Empirical Evaluation of the Communicative Effectiveness of an Automatic Speech-to-Speech Translation System

Xinhao Wang[†]*, Keelan Evanini*[‡]

Educational Testing Service R&D
[†]90 New Montgomery St., Suite 1500, San Francisco, CA, USA
[‡]660 Rosedale Rd., Princeton, NJ, USA
{xwang002,kevanini}@ets.org

Abstract

In this study we evaluate the ability of a state-of-the-art speech-to-speech machine translation system, Skype Translator, to preserve the communicative effectiveness of its input speech. Specifically, we elicited Spanish spoken responses from 24 native speakers of Spanish in the following two scenarios: academic speaking (in the context of 6 questions from a standardized assessment of academic speaking proficiency) and daily communication (in the context of 6 map-based directions tasks). Skype Translator was then used to translate these Spanish responses into English, and expert human raters subsequently provided speaking proficiency scores using both holistic and analytic scoring rubrics with a 1-4 range for both the original Spanish responses and the corresponding automatically translated English versions. The results show that the average holistic scores for the academic speaking task decreased from 3.5 (for the original Spanish responses) to 1.9 (for the English translations) on a 4-point scale, and the average task completion scores for the map-based tasks decreased from 3.4 to 2.0. While these results show that automated speech-to-speech machine translation does not yet fully meet the communicative demands of these speaking tasks, they also indicate how far the technology has improved in recent years and can serve as a benchmark for tracking further improvements in the future.

Index Terms: speech-to-speech machine translation, human scoring, communicative effectiveness

1. Introduction

Automatic speech-to-speech translation is a complex task that joins together multiple NLP and speech processing systems, including automatic speech recognition (ASR), machine translation (MT), and text-to-speech synthesis (TTS). Early speech-to-speech translation systems used speech recorded under controlled conditions with restricted vocabularies and were therefore limited to specific topic domains and speaking styles. During the past two decades there have been several large-scale speech-to-speech translation research initiatives involving more authentic and complex spoken language, including TRANSTAC and BOLT (both funded by DARPA in the U.S.) and TC-STAR (funded by the European Commission); these efforts, in addition to the recent application of deep learning approaches to ASR and MT, have helped to spur on dramatic improvements in the quality of automatic speech-to-speech translation. Recently, a few companies have publicly released free speech-to-speech translation systems that have received widespread media attention, including Google Translate[1] and Skype Translator[2]. Given the continually increasing accuracy of these systems across an increasingly diverse set of conversational topics and situations, one may wonder whether they will someday be sufficient for enabling successful communication between speakers of two different languages across a wide range of situations, thus potentially rendering foreign language learning unnecessary in some cases. In this study, we explore this question by examining how well a state-of-the-art automatic speech-to-speech translation system performs in two different communicative situations.

Specifically, we employed the Skype Translator system, a state-of-the-art speech-to-speech translation system developed by Microsoft, which currently enables near-real-time translation to and from the following 9 languages: English, French, German, Chinese, Italian, Spanish, Portuguese, Arabic, and Russian. English and Spanish were selected as the target language pair for this study, due to the fact that this was the initial language pair available in the first release of Skype Translator. We empirically examined how effectively Skype Translator translates spoken Spanish into spoken English in two different communicative scenarios. One set of tasks simulated simple language that would be used with tourists; in these tasks, speakers provided directions in Spanish to locations on a map. The second set of tasks elicited more complex speech similar in nature to speech that would be used in an academic environment, such as at a university. The main research question targeted by this study is to determine to what extent the English speech-to-speech translation output based on the Spanish input is intelligible and meets the communicative demands of the two types of speaking tasks. This question was addressed by obtaining expert human scores across a range of speaking proficiency dimensions for both the spoken Spanish responses and corresponding English output produced by the speech-to-speech translation system and then comparing the two sets of scores to determine the strengths and weaknesses of the system.

