Designing User Experience for Machine Translated Conversations

Tanvi Surti tsurti@microsoft.com

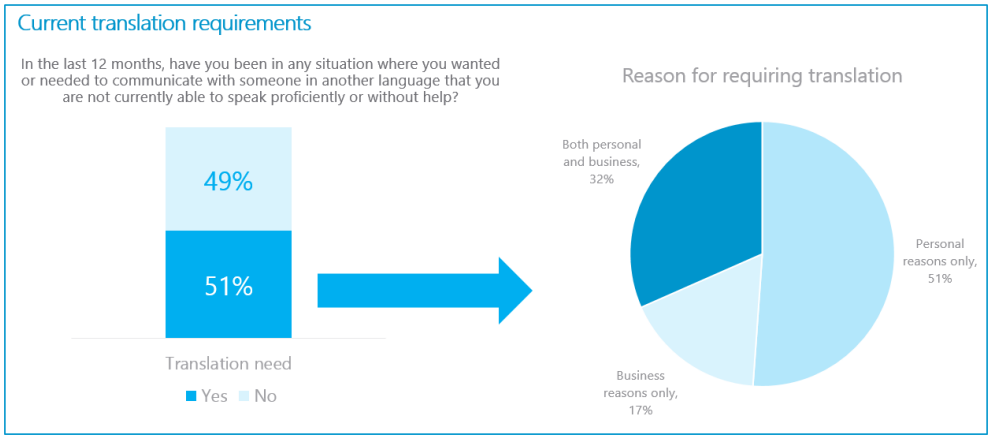
Abstract

Speech Translation technology in Skype enables users to have a live translated conversations across language barriers. From data collected from usability studies and thousands of Skype users, we've uncovered unique user experience challenges of a translated call that dissuade users from having conversations. This talk summarizes these findings and details how we iterate our designs to maintain a semblance of normalcy in translated conversations.

1. Introduction

The advent of deep neural networks in Automatic Speech Recognition (ASR) enabled researchers to reduce the word error rate in recognized speech by a third, and made it feasible to use ASR beyond the limited scope of SMS dictation, personal assistants, voice navigation, and made it applicable to the wider domain of every-day conversational speech. And by chaining together the ASR models with existing text translation, it now became possible to build automatic speech translation software for human conversations with previously unachievable accuracy.

This breakthrough in ASR resulted in the Skype Translator project, released in December 2014, enabling users to have automatic translated conversations over Skype. Skype Translator logged over 700 thousand app downloads over 9 months and clocked hundreds of hours in call time. There was clearly a need and interest in automatically-translated speech conversations, both in the personal and business sphere.



Yet, when users and reviewers tried Skype Translator for the first time, feedback for improvement was surprisingly equally concentrated around the experience of using Skype Translator as it was on the quality of translation on Skype Translator. In fact, users were more willing to forgive translation mistakes, acknowledging that speech translation was nascent technology; and were less patient with user experience issues as evidenced by the following excerpts taken from a usability study in February 2015 from first-time Skype Translator users –

"It (the call) was very chaotic."

"I zoned out waiting for the translations."

"I tried listening to the voice in the beginning, and when it wasn't working, I turned to the text."

"A bad translation is a conversation killer"

"I know that this is a monumental task and will revolutionize technology… but there isn't a flow in communication …"

"I felt like four people were speaking - two in English and two in Spanish"

Skype Translator Usability Lab, Mountain View – February 3rd 2015

2. User Experience areas of focus

Based on our usability studies and data from real-world users using Skype Translator, we identified that user experience of a translated call was a top pain-point for Skype Translator users, and over several design iterations, here are the top aspects of translated call experience we've addressed with some success -

2.1. First-Run and Learning Curve

Early usage data for the Skype Translator showed that ~40% of users had not made more than two calls on Skype Translator and that most calls on Skype Translator were under several minutes. This can be attributed to several issues such as poor translation quality and connectivity problems; but one underlying issue that emerged was that users didn't

know how to conduct a translated call. Having a translated Skype call was dissimilar to a normal Skype call, because included learnt behaviours such as waiting for the translation audio to play, remembering to pause between sentences and avoiding interruptions.

This first-run issue was addressed with two UX solutions.

Solutions

- User Education Video – all first-run users were taken through a two-minute explanation video to walk them through how to conduct a translated call on the first use of Skype Translator
- Tooltips – first-run users were given useful tips during their first translated call which provided context for how to have a successful translated call such as reminders to wear a headset.

2.2. Sensory overload

After his first call on Skype Translator, one male user study participant sat back and proclaimed - "You have to be a woman to be able to multitask in this thing…"

The sentiment he expressed referred to the multitude text and audio output the user receives during a translated call. First, there are four voices in the call – the caller's, the caller's translated voice, the callee's and the callee's translated voice. This gets cacophonous, especially when sequences of utterances are said in quick succession. Secondly, along with the audio, the user is also reading along to the translated transcript for her utterance and her partner's utterance in both languages. Many users complained that this was a lot of feedback to follow at once while trying to conduct a normal conversation.

Solution

- Audio Ducking – A technique used on radio, where if two audio clips are played at the same time, the volume is lowered on the less relevant once. Similarly, for Skype Translator, if translated audio and original audio is played at the same time, audio ducking is used to reduce the volume on the audio in the foreign language.

2.3. Perceived Translation Speed

Another frequently heard area of feedback from users was around the slowness of translation. Users felt they had to wait a long time to hear and read their partner's translated utterance and therefore made the conversation seem stretched out and awkward.

To a large extent, this delay is a *perceived* speed issue, on account of the fact that a user's translated audio could not be played until the user had completed their utterance, so as to not interrupt the user in the middle of their speech.

Solutions

- Partial recognition – Partial recognition enabled Skype to return partially understood utterances before the user had finished speaking. These "partials" are displayed in the transcript pane so that the user could follow along minimally to what their partner is saying.

- Silence interval – This advanced-user setting let users changed the value of the amount of time Skype would wait before translating their utterance. This allowed for users with a faster cadence of speech to set a low silence interval value that allowed their speech to be translated quicker.

2.4. Misrecognitions and Mistranslations

The most frequent problem users encounter during a translated conversation is misrecognitions and mistranslations. Some users see misrecognitions more than others, usually users with regional accents or children because of the lack of training data for these types of speech.

We reviewed several unsuccessful approaches to equip users to address misrecognitions and mistranslations. In the first iteration of the design, we tried to get users to cancel out wrong recognitions by clicking on a cancel button which their partner would also be able to see. In another iteration, we attempted to get users to correct the mistranslation by typing in the correct recognition instead by clicking on an Edit button. However, subsequent user studies demonstrated that users were generally unwilling to switch modalities from speaking to typing and clicking.

Solutions

- Basic user education – During the first-run setup video, users were told to repeat themselves when they were misrecognized or to rephrase their statement.
- IM prompt – Skype Translator tracked the confidence scores in the last five user utterances. If Skype Translator saw repeated low-confidence recognitions from users, the user was told that they should use the chat window to type to communicate instead. Therefore users with consistently bad recognitions were prompted towards a workaround.

3. Conclusion

Our research around Skype Translator revealed the importance of good user experience and design during a translated speech conversation. Users can be taught, over time, how best to leverage translation capabilities without expecting perfection, if the translation software sets the right context for them. Over time, users can learn to use speech translation tools in day-to-day communication along with a healthy caution to not expect perfection.