Mobile Speech Translation for Multilingual Requirements Meetings: A Preliminary Study

Fabio Calefato, Filippo Lanubile, Damiano Romita
Dipartimento di Informatica
Università di Bari
Bari, Italy
fabio.calefato@uniba.it, filippo.lanubile@uniba.it,
d.romita@gmail.com

Rafael Prikladnicki, João Henrique Stocker Pinto,
Pontifícia Universidade Católica do Rio Grande do Sul
PUCRS
Porto Alegre, Brazil
rafaelp@pucrs.br, joaohenrique.jhsp@gmail.com

***Abstract*—Communication in global software projects usually occurs between native and non-native English speakers with the drawback of an unequal ability to fully understand and contribute to discussions. In this paper, we investigate the adoption of combining speech recognition and machine translation in order to overcome language barriers among stakeholders who are remotely negotiating software requirements. We report our findings from a simulated study where stakeholders communicate speaking three different languages with the help of the Google mobile speech translation service.**

Keywords—speech translation; language distance; distributed development; requirements engineering; simulation

I. INTRODUCTION

Communication in global software projects usually occurs between native and non-native English speakers with the drawback of an unequal ability to fully understand and contribute to discussions. Currently, machine translation technology is typically available in the form of cross-language Web services, which can be embedded into multiuser and multilingual chats, but they are mostly text-based. In our previous work, we have studied how machine translation affects text-based chats for complex communication tasks [3][4][5][17]. We found that, despite far from 100% accurate, real-time machine translation is not disruptive of the conversation flow, is accepted with favor, and grants a more balanced discussion.

In this paper, we investigate the adoption of combining speech recognition and machine translation in order to overcome language barriers among stakeholders who are remotely negotiating software requirements. We expect a bigger impact when speech is involved rather than text-based conversations because, when hearing, participants have less time to deliberate on the supposed meaning of foreign words and sentences. We are aware that real-time speech recognition also comes with costs that derive from both inaccurate transcriptions and inadequate translations. Hence, research needs to assess in measurable terms whether the costs of recovering from incorrect speech translations outweigh the benefits of communicating in native language. Besides, the recent technological progress in the field of automatic speech recognition has also found its way in mobile devices, something that definitely calls for further investigation, especially in combination with machine translation [8]. In fact, the exponential growth in both the development and use of mobile applications has in turn increased the popularity of bring your own device (BYOD) policies in the workplace, thus opening new possibilities of participating in projects meetings also when on the move.

Considering the rather exploratory nature of this study, we run a simulation in which we used *Google Translate for mobile* as real-time speech translation service to translate a test set of chat log entries collected from requirements engineering workshops. We specifically selected requirements engineering as the appropriate domain for this simulation because it is the most communication-intensive activity of software development and thus the one that is alleged to suffer more from language difficulties.

The remainder of this paper is structured as follows. In Section II we briefly overview previous research in the fields of machine translation and speech recognition research field, as well as a review of some of the speech translation solutions currently available. Section III describes our simulation procedure run in order to explore the performance of in real-time speech translator in mobile settings. The findings from our simulation are presented and discussed, respectively, in Section IV and V. Threats to validity are presented in Section VI. Finally, we conclude in Section VII.

II. BACKGROUND ON SPEECH TRANSLATION

In this section, we review the background on the two building blocks of speech translation systems, that is, speech recognition and machine translation.

A. Speech Recognition

Speech recognition 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 [19]. Previous studies have mostly investigated automatic speech recognition as a way to augment electronic meeting systems with features like note taking facilitations [15], annotated recordings [7], and summarization [20]. Overall, according to Ranchal et al. [18], the research from the past decade has shown evidence that the speech

recognition technology available was appropriate for dictation with punctuation, but it was unsuitable for providing real-time captioning or transcription of speech. In fact, word error rate in speech recognition systems typically increase when switch from read speech to conversational speech. Over the last 40 years, however, there has been a substantial advance in the field to a point that we had not imagined back than [19]. Therefore, in spite of all the accomplishment in speech recognition, additional challenges remain, as complex as those overcome so far [12].

In the following, we review some of the available technologies for speech recognition. Microsoft Speech SDK is part of the .NET Framework package and incorporates the native API for Windows, Microsoft Speech API (SAPI)[1]. However, it is typically used by developers to let applications recognize spoken, predefined commands instead of complex phrases. CMU Sphinx[2] 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 it needs to recognize. The Sphinx group has made available models for English, Chinese, French, Spanish, German, and Russian languages. Dragon Naturally Speaking[3] by Nuance Communications is a commercial application suite for speech recognition, supporting several languages other than English, including French, German, Italian, and Dutch. 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[4] is an example of a speech-recognition app powered by Nuance technology. In early 2013, Google added to Chrome browser[5] the support for speech recognition though the Web Speech API [22]. 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.

