An Empirical Simulation-based Study of Real-Time Speech Translation for Multilingual Global Project Teams

Fabio Calefato, Filippo Lanubile
Dipartimento di Informatica
Università di Bari
Bari, Italy
fabio.calefato@uniba.it,
filippo.lanubile@uniba.it

Rafael Prikladnicki, João Henrique S. Pinto,
Computer Science School
PUCRS
Porto Alegre, Brazil
rafaelp@pucrs.br,
joaohenrique.jhsp@gmail.com

ABSTRACT

Context: Real-time speech translation technology is today available but still lacks a complete understanding of how such technology may affect communication in global software projects.
Goal: To investigate the adoption of combining speech recognition and machine translation in order to overcome language barriers among stakeholders who are remotely negotiating software requirements.
Method: We performed an empirical simulation-based study including: Google Web Speech API and Google Translate service, two groups of four subjects, speaking Italian and Brazilian Portuguese, and a test set of 60 technical and non-technical utterances.
Results: Our findings revealed that, overall: (i) a satisfactory accuracy in terms of speech recognition was achieved, although significantly affected by speaker and utterance differences; (ii) adequate translations tend to follow accurate transcripts, meaning that speech recognition is the most critical part for speech translation technology.
Conclusions: Results provide a positive albeit initial evidence towards the possibility to use speech translation technologies to help globally distributed team members to communicate in their native languages.

Categories and Subject Descriptors

D.2.9 [**Management**]: Programming teams

H.4.3 [**Communications Applications**]: Computer conferencing, teleconferencing, and videoconferencing.

I.2.7 [**Natural Language Processing**]: Machine translation.

General Terms

Experimentation, Human Factors.

Keywords

Controlled experiment; global software engineering; machine translation; requirements meetings.

ESEM'14, September 18–19, 2014, Torino, Italy.
Copyright 2014 ACM 978-1-4503-2774-9/14/09...$15.00

1. INTRODUCTION

Opportunities for global software development are limited in those countries with a lack of English-speaking professionals. For this reason, communication often occurs between native and non-native English speakers with the drawback of an unequal ability to fully understand and contribute to discussions. Being one of the most communication-intensive activities in software development, requirements engineering suffers much from language difficulties in global software projects [8][9][13]. When participants to requirements meetings are weak in listening and speaking English, they might take a great advantage to use their mother language, thus ensuring equality of participation in meetings and reducing those risks due to language skill disparities, such as conversation domination, social consensus, and peer/group pressure.

Speech translation technologies are today experiencing a tremendous growth of interest because of advances in the fields of automatic speech recognition as well as machine translation [10]. In this paper, we report from an empirical simulation-based study where we investigated the adoption of combining speech recognition and machine translation in order to overcome language barriers among stakeholders who are remotely negotiating software requirements.

In our previous work, we begun investigating how machine translation affects distributed group communication for complex tasks. We first run a simulated study, which proved that state-of-the-art machine translation services, such as Google Translate, could be embedded into synchronous text-based chat with a negligible extra time [5]. Then, we conducted a controlled experiment [6] and a replication [7] to investigate whether real-time machine translation could be successfully used instead of English in distributed multilingual requirements meetings, with non-native speakers with different level of proficiency. We could observe that, despite far from 100% accuracy, real-time machine translation is not disruptive of the conversation flow, is accepted with favor, and grants a more balanced discussion. However, since typing is slower than speaking, even low proficient subjects had sufficient time to elaborate on the meaning of sentences written in English and participate in the discussion. As such, the effectiveness of real-time machine translation is expected to become even more evident when audio is involved [18].

The remainder of this paper is structured as follows. Section 2 provides some background on speech recognition and machine translation technologies. Section 3 describes the experiment, including the design, the variables, the instrumentation and execution.