2. Prior Work

Several studies have investigated the impact of MT technology on foreign language learning, in particular, how it can be used as a pedagogical tool. For example, one potentially beneficial approach is to use an MT system to translate a text in the target language into a learner's native language and then have the learner revise the translated text by referring back to the original text [1]. This process of correcting the MT errors leads learners to focus on linguistic differences between the original text and the MT output. A similar approach is to use MT sys-

[1] https://translate.google.com/

[2] https://www.skype.com/en/features/skype-translator/

tems in a bi-directional manner, i.e., by translating texts into both the learner's native language and target language and having the learner review and correct both sets of MT output either by using their own native-speaker judgments (for translations into their native language, as in [1]) or reference materials (for translations into the target language) [2]. [3] discusses these approaches along with others in a comprehensive overview of different potential classroom uses for MT tools and their associated advantages and disadvantages. While most of the studies using MT technology in a foreign language learning environement have focused on text MT systems (since high-quality speech-to-speech machine translation systems have only been made freely available for widespread use in recent years), a few have also looked at speech-to-speech translation systems. For example, [4] applied a speech-to-speech translation system in the context of a multilingual dialog system for queries about the weather; in this application, language learners were able to pose a question in their native language (English) and obtain an automated translation in the target language (Mandarin Chinese) that they could then use as a model for their spoken response.

Empirical evaluations of the quality of speech-to-speech translation systems typically employ metrics that can be computed automatically based on references (such as Word Error Rate for the ASR component and BLEU for the MT component) as well as metrics based on human annotation of important characteristics of the spoken utterance, such as fluency and adequacy [5]. Since these standard evaluation metrics may not fully capture the effectiveness of the speech-to-speech translation system in accomplishing the speaker's communicative goals, some studies have instead emphasized the importance of evaluating the ability of the speech-to-speech translation output to meet the communicative demands of the speaking task. For example, [6] demonstrates how conventional measures of speech-to-speech translation accuracy underestimate the functional accuracy of a commercial system that produced translations in the tourism domain. Similarly, the current study focuses on the ability of a speech-to-speech translation system to meet the communicative demands of a range of speaking tasks, in particular, by investigating to what extent the quality of the automated translations deteriorates from the original responses provided by native speakers.

3. Data

This section provides details about the speaking tasks that were used in this study, the methodology for recruiting participants and collecting spoken responses from them, the procedure for processing the responses using Skype Translator, and the human scores that were obtained for the Spanish spoken responses and English Skype Translator output.

3.1. Speaking Prompts

Speaking tasks from two different domains, academic speaking and daily communication, were included in this study in order to evaluate the performance of automatic speech-to-speech translation systems with different types of spoken language. For the academic speaking domain, six speaking tasks from a standardized assessment of English proficiency for academic purposes, the TOEFL iBT®, were included.[3] All six tasks were drawn from a single TOEFL iBT test, and thus included two Independent tasks (in which test takers are asked to speak about a familiar topic or provide a personal opinion) and four Integrated tasks (in which test takers are provided with reading and/or listening passages on university-related and academic topics and are asked to answer a question about the content of these passages). The six TOEFL tasks (including the reading and listening passages) were translated into Spanish for this study; Spanish versions of the listening passages were recorded by native speakers of Spanish. For the daily communication domain, six quesions eliciting directions based on two different maps that depict landmarks and streets in fictitious cities were included; an example of one of these Map-based tasks is as follows: *¿Cómo puedo llegar al hospital desde el estadio?* (*How can I get from the hostpital to the stadium?*)

[3]See the following document for sample TOEFL iBT speaking tasks: `https://www.ets.org/Media/Tests/TOEFL/pdf/SampleQuestions.pdf`.

3.2. Data Collection

24 adult participants (12 males and 12 females) with Spanish as their native language were recruited for this study and were compensated $50 for their participation. The average age of the participants was 29.8 (std. dev. = 8.2, min. = 20, max. = 46). The education levels of the participants were as follows: 3 participants completed post-graduate studies, 15 participants graduated from college, 5 participants completed high school, and one participant completed elementary school.

The speaking tasks (including stimulus materials such as reading passages, maps, etc.) were presented to the participants in Microsoft PowerPoint slides, and the built-in Windows sound recorder was used to capture their spoken responses via a headset microphone. The participants were given fixed response times for the Academic speaking tasks corresponding to the response times in the TOEFL iBT assessment (45 seconds for the Independent speaking tasks and 60 seconds for the Integrated tasks). For the Map-based tasks, there was no fixed response window; participants were instructed to advance to the next task when they had completed their response to the previous one; typical responses to the Map-based tasks consisted of approximately 2-4 short sentences, and the average response duration was 26 seconds. The participants responded to the twelve questions in Spanish, and the audio file from each participant's recording session was manually segmented into 12 separate audio files, one for each Spanish spoken response; a total of 287 spoken responses were collected (one participant did not provide a response for one of the Map-based tasks).