B. Machine Translation

Machine Translation is defined as the use of a computer to translate a text from one natural language to another one [1]. Machine translation is difficult mainly because natural language is highly ambiguous and thus, its translation involves a huge amount of human knowledge to be encoded in a machine-processable form. Yet, machine translation is particularly appealing because it is quicker, more convenient, and less expensive than human translators are.

An interesting research conducted by Yamashita et al. [21] has investigated the effects of machine translation on mutual understanding. The study found that shared understanding is affected by the asymmetry of machine translation since the sender of a message does not know how well it has been translated into the target language. Wang et al. [23] suggest to take advantage of this asymmetry and use machine translation technology only to translate messages from a native language to English, while leaving native English speakers' messages untranslated, thus avoiding comprehension problems introduced by machine translation.

A common limitation of the studies in the field of machine translation is the employment of experimental tasks like picture description or idea exchange, often in one to one chat, a setting which is likely to miss out some the facets and subtleties of realistic, communication-intensive tasks [9], such as those performed by software teams.

In the following, we review some of the technologies currently available for machine translation. 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[6], 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 costly. Conversely, *corpus-based* MT systems, such as Google Translate[7], 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, such type of systems require 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 [2].

III. THE SIMULATION

The overall goal of this preliminary study is to evaluate the feasibility of adopting a real-time speech translation service for supporting cross-language communication in multilingual requirements meetings.

In our previous works on machine translation [3][4][5], we found that state-of-the-art machine translation technology is still far from perfect. Yet, so is speech recognition technology [12]. Therefore, we want to assess how inaccuracies in the recognition of speech affect the resulting adequacy of the translated outcome.

RQ1 – *How do errors in the speech recognition process affect the resulting machine translation outcome?*

In a multilingual group, some participants may have limited communication and comprehensions skills in English. To contribute a message to the group, these participants would definitely benefit from using a speech translation system and their native language as the source language. Yet, does this hold true also for those group members who are fluent in English as well? In fact, group participants with better English communication skills might contribute a message in either

[1] http://msdn.microsoft.com/en-us/library/ee125663.aspx
[2] http://cmusphinx.sourceforge.net
[3] http://www.nuance.com/dragon
[4] http://www.apple.com/ios/siri
[5] https://www.google.com/intl/en/chrome/demos/speech.html
[6] http://www.apertium.org
[7] http://translate.google.com

English or their native language. However, one of the findings from existing research is that using English as a *lingua franca* to overcome the linguistic barrier does have the drawback of granting English native speakers better abilities to steer communication. In fact, being not as fluent, non-native English speakers suffer the inability to contribute to a discussion to the same extent. This finding would suggest that the use of the native language is a better option indeed. However, in our previous works on machine translation we found these systems to achieve better results when translating both *into* and *from* English. In other words, to date we do not know what source and target languages work best with current speech translation technology. Therefore, we define the second research question as follows:

RQ2 – *Which source language works better for non-native English speakers with state-of-the-art speech translation technology?*

Finally, as stated before, machine translation systems achieve better results when translating both into and from English. Instead, as per speech translation technology, to date we do not know what source and target languages work best. Therefore, we define the fourth and final research question as:

RQ3 – *How well does state-of-the-art speech translation technology perform when English is not used as either the source or the target language?*

We have investigated these research questions by means of a simulation described in the following. To run the simulation, we took into account three different languages, that is, Italian (IT), English (EN), and Brazilian Portuguese (PT) used as both source and target language with speech translation technology, thus generating six language pairs combinations (see Table I). We used the following notation to distinguish the various combinations. For instance, $EN_{(PT)} \rightarrow IT$ means that Italian is the target language, whereas English is the source language of the test sentence that was read in by a Brazilian Portuguese native speaker. Likewise, $IT_{(IT)} \rightarrow EN$ means that a test sentence in Italian is read in by an Italian native speaker and translated into English.

TABLE I. THE LANGUAGE PAIRS COMIBNATIONS AND HOW THEY MAP TO THE RESEARCH QUESTIONS DEFINED.