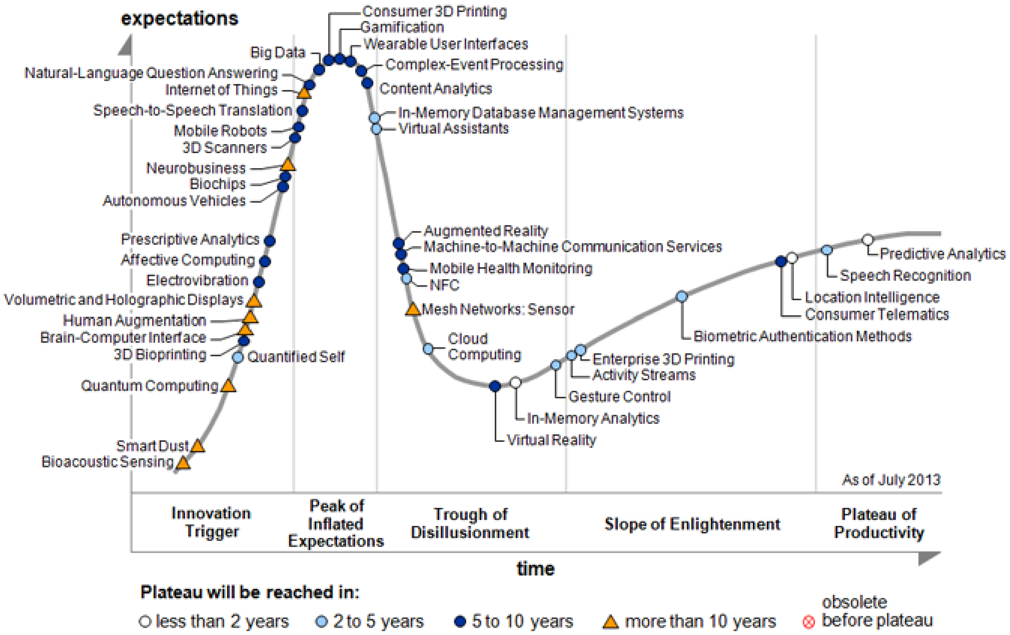


Figure 1. Hype Cycle for Emerging Technologies, 2013 (from www.gartner.com/newsroom/id/2575515)

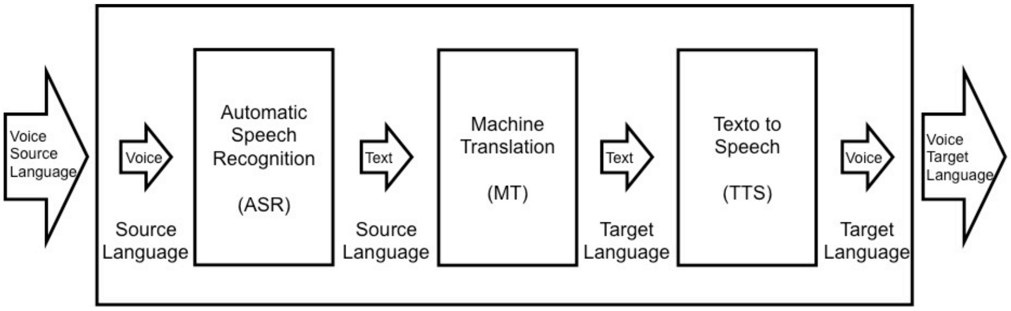


Figure 2. Speech-to-speech translation

Findings are respectively presented and discussed in Section 4 and 5. Threats to validity are described in Section 6. Finally, conclusions and future research activities are presented in Section 7.

2. SPEECH TRANSLATION

The Gartner Hype Cycle for emerging technologies (Figure 1) provides an assessment of the maturity, business benefit and future direction of more than 2,000 technologies, grouped in 98 areas. The 2013 version reports an important trend in speech-to-speech translation. Speech-to-speech translation has three components: automatic speech recognition (ASR), machine translation (MT), and voice synthesis (or text to speech; TTS). As shown in Figure 2, the ASR component processes the voice in its original language, creating a text version of what the speaker said. This text in the original language goes through the MT component, which translates it to the target language. Finally, this translated text goes through the TTS component, which "speaks" the text using a synthesized voice in the target language. In this section, we review the background on the two most critical building blocks of speech translation systems, that is, speech recognition and machine translation, which are also the focus of our paper.

2.1 Speech Recognition

Automatic speech recognition (ASR) is defined as the transcription of spoken words into text [11]. Producing a transcript from a continuous and unbroken stream of text, as in the case of extemporaneous speech is challenging. Research on speech recognition dates back to the early 1970s [15]. In fact, the automatic translation of spoken words into text has a broad range of applications, from captioning video for the hearing-impaired to voice controlled computer operation and dictation. Moreover, several contrasting approaches are available, such as template matching, the rule-based approach, and the statistical approach. Over the last years, however, there has been a substantial advance in the field [11]. Nevertheless, many challenges are still open, like speaker and language variability or the size of the vocabulary to be recognized.

Some of the available technologies for automatic speech recognition are: Microsoft Speech SDK, which is part of the .NET Framework package and incorporates the native API for Windows, Microsoft Speech API (SAPI). It is typically used by developers to let applications recognize spoken and predefined commands instead of complex phrases. CMU Sphinx is an open source toolkit for speech recognition from Carnegie Mellon University. Sphinx framework is language independent, so developers can use it to build a system that recognizes any language. However, Sphinx requires a model for the language that is going to be recognized. Dragon Naturally Speaking by Nuance Communications is a commercial application suite for speech recognition, supporting several languages. It is available as a desktop application for PC and Mac and as a mobile app for Android and iOS. Nuance also provides software development kits (SDKs) for enabling speech recognition in third-party applications. Apple's Siri is an example of a speech-recognition app powered by Nuance technology. In early 2013, Google added to Chrome browser the support for speech recognition though the Web Speech API[1]. This new API is a JavaScript library that lets developers integrate speech recognition to their Web applications. Although this technology can only be used in the Chrome browser, Google also support speech recognition on mobile devices through voice input.

