Skype Translator: Breaking Down Language and Hearing Barriers
A Behind the Scenes Look at Near Real-Time Speech Translation

William D. Lewis
Microsoft Research
One Microsoft Way
Redmond, WA 98125
wilewis@microsoft.com

Abstract

In the Skype Translator project, we set ourselves the ambitious goal of enabling successful open-domain conversations between Skype users in different parts of the world, speaking different languages. Building such technology is more than just stitching together the component parts; it also requires work in allowing the parts to talk with one another. In addition to allowing speech communication between users who speak different languages, these technologies also enable Skype communication with another class of users: those who have deafness or hard of hearing. Accommodating these additional users required design changes that benefited all users of Skype Translator. Not only does Skype Translator promise to break down language barriers, it also promises to break down the hearing barrier.

1 Introduction

In 1966, Star Trek introduced us to the notion of the Universal Translator. Such a device allowed Captain Kirk and his crew to communicate with alien species, such as the Gorn, who did not speak their language, or even converse with species who did not speak at all (e.g., the Companion from the episode *Metamorphosis*). In 1979, Douglas Adams introduced us to the "Babelfish" in the *Hitchhiker's Guide to the Galaxy* which, when inserted into the ear, allowed the main character to do essentially the same thing: communicate with alien species who spoke different languages. Although flawless communication using speech and translation technology is beyond the current state of the art, major improvements in these technologies over the past decade have brought us many steps closer. Skype Translator puts together the current state of the art in these technologies, and provides a speech translation service in a Voice over Internet (VoIP) service, namely Skype. With Skype Translator, a Skype user who speaks, say, English, can call a colleague or friend who speaks, say, Spanish, and be able to hold a bilingual conversation mediated by the translator.[1]

In the Skype Translator project, we set ourselves the ambitious goal of enabling successful open-domain conversations between Skype users in different parts of the world, speaking different languages. As one might imagine, putting together error-prone technologies such as speech recognition and machine translation raises some unique challenges. But it also offers great promise.

The promise of the technologies is most evident with children and young adults who accept and adapt to the error-prone technology readily. They understand that the technology is not perfect,

[1] It is important to note that the Speech Translation service described here is not the first of its kind. There have been a number of Speech Translation projects over the past couple of decades, e.g., VERBMOBIL (Wahlster 2000) and DARPA GALE (Olive et al 2011). See Kumar et al (2014) for more background. Crucially, however, Skype Translator is the first of its kind integrated into a VoIP service available to hundreds of millions of potential consumers.

yet work around and within these limitations without hesitation. The ability to communicate with children their own age, irrespective of language, gives them access to worlds that fascinate and intrigue them. The stunning simplicity of the questions they ask, e.g., "Do you have phones?" or "Do you like wearing uniforms in school?", shows how big the divide can be (or is perceived to be), but it also shows how strongly they wish to connect. Because they also readily adapt the modality of the conversation, e.g., using the keyboard when speech recognition or translation may not be working for them, means they also readily accept the use of the technology to break down other barriers as well. Transcriptions of a Skype call, a crucial cog in the process of speech translation, are essential for those who do not hear, as are the text translations of those transcripts. Freely mixing modalities and readily accepting them offers access to those who might otherwise be barred access. Adjusting the design of Skype Translator to accommodate those with deafness or hard of hearing added features that benefited all users. The technologies behind Skype Translator not only break down language barriers, they also break down the hearing barrier.

2 Breaking down the Language Barrier: Technologies Behind Skype Translator

Underlying Skype Translator is a speech-to-speech (S2S) pipeline. The pipeline consists of three primary components:[2]

A. Automated Speech Recognition (ASR)
B. Machine Translation (MT) engine
C. Text-to-Speech (TTS)

The first, ASR, converts an input audio signal into text, essentially "transcribing" the spoken words into written words. Each language must have its own custom built engine, and it generally requires hundreds to thousands of hours of human-transcribed content in order to train a robust ASR engine. Machine Translation (MT), the second component, maps words and phrases in one language to words and phrases in the second. Most modern MT is statistically based (e.g., Microsoft Translator and Google Translate use statistical engines), and learn from *parallel* data (i.e., documents sourced in one language and translated into another) a probabilistic mapping between words and phrases in one language to translations and those in the other. Statistical MT is often trains over millions, and sometimes billions, of words of parallel text. Finally, Text-to-Speech (TTS) maps text in a language to a spoken form, and is generally trained on carefully recorded audio and transcripts from one native speaker.