Language pairs combinations	**Help to answer the research question…**
1. $IT_{(IT)} \rightarrow EN$ 2. $PT_{(PT)} \rightarrow EN$	RQ1 How errors in speech recognition affect resulting translations
3. $EN_{(IT)} \rightarrow PT$ 4. $EN_{(PT)} \rightarrow IT$ 5. $IT_{(IT)} \rightarrow PT$ 6. $PT_{(PT)} \rightarrow IT$	RQ2 Which source language works better for non-native English speakers with state-of-the-art speech-translation technology
5. $IT_{(IT)} \rightarrow PT$ 6. $PT_{(PT)} \rightarrow IT$	RQ3 How languages other than English are supported by state-of-the-art speech-translation technology

A. *Instrumentation*

In order to run the simulation, we used the Google Translate mobile app (ver. 3.0.4), with voice input enabled in order to pipe the speech recognition output into the machine translation (see Fig. 1). The tests were executed on the following devices: a Galaxy Note 3, running Android 4.3, at the Italian site; a Galaxy Note GT N7000, running Android 4.1.2, at the Brazilian site. Besides, all the tests were executed on the two devices using a Wi-Fi connection.

As the test set, we selected 51 sentences of growing length (word count, min. 4, max. 94). The sentences were selected from real chat logs in English, collected from five requirements workshops run as part of an experiment on the effects of text-based communication in distributed requirements engineering [6]. Participants in each workshop ranged from five to eight undergraduate students attending a requirements engineering course at the University of Victoria, Canada. During a workshop, the participants, either acting as a client or as a developer, had first to elicit the requirements specification of a web application (first session); then, they had to negotiate and reach closure on the previously collected requirements (second session). Table II contains a few examples of sentences of growing length, opportunistically selected from the chat logs.

B. *Evaluation of speech translation quality*

The evaluation process in our simulation is challenging because errors in the speech recognition process negatively affect the outcome of machine translation, a service that we have already proved to be far from perfect on its own [3]. Therefore, speech recognition quality was taken in account separately and evaluated before machine translation. Besides,

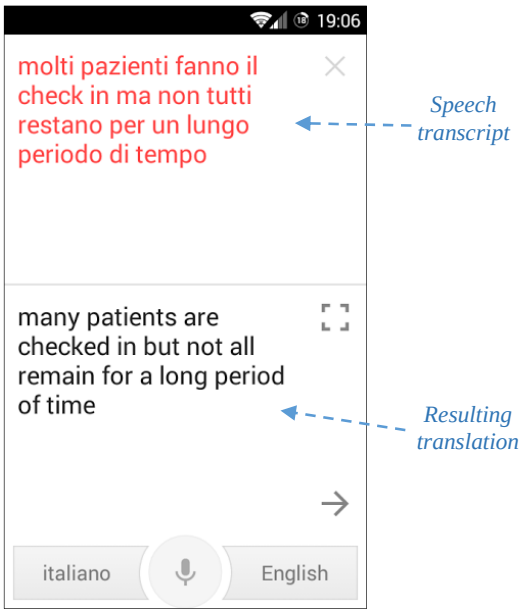

Fig. 1. A screenshot of the app accurately recognizing and adequately translating from Italian to English with voice input.

TABLE II. SOME EXAMPLES OF SENTENCES SELECTED FROM THE LOGS.

Message	Length (word count)
Many patients check in but not all stay for a long period of time.	14
How will web users be authorized to access patient information on the hospital website? Or will there be no information they can access.	23
Welcome to the first elicitation meeting. We are going to get started here shortly, but in the meantime feel free to state your name and role. Is there anyone missing that requires an invite?	34
I was still trying to focus on these terminals, which are located inside of the hospital. That said, it would be my understanding that there is no need for a map of the location of the hospital on the terminals, but perhaps floor plans would be a better idea.	49
Welcome to the first elicitation meeting. The goal of this first meeting between St. Peter's Hospital and Tri-Systems is to bring both parties to collaborate and aid Tri-Systems understand the requirements of system outlined in the RFP. Jane and I are software developers for Tri-Systems. I am John and I will be moderating this meeting. Odysseus will be recording all decisions in the Decision place at the right side of your screen. If at any time you do not agree with one of the decisions please send a message and we will discuss further.	94

while the effectiveness of a machine translation service relates to the fluency and fidelity of the translated output, the effectiveness of a speech recognition system relates to the correct number of words recognized in a spoken sentence.