2.2 Machine Translation

Machine translation (MT) is a subfield of Natural Language Processing in which software is used to automatically translate a text from one natural language, the source language, into another one, the target language [1]. MT systems can be broadly classified into two main categories, corpus-based and rule-based, according to the nature of the linguistic knowledge being used. The *rule-based* MT systems, such as Apertium, use knowledge in the form of rules explicitly coded by human experts, which attempt to codify specific linguistic knowledge (e.g., morphological and bilingual dictionaries, lexical and structural transfer rules) that automatic systems can process. This approach is however human intensive. Conversely, *corpus-based* MT systems, such as Google Translate, use large collections of parallel texts (i.e., pairs consisting of a text in a source language and its translation into a target language) as the source of knowledge from which the engine learns how to perform translations without direct human intervention. Although cheaper (i.e., no specific linguistic resource needs to be coded by humans), such type of systems requires huge amounts of training data, which may not be available for all languages and domains. Since both MT paradigms have different strengths and shortcomings, recently hybrid approaches have also emerged [4].

MT is difficult mainly because translation *per se* involves a huge amount of human knowledge that must be encoded in a machine-processable form. In addition, natural languages are highly ambiguous, as two languages seldom express the same content in the same way [1]. Accurate computer translation is particularly appealing because it is quicker, more convenient, and less expensive than human translators are.

3. THE SIMULATION

The overall goal of this simulation is to assess the usage of real-time speech translation to support communication in multilingual requirements meetings. This simulated study is a necessary preliminary step towards the design of future experiments that will involve real-time communication among individuals, augmented with speech translation.

Research from the past decade has shown evidence that the speech recognition technology available was unsuitable for providing real-time captioning or transcription of speech [14]. Although commercial speech recognition tools available today claim to achieve a word recognition accuracy as high as 99%, they have been developed for dictation rather than to produce a transcript from a continuous and unbroken stream without any punctuation [2][3][18]. Therefore, we refine our general research goal as:

RQ1 – *How well does speech translation work for continuous speech in global software projects?*

When stakeholders communicate during requirements meetings, many technical words are used. On top of that, technical words might even be in a language different from the one used by speakers. For instance, lawyers sometimes use Latin jargon; computer scientists typically use technical words in English. As such, technical jargon is less likely to occur both in real communication and in training sets used to build language models for speech recognition engines. Therefore, it has been previously observed that speech recognition errors are more likely to occur in words given a very low probability by the language model [16]. We thus define the second research question as:

RQ2 – *How does technical jargon affect speech translation in global software projects?*

We have investigated these two research questions by means of a simulation-based study described next.

3.1 Design and Execution

The simulation involved eight software engineering professionals as speakers, half from Brazil (selected by PUCRS's researchers) and half from Italy (selected by UniBari researchers). The speakers (7 males, 1 female) were divided into two groups of four according to the spoken language, either Italian or Brazilian Portuguese, which will be used as source language in the simulation.

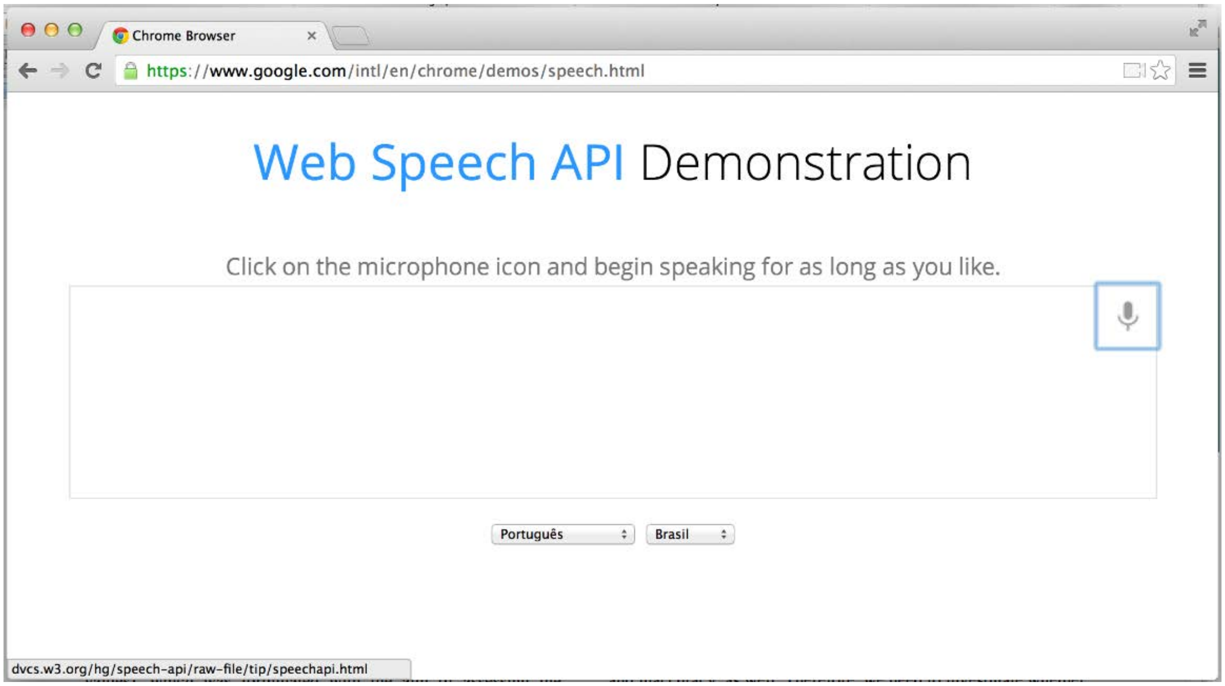
The unit of analysis is the utterance, which is a continuous piece of speech beginning and ending with a clear pause. We selected a test set of 60 utterances, ranging from 5 to 30 words, which speakers had to read aloud thus simulating the participation to a remote meeting. The original utterances used for simulation were in English and came from the logs of several real requirements meetings. In particular, the requirements meetings served to elicit and negotiate the requirements for six different systems, such as a system to keep track of supplies, equipment, and patients in hospitals, a bus tracking system, an educational game, and a system to book room resources around campus. After having selected the sample of utterances, they were manually translated by two of the researchers from English into both Italian and Brazilian Portuguese. The set of original utterances in English together with the manual translation in Italian and Portuguese formed the experimental sample. During the translation process, we were extremely careful at maintaining both the original meaning and the interaction style intact. Half of the selected utterances contained jargon, that is, one or more technical terms or characteristic acronyms used by software developers. The remaining half, we call them generic utterances, contained only words that are included in an Italian or Portuguese dictionary. Table 1 lists a few examples of jargon and generic utterances of different lengths, in English (i.e., before manual translation).