3.3. Skype Translator Processing

Each Spanish response was input manually into Skype Translator on one computer by setting Stereo Mix as the recording input and playing the audio file. The following three types of output from Skype Translator were then captured on another computer that was connected to the first one via Skpye: automatic speech recognition, automatic machine translation, and text-to-speech synthesis. The text to speech synthesis output was subsequently used for the human scoring experiment, while the other two types of output were analyzed to obtain a better understanding of the performance of the Skype Translator system.

3.4. Human Scoring

Human raters were recruited from the authors' organization to provide scores for the Spanish spoken responses and English Skype Translator output; separate rater pools were recruited for scoring the two different languages. A total of 13 raters, all

with college degrees and experience scoring spoken constructed responses, were recruited to score the English responses; 9 of them scored the Academic responses exclusively, and the other 4 scored the Map-based responses. 12 raters, all native speakers of Spanish with college degrees and 7 with prior experience scoring spoken constructed responses (in English), were recuited to score the Spanish responses; 7 of them scored the Academic responses exclusively, 3 of them scored the Map-based responses exclusively, and the remaining 2 scored both Academic and Map-based responses.

The Academic responses (both Spanish and English versions) were scored using both holistic and analytic scoring rubrics. The standard TOEFL iBT scoring rubrics[4] were used to obtain the holistic scores; analytic scores were obtained by adapting the holistic rubrics to focus on the following six specific components of speaking proficiency: pronunciation, rhythm and intonation, lexical range and accuracy, grammatical range and accuracy, content appropriateness and accuracy, and cohesion. The holistic scoring rubrics employed a 4-point rating scale (with 1 represnting the lowest proficiency and 4 representing the highest) and the analytic scoring rubrics employed a 5-point rating scale with the score point 0 reserved for responses that contained no intelligible speech (this was included since a few of the English responses produced by the automated speech translation system were completely unintelligible).

The Map-based tasks (both Spanish and English versions), on the other hand, were only scored using the analytic scoring rubrics, since pre-existing scoring rubrics did not exist for this task and since the analytic rubrics provide more detailed information that can help highlight the strengths and weaknesses of the speech-to-speech translation system. The analytic scoring rubrics for the Map-based tasks were identical to the ones that were used for the acadmic speaking tasks for the following four components of speaking proficiency: pronunciation, rhythm and intonation, lexical range and accuracy, and grammatical range and accuracy. However, the content appropriateness and accuracy and cohesion components were not included in the scoring rubrics for the Map-based tasks, since they are relevant only to the Academic speaking tasks; instead, the analytic rubrics for the Map-based tasks included a component about task completion, i.e., indicating whether the directions provided by the speaker were precise and easy to follow.

As an example of the analytic scoring rubrics that were used, Table 1 presents the scoring rubrics for the lexical range and accuracy dimension of speaking proficiency. These rubrics for lexical range and accuracy were part of the section of the scoring rubrics that also included grammatical range and accuracy and was prefaced by the following general description: *This domain relates to the range and precision of the vocabulary and grammatical structures used. Control of a range of vocabulary as well as comfort with a variety of grammatical forms and structures result in efficient and effective expression of ideas.* As shown in Table 1, the rubrics emphasize communicative effectiveness through their focus on how the speaker's lexical range and accuracy impact the intelligibility of their response; the other analytic scoring rubrics contain a similar emphasis on intelligibility and communicative effectiveness.

Benchmark responses were selected for both the holistic and analytic scoring rubrics to exemplify how the rubrics should be applied. Raters first reviewed the rubrics and benchmarks and then proceeded to rate responses that were randomly assigned to them based on the following constraints:

- All responses were double scored
- The same rater could not provide both ratings for a response
- Raters who scored the Academic tasks first scored responses using the holistic rubrics and then scored responses using the analytic rubrics
- Raters who scored the Academic tasks could not provide both the holistic score and the analytic score for the same response

The raters who scored the English responses produced by Skype Tranlsator were informed of the nature of the study and were aware that the spoken responses were produced by an automatic speech to speech translation system based on original recordings of Spanish responses. They were instructed to score the responses in an identical manner as they would typically score English spoken responses, and to not modify their interpretation of the rubrics based on the fact that they were produced by a speech-to-speech translation system; this point was discussed during the training phase and the raters reviewed the benchmark responses until they agreed about the interpretation of the rubrics for the English responses.