Therefore, although quite similar, the scoring schemes proposed for speech recognition accuracy and translation adequacy are not the same. They both consist of a 4-item Likert scale (see Table III), anchored respectively with values *4=completely accurate/adequate* and *1=completely inaccurate/inadequate*. Such scales are adapted from the intelligibility scale proposed in [1]. We judged it appropriate to our goal because: (a) it is not too fine grained (i.e., does not consist of too many values); (b) it can be easily applied as descriptions are well defined (i.e., it can be uniformly interpreted by raters); (c) and there is no middle value (i.e., it helps to avoid central tendency bias in ratings by forcing raters to judge the output as either adequate or not) [10][14].

With respect to the evaluation of speech recognition accuracy, the standard metric adopted is the word error rate [12], which is defined as:

$$W_{err} = \frac{\#\ recognized\ words - \#\ missing\ words}{\#\ words\ in\ sentence}$$

Then, accuracy becomes:

$$W_{acc} = 1 - W_{err}$$

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 [16]. As per our own accuracy scale, the recommended acceptable rate suggests to aggregate the scale values into two broader categories mapped to word accuracy intervals as follows:

- Accurate transcript (categories 3, 4) ⇔ [0, 25]% W_{acc}
- Inaccurate transcript (categories 1, 2) ⇔ [26, 100]% W_{acc}

In order to assess the quality of speech translation we used a two-step approach. During the simulation, we entailed as human raters two MSc students selected at each site. The two raters completed the evaluation of the test set independently, in a couple of days. In the first step of the evaluation, they read aloud the sentences from the test set in a mobile device, which showed the transcript along with the resulting translation in the target language. After that, both the transcripts and the resulting translations were reported in a spreadsheet. The captured transcripts were then presented to the raters, along with the original body of sentences. They assessed the accuracy of the recognition process by assigning a score to each transcript, judging whether they contained the same words as the original ones. In the second step, instead, the two raters evaluated the adequacy of the translations by assigning a second score that valued the extent to which the translations reflected the meaning of the original input.

Finally, we note that the Italian rater read the test sentences in Italian and English, whereas the Brazilian rater read them in Brazilian Portuguese and English. Because the entire test set was in English, two of the researchers manually translated the sentences to Italian and Brazilian Portuguese (i.e., their own native language). We also note that in the cases of $IT_{(IT)}$→PT and $PT_{(PT)}$→IT language pairs, the rater who read the sentence in is not the rater that assessed the translation quality. In other words, for the pair $IT_{(IT)}$→PT, the Italian rater read in the sentences that were translated into Brazilian Portuguese; then

TABLE III. ACCURACY/ADEQUACY SCALES FOR SPEECH RECOGNITION/MACHINE TRANSLATION QUALITY ASSESSMENT

Value	Description
4	***Completely accurate|adequate*** The *transcript|translation* clearly reflects the information contained in the original sentence. It is perfectly clear, intelligible, grammatically correct, and reads like ordinary text.
3	***Fairly accurate|adequate*** The *transcript|translation* generally reflects the information contained in the original sentence, despite some inaccuracies or infelicities in the text. It is generally clear and intelligible and one can (almost) immediately understand what it means.
2	***Somewhat accurate|adequate*** The *transcript|translation* poorly reflects the information contained in the original sentence. It contains grammatical errors and/or poor word choices. The general idea of the text is intelligible only after considerable study.
1	***Completely accurate|adequate*** The *transcript|translation* is unintelligible and it is not possible to obtain the information contained in the original sentence. Studying the meaning of the text is hopeless and, even allowing for context, one feels that guessing would be too unreliable.

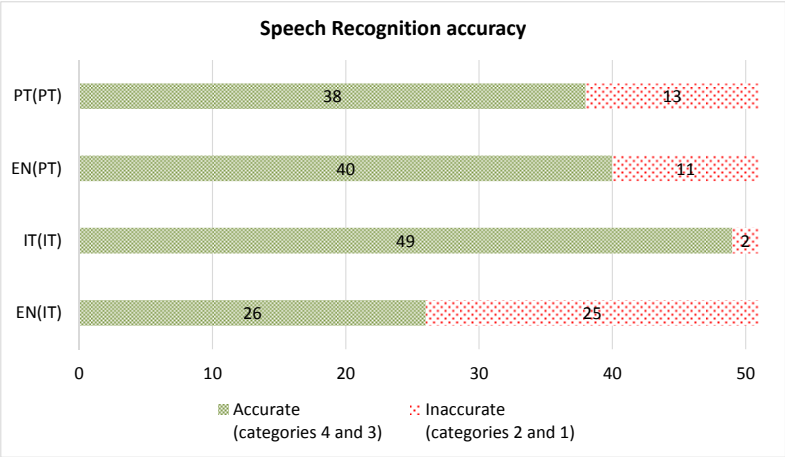

Fig. 2. Results from the evaluation of the speech recognition accuracy.

the resulting translations were captured and sent to the remote site, where the Brazilian rater evaluated their adequacy with respect to Brazilian Portuguese. The opposite happened for the language pair PT→IT. Instead, all the pairs including translations from and into English were rated locally by the same rater who read in the sentences.