Each speaker read aloud 30 generic and 30 jargon utterances, for a total of 60 spoken utterances. In order to avoid biases, utterances were spoken alternating one generic utterance and one jargon utterance. As a speech recognizer, we have used the Google demo version of the Web Speech API available for the Chrome browser (Figure 3), whereas the translation service was provided by Google Translate (http://translate.google.com), which is the most popular corpus-based MT system.

[1] https://dvcs.w3.org/hg/speech-api/raw-file/tip/speechapi.html

Table 1. A few examples selected from the utterance sample (before translation).

Message	Lexicon	Length (word count)
The RFP document said you want to know where doctors and nurses are… [pause]… that would imply some sort of tracking.	Jargon	20
How about daily recurring bookings?	Generic	5
Is there any calendar formats your organization uses, like iCal?	Jargon	10
We would like a view to see at a glance when a bus will arrive at a particular stop. It should have all other stops on that route listed too.	Generic	30


Chrome Browser
https://www.google.com/intl/en/chrome/demos/speech.html
Português
Brasil
dvcs.w3.org/hg/speech-api/raw-file/tip/speechapi.html

Web Speech API Demonstration

Click on the microphone icon and begin speaking for as long as you like.

Figure 3. Google Web Speech API Demo (http://www.google.com/intl/en/chrome/demos/speech.html)

For each of the 60 utterances in the sample, a speaker started by clicking on the microphone icon and began speaking until the end of the utterance. Participants spoke in a colloquial style at their own pace. If the researcher realized that the spoken utterance was different from the original content or the speaker stopped before arriving to the end of the utterance, then the researcher invited the speaker to try again. On average, a speaker finished the simulation in about 30 minutes, at a pace of two utterances per minute.

Once the transcript appeared in the text field, it was copied and pasted by one of the researchers into a spreadsheet. Upon completing a speech recognition session, one of the researchers at each side performed machine translation on the respective transcripts recorded. Using Google Translate, each transcript was translated from Italian (IT) into both English (EN) and Portuguese (PT), and from Portuguese (PT) into both English (EN) and Italian (IT). Translations results were then copied in the spreadsheet.

Once all the subjects completed their tasks, one researcher at UniBari rated the quality of translations of all the translations from Italian to English (IT->EN) and from Portuguese to Italian (PT->EN); likewise, one researcher at PUCRS rated the translations from Brazilian Portuguese to English (PT->EN) and from Italian to Brazilian Portuguese (IT->PT).

3.2 Variables and Levels of Measurement for Speech Recognition

Variables differ according to the stages of speech translation: speech recognition and machine translation. As regards speech recognition, we have the following independent variables (factors):

1. Source **Language** (levels: *Italian* and *Brazilian Portuguese*)

The language used as a source for speech translation. The source language is the native language of the speakers, that is, the language spoken since birth.

2. **Speaker** (levels: *speaker1* - *speaker4*, under each language level)

There are four speakers at UniBari speaking in Italian and 4 speakers at PUCRS speaking in Brazilian Portuguese. The factor *speaker* is nested in factor *language* because each level of factor *speaker* occurs in conjunction with only one level of factor *language*.

3.**Lexicon** (levels: *generic* and *jargon*)

A generic utterance contains only words that are included in a source language dictionary. A jargon utterance contains one or more technical terms or characteristic acronyms used by software developers. Lexicon is a fixed effect factor.

4. **Replication** (levels: *replication1 - replication 30*, under each lexicon level)

Speech recognition tasks are repeated 30 times under each lexicon level, for a total of 60 utterances to be spoken per subject. The 30 replications under each lexicon level are considered a random factor nested under lexicon.

