Evaluating Automatic Speech Recognition in Translation

Evelyne Tzoukermann tzoukermann@mitre.org
Corey Miller camiller@mitre.org
The MITRE Corporation, 7525 Colshire Dr, McLean, VA 22102

Abstract

We address and evaluate the challenges of utilizing Automatic Speech Recognition (ASR) to support the human translator. Audio transcription and translation are known to be far more time-consuming than text translation; at least 2 to 3 times longer. Furthermore, time to translate or transcribe audio is vastly dependent on audio quality, which can be impaired by background noise, overlapping voices, and other acoustic conditions. The purpose of this paper is to explore the integration of ASR in the translation workflow and evaluate the challenges of utilizing ASR to support the human translator. We present several case studies in different settings in order to evaluate the benefits of ASR. Time is the primary factor in this evaluation. We show that ASR might be effectively used to assist, but not replace, the human translator in essential ways.

1. Introduction

Advances in deep learning have had a major impact on human language technologies and the results have been visible in Neural ASR and in Neural Machine Translation. As such, we can now reevaluate the benefits and challenges of using improved ASR systems to facilitate the translation of audio documents.

Translating audio documents involves the use of a media player specifically developed for performing transcription and translation; that is, a media player capable of going back and forth in the audio stream, looping over segments in order to listen and re-listen to unclear sections, and slowing down the audio in order to capture the content. These tools are for the most part available to government linguists. The problem lies in the nature of the incoming files which may be very noisy. Numerous factors can be the source of the noise in the audio files. Multiple conditions can be present, such as inside/outside noise, landline, cellular, and Voice over Internet Protocol (VoIP). Each of these conditions can in turn be associated with a diversity of noise, such as overlapping voice with other voices, music, street noise, static on the line, etc. As a result, government linguists who are given the task of translating audio files spend a considerable amount of time translating audio input[1].

[1] According to professional industry standards, for each minute of audio, an average of 4 times the length is required to translate. Thus, for example, for a short fifteen minutes of speech, one hour of transcription time is required. Furthermore, for noisy recorded audio, the same operation can take several more hours. See http://www.confidentialtranscription.co.nz/cost.htm.

Given this variability in recording conditions and incoming files, we have decided to isolate the problem of "noisy" files, and deal only with "clean" files or clearly recorded files in order to investigate the integration of ASR technologies in the workflow of audio translation. This reduces factors impacting performance so we can rigorously test without confounding factors.

The next section presents related research particularly as it relates to operational settings. Section 3 shows three different tasks on which we applied ASR and scoring. Section 4 shows the results of the experiments. We then conclude in offering recommendations for decision makers.

2. Related Research

Academic literature abounds in research and evaluation on ASR, speech translation, and all the applications that include speech recognition and machine translation (MT). However, there is less work addressing the issues that we are bringing up in this paper, which is the integration of ASR into the linguist workflow. From a research perspective, work on ASR for translation has been studied and presented within the lens of speech translation – that is, audio input in a given source language translated into text of a target language, or speech-to-speech translation, which is the same as speech translation but the target language output is spoken.

Stüker et al. (2007) describe the operational settings of the European Union where European Parliament speeches are translated in 21 languages, and the need for combining ASR and MT is required. Stüker et al. and Paulik et al. (2005) report on the benefits of smoothly coupling the two technologies and refer to a speech translation enhanced ASR system (STE-ASR). They demonstrate how the quality of ASR influences the overall quality of the output and how by adapting the ASR models to the task at hand, the Word Error Rate (WER) is lowered by 3% to 4.8%, providing more accurate results.

From a practical perspective, ASR offers a variety of advantages as well as challenges to translators. Ciobanu (2014) surveys the advantages and disadvantages of using ASR in translation services. The outcome of the survey demonstrates that the advantages outweighed the disadvantages and that "professional translators are essentially missing out by not engaging with such technologies more". In his later work, Ciobanu (2016) conducted research at University of Leeds Centre for Translation Studies to study the benefits of inserting ASR, and presented the challenges of ASR in the language services industry by concluding that "ASR has the potential to increase the productivity and creativity of the translation act, but the advantages can be overshadowed by a reduction in translation quality unless thorough revision processes are in place." (p.124)

Other academic research (e.g. Zapata 2012 and 2016) explores the benefits of interactive translation dictation, a translation technique that involves interaction with multimodal interfaces equipped with ASR throughout the entire translation process. In this work, Zapata demonstrates the range of interaction between humans and machines in translation processes and claims that a new turn in translation technology is needed, with the human translator as the central axis of investigation. Zapata's work provides a basis for well-grounded research on translator-computer and translator-information interaction, particularly for the design and development of interactive translation dictation environments. These interactive systems are expected to support professional translators' cognitive functions, performance, and workplace satisfaction.