IV. Results

In this section, we report our findings from the analysis of the effectiveness of the speech recognition and machine translation processes.

A. Speech Recognition Results

In order to assess the quality of speech recognition, we first evaluated how many sentences were evaluated as accurate (i.e., belonging to category 4 or 3) and inaccurate (i.e., belonging to category 2 or 1). Fig. 2 shows the results of the recognition process for the three languages. We remind the reader that, according to our notation, for example $EN_{(PT)}$ means that a test sentence in English is read by a Brazilian Portuguese native speaker.

As per $IT_{(IT)}$ case, the large majority of the sentences (49 out of 51, 96%) has been rated adequate. We also found that the mean and median values in this case are, respectively, 3.6 and 4. With respect to $PT_{(PT)}$, 38/51 (74%) sentences have been rated adequate. The mean and median values are 3.2 and 3, respectively. As far as using English as the source language is concerned, we observed a remarkable disparity between the $EN_{(IT)}$ and $EN_{(PT)}$ cases, that is, when English sentences were read, respectively, by the Italian rater and the Brazilian rater. In fact, for the $EN_{(IT)}$ case, only a half of the test set (26 sentences out of 51, 51%) has been rated adequate. The mean and median are, respectively, 2.59 and 3 in this case. As per the $EN_{(PT)}$ case, instead, 40/51 (78%) sentences have been judged adequate, with a mean value of 2.84 and a median value of 3.

Given that the $EN_{(IT)}$ received considerably worse scores than the $EN_{(PT)}$ counterpart, we thought that these results might be caused by the Italian rater's characteristics, such as accent, inappropriate pronunciation, or even gender. Consequently, we opportunistically performed a second accuracy evaluation of the $EN_{(IT)}$ combination, using as rater a woman with excellent English communication skills (she worked for three years in Germany as member of an international team of physicists and then moved to UK two years ago). The results, however, were consistent with the previous ones: 22/51 (43.1%) sentences judged adequate, mean 2.39, and median 3. Therefore, we excluded that the low quality in the recognition of Italian speech was related to the rater.

B. Speech Translation Results

Analogously to the speech recognition quality assessment, to evaluate the quality of the resulting translations we first calculated how many sentences were evaluated as adequate (i.e., belonging to category 4 or 3) and inadequate (i.e., belonging to category 2 or 1). Fig. 3 shows the results.

The best results have been achieved when Italian is used as the source languages, as in the cases $IT_{(IT)}$→EN, for which 48/51 (94%) translations were judged adequate, and $IT_{(IT)}$→PT, for which the speech translation results were as high as 40/51 adequately translated sentences (78%). The performances have been slightly worse using the pair $PT_{(PT)}$→IT, with 36/51 (71%) adequate translations. Average results have been achieved employing the pair $PT_{(PT)}$→EN, for which the results were 22/51 (43%) adequate speech translations, respectively. Instead, the worst results have been achieved with pairs $EN_{(IT)}$→PT (19/51, 37%) and $EN_{(PT)}$→IT (10/51, 20%).

Finally, in order to assess how speech recognition accuracy affects speech translation adequacy, we computed contingency tables for two language pairs, namely $IT_{(IT)}$→EN and $PT_{(PT)}$→EN (see Table IV). The χ^2 tests performed were both significant at the 0.01 level. As largely expected, χ^2 tests confirmed that inaccurate transcriptions always result in inadequate translations. More interestingly, instead, the χ^2 tests also showed that accurate transcriptions do result in adequate translations. Only in the case of the $PT_{(PT)}$→EN language pair,