Armed with these three technologies, it would seem that all you would need to do is stitch one to the other in order to build a working S2S pipeline: ASR outputs words in text, MT converts text in one language to text in another, and TTS outputs the audio of the words in the target language. However, it is not quite that simple. The problem starts with the users: most language speakers assume they are talking fairly fluently when they speak, but often, what is being said is quite different than what a person *thinks* is being said. Here's an example from a corpus of transcribed telephone conversations:[3]

a. Yeah, but um, but it was you know, it was, I guess, it was worth it.

---

[2] For a technical overview of a Speech Translation pipeline, see Kumar et al (2014).

[3] This example is drawn from CALLHOME, a corpus of audio and transcripts of telephone conversations. It is one of the most commonly used corpora used by the speech research community to train ASR engines. It is available through the Linguistic Data Consortium (LDC, http://www.ldc.upenn.edu/), LDC corpus ID # LDC97S42.)

The user likely intended to, and probably thought, he said the following:

b. Yeah. I guess it was worth it.

When translation is applied, translating the first (a) can result in "word salad", something that the recipient of the translation would likely not understand. When cleaned up, however, such as in (b), the translation may be perfectly understandable. For example, here are translations to German for both the original (a) and the cleaned up (b) version:

a. Ja, aber ähm, aber es war, weißt du, es war, ich denke, es hat sich gelohnt.
b. Ja. Ich denke, es hat sich gelohnt.

But the issue is even more complicated than that. Current MT technology is based on translating grammatical, well-formed, and well-punctuated sentences. The problem is that people do not talk in sentences, nor do they insert punctuation when they talk (unless for dramatic effect), nor is the output necessarily grammatical (per (a) above). As it turns out, there is a lot of work in "repairing" ASR so that its output is more favorable to MT. Take, for example, the following utterance by a Spanish speaker using Skype Translator. Note the varying translations depending on how the input is punctuated. (e) is probably the closest to the intended punctuation and meaning:

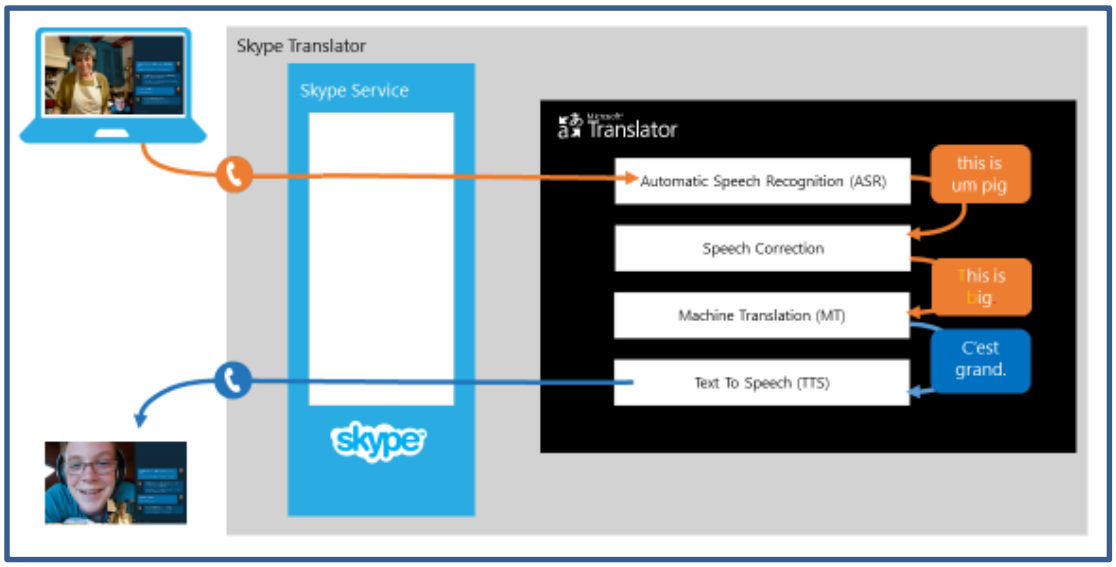
c. claro también es verdad sí eso es cierto → also clear is true yes that is true
d. claro. también es verdad. sí. eso es cierto. → of course. is also true. yes. that is true
e. claro. también, es verdad. sí. eso es cierto. → of course. also, it is true. yes. that is true.