As dependent variable, we evaluate the performance of speech recognition in terms of **transcript accuracy**, computed as:

$$T_{acc} = \frac{\#\, recognized\ words - \#\, errors}{\#\, words\ in\ utterance}$$

Errors include missing and wrong words. Here we do not consider additional words since we observed no such case in our simulation where the system recognized accurately all the words in an utterance, plus additional erroneous words. As such, $T_{acc} \in [-1, 1]$. Therefore, in order to show values as percentages, we normalize it as $T'_{acc} = (T_{acc} + 1)/2$. We note that, as defined, the transcript accuracy handles the differences in length of the utterances in our experimental sample.

3.3 Variables and Levels of Measurement for Machine Translation

With respect to machine translation, instead, we have the following independent variables:

1.**Language Pairs** (levels: *IT->EN; IT->PT; PT->EN; PT->IT*)

A language pair represents a couple of a source and a target language in the automatic machine translation process. The utterance in input is a transcript returned as an output of the speech recognition process.

2.**Lexicon** (levels: *generic* and *jargon*)

The operational definition is the same of the previous section.

As dependent variable, we have **translation adequacy**. Two raters (one from each side) assessed the adequacy of translations assigning scores to output sentences. More specifically, the raters assessed whether each translation contained the information that existed in the original sentence. The scoring scheme adopted is a 4-point Likert scale as follows.

4 = Completely adequate. The translation clearly reflects the information contained in the original utterance. It is perfectly clear, intelligible, grammatically correct, and reads like ordinary text.

3 = Fairly adequate. The translation generally reflects the information contained in the original utterance, despite some inaccuracies in the text. It is generally clear and intelligible and one can (almost) immediately understand what it means.

2 =Somewhat adequate. The translation poorly reflects the information contained in the original utterance. It contains grammatical errors and/or poor word choices. The general idea of the text is intelligible only after considerable study.

1 = Completely inadequate. The translation is unintelligible and it is not possible to obtain the information contained in the original utterance. Studying the meaning of the text is hopeless and, even allowing for context, one feels that guessing would be too unreliable.

We have already used this scale in other studies before (e.g., see [7]) because it offers several advantages. First, it is not too fine grained, i.e., it does not consist of too many values. Second, it can be easily applied as the descriptions are well defined, i.e., raters can uniformly interpret them. Finally, there is no middle value, which helps to avoid central tendency bias in ratings by forcing raters to judge the output as either adequate or not.

Translation adequacy was measured post-facto by two researchers once that all the translations were available. While transcript accuracy is the dependent variable for the speech recognition stage, here it may affect the performance of MT and then its effect on translation adequacy must be monitored. Therefore, **transcript accuracy** is considered a control variable.

4. RESULTS

In this section, we report the results from the analyses on the accuracy of speech recognition and the adequacy of machine translation.

4.1 Speech Recognition Results

Table 2 reports the mean values of the transcript accuracy measured by language and lexicon. In both cases, we observe minimal differences. In fact, the mean accuracy for utterances spoke in Italian is 81%, whereas for Brazilian Portuguese it is 75%. Likewise, slightly better accuracy results were achieved on average for generic utterances (80%) as compared to jargon utterances (77%). Table 3, instead, reports the average accuracy per speaker. In addition, in this case we cannot observe large differences. The only noticeable result is the performance of Brazilian subject PT-Speaker2, who achieved the lowest accuracy of 68% while all the other speakers' accuracy levels are equal to or above 73%. The best accuracy was achieved by the Italian subject IT-Speaker2 who achieved 88% of accurate transcript.

Finally, we run a univariate analysis of variance (UNIANOVA procedure) in SPSS to test for differences in the accuracy of transcripts produced by the factors and their interactions.

The results, reported in Table 4, show that the speaker factor (nested within language) significantly affected the accuracy of the speech recognition service (F=12.91, p=.003). Besides, also replication (nested within lexicon) was found to significantly influence the quality of the recognition process (F=1.74. p=.018). Instead, neither the source language nor the lexicon were found to affect transcript accuracy.

Table 2. Transcript accuracy means by language and lexicon

		Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Language	IT	.81	.008	.793	.825
	PT	.75	.008	.731	.763
Lexicon	Generic	.80	.008	.779	.810
	Jargon	.77	.008	.746	.778

Table 3. Transcript accuracy means by speaker

Speaker	Language	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
PT-Speaker1	PT	.78	.016	.753	.817
PT-Speaker2	PT	.68	.016	.650	.714
PT-Speaker3	PT	.79	.016	.762	.826
PT-Speaker4	PT	.73	.016	.697	.760
IT-Speaker1	IT	.76	.016	.724	.788
IT-Speaker2	IT	.88	.016	.845	.909
IT-Speaker3	IT	.78	.016	.750	.814
IT-Speaker4	IT	.82	.016	.791	.854

Table 4. Differences in the accuracy of transcripts produced by the factors and their interactions