In contrast, our current approach explores the extent to which resources should be expended at improving ASR transcripts prior to either human or machine translation. In the case of both speech translation and machine translation, we factor in the time that must be expended to correct its output. These measurements are considered with respect to three different possible workflows for combining ASR and translation.

3. Method

This section addresses the selection of languages and files, and explains the way audio files are processed, timed, and scored for accuracy. In addition, we describe the human particpants in our experiment.

3.1. Language Selection and File Selection

For this experiment, we selected the following languages: French, Spanish, and Serbian. We used two of the latest systems that are publicly available online: IBM Bluemix ASR[2] and Google Translate ASR[3]. For the French experiment, we selected excerpts from the speech that French President Emmanuel Macron delivered during his inauguration on May 15, 2017[4]. We downloaded the files from YouTube and used an online converter to convert the files into *.wav format. We then used Audacity open source software[5] for recording and editing and selected two speech segments of one to two minutes each. For the Spanish and Serbian experiments, similar political speeches by Mexican President Enrique Peña Nieto[6] and Serbian President Aleksandar Vučić[7] were selected and Praat software[8] was used to navigate and listen in order to make transcript corrections. As additional Spanish data, we used a television interview of the Bolivian Minister of the Economy, Luis Arce[9]. The files were originally recorded at a very clear high-quality stereo 44100Hz, PCM 16 bit, and this naturally yields better results. The experiments were performed by four linguists, one French, two Spanish, and one Serbian. Note that the Serbian linguist is a professional translator whereas the other three linguists are only fluent in the language.

3.2. Running and Analyzing ASR

For each of the tasks, files were run through an ASR engine. Each audio file was run through IBM ASR and Google ASR. Note that since Serbian is not available among the IBM ASR languages, the Serbian files were run only through the Google ASR system. While the IBM system allows file upload, the Google system does not. For the IBM system, we used the French and Spanish 16KHz broadband models. The Google Translate system provides a microphone icon that one can click in order to provide live audio input via a microphone, rather than typing text in. We employed a software package called Virtual Audio Cable 4.5[10] that allowed us to redirect file-based audio through the microphone so that it could serve as input to Google Translate. Note that when providing audio to Google Translate, there are two outputs: on the left side

2 https://speech-to-text-demo.ng.bluemix.net/ (as of February 2, 2018)
3 https://translate.google.com/ (as of February 6, 2018)
4 https://www.youtube.com/watch?v=K8Ea_RXktgg
5 https://www.audacityteam.org/
6 https://www.youtube.com/watch?v=qUeqwMl-_U0
7 https://www.youtube.com/watch?v=kGz9diiTV-M
8 www.fon.hum.uva.nl/praat/
9 https://www.youtube.com/watch?v=pxqw4TaqK1A
10 http://software.muzychenko.net/eng/vac.htm

is the ASR output in the source language, and on the right side is the translation in the target language; what we are calling "speech translation".

Figures 1a and 1b below show a sample of IBM ASR and Google ASR output for French. Overall, both transcriptions are of good quality in that the reader can get a gist of the speech. Both transcriptions are close to each other. The main differences in Figures 1a and 1b are highlighted in yellow. Note that the figures show the raw output of the recognition, and thus contain errors, such as agreement errors, erroneous words, missing words, substituted words, etc. The IBM system differs from Google in generating sentence boundaries, and as a consequence, adds punctuation and capitalization. In the IBM system, sentence boundaries appear to be based essentially on pauses and since Macron's speech is well articulated, the program adds too many periods, such as "L'audace de la liberté. Les exigences de l'égalité. La volonté de la fraternité." As a reference and comparison, Figure 1c shows the official transcript of the president's speech. One can notice the differences in orthographic realization (also mentioned in Section 3.3) with the representation of numbers, such as "*sept mai*" and "*7 mai*".

L'audace de la liberté. Les exigences de l'égalité. La volonté de la fraternité. Or depuis des décennies la france doute d'elle-même. Elle se sent menacée dans sa culture quand son modèle social dans ces croyances profondes et les doutes. De ce qui le fait. Voila pourquoi mon mandat sera guidée par des exigences. La première sera de rendre aux français. Cette confiance en eux. Depuis trop longtemps a faibli. Oh je vous rassure. Je n'ai pas pensé une seule seconde. Quel sera instauré comme par magie le soir du sept mai. Ce sera un travail long. Exigeant. Mais indispensable. Il appartiendra de convaincre les françaises et les français que notre pays.