Likewise, punctuating incorrectly can result in seriously embarrassing output, so the cost of getting it wrong can be high:

f. tienes una hija ¿no? es muy preciosa → you have a daughter right? is very beautiful
g. tienes una hija no es muy preciosa → you have a daughter is not very beautiful

So, a crucial component in an S2S pipeline is one that processes the output from the ASR (what we might call "Speech Correction"). It needs to remove disfluencies of varying sorts (e.g., ums, uhs, pauses, restarts), punctuate the input correctly, and reformat the text so that it is in the more "formal" form expected by the MT engine. And, in the context of a conversation, it needs to do it in real-time, as the person is speaking, all the while translating into the target language *as the person speaks*. It is truly a daunting task. The following diagram shows the Skype Translator S2S pipeline, including Speech Correction.[4]

[4] Notably, Kumar et al (2014), do not use "Speech Correction" component, what our team calls *TrueText*. Instead, they train their MT on parallel data consisting of noisy transcripts mapping to clean target language data. The downside of this approach is finding parallel data that is so configured.



In addition to correcting the output of ASR, MT needs to be trained on data that is less formal and more conversational so that it expectations more closely match what it is being output by the ASR engine. Most of the parallel content that is available and used to train MT engines is far too formal for the conversational context. Compare the following two excerpts, one from CALLHOME, the other from transcriptions of the European Parliament. The latter is data that is often used to train MT engines. You can see how different the two types of data are.

h. He ain't my choice. But, hey, we hated the last guy.
We're going to hit it and quit it.
Boy, that story gets better every time you hear it.
I swear to God I am done with guys like that.

i. Mr President, Commissioner, Mr Sacconi, ladies and gentlemen, as the PPE-DE's coordinator for regional policy, I want to stress that some very important points are made in this resolution.
I am therefore calling for integrated policies, all-encompassing policies that we can adapt to society, which must listen to our recommendations and comply with them.

In training the MT engines used by Skype Translator, it was necessary to find or create new sources of parallel data, specifically content that was conversational in nature. MT, however, requires that the sources be parallel, since statistical MT can only learn from the mapping of words and phrases between languages. Precious little parallel, *conversational* data exists, and that which does exist is difficult to find. Our team had to be creative in both finding and creating parallel conversational content, which itself relied on a variety of technologies.

Finally, the Speech Translation pipeline, composed of all of these technologies, needs to run in real-time. It is not possible to have bilingual conversations through a speech translator if the translator takes minutes to do its work. The speech translator must operate in real-time, translate as the person speaks, and must also operate at scale: millions of users use Skype.

So, in summary, although Speech Translation relies on the three technologies described above, namely, ASR, MT and TTS, it is not enough to blindly stitch these three components together. ASR tends to produce difficult to translate output since it is often conversational, disfluent, and noisy. Likewise, MT needs to be trained on more conversational, and less grammatical, content in order to perform better. By adding in components that more seamlessly pair each component, and creating an infrastructure that can operate in near real-time, which is then integrated into an existing (or new) VoIP tool, such as Skype, we result in a workable product.[5]

