES [14]. To create references, we had a human translator create translation results independently of the original TED subtitle translations. Japanese was segmented with KyTea [15] prior to evaluation. The accuracy of each system is shown in Table 1.

5.1.3. Delay

Given these translation results, we next generate videos with these results delayed by a certain number of seconds, following the procedure in Section 3.2. Specifically, we use 7 varieties of delay: $\boldsymbol{D} = \{0, 1, 2, 3, 5, 7, 10\}$.

5.1.4. Training and Evaluation

As a classifier to solve the ranking problem in Section 4, we used LIBLINEAR [16] with the default settings.[3] To evaluate the quality of the learned evaluation measure, we perform 20-fold cross validation, holding out one of the videos as test data and using the other 19 as training data. Given this classifier, we would like to measure its quality. To do so, measure the accuracy of each pairwise decision in the ranking problem, which gives a chance rate of 50%. If we find an accuracy significantly higher than the chance rate, we can say that the features being used in evaluation are effective in discriminating between good and bad translations according to human judgements.

5.2. Experimental Results

First, to examine the usefulness of considering delay and accuracy in the evaluation of simultaneous speech translation results, we show in Table 2 the accuracy of evaluation, for both the text and speech modalities. Starting at the top of the table, each system uses as features: delay only (row 1), accuracy only (rows 2-4), or both delay and accuracy (rows 5-7).

The first thing we can observe from this table is that in the

[3] Attempts to tune the parameters did not result in a significant gain in accuracy.

Table 2: Pairwise evaluation accuracy using each feature set for the text and speech modalities.

Feat.	Measure	Text	Speech
Del.	-	0.58	0.54
Acc.	BLEU+1	0.52	0.55
	RIBES	0.57	0.61
	Adeq.	0.65	0.70
Del. + Acc.	BLEU+1	0.62	0.60
	RIBES	0.64	0.61
	Adeq.	0.68	0.72

Table 3: Weights of each feature and the ratio of accuracy to delay for each modality.

Modal.	Measure	Delay	Acc.	Ratio
Speech	BLEU+1	-0.041	1.19	28.9
	RIBES	-0.038	0.99	26.2
	Adeq.	-0.040	1.27	31.9
Text	BLEU+1	-0.013	2.03	155
	RIBES	-0.018	1.51	86.6
	Adeq.	-0.018	1.99	114

majority of cases, considering both delay and translation accuracy results in more accurate evaluation than considering the factors independently. This confirms our hypothesis that both speed and accuracy have an effect on subjective impressions, and that the proposed evaluation method is able to take advantage of this fact for better evaluation.

The second thing we can observe from this table is that the trends in the two modalities are noticeably different. In the case of text presentation, we can see that delay plays an important role, with a classifier using delay alone achieving higher accuracy than that of the automatic evaluation measures. On the other hand, for speech presentation, delay is relatively unimportant, underperforming all accuracy measures, and only contributing a small amount when combined with them. We hypothesize that this is due to fact that when translation results are presented through subtitles, the original speech is played at a relatively loud volume, and thus the user becomes more aware of the difference in timing between the original speech and presentation of the results, and thus places more emphasis on delay when making their subjective judgements.[4]

Next, to explicitly examine the relative importance of delay (in seconds) and accuracy, in Table 3 we show the weights learned by each classifier, along with the ratio between the weights, which shows the relative importance of accuracy compared to delay. Based on these results, we can observe that the ratio between accuracy and delay is not affected much by the measure used to evaluate translation accuracy, but it is affected significantly by modality of presentation.

If we focus on human adequacy, and divide each ratio by 4 to map back from 0-1 scaled scores to the original 1-5 adequacy scores, we can see that in the case of text a single point of adequacy is judged equivalent to a reduction of $31.9/4 = 8.0$ seconds of delay, while in the case of speech it is $114/4 = 28.5$

[4]Of course, this, like other results in this paper is dependent on the genre of the speech, and thus it is likely that different results would be seen for different genres such as dialog.