Table 1: *Analytic scoring rubrics for the lexical range and accuracy dimension of speaking proficiency*

Score	Description
4	Word choice is mostly precise. Speaker shows good command of a range of vocabulary and idiomatic expressions with only minor errors.
3	Word choice shows some range but lacks precision at times. Attempts at idiomatic speech may be flawed. Minor lexical errors are evident but do not impact overall meaning.
2	Word choice is sometimes vague, imprecise, and/or repetitive. Lexical errors obscure meaning at times.
1	The response is marked by frequent repetitions of words and phrases. Mostly simple routine vocabulary with noticeable errors. Lexical errors may often obscure meaning.
0	No intelligible speech.

3.5. Sample Responses

Table 2 provides transcriptions of two sample Spanish responses that were collected in this study, one for an Academic task and one for a Map-based task, along with the resulting Skype Translator automatic speech recognition and machine translation output and the human scores that were obtained for each response (the scores shown are the average of the two human scores for each of the rubrics). Note that the raters provided scores in all cases based solely on the audio files (the original Spanish response or the audio file produced by Skype Translator's TTS engine based on the automatic translation); the text transcriptions are presented in Table 2 to exemplify the procedure, but the raters did not have access to them.

[4] https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf

Table 2: *Two sample responses along with their associated Skype Translator output (automated speech recognition and machine translation) and human scores based on the average of two human scores for each of the rubrics*

	Spanish transcription	Skype Translator Spanish ASR	Skype Translator English MT	Scores (Spanish, English)
Academic	Creo que el lugar adecuado para estudiar es el lugar que te de la tranquilidad, que te genere un ambiente propicio para, para tener la concentración y los recursos que necesites para tu estudio, entonces tienes que tener una buena iluminación eh unos buenos recursos, si necesitas libros, computadores, eh un ambiente apropiado con poco ruido, eh una silla cómoda para que te sientes bien y y bueno sobre todo el entorno que te genere tranquilidad y y no distracciones para tener la concentración suficiente para estudiar.	Creo que el lugar adecuado para estudiar el lugar que te la tranquilidad que te genera un ambiente propicio para. Para tenerla con su bueno, me voy. No bueno, iluminación en los buenos recursos. Si necesitas libros computadores. La un ambiente apropiado con poco ruido en una silla cómoda para que te sientes bien, y y bueno, sobre todo el entorno que te genere tranquilidad, y, y no distracciones para tener la concentración suficiente para estudiar.	I think the right place to study the place you the peace of mind that generates an environment conducive to. To have her with her well, I'm off. Not good, lighting in the good resources. If you need computer books. The appropriate atmosphere with little noise in a comfortable chair so that you feel good, and and well, especially the environment that you generate peace of mind, and, and no distractions to have enough concentration to study.	holistic: 4.0, 2.0 pronunciation: 4.0, 2.5 prosody: 4.0, 2.5 lexicon: 4.0, 2.0 grammar: 4.0, 2.0 content: 3.5, 1.5 cohesion: 3.0, 2.0
Map-based	Desde la tienda, sales hacia el restaurante a mano derecha, llegas al centro comercial, volteas a la izquierda y ahí encuentras la piscina después del centro comercial.	Desde la tienda salida hacia el restaurante mano derecha llegado centro comercial volteado a la izquierda. Y ahí encuentra la piscina del centro comercial.	From the store exit to the restaurant right hand arrived Mall flipped over to the left. And there is the pool of the Mall.	pronunciation: 4.0, 1.5 prosody: 4.0, 2.0 lexicon: 4.0, 1.0 grammar: 4.0, 1.0 completion: 4.0, 2.0

4. Results

4.1. Speech Recognition Accuracy

In order to evaluate the accuracy of the Skype Translator's Spanish automatic speech recognition system, manual orthographic transcriptions were obtained from a native Spanish speaker for the Spanish spoken responses that were used as input to Skype Translator. To compute the WER between the transcriptions and the Skype Translator Spanish ASR output, punctuation and filled pauses (*um*, *eh*, etc.) were removed from both the transcriptions and the Skype Translator output. The overall WER across all Spanish responses was 0.310, and the WERs for the 12 different speaking tasks ranged from 0.268 to 0.340.