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	Hypothesis	290.785	1	290.785	587.408	.017	.998
	Error	.573	1.157	.495a			
Language	Hypothesis	.460	1	.460	2.948	.144	.359
	Error	.822	5.266	.156			
Speaker(Language)	Hypothesis	.996	6	.166	12.907	**.003***	.928
	Error	.077	6	.013			
Lexicon	Hypothesis	.125	1	.125	3.285	.104	.272
	Error	.334	8.794	.038			
Replication(Lexicon)	Hypothesis	4.770	58	.082	1.740	**.018***	.635
	Error	2.741	58	.047			
Language * Lexicon	Hypothesis	.003	1	.003	.068	.797	.002
	Error	1.310	29.507	.044f			
Language * Replication(Lexicon)	Hypothesis	2.741	58	.047	3.004	.000	.334
	Error	5.473	348	.016			
Lexicon * Speaker(Language)	Hypothesis	.077	6	.013	.817	.557	.014
	Error	5.473	348	.016			

**** results significant t the 5% level***

4.2 Speech Translation Results

In order to identify differences in the quality of translation produced according to the various combinations of language pairs and lexicon, we first evaluated how many sentences were rated adequate (i.e., categories 4 and 3) and inadequate (i.e., categories 1 and 2). Figure 4 shows a breakdown of the translation results by language pairs. We can observe a similar behavior for all the combinations, with a minimum of 75 (PT->IT, 31%) and a maximum of 99 (IT->EN, 41%) adequately translated utterances out of 240. Instead, Figure 5 shows a breakdown of the translation results also by lexicon. It shows that, for all the four language pairs, the inadequate translations outnumber the others categories regardless of the lexicon. In other words, generic utterances were translated no more adequately than jargon utterances.

Afterwards, we computed Spearman's rank correlation coefficient to measure the interdependence between transcript accuracy and resulting translation adequacy. Results are reported in Table 5 and Table 6, grouped by language source (Brazilian Portuguese and Italian, respectively). The results in the two tables show that, regardless of both the lexicon and the language pairs, there is a moderate positive correlation between transcription accuracy and translation adequacy. In other words, when the speech recognition component produced an inaccurate transcription, the machine translation tended to produce a less adequate translation.

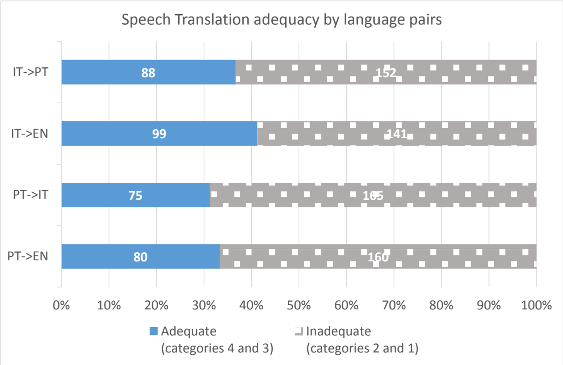


Figure 4. Speech translation adequacy by language pairs

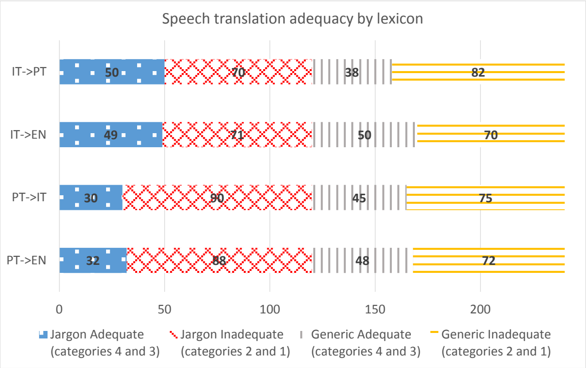


Figure 5. A breakdown of speech translation adequacy by lexicon

Table 5. The correlation between transcript accuracy and translation adequacy with Brazilian Portuguese as the source language

Spearman's rho		PT->IT				PT->EN			
		Jargon		Generic		Jargon		Generic	
		Translation adequacy	Transcription accuracy	Translation adequacy	Transcription accuracy	Translation adequacy	Transcription accuracy	Translation adequacy	Transcription accuracy
Translation adequacy	**Correlation**	1.0	**.55***	1.0	**.54***	1.0	**.63***	1.0	**.59***
	Sig.	.	.000	.	.000	.	.00	.	.00
	N	120	120	120	120	120	120	240	240
Transcription accuracy	**Correlation**	**.55***	1.000	**.54***	1.0	**.63***	1.0	**.59***	1.0
	Sig.	.00	.	.00	.	.00	.	.00	.
	N	120	120	120	120	120	120	120	120

***** Correlation is significant at the 0.01 level (2-tailed).*

Table 6. The correlation between transcript accuracy and translation adequacy with Italian as the source language

Spearman's rho		IT->PT				IT->EN			
		Jargon		Generic		Jargon		Generic	
		Translation adequacy	Transcription accuracy	Translation adequacy	Transcription accuracy	Translation adequacy	Transcription accuracy	Translation adequacy	Transcription accuracy
Translation adequacy	**Correlation**	1.0	**.71***	1.0	**.55***	1.0	**.72***	1.0	**.61***
	Sig.	.	.00	.	.00	.	.00	.	.00
	N	120	120	120	120	120	120	120	120
Transcription accuracy	**Correlation**	**.71***	1.0	**.55***	1.0	**.72***	1.0	**.61***	1.0
	Sig.	.00	.	.00	.	.00	.	.00	.
	N	120	120	120	120	120	120	120	120

***** Correlation is significant at the 0.01 level (2-tailed).*

5. DISCUSSION

In this section, we discuss the results reported in the previous section. Our analysis focused on (1) evaluating the effectiveness of the speech translation process, including speech recognition without interruptions and how it affects machine translation, and (2) the adequacy and accuracy of speech translation outcome (both transcription and translation), taking into account the use of technical jargon.