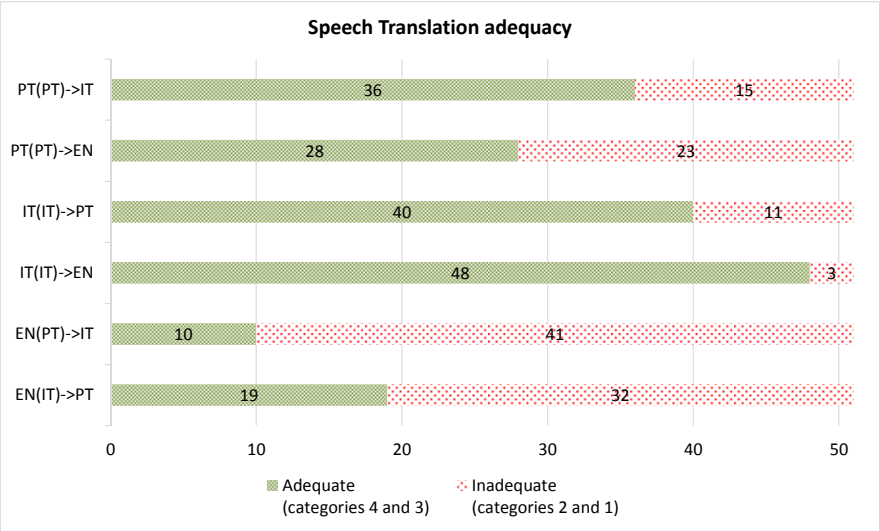

Fig. 3. Results from the evaluation of the machine translation adequacy

TABLE IV. CONTINGENCY TABLES TO ASSESS THE EFFECT OF TRANSCRIPTION INACCURACIES ON TRANSLATION ADEQUACY.

		$IT_{(IT)}$→EN			**$PT_{(PT)}$→EN**		
		Translations		***Total***	***Translations***		***Total***
		Inadequate	***Adequate***		***Inadequate***	***Adequate***	
Transcriptions	***Inaccurate***	2	0	2	13	0	13
	Accurate	1	48	49	10	28	38
Total		3	48	51	23	28	51
		χ^2**(1, N=51)=*31.31, p=.000***			χ^2**(1, N=51)=*21.24, p=.000***		

we observe that 10 accurate transcriptions resulted in translations judged inadequate by the two raters.

V. DISCUSSION

In this section, we discuss the results reported in the previous section. Our analysis focused on (1) evaluating the effectiveness of the speech recognition process, in terms of the accuracy of the transcriptions captured, and (2) the adequacy of the speech translations, in terms of accuracy of the translations produced as the final outcome.

A. *RQ1 – How do errors in the speech recognition process affect the resulting machine translation outcome?*

The χ^2 tests performed on the contingency tables in Table IV show that, when an input speech is accurately captured, an adequate translation is generally provided. This finding suggests that the speech recognition component is critical and speech translation should be chosen only for those languages that guarantee a high recognition accuracy. Further advances in the state of the art speech translation technology appear necessary to produce accurate transcripts from a continuous and unbroken streams of text, as in the case group meetings. Of course, in a real setting, other issues may arise, such as noise or speech overlapping. However, these are challenges the extent of which cannot be assessed by means of a mere simulation.

B. *RQ2 – Which source language works better for non-native English speakers with state-of-the-art speech translation technology?*

To answer this research question, we compared the performances of language pairs that use English as the source language against those that use either Italian or Portuguese. More specifically, we compared $EN_{(IT)}$→PT against $IT_{(IT)}$→PT, and $EN_{(PT)}$→IT against $PT_{(PT)}$→IT.

A remarkable difference can be observed in these two comparisons. The speech translation results with pairs that use English as the source languages, i.e., $EN_{(IT)}$→PT and $EN_{(PT)}$→IT, are respectively 37% and 20% of adequate speech translations. Instead, performance with the other two language pairs $IT_{(IT)}$→PT and $PT_{(PT)}$→IT are much better, since they have achieved 78% and 71% of adequate speech translations, respectively.

RQ1 has confirmed that inaccuracies in the speech translations do affect accuracy of the resulting translations as expected. Therefore, in order to understand whether these performance discrepancies are due to poor recognition rather than translation performance, we looked at the intermediate speech recognition results only. We observed that performance in the case of English input was good in one case, $EN_{(PT)}$, with 78% accurate transcriptions, and average in the other one, $EN_{(IT)}$, with 51% accurate transcriptions. Therefore, this difference in the speech recognition results suggests that, no matter how accurate the transcription is, the resulting speech translation performance tend to be inadequate when English is the source language. Therefore, we can conclude that current speech translation technology works better when non-native speakers can use their native languages as the source language.

C. *RQ3 – How well does state-of-the-art speech translation technology perform when English is not used as either the source or the target language?*

RQ2 confirmed that English is not a good choice for non-native speakers to provide input to current speech translation systems. Therefore, here we consider only the pairs that do not involve the English language, i.e., only those with Italian and Brazilian Portuguese used as either the source or the target language. We found that both $IT_{(IT)}\rightarrow PT$ and $PT_{(PT)}\rightarrow IT$ performed well with 78% and 71% of adequate speech translations, respectively.