4.2. Comparison of Proficiency Scores

In order to evaluate to what extent the Skype Translator system could meet the communicative demands of the speaking tasks, we compared the scores provided for the original Spanish responses with the scores provided for the English Skype Translator output. Since each response was double scored, the average of the two expert scores was used as the final score for each response; for this analysis, the benchmark responses with their associated gold standard scores were also included.

Table 3 presents descriptive statistics (mean, standard deviation, minimum, and maximum) for the average of the two human scores for both the Spanish and English responses across all of the scoring rubrics. In addition, the table also presents descriptive statistics for the difference between the the average of the two human scores. The difference was calculated by subtracting the English score (for the Skype Translator output) from the Spanish score (for the original response). Since the scores for the Spanish responses were nearly always higher than the corresponding scores for the English responses, the differences are nearly always positive; however, there were a few cases in which the English response received a higher score than the Spanish response, as exemplified by the few negative values in the column displaying minimum score differences in Table 3.

As shown in Table 3, the scores for the Spanish responses are skewed towards the high end of the score range, and the average scores for all scoring dimensions are close to the maximum value of 4.0; however, not all speakers received perfect scores for all scoring rubrics, as might be expected given the fact that they are native speakers of Spanish. Since the scoring rubrics are based on how the different aspects of speaking proficiency contribute to how effectively the spoken response meets the communicative demands of each specific speaking task (as shown in Table 1), native speakers are not expected to always receive perfect scores. In particular, Table 3 shows that the analytic scores for the Spanish responses for dimensions that focused primarily on the content of the response (content and cohesion for the Academic tasks and task completion for the Map-based tasks) were slightly lower than the analytic scores based on rubrics that focused primarily on delivery and language use (lexicon and grammar).

The main result shown in Table 3 is that the proficiency scores for the translated English responses are consistently lower than the scores for the original Spanish responses. All of the average differences between the two sets of scores except one are 1.2 or greater, with the largest average difference

being 1.9. The mean scores for the English responses cluster mostly around 2.0 across the different rubrics for both the Academic and Map-based responses. These relatively low mean scores for the English responses and relatively large differences between the scores for the Spanish and English responses indicate that the automatically translated output produced by Skype Translator is considerably less effective at meeting the communicative demands of the speaking tasks. In particular, the large average differences between the Spanish and English holistic scores for the Academic task (1.7) and between the Spanish and English task completion scores for the Map-based task (1.4) indicate that the quality of the translated responses in meeting the commnicative demands of the speaking task was substantially degraded.

Table 3 shows that the score differences between the Spanish and English responses were lowest for the pronunciation scoring rubric for both task types, with an average difference of 0.5 for the Academic tasks and 1.2 for the Map-based tasks. This result indicates that the text-to-speech component of Skype Translator generally produces translated English speech that is intelligible and natural sounding. However, the larger degradations in scores for the other analytic rubrics demonstrate how the English responses contain issues with other aspects of English speaking proficiency that obscure the communicative intent of the original Spanish responses.

A total of 17 different English responses received a score of 0 from at least one rater for at least one of the analytic scoring rubrics. This indicates that the human raters were unable to understand any substantial amount of meaningful speech for these responses. 16 of these responses were for Map-based tasks and one was for an Academic task. Since the Map-based responses were much shorter in duration than the Academic responses, it makes sense that the frequency of 0 scores was much higher for the Map-based responses. None of the Spanish responses received a score of 0.

Table 4 presents the Pearson correlations between the scores provided by the raters based on the different scoring rubrics for the Spanish spoken responses and the automatic spoken English translations produced by Skype Translator. As shown in the table, the correlations are highest for the holistic scores for the Acadmic tasks (0.419) and for the task completion scores for the Map-based tasks (0.456) and are quite low for the pronunciation and prosody scores for both tasks. This result is consistent with the distributions of scores as shown in Table 3: the Spanish holistic and task completion scores are somewhat lower and have relatively larger standard deviations whereas the Spanish pronunciation and prosody scores are nearly all 4.0 (which contributes to the low correlations).