5.1 RQ1: How well does speech translation work for continuous speech?

As regards RQ1, the first step in our simulation was the speech recognition process from a source language, either Italian or Portuguese. Considering that our study is a simulation of conversation transcripts, we found our setup similar to the setting of speech recognition applied to the automatic generation of transcripts from webcast lectures, where acceptable error rates are equal or less than 25%, that is, 75% of word accuracy [12]. Our results are in line with such baseline.

When transcriptions were translated to the target language, we found only a moderate correlation between accuracy and adequacy of translation results. This is because while there may be cases where inadequate translation occurred with accurate transcriptions, the opposite, adequate translation from inaccurate transcriptions, never happened.

5.2 RQ2: How does technical jargon affect speech translation?

As regards RQ2, our simulation took into account the difference between technical (jargon) and generic words, in order to evaluate if using a jargon would affect the outcome of the speech translation process. In our results, we did not find evidence that the use of jargon generates a worse transcription and consequently an even worse translation if compared to the use of generic words. However, we found that albeit all software development professionals know how to read jargon, different professionals speak jargon differently. In other words, we could observe significant differences among speakers, and among spoken utterances. For instance, in the case of SQL term, some speakers read it spelling letters in their native language, some spelled it in English, and some others read it as SEQUEL. The speech recognition handled it differently and was unable to capture SEQUEL correctly.

6. THREATS TO VALIDITY

One of the key issues in experimentation is evaluating the validity of results [17]. In this section, we discuss the potential threats that are relevant for our study and how they are addressed.

The utterances in the experimental sample were opportunistically selected from a larger set of over 2000 utterances in English. Specifically, we selected two sets of 30 utterances with and without technical jargon among those that had clear meaning and errors. Two researchers manually translated the selected sample into Italian and Brazilian Portuguese. We acknowledge that this selection and manual translation make our setting more artificial in that, when talking, people do make mistakes (e.g., repeat words) which sometimes have to be repaired (e.g., mispronounced words).

We also identified a couple of threats related to construct validity. First, with respect to the construct of *transcription accuracy*, we computed the metrics as the ratio between the # of correct words minus the # of errors divided by the total # of words. As per the errors, in order to refine the assessment of speech recognition technology, we acknowledge the need to distinguish the # of words that are wrongly recognized from those that are completely missing. Second, with respect to the construct of *translation adequacy*, two of the researchers (one for each side) acted as raters. In particular, the Italian rater evaluated the utterances for the two language pairs IT->EN and PT->IT, whereas the Brazilian researcher rated the utterances from the two language pairs PT->EN and IT->PT. Because the set of sentences they rated are disjoint, no Cohen's k-index could be computed to assess their inter-rater agreement level. As such, the adequacy scale may not have been applied homogenously. Nevertheless, we have employed the same scale for several previous experiments and, as such, the two researchers have a considerable amount of training in using it.

Finally, we identified a few threats related to external validity. First, in our simulation we only used one speech translation system (Google Web Speech demo combined with Google Translate). Therefore, findings might not extend to other existing speech translation technologies available. As such, we acknowledge the need to compare the performance of more systems in our future work. Second, the speakers were mostly male (7 out of 8) and from Brazil and Italy only. As such, we cannot exclude that differences in accent and pronunciations due to the gender and nationality had any effect on the speech translation accuracy. Finally, the study reported here is a simulation during which several professional developers read several utterances unrelated to each other into a speech translation system. Although collected from several real requirements meetings, such set does not fully represent an example of real requirements workshop meeting augmented with speech translation. In fact, our simulation does not take into account factors like task completion, communication flow, context and grounding. Therefore, we acknowledge the need to perform future controlled experiments that involve cross-language group communication augmented with speech translation. Finally, our findings showed that accuracy in terms of speech recognition was significantly affected by speaker and utterance differences. As such, we acknowledge that the limited numbers of speakers (4 for each of the two source languages) and utterances (30 for each of the two kinds of lexicon) are not ideal from the statistical point of view. Such limitations will be addressed in future replications.

7. CONCLUSIONS

In this paper, we report from an empirical simulation-based study where we investigated the adoption of combining speech recognition and machine translation in order to overcome language barriers among stakeholders who are remotely negotiating software requirements. We have used Google Web Speech API and Google Translate and involved two groups of four subjects from Italy and Brazil, speaking their native language. The simulation was executed with a test set of 60 technical and non-technical utterances.