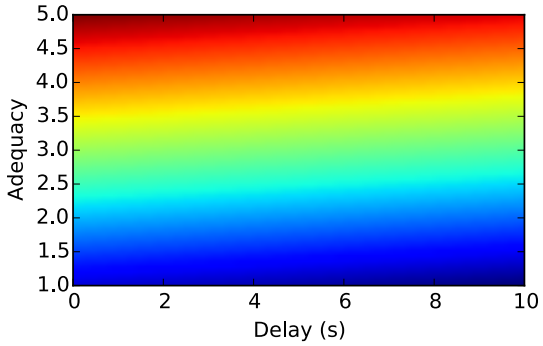

Figure 3: A visualization of the learned function for speech

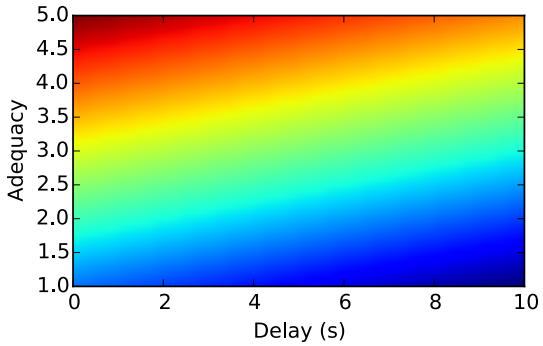

Figure 4: A visualization of the learned function for text

seconds of delay. In addition, in Figures 3 and 4, we show a visualization of the evaluation functions for speech and text respectively. These numbers and figures further demonstrate the relative importance of delay when translating into text, as opposed to speech.

6. Conclusion

In this paper, we performed an examination of the relative importance of speed and accuracy in simultaneous speech translation. As a result we found that considering both speed and translation accuracy in the evaluation of simultaneous speech translation systems results in more effective evaluation. We also found that speed was relatively important when presenting results by text, at least in the domain of TED talks that we examined in this paper.

The most relevant future work is the actual application of this measure to the design of speech translation systems. For example, the metric can be used to optimize the parameters of a simultaneous speech translation system to achieve the optimal balance of expeditious translation and accuracy. We also plan on refining the metric by extending the linear model in this paper to non-linear models that can learn more flexible evaluation functions. In addition, while the results presented here are applicable to TED Talks in English-Japanese translation, we plan on examining results for other genres and language pairs.

Acknowledgements: Part of this work was supported by JSPS KAKENHI Grant Number 24240032.

7. References

[1] E. Matusov, A. Mauser, and H. Ney, "Automatic sentence segmentation and punctuation prediction for spoken language translation," in *Proc. IWSLT*, 2006, pp. 158–165.

[2] C. Fügen, A. Waibel, and M. Kolss, "Simultaneous translation of lectures and speeches," *Machine Translation*, vol. 21, no. 4, pp. 209–252, 2007.

[3] S. Bangalore, V. K. R. Sridhar, P. K. L. Golipour, and A. Jimenez, "Real-time incremental speech-to-speech translation of dialogs," in *Proc. NAACL*, 2012.

[4] T. Fujita, G. Neubig, S. Sakti, T. Toda, and S. Nakamura, "Simple, lexicalized choice of translation timing for simultaneous speech translation," in *Proc. 14th InterSpeech*, 2013.

[5] Y. Oda, G. Neubig, S. Sakti, T. Toda, and S. Nakamura, "Optimizing segmentation strategies for simultaneous speech translation," in *Proc. ACL*, 2014.

[6] A. Grissom II, H. He, J. Boyd-Graber, J. Morgan, and H. Daumé III, "Don't until the final verb wait: Reinforcement learning for simultaneous machine translation," in *Proc. EMNLP*, 2014, pp. 1342–1352.

[7] E. Franco, A. Matamala, and P. Orero, "Voice-over translation." Peter Lang Pub Inc, 2013.

[8] DARPA, "Linguistic Data Annotation Specification:Assessment of Fluency and Adequacy in Arabic-English and Chinese-English Translations," 2002.

[9] R. Herbrich, T. Graepel, and K. Obermayer, "Support vector learning for ordinal regression," in *Proc. Artificial Neural Networks*, 1999.

[10] H. Shimizu, G. Neubig, S. Sakti, T. Toda, and S. Nakamura, "Collection of a simultaneous translation corpus for comparative analysis," in *Proc. LREC*, 2014.

[11] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proc. ACL*, 2007, pp. 177–180.

[12] G. Neubig, "Travatar: A forest-to-string machine translation engine based on tree transducers," in *Proc. ACL*, 2013, pp. 91–96.

[13] C.-Y. Lin and F. J. Och, "A method for evaluating automatic evaluation metrics for machine translation," in *Proc. COLING*, 2004, pp. 501–507.

[14] H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada, "Automatic evaluation of translation quality for distant language pairs," in *Proc. EMNLP*, 2010, pp. 944–952.

[15] G. Neubig, Y. Nakata, and S. Mori, "Pointwise prediction for robust, adaptable Japanese morphological analysis," in *Proc. ACL*, 2011, pp. 529–533.

[16] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, 2008.