4.3. Human Agreement

The human-human agreement statistics for the double ratings are presented in Table 5 in terms of quadratically weighted kappa and percentage of exact match (the benchmark samples are excluded from this analysis). As shown in the table, the kappa values are quite low for many of the analytic scoring rubrics, especially for the scores for the Map-based tasks. This is likely an indication of the difficulty of providing reliable scores for specific aspects of speaking proficiency for such short spoken responses, since the responses contain only minimal evidence for each of the proficiency aspects. Also, the low kappa values are caused in some cases by the high levels of chance agreement due to the fact that a large number of the Spanish responses received scores of 4; this is demonstrated, for example, by the exact match percentage of 0.943 for the pronunciation scores for the Map-based tasks. The fact that a large number of the Spanish responses received scores of 4, especially for the Map-based tasks, led to less variability among the scores for the Spanish responses and can also explain why the exact match rates were generally higher for the Spanish responses than for the English responses (as shown in Table 3, the standard deviations for the English ratings were generally higher than for the Spanish ratings).

5. Discussion and Conclusion

In this study, we recruited 24 native speakers of Spanish and elicited their Spanish spoken responses for 12 speaking tasks, 6 drawn from an assessment of academic speaking proficiency and 6 targeted to elicit map-based directions. Skype Translator, a state-of-the-art speech-to-speech machine translation system, was used to translate the Spanish spoken responses into spoken English. Expert human raters provided analytic scores covering a range of speaking proficiency dimensions for all responses as well as holistic scores for the responses to the Academic tasks. Analyses were then conducted to compare the differences between the scores for the original Spanish spoken responses and their automatically translated English counterparts. The results show that the quality of the translated English spoken responses is substantially degraded across all aspects of speaking proficiency, with an average difference of 1.7 for the holistic scores for the Academic tasks and average differences ranging from 1.2 - 1.9 for the analytic rubrics for the Map-based tasks; the magnitude of these degradations corresponds to approximately half of the score range (1-4). These large differences between scores for the original Spanish responses and English speech-to-speech translation output demonstrate that the translated responses do not adequately meet the communicative demands of the speaking tasks.

Contrary to initial expectations, there was not a noticeable difference in quality between the English translations for the Academic tasks and the Map-based tasks. Both resulted in similar degradations from the scores for the original Spanish responses despite the fact that the language used in the Map-based tasks was less complex and therefore expected to be more amenable to automated translation.

In addition, this study showed that the human rating task was challenging, as demonstrated by the relatively low inter-annotator agreement results presented in Table 5. One of the main factors contributing to the low kappa agreement values is likely the skewed distribution of the human scores towards the high end of the proficiency scale, especially for the Spanish responses, as discussed in Section 4.3. However, other factors related to the novelty of the rating task may also have contributed to lower agreement. For example, this study was the first time these particular analytic scoring rubrics had been used, so additional work may be necessary to refine the rubrics to improve agreement. Also, none of the raters had scored Spanish spoken responses previously and 5 of the 12 Spanish raters had no prior experience scoring spoken responses. Finally, the task of scoring machine translation output was novel and additional training and discussion may be required to achieve higher inter-annotator agreement for this task.

While the results of this study indicate that automatic speech-to-speech translation systems are not imminently poised to completely meet the communicative demands in speaking tasks that involve open-domain, complex, spontaneous speech, follow-up studies should be conducted on a regular basis, since

Table 3: *Descriptive statistics for the human scores for the Spanish responses, the English Skype Translator output, and the response-level score differences*

	Rubric	Spanish				English				Difference			
		Mean	S.D.	Min.	Max.	Mean	S.D.	Min.	Max.	Mean	S.D.	Min.	Max.
Academic	holistic	3.5	0.5	1.5	4.0	1.9	0.6	1.0	4.0	1.7	0.6	0.0	3.0
	pronunciation	3.9	0.3	2.5	4.0	3.4	0.5	2.0	4.0	0.5	0.6	-1.0	2.0
	prosody	3.7	0.4	2.4	4.0	2.4	0.5	1.0	4.0	1.3	0.6	-0.5	3.0
	lexicon	3.7	0.5	2.0	4.0	2.1	0.5	1.0	3.5	1.6	0.6	0.0	3.0
	grammar	3.7	0.5	2.0	4.0	2.1	0.4	1.0	3.0	1.6	0.6	0.0	3.0
	content	3.5	0.6	1.0	4.0	2.0	0.6	0.5	4.0	1.5	0.7	0.0	3.0
	cohesion	3.6	0.6	1.5	4.0	2.1	0.6	0.5	3.5	1.4	0.7	0.0	3.0
Map-based	pronunciation	4.0	0.1	3.0	4.0	2.8	0.5	1.0	4.0	1.2	0.6	0.0	3.0
	prosody	3.9	0.2	3.0	4.0	2.2	0.5	0.5	3.5	1.8	0.5	0.5	3.0
	lexicon	3.8	0.3	2.5	4.0	1.9	0.7	0.5	3.5	1.9	0.7	0.0	3.5
	grammar	3.8	0.3	2.5	4.0	1.9	0.6	0.5	3.0	1.9	0.6	0.0	3.5
	task completion	3.4	0.7	1.0	4.0	2.0	0.8	0.5	3.5	1.4	0.8	-0.5	3.5