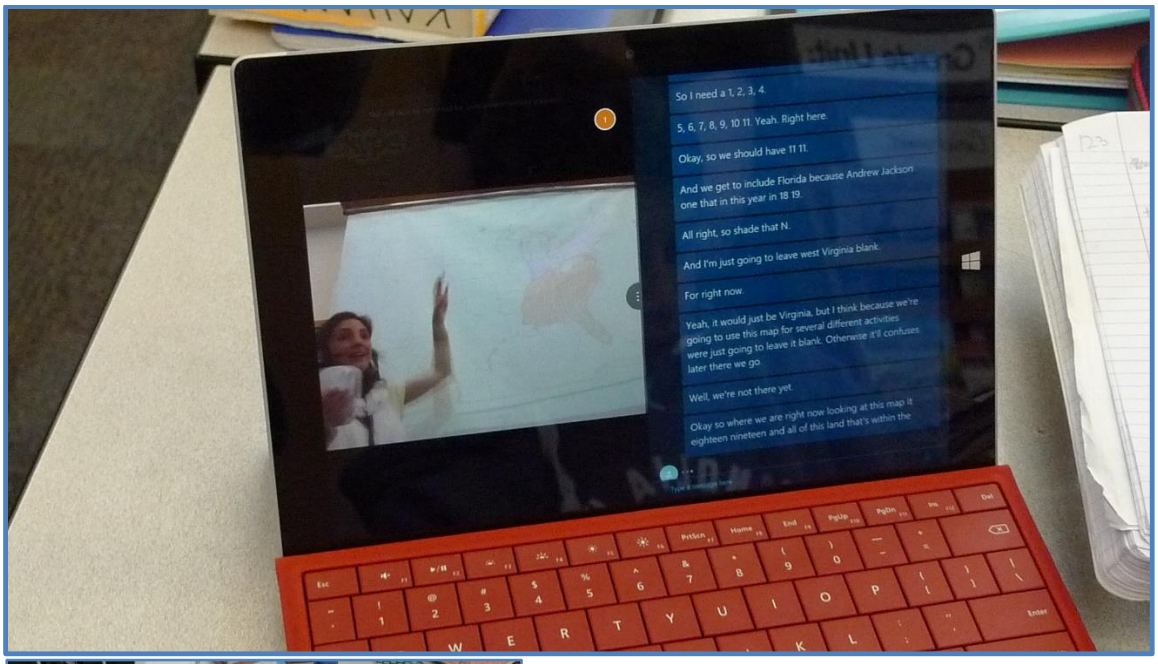
3 Breaking down the Hearing Barrier

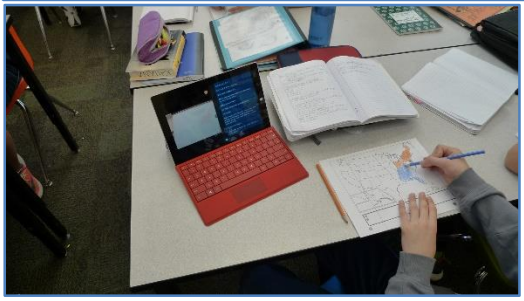
Ted Hart, a senior developer for Microsoft Research, is profoundly deaf, having lost his hearing at the age of thirteen due to the mumps. When he first started working with the earliest versions of Skype Translator, he immediately recognized the impact the technologies could have on his life. Ted doesn't make unaided phone calls. He can't. Even the simple task of making a phone call, say, to cancel a doctor's appointment or order a pizza, is not within his reach without engaging a third party. With reasonably robust speech recognition embedded in a phone client such as Skype, however, Ted can act on his own: *he* can make the call, *he* can cancel the appointment, *he* can order that pizza.

In the fall of 2014, Ted made a call to his wife on Skype. Ted was using Skype Translator, his wife, who is hearing, was running Skype on her iPhone. For Ted and his wife, this was the first unaided call they had ever had in their 18 years of marriage. The simplicity of what was discussed in that first call underlies the true benefits of the technology, and the joy that both had in even being able to have the call at all: "How's it going? Are the kids joining us for dinner? What are we having? Please stop by the store and pick up some milk on the way home." What seems so ordinary to most of us becomes extraordinary to those who are otherwise blocked from access.

So too in the schools. In the spring of 2015, Jean Rogers, Chief Audiologist and Liz Hayden, then Teacher for the Deaf, of Seattle Public Schools, started testing Skype Translator in the classroom. Their configuration was fairly simple: setup a teacher workstation with a camera at the front of the classroom, install Skype, and instrument the teacher with a Bluetooth headset linked to the computer. Then setup a tablet at a student's desk running Skype Translator, connect the two computers via a Translated call, turn off any voice recording or playback on the tablet, and voila, you have an automated captioning device. The following two pictures show a student's tablet running Skype Translator in the classroom. The picture on the top shows the video image of the front of the classroom and transcript of the lecture and discussion. Although the transcript isn't perfect—there are at least four errors—all the errors are easily surmountable, and nothing in the transcript prevents the student from understanding what is being said. The picture on the bottom shows the student at his desk, acting on the teacher's instructions and following along with all of his hearing cohorts.