Figure 1a. IBM ASR output of excerpt 1 from French President Emmanuel Macron

l'audace de la Liberté l'exigence de l'Égalité la volonté de la fraternité or depuis des décennies la France doute d'elle-même elle se sent menacé dans sa culture dans son modèle social dans ses croyances profondes elle doute de ce qu'il a fait voilà pourquoi mon mandat sera guidé par deux exigences la première sera de rendre au français cette confiance en eux depuis trop longtemps affaiblit oh je vous rassure je n'ai pas pensé une seule seconde qu'elle se restaurer comme par magie le soir du 7 mai ce sera un travail non exigeant mais indispensable il m'appartiendra de convaincre les Françaises et les français que notre pays

Figure 1b. Google ASR output of excerpt 1 from French President Emmanuel Macron

...l'audace de la liberté, l'exigence de l'égalité, la volonté de la fraternité. Or, depuis des décennies, la France doute d'elle-même. Elle se sent menacée dans sa culture, dans son modèle social, dans ses croyances profondes. Elle doute de ce qui l'a faite. Voilà pourquoi mon mandat sera guidé par deux exigences. La première sera de rendre aux Français cette confiance en eux, depuis trop longtemps affaiblie. Je vous rassure, je n'ai pas pensé une seule seconde qu'elle se restaurerait comme par magie le soir du 7 mai. Ce sera un travail lent, exigeant, mais indispensable. Il m'appartiendra de convaincre les Françaises et les Français que notre pays…

Figure 1c. Excerpt of the Official Transcript of French President Emmanuel Macron

3.3. Preparing files for ASR scoring

The object of ASR scoring is to establish a word error rate (WER) with respect to a given reference transcription and hypothesis transcription. We use NIST sclite[11] scoring software, which is a tool for scoring and evaluating the output of speech recognition systems.

In this project, we have two levels of transcript correction: basic and full, as explained in Section 3.4. For the purpose of ASR scoring, we have decided to focus initially on the basic correction.

[11] See https://www.nist.gov/sites/default/files/documents/itl/iad/mig/KWS15-evalplan-v05.pdf

The reason for this is that it can provide the most optimistic WER. For example, if ASR confuses two French homophones, e.g. "*parle*" and "*parles*", we don't necessarily want to penalize the recognizer for confusing these. At the basic correction step, if we encounter *parles*, we also consider *parle* to be correct, since the two words are homophones and one is an inflectional variant of the other.

The scoring software compares the machine-generated transcription of audio segments to a human transcription of the same segment. Usually, the segments are aligned using time stamps associated with the audio file. Since Google ASR does not provide us with timing information, we added line breaks in the reference and hypothesis files corresponding to sentences or intonation units and tagged the lines in parallel, according to sclite's *trn* format. The purpose of separating the text into intonation units is to give the scoring mechanism, which is based on dynamic programming, a chance to reset and compare the same segments. This way, the exact boundaries do not matter[12], as long as they are consistent between reference and hypothesis.

We removed punctuation from both reference and hypothesis files. At this stage, punctuation accuracy is not scored. We also removed capitalization in both files. Theoretically, the output of step 1 (i.e. the basic transcript correction) for Google and IBM would be the same. However, in practice, there may be multiple cases where the orthographic realizations differ and thus, would require full normalization. For example, if Google outputs *50* and IBM outputs *fifty*, basic correction does not require this to be normalized. So, creating separate basic ASR scoring transcripts for Google and IBM was deemed the best way forward to ensure that neither engine is penalized for confusing things like *50* with *fifty*.

Table 1 shows the results of the IBM ASR and the Google ASR for French, Spanish, and Serbian. Note that since Serbian is not among the languages available in IBM ASR system, only the Google results are presented here.

	IBM ASR		**Google ASR**	
	Correct	**WER**	**Correct**	**WER**
French 1	87.2	12.8	96.7	3.3
French 2	82.0	22.3	93.9	6.1
Spanish 1	94.4	6.5	95.6	4.4
Spanish 2	80.3	22.2	81.7	18.3
Serbian	n/a	n/a	90.4	10.4

Table 1. Accuracy of IBM and Google ASR systems

Overall, the Google ASR system performs distinctly better than the IBM system. For the French documents, the results are on average 10% better with Google than they are with IBM. All the documents processed by Google are over 90% for Serbian and close to, or above 95% for the other languages. The difference in performance between the two Spanish documents has to do with their contents. "Spanish 1" is the Mexican president's speech, while "Spanish 2" is a television interview with the Bolivian economics minister. The speech is much more articulated and deliberate in contrast to the interview.