Table 4: *Correlations between scores provided based on different scoring rubrics for Spanish spoken responses and English Skype Translator output*

	Academic	Map-based
holistic	0.419	N/A
pronunciation	0.057	0.014
prosody	0.065	0.198
lexicon	0.239	0.157
grammar	0.248	0.155
content	0.399	N/A
cohesion	0.367	N/A
task completion	N/A	0.456

Table 5: *Human-human agreement represented by quadratic weighted kappa and exact match percentage*

	Rubric	kappa		Exact match	
		Spanish	English	Spanish	English
Academic	holistic	0.369	0.535	51.4%	62.3%
	pron.	0.296	0.329	84.1%	55.8%
	prosody	0.431	0.163	72.5%	51.4%
	lexicon	0.211	0.269	65.2%	50.0%
	grammar	0.311	0.19	65.9%	51.4%
	content	0.344	0.474	55.1%	52.9%
	cohesion	0.464	0.43	64.5%	54.3%
Map-based	pron.	-0.019	-0.054	94.3%	39.3%
	prosody	-0.088	0.015	82.9%	37.1%
	lexicon	-0.053	0.362	70.7%	40.7%
	grammar	0.014	0.206	72.1%	41.4%
	completion	0.396	0.505	53.6%	47.1%

the technology is improving rapidly. It is expected that the ASR performance will continue to improve in the near-term with the application of more sophisticated Deep Neural Network-based acoustic models and the availability of greater amounts of training data, thus leading to substantially lower word error rates. In addition, developments in Neural Machine Translation models continue to improve the quality of MT systems. With sufficiently accurate ASR and MT components, it is conceivable that an automatic speech-to-speech translation system could successfully translate spoken language to meet the communicative needs of a wide range of speaking tasks in the future, thus potentially reducing the need to learn foreign languages for certain situations. Future research should therefore continue to investigate the performance of these systems and their impact on foreign language learning and pedagogy. In addition, subsequent studies should investigate additional language pairs as well as the performance of speech-to-speech translation systems when translating from English into another language.

6. Acknowledgments

The authors would like to thank several colleagues who contributed to this project, including Melissa Lopez, Florencia Tolentino, and Ayana Stevenson for the data collection; Florencia Tolentino, Hillary Molloy, and Ben Leong for the data processing; Pamela Mollaun and Molly Palmer for leading the scoring effort; and all of the raters who scored the spoken responses.

7. References

[1] H. Somers, "Three perspectives on MT in the classroom," in *Proceedings of the MT Summit VIII Workshop on Teaching Machine Translation*, 2001, pp. 25–29.

[2] D. D. Anderson, "Machine translation as a tool in second language learning," *CALICO Journal*, vol. 13, no. 1, pp. 68–97, 1995.

[3] A. Niño, "Machine translation in foreign language learning: Language learners' and tutors' perceptions of its advantages and disadvantages," *ReCALL*, vol. 21, no. 2, pp. 241–258, 2009.

[4] C. Wang and S. Seneff, "High-quality speech-to-speech translation for computer-aided language learning," *ACM Transactions on Speech and Language Processing*, vol. 3, no. 2, pp. 1–21, 2006.

[5] O. Hamon, C. Fügen, D. Mostefa, V. Arranz, M. Kolss, A. Waibel, and K. Choukri, "End-to-end evaluation in simultaneous translation," in *Proceedings of the 12th Conference of the European Chapter of the ACL*, 2009, pp. 345–353.

[6] J. Bernstein and E. Rosenfeld, "Evaluating an automatic speech-to-speech interpreter in context," Poster presented at the 146th Meeting of the Acoustical Society of America, Kansas City, MO, 2012.