Hence, is English to be avoided altogether by non-native speakers using a speech translation system? To answer this, we complement RQ2 findings by looking at the translation adequacy of the pairs having English as the target language. We observe that $IT_{(IT)}\rightarrow EN$ is the best performing language pair, with 94% of adequate translations provided. Instead, with the $PT_{(PT)}\rightarrow EN$ pair the system achieved an average performance with 55% of sentences adequately translated. Therefore, our simulation suggests that: (i) with respect to languages with roots in Latin, like Italian and Portuguese, speech translation systems perform well when translating from one another; (ii) if English is to be used, it is better to use it as the target language rather than the source.

VI. Threats to Validity

The generalizability of the results from our study is limited by being a simulation. We identified the following major threats.

First, in our simulation we only used one speech translation system (Google Translate mobile). Therefore, findings might not extend to other existing speech translation technologies available. We acknowledge the need to compare the performance of more systems in our future work. In addition, as per the evaluation procedure, we involved only two raters, one Italian and one Brazilian. However, the test set included sentences in Italian, Brazilian Portuguese, and English too. Since English sentences were not read in by any native speaker, we acknowledge the possibility that speech recognition results might have been negatively influenced in such cases (i.e., $EN_{(IT)}$, and $EN_{(PT)}$). Yet, even poorer accuracy was obtained in the $EN_{(IT)}$ combination when we employed as a rater an Italian woman who lives in UK for two years and has excellent communication skills. Furthermore, we consider the case of group communication involving non-native speakers who use English a more representative scenario for global software engineering.

Second, we evaluated speech translation exclusively in terms of adequacy, that is, the raters judged the comprehensibility of a translated sentence only with respect to the original sentence and not in the context provided by the chat log. When evaluating translation quality other dimensions are often explicitly taken into account, such as fluency and fidelity. However, as Hutchins and Somers noted [13], style matters only when a translation is adequate and intelligible. On the contrary, it is more efficient to analyze just those cases where the output is rated incomprehensible, leading one to suppose something has gone wrong. As per speech recognition, we assessed the accuracy of transcripts using the sentence as the unit of analysis. Instead, word accuracy (W_{acc}), the standard measure for accuracy, works at word level. Although we provided a clear mapping between our categories and W_{acc} intervals, using word accuracy in our future work will allow us to have a more precise breakdown of the number of errors occurring in each sentence.

Third, as per the evaluation of the efficiency, i.e., to the amount of *time* necessary to the solution tested in order to capture and translate the input text, we were not able to capture the response times precisely because the app does not provide this piece of information. Anyway, we directly observed that response times were usually instantaneous at both sites, definitely within 1 second.

Finally, our simulation analyzed a selection of entries from a number of chat logs collected from requirements meetings that were conducted in English, without speech translation. Hence, albeit results from our simulation are somewhat encouraging, at this moment we can by no means hypothesize whether the speech translation quality of Google Translate mobile service would be good enough to allow participants to complete a group task successfully – in our scenario, allow stakeholders to define and negotiate software requirements for web applications. Our previous works with machine translation (e.g., [5]) show that employing machine translation does not prevent stakeholders to complete such tasks, although it slows meetings down. Yet, one can reasonably argue that even greater issues would arise switching from machine translation to speech translation, due to the further inaccuracies brought by speech recognition. Requirements meetings are complex, communication-intensive tasks that require specialized knowledge and techniques to be applied. As such, low-quality speech translations might worsen or even cause misunderstandings during their execution, thus possibly generating defects in the resulting requirements specifications.

VII. Conclusions & Future Work

In this paper, we investigate whether mobile speech translation technology is ready to be used in the context of global software development to support communication-intensive activities, such as remote multilingual requirements meetings. In particular, we started by assessing the performance of the Google mobile speech translation service. We run a simulation using a test set of entries selected from

real logs of distributed requirements meetings. The transcripts and the resulting translations were assessed in terms of accuracy and adequacy, respectively. Obviously, a definitive answer cannot come from an initial simulation. Yet, a few interesting points can be made.