[5] Not covered here is the design of the User Interface (UI) and User Experience (UX) for such a product. Questions that should be asked are: how should transcriptions and translations be displayed (e.g., in chunks, or rendered progressively), where should they be displayed (e.g., as captions, or to the side in IM), what input should users have to make corrections or to retry, how do we aid users in avoiding unproductive "loops" in conversations when insurmountable errors are encountered, etc. See Surti (2015) for an exegesis on the User Experience aspects of Speech Translation.



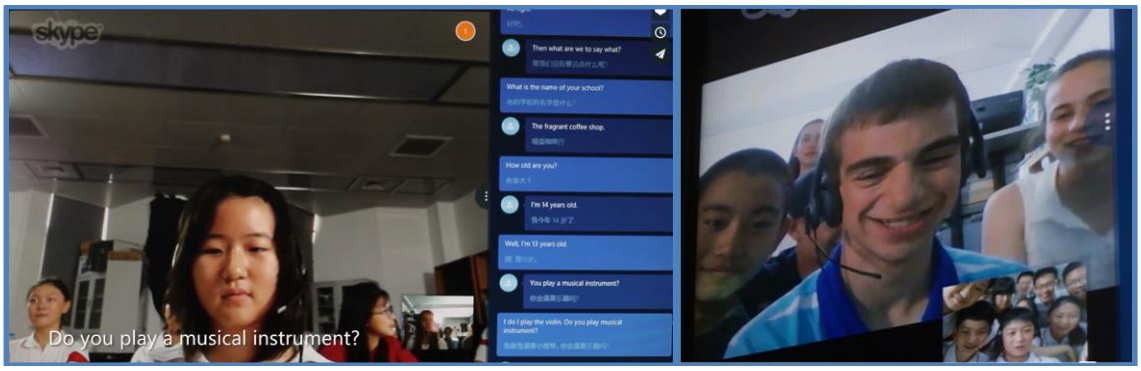


Seattle Public Schools has also been testing the use of Skype Translator in the context of *Mystery Skype*. *Mystery Skype* is a question answering and guessing game where kids learn about geography and culture of other children all over the world.[6] *Mystery Skype* is usually conducted between classrooms whose students speak the same language, e.g., English-speaking classrooms call other English-speaking classrooms. In its standard form, it is also not possible for deaf or hard of hearing kids to participate.

Speech transcription and translation opens the door to many more connection possibilities in *Mystery Skype*, since the languages being spoken are no longer a restriction, nor is the ability to hear. The relatively well known video of English-speaking children in Tacoma, Washington speaking with Spanish-speaking children in Mexico City via Skype Translator demonstrates the possibilities of the technology.[7] Seattle Public Schools extended the *Mystery Skype* engagement to include deaf and hard of hearing kids, who talked with their hearing cohorts in Beijing, China. See the pictures on the next page. The picture on the left shows the students in China who are speaking Mandarin, and the transcription and translation of the call. The picture on the right shows one of the kids who has hard of hearing who participated in the call. What one of the hard of hearing kids said says it all: "I was able to be with all of my friends and talk with someone in China who was speaking a different language than me and I

[6] For more on Mystery Skype, see the educational materials provided here: https://education.microsoft.com/connectwithothers/playmysteryskype?
[7] https://m.youtube.com/watch?v=G87pHe6mP0I

could see what they were saying on the screen so I could perfectly understand what they were telling me."[8]



4 Changing the User Experience to Support those with Deafness and Hard of Hearing

Skype Translator originally was not designed to support those with deafness and hard of hearing. It was Ted Hart's epiphany that led us down that path. Crucial to someone who does not hear are the following features. By including these features in the design, however, we not only benefited those with deafness and hard of hearing, but *all* Skype Translator users:

1. Near real-time transcripts: In the original implementations of Skype Translator, the transcripts were only displayed in chunks, after each utterance was complete. By "progressively rendering" the transcript, the non-hearing participant can see the display of the text in close to real-time. The progressive rendering change also aided hearing participants, especially when translation was engaged, since the translation itself was progressively rendered. Rather than waiting for each utterance to be completed before a translation was provided, each participant can see the transcript and translation unfold in near real-time. In user studies, we found that most preferred this.
2. Support for IM-to-speech: Speech technology is useless for those who are unable to speak or have difficulty speaking. However, if such users are able to type, enabling a "voice" for what they type gives them the ability to engage in a call over Skype with any device. Instant Messaging (IM)-to-speech in Skype Translator was added to allow those with this disability to participate, whether or not they are deaf. The IM-to-speech change also proved useful to hearing and speaking participants, specifically those who are either in a situation where they are not be able to speak (e.g., in a noisy environment where speech recognition is failing) or do not want to (e.g., in an environment where speaking may be disruptive to others, such as on a public bus).
3. Disabling speech recognition: For those users whose accent is difficult for the ASR to process, such as those with a strong deaf "accent", current speech recognition technology is ineffective and distracting. Allowing these users to disable speech recognition allows them to speak freely, without being distracted by their own transcript. Yet they still benefit from the transcript of the other user.
4. Disabling text to speech: Although not as important as 1-3, for a deaf or hard of hearing user who cannot hear the voice being uttered, turning off text-to-speech can lessen the distraction to others (it is also unnecessary for them). This feature also enabled a

[8] Quote and images from the short documentary film *Inclusive*. The film can be viewed here: https://vimeo.com/138671443.

unique feature for hearing participants who are partially bilingual. Rather than waiting for the "translated voice" of the remote user to be finished before responding, they can just read the translated transcript. If they mostly understand the other language, they can focus on those words that they do not understand in the source, and respond freely in their own language in real-time (e.g., they can interrupt and interject, as they might do in a monolingual conversation).

By enabling these features, we created a user experience that was positive for those who could not hear or had trouble hearing, and which allows them to make and participate in calls over Skype. The features aided hearing users as well. Our tests have been generally positive, both in monolingual settings—e.g., hearing users talking with deaf or hard of hearing counterparts—and bilingual settings—the same, but across spoken languages as well, e.g., English to and from Spanish, with deaf or hard of hearing users on one side or the other. Some notable vignettes from our testing: One deaf tester was troubled that the person he was speaking with kept "typing to him". Ultimately, it was made clear that what he was seeing was transcripts of the other user *talking* with him; she was not typing. Another tester was happy with the English transcript translations provided of the remote user who was speaking Spanish, and wondered how the person doing the translations could translate so quickly. It was explained to him that there was no "person in the loop". In both cases, the quality of the transcripts and translations were clearly good enough that the users were not aware they were automated. This then suggests sufficient quality to be used in real-life situations.

5 Overview and Conclusion

Although we have some ways to go to achieve fully seamless, real-time spoken translation, we see in Skype Translator the potential for real-time, open-domain, cross-lingual conversations. One can witness this in the excitement that children experience when they are first exposed to the technology and have their first translated call, when they first interact with children in some other part of the world who do not speak or understand their language. Seeing them use the technology is infective, yet at the same time, it is also incredibly touching. Intuitively and viscerally we understand that without a language barrier we can step outside ourselves, and make a connection and have a conversation with those whose world view may at first seem so much unlike ours, but, over time we realize is very much the same. At the same time, we see these technologies opening doors between communities that are differently enabled, breaking through another barrier—the hearing barrier—one that is also not so easily breached. Breaking through these barriers presents great challenges, but also promises great hope. The goal is the same: facilitating unfettered communication between our fellow human beings.

References

Kumar, Gaurav, Matt Post, Daniel Povey and Sanjeev Khudanpur (2014). "Some Insights from Translating Conversational Telephone Speech," in *Proceedings of the 39th International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*. Florence, Italy.

Olive, Joseph, Caitlin Christianson, and John McCary, Eds. (2011). *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer, Mar. 2011

Surti, Tanvi (2015). "User Experience in Skype Translator," in *Proceedings of MT Summit XV*. Miami, Florida.

Wahlster, Wolfgang (2000). *Verbmobil: Foundations of speech-to-speech translation*. Springer, Sept. 2000.