[12] We have noticed that using particularly long sentences can result in higher WER, presumably as a side effect of dynamic programming. This warrants further study.

We found that the performance of ASR is a strong indicator of how much effort will be required for humans to edit a document so that it is accurate. High ASR performance, e.g. low WER results, suggests that the human effort is minimized, even for more difficult transcriptions, such as the television interview. The following section shows the setup of the experiments where linguists correct the transcripts to prepare them for the follow-on processes.

3.4. Setting the Different Tasks

In order to measure the benefits of integrating ASR in the linguist's workflow, we designed three different task scenarios where each of the task components is measured in time. Each of these tasks represents a different possible workflow for combining translation and ASR.

Task 1. **BASIC** – ASR followed by Human Translation: the linguist is given an audio document and the file is run through the ASR system(s). The task consists of (i) correcting the ASR output, and (ii) translating the document into English. Note that correcting the ASR output at this level is time consuming. It involves using a media player, listening the audio, comparing it with the ASR output, and going forward and backward with the player to add, substitute, and replace words. At the same time, since the human will translate the ASR transcript, the output does not need to be perfectly accurate on sentence boundaries, capitalization, punctuation, and word agreement. However, the transcript should be accurate enough so that somebody using the resulting edited ASR transcript for translation purposes will **not** need to consult the audio file. This is the reason we call this correcting task "basic".

Task 2. FULL – ASR followed by Machine Translation (MT): the linguist is given an audio document as well as the corrected ASR output from Task 1. This task consists of fully correcting the ASR output after it has already undergone the basic level correction. Capitalization, agreement, and accents need to be corrected. Sentence boundaries need to be inserted for some systems, such as Google ASR and corrected for others, such IBM ASR. This stage consists of a complete and thorough transcript correction so that it is presentable to an MT engine. This is what we call "full" corrected output. This output is then submitted to Google Translate and the linguist times how long it takes him/her edit. This is a standard workflow in machine translation and is referred to as post-edited machine translation (PEMT).

Task 3. Speech to Text Translation: This step consists of using an end-to-end speech-to-text translation system, such as Google Translate, which takes audio input in our three selected languages and returns the text translated in English. The linguist takes the English output and times how long it takes them to edit and correct it. We refer to this operation as post-edited speech translation (PEST).

4. Results

Table 2 presents the time measurements (in minutes) of the transcription editing parts of Task 1 and Task 2, along with the word counts and the duration of each file. It is interesting to point out that it takes two to three times as long to correct the IBM transcript at a basic level as it takes to correct the Google transcript. For both IBM and Google ASR, and except for French 1, the time it takes to complete the full correction is minimal compared to the time necessary for achieving the basic level correction. Also, French 2 had more errors than French 1, yet it took less time to correct the errors—perhaps these results are attributable to priming effects since the linguist worked on French 1 followed by French 2.

The numbers in Table 2 clearly demonstrate that the quality of Google ASR is markedly better than that of IBM ASR in this use case. It also shows the correlation between ASR performance and the human time needed to correct the transcripts. As mentioned above, high ASR performance yields human time saving for transcript correction. At the same time, it shows that the overall timings exceed 4 times the duration of the cuts, as mentioned by industry standards. This is probably due to the experimental nature of the tasks (small amount of data) and to the fact that the linguists lacked experience in these particular tasks. More data need to be evaluated and run in similar experiments to understand these numbers in a clear fashion.

			Word count	**IBM ASR**			**Google ASR**		
Language	**Doc**	**Time**		**Basic**	**Full**	**Total**	**Basic**	**Full**	**Total**
French	1	1:22	180	20	16	**36**	7	3	**10**
	2	1:11	145	15	7	**22**	4	3	**7**
Spanish	1	2:42	343	18	5	**23**	6	5	**11**
	2	2:00	203	35	5	**40**	18	7	**25**
Serbian	1	3:30	516	n/a	n/a	n/a	16	6	**22**

Table 2. Times in Minutes for Correcting Basic and Full ASR Transcripts (Task 1 and Task 2)

Table 3 shows the time it takes to perform additional components of the Tasks. In Task 1, once the basic level of correction is completed, the linguist performs the manual translation of the file (Human Trans). For Task 2, once the full level of correction is completed, the linguist ingests the file into Google Translate, then checks and improves the accuracy of the machine translation by post-editing the document (PEMT). In Task 3, the linguist simply performs post-editing of the speech translation (PEST). Human translation is clearly the slowest of the three. Note that from a translation quality standpoint, PEMT is very reliable and very quick for French and Spanish.