First, in general speech translation performance varies, with adequacy of translations ranging between 94% and 20%. Such variation is mostly influenced by the accuracy of the speech recognition component. Second, languages are not supported at the same level. Our initial results show that speech translation works better when English is the target language, not the source. We do not know the extent to which speech recognition is sensitive with respect to pronunciations and accent. Finally, we found that, when speech transcriptions are correct, so are translations. Then, we conclude that the speech recognition component, the accuracy of which may vary with the source language, is the critical part that makes speech translation a viable solution for multilingual communication.

As future work, we intend to seek for confirmation of these initial results. In particular, we will run a controlled experiment in order to compare groups of people who communicate through a speech translations system, using either English or their native languages, to complete communication-intensive tasks in the context of globally distributed development teams.

ACKNOWLEDGMENT

This work fulfils the research objectives of the PON 02_00563_3470993 project "VINCENTE - A Virtual collective INtelligenCe ENvironment to develop sustainable Technology Entrepreneurship ecosystems" funded by the Italian Ministry of University and Research (MIUR). This research is also partially funded by the Rio Grande do Sul State funding agency (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] 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.

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

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

[5] 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.

[6] F. Calefato, D. Damian, and F. Lanubile, "Computer-Mediated Communication to Support Distributed Requirements Elicitations and Negotiations Tasks", Empirical Software Engineering Journal, 2012, Vol. 17, No. 6, pp. 640-674, ISSN: 1573-7616, DOI: http://dx.doi.org/10.1007/s10664-011-9179-3.

[7] L. Dib, D. Petrelli, and S. Whittaker. "Sonic Souvenirs: Exploring the Paradoxes of Recorded Sound for Family Remembering." *In Proc. Conf. on Computer Supported Cooperative Work (CSCW'10)*, Savannah, GE, USA, pp. 391-400, Feb. 6–10, 2010.

[8] T. Duarte, R. Prikladnicki, F. Calefato, and F. Lanubile, "Speech Recognition for Voice-Based Machine Translation", *IEEE Software*, 31(1), Jan./Feb. 2014.

[9] G. Gao, H-C. Wang, D., and S.R. Fussell. "Same translation but different experience: the effects of highlighting on machine-translated conversations." *In Proc. SIGCHI Conference on Human Factors in Computing Systems* (CHI '13), Paris, pp. 449-458, Apr.–May 27, 2013.

[10] R. Garland. "The Mid-Point on a Rating Scale: Is it Desirable?", *Marketing Bulletin*, Vol. 2, 1991, pp. 66-70.

[11] X. Huang, A. Acero, H.W. Hon, "*Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*." Prentice Hall PTR, Upper Saddle River, NJ, USA, 2001.

[12] 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.

[13] W.J. Hutchins and H.L. Sommers, "*An Introduction to Machine Translation*", Academic Press, 1992

[14] R. Johns. "One Size Doesn't Fit All: Selecting Response Scales For Attitude Items." *Journal of Elections, Public Opinion, and Parties*, Vol. 15, No. 2, 2005, pp. 237-264.

[15] V. Kalnikaitė, P., and S. Whittaker. "Markup as you talk: establishing effective memory cues while still contributing to a meeting". *In Proc. Conf. on Computer Supported Cooperative Work (CSCW'12)*, Seattle, WA, USA, pp. 349-358, Feb. 11-15, 2012.

[16] 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.

[17] R. Prikladnicki, T. Duarte, T. Conte, F. Calefato and F. Lanubile, "Real-Time Machine Translation for Software Development Teams", *Software Engineering Innovation Foundation (SEIF'13)*, Rio de Janeiro, Brazil, Nov. 25–26, 2013

[18] 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, pp. 299-311, 2013.

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

[20] S. Tucker and S. Whittaker. "Time is of an Essence: an Evaluation of Temporal Compression Algorithms." *In Proc. Conf. on Human Factors in Computing Systems (CHI'06)*, Montréal, Canada, pp. 329-338, Apr. 24-27, 2006.

[21] N. Yamashita, R. Inaba, H. Kuzuoka, and T. Ishida. "Difficulties in establishing common ground in multiparty groups using machine translation." *Proc. 27th Int'l Conf. on Human Factors in Computing Systems (CHI '09)*, Boston, USA, pp. 679-688, Apr. 4-9, 2009.

[22] W3C Web Speech API Specification, https://dvcs.w3.org/hg/speech-api/raw-file/tip/speechapi.html (last accessed Feb. 24, 2014).

[23] H-C. Wang, S. Fussell, and D. Cosley. "Machine translation vs. common language: effects on idea exchange in cross-lingual groups." *Proc. Conf. Computer supported cooperative work (CSCW '13)*, San Antonio, TX, USA, pp. 935-944, Feb. 23-27, 2013.