Our empirical results confirmed the possibility to use speech translation technologies to help globally distributed team members to communicate in their native languages. We found the accuracy of speech recognition to be affected by speaker and utterance differences. Yet, the accuracy level measured was acceptable (≈75%) in that it is in line with previous findings [12]. We also found that speech recognition is the most critical part aimed at speech translation technology, as adequate translations tend to follow accurate transcripts.

As future work, we intend to execute a controlled experiment instead of a simulation, in order to compare subjects who communicate through a speech translation system, using their native language and subjects that communicate by English, using, for example, voice-based chat services.

ACKNOWLEDGMENTS

This work fulfils the research objectives of the PON 02_00563_3470993 project "VINCENTE - A Virtual collective INtelligenCeENvironment to develop sustainable Technology Entrepreneurship ecosystems" funded by the Italian Ministry of University and Research (MIUR). This research is also partially funded by FAPERGS, projects 11/2022-3 and 002061-2551/13, and CNPq (project 309000/2012-2).

REFERENCES

[1] D. Arnold and L. Balkan and R.L. Humphreys and S. Meijer and L. Sadler, *Machine Translation: an Introductory Guide*, NCC Blackwell, 1994.

[2] K. Bain, S. Basson, and M. Wald, "Speech recognition in university classrooms: liberated learning project," *Proc. The Fifth International ACM SIGCAPH Conference on Assistive Technologies (ASSETS)*, pp. 192-196, 2002.

[3] K. Bain, S. Basson, A. Faisman, D. Kanevsky, "Accessibility, transcription, and access everywhere," *IBM Systems Journal*, vol. 44, no. 3, pp. 589-604, 2005.

[4] A. Burchardt, C. Tscherwinka, A. Eleftherios, and H. Uszkoreit, "Machine Translation at Work." in *Computational Linguistics*, Studies in Computational Intelligence Vol. 458, pp 241-261, Springer, 2013.

[5] F. Calefato, F. Lanubile, and P. Minervini, "Can Real-Time Machine Translation Overcome Language Barriers in Distributed Requirements Engineering?", *Proc. 5th Int'l Conference on Global Software Engineering (ICGSE'10)*, Princeton, NJ, USA, Aug. 23-26, pp. 257-264, 2010.

[6] F. Calefato, F. Lanubile, and R. Prikladnicki, "A Controlled Experiment on the Effects of Machine Translation in Multilingual Requirements Meetings", *Proc. 6th Int'l Conference on Global Software Engineering (ICGSE'11)*, Helsinki, Finland, August 15-18, 2011.

[7] F. Calefato, F. Lanubile, T. Conte and R. Prikladnicki, "Assessing the Impact of Real-Time Machine Translation on Requirements Meetings: A Replicated Experiment", *6th Int'l Symposium on Empirical Software Engineering and Measurement (ESEM'12)*, Lund, Sweden, Sep. 19–20, 2012.

[8] D. Damian and D. Zowghi, "Requirements Engineering Challenges in Multi-Site Software Development Organizations", *Requirements Engineering Journal*, 8-3, 2003, pp. 149-160.

[9] D. Damian, "Stakeholders in Global Requirements Engineering: Lessons Learned from Practice", *IEEE Software*, 24-2, 2007, 21-27.

[10] P. Hyman, "Speech-to-speech translations stutter, but researchers see mellifluous future", *Commun. ACM* 57, 4 (April 2014), 16-19.

[11] X. Huang, J. Baker, and R Reddy, "A historical perspective of speech recognition," *Commun. ACM*, vol. 57, no. 1, January 2014, pp. 94-103. DOI=10.1145/2500887.

[12] C. Munteanu, R. Baecker, and G. Penn, "Collaborative editing for improved usefulness and usability of transcript-enhanced webcasts." *In Proc. Conf. on Human Factors in Computing Systems* (CHI '08), Florence, Italy, Apr. 5-10, 2008, pp. 373-382, DOI=10.1145/1357054.1357117.

[13] B. Nuseibeh, and S. Easterbrook, “Requirements engineering: a roadmap,” *Proc. Int’l Conf. on the Future of Software Engineering* (ICSE ’00), pp. 35-46, June 2000.

[14] R. Ranchal, T. Taber-Doughty, Y. Guo, K. Bain, H. Martin, J.P. Robinson, B.S. Duerstock, "Using Speech Recognition for Real-Time Captioning and Lecture Transcription in the Classroom," *IEEE Transactions on Learning Technologies*, vol. 99, 2013.

[15] D.R. Reddy, “Speech Recognition by Machine: A Review,” *Proceedings of the IEEE*, Vol. 64, No. 4, 1976, pp. 501-531.

[16] Rosenfeld, R. "Two decades of statistical language modeling: where do we go from here?", *Proc. of the IEEE,* pp: 1270 - 1278 Volume: 88, Issue: 8, Aug. 2000.

[17] C. Wohlin, P. Runesson, M. Höst, M.C. Ohlsson, B. Regnell, and A. Wesslén. *Experimentation in Software Engineering, An Introduction*. Kluwer Academic Publishers, 2000.

[18] M. Wald, K. Bain, “Universal access to communication and learning: the role of automatic speech recognition,” *Universal Access in the Information Society*, vol. 6, no. 4, pp. 435-447, 2008.