Lan-guage	**Doc**	**Word count**	**Human Trans (Task 1)**	**PEMT (Task 2)**	**PEST (Task 3)**	**Google Basic + Human-Trans (Task 1)**	**Google Basic + Full + MT+ PEMT (Task 2)**
French	1	180	12	**2**	3	19	12
	2	145	7	**3**	3	11	10
Spanish	1	343	13	**2**	5	19	13
	2	203	19	**4**	7	37	29
Serbian	1	516	n/a	**23**	33	n/a	45

Table 3. Times in Minutes for Human Translation, Post-Editing Machine Translation, and Post-Editing Speech Translation

The final three columns of Table 3 represent our three tasks or workflow scenarios, and the total time required to achieve a correct translation using them[13]. Task 3, using end-to-end speech to text translation, allows one to ignore the transcription correction process and proceed directly to post-editing the speech translation, and this seems to be the shortest way to a correct translation for the audio documents studied here. Task 1, ASR with basic correction followed by human translation, appears to be the slowest. Task 2, ASR followed by full transcript correction, followed by MT and PEMT, appears to be the second fastest method. The timing for Task 3 appears very competitive, which leads us to conclude that this approach is very compelling and promises to aid in creating more efficient audio translation workflows.

5. Conclusion and Future Work

In this paper, we addressed the challenges of using ASR to support the linguist translating audio files. Since there is large variability in incoming audio files, we experimented solely with clearly recorded files to control for extraneous variables and ensure reliable results. We implemented three different tasks where we measured the time it takes for linguists to achieve correct translations following different paths. We combined these timings with post-editing measures and we demonstrated that, except for the French documents generated by IBM ASR, there is a correlation between the quality of the automatic recognition and the amount of work that is necessary to edit the transcripts or a speech translation.

While it appears that time can certainly be saved by restricting transcription editing to a "basic" level, and that this is sufficient for subsequent audio-free human translation, we are not sure whether such a transcript is sufficient to generate reasonable MT or if would incur PEMT costs down the line. We need to compare PEMT on basic and full transcripts, since in this study we only measured PEMT based on the full transcript.

Based on these preliminary results so far, it appears that speech translation, coupled with PEST, may offer the fastest route to correct translation. The advantage to this approach is that it obviates the need for transcription correction. However, we need to examine this more closely and on more data in a variety of acoustic conditions, since the quality of speech translation is obviously especially sensitive to the quality of the input audio.

All in all, we can conclude that ASR has the potential to increase linguist productivity. This concurs with the outcomes of Ciobanu's survey. These results can be used as recommendations for decision makers who face the need to modernize their processes and increase the productivity of their workforce.

In future work, we are planning on making more extended tests and more fine-grained tests so that we can estimate the limitations of ASR in various domains and genres. Additionally, as research is making progress in the areas of adaptation and customization, we would like to explore how customized models for such domains and genres can improve recognition, and consequently reduce transcription adjustment time, leading to more efficiently produced translations of audio.

[13] Note that Serbian appears to be the language that is the most time consuming. The translations were processed by a professional translator as opposed to the other linguists who are fluent but not professional translators, and this possibly explains the discrepancy.

Acknowledgements

We thank Vanesa Jurica and Jason Duncan for their linguistic expertise as participants in this research.

References

Ciobanu, D. (2014). Of Dragons and Speech Recognition Wizards and Apprentices. Revista Tradumàtica, (12), 524–538.

Ciobanu, D. (2016). Automatic Speech Recognition in the Professional Translation Process. Translation Spaces, 5(1): 124–144.

Paulik, M., S. Stüker, C. Fügen, T. Schultz, Thomas Schaaf, and A. Waibel (2005), Speech Translation Enhanced Automatic Speech Recognition, in ASRU, San Juan, Puerto Rico.

Stüker, S., M. Paulik, M. Kolss, C. Fügen, and A. Waibel (2007), Speech Translation Enhanced ASR for European Parliament Speeches - On the Influence of ASR Performance on Speech Translation, in Proc. ICASSP, Honolulu, Hawaii.

Zapata, J. (2012). Traduction dictée interactive : intégrer la reconnaissance vocale à l'enseignement et à la pratique de la traduction professionnelle. M.A. thesis. University of Ottawa. Canada.

Zapata, J. (2016). Translators in the Loop: Observing and Analyzing the Translator Experience with Multimodal Interfaces for Interactive Translation Dictation Environment Design. Ph.D. thesis. University of Ottawa. Canada.