INTERSPEECH 2019: Show & Tell Contribution
September 15–19, 2019, Graz, Austria

Unbabel Talk - Human Verified Translations for Voice Instant Messaging

Luís Bernardo[1], *Mathieu Giquel*[1], *Sebastião Quintas*[2,3], *Paulo Dimas*[1], *Helena Moniz*[1,3], *Isabel Trancoso*[2,3]

[1]Unbabel, Lisboa, Portugal
[2]Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal
[3]Spoken Language Systems Laboratory, INESC-ID Lisboa, Lisboa, Portugal

luis.bernardo@unbabel.com, mathieu@unbabel.com, sebastiao.quintas@tecnico.ulisboa.pt, pdimas@unbabel.com, helena.moniz@inesc-id.com, isabel.trancoso@inesc-id.pt

Abstract

Unbabel Talk is a speech-to-speech translation application that provides human certified translations for voice instant messaging (IM) in multilingual scenarios. By combining Unbabel's translation pipeline with state-of-the-art automatic speech recognition (ASR) and text-to-speech (TTS) models, Unbabel Talk can be used to send a voice message in a language of choice through popular messaging platforms. The app further ensures that translations have high quality, either by certifying them through Unbabel's own quality estimation (QE) tool and/or through Unbabel's community of translators. There are two versions of the app. On version 1, the app synthesizes audio that can be delivered with male or female standard voices. Version 2 has features that are currently being developed, such as voice morphing and transcription correction through Unbabel's community.

Index Terms: automatic speech recognition, translation, text-to-speech, voice instant messaging

1. Introduction

Instant messaging (IM) is a type of online chat where two or more people communicate in real time over the internet. Text-based messages have been the standard choice of communication in IM for many years, but in recent years voice messaging has started to gain popularity. Voice messages can be more convenient for the sender, who can record a message during the commute to work or just to avoid typing. It is also easier to convey the correct tone and emotion associated with the message, mitigating misunderstandings of the sender's intentions. However, voice messages come with their own disadvantages, more noticeably for the recipient of the message. The recipients might need to use headphones to hear the message, either because of ambient noise or for privacy reasons, and might spend more time repeating the received messages if they do not understand them correctly the first time [1].

Nonetheless, using voice in IM has been a common phenomenon for the past years in countries such as China, where around 6.1 billion voice messages were sent through WeChat, the Chinese competitor of WhatsApp[1]. Instagram has recently added voice messaging as an option in its direct messages, in both private and group chats[2], and other apps, such as Telegram and Facebook Messenger, also have it as an option. According to Facebook, voice messages are the most popular form of shared media, after photos, in the Messenger app[1].

Unbabel Talk, an app developed by Unbabel Labs, intends to tackle the voice IM issue regarding multilingual communication. A possible use case is for customer support, where an agent must interact with clients all over the world. Unbabel Talk is a multilingual communication tool that can be used to send synthesized messages in the language of a recipient, independently of the language of the sender. The user says the message out loud to the app, which transcribes it using automatic speech recognition (ASR). If there are errors, the transcription can be easily edited by the user. The transcription is then translated through Unbabel's translation pipeline, described in the next section, which will deliver to the app the translated text with human certification. The user can then send the translated message through instant messaging apps (WhatsApp, Facebook Messenger, etc.), either in text or audio format, where the synthesized audio uses one of the standard Amazon Polly voices.

Version 1 of Unbabel Talk will be launched in the iOS app store in June with the features already described. In parallel, we are currently working on version 2. One of the features of this version is the correction by the translators of the transcribed text (the user no longer needs to edit the text if there are errors in the transcription). Another feature of version 2 that is currently under development is voice morphing that allows to synthesize the text in the user's own voice. On the following Sections we describe the Unbabel Translation Pipeline, and the two Unbabel Talk versions sequentially, in a more detailed fashion.

2. Unbabel's Translation Pipeline

Figure 1 shows a block diagram of Unbabel's text translation pipeline, in order to illustrate how to deliver certified translations. In the first stage, a user provides a text to be translated through a proper API. This text is then fed through a machine translation (MT) system, which will output the translated text. The translation is verified by Unbabel's quality estimation (QE) tool [3]. If the translation has good quality, it is delivered to the user. Otherwise, the translation is posted as a job to Unbabel's translators community. A member of this community receives both the original text and the MT translation and edits the latter. This ensures a human quality translation that is ready to be returned to the source.

Figure 1: *Diagram of Unbabel's translation pipeline.*

3. Unbabel Talk - Version 1

Unbabel Talk enables multilingual communication through Unbabel's translation services. The focus of the app falls on facilitating the transmission of information in other languages, either through text or audio. It is also fairly straightforward to use. Firstly, the user has to select two languages: a source language, chosen at the initial setup of the app, and a target language, which is easily altered through the main window. The user then presses the recording button and starts speaking. The system uses ASR to recognize the message, and presents the transcription to the user. If there are transcription errors, the user can edit them as in the case of a text message. The transcribed text is then sent to Unbabel's translation pipeline. When the translation is verified, using Unbabel's QE tool alone or with a human in the loop, it is saved in the app's main screen. One can later share the translated text or send it in audio format through the installed IM apps (WhatsApp, Telegram, WeChat, etc.). To synthesize the audio, Unbabel Talk uses an Amazon Polly generated voice. The interaction between all the modules of the Unbabel Talk system can be observed in Figure 2.

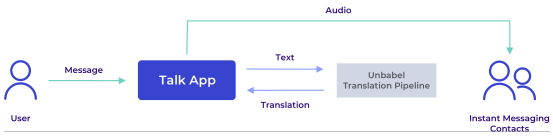

Figure 2: *Diagram of the Unbabel Talk system.*

The accompanying video shows Version 1 as it will be deployed, but omitting the time taken by human correction. Without human verification, the pipeline takes around 4 seconds. The human intervention depends heavily on the translator community for the language pair. In the best scenario, this process takes around 2 to 3 minutes.

4. Unbabel Talk - Version 2

Unbabel Talk Version 2 will include two features that are currently under development: transcription correction by Unbabel's community; voice morphing for audio message synthesis.

Transcription correction by humans. The intervention of Unbabel's community for transcription correction will occur in a similar way to what happens currently with translations. When the user requests human help, a translator receives both the audio and the transcription, which is then corrected directly on the translator's API. The process then follows the previously described procedure where the transcription is translated.

Voice morphing for audio message synthesis. In an ideal scenario for multilingual communication, all participants would be able to understand each other while still speaking in their mother tongue. Unbabel Talk version 1 already allows the user to send messages in a target language with a standard synthesized voice, which, although enough for communicating, will not come as natural as hearing the user's voice. To address this issue, we are currently working on the adaptation of both models in a Tacotron + Wavenet TTS system [4], [5], a procedure which allows the synthesis of a specific voice.

To adapt the TTS models, a minimum set of around 100 sentences in that voice is currently required, although a larger set would of course yield a better quality model. Ideally, the speech material could be obtained from the usage of the Unbabel Talk app, with the user's permission, and following Unbabel's Code of Ethics. However, these recordings would be in the mother tongue of the user and would therefore not cover a number of xenophones that only exist in the target language. The problem of cross-lingual voice morphing [6], [7] is one of the challenges of this version. Another challenge is the time interval that the Tacotron+Wavenet system currently takes to synthesize audio. Both challenges are currently being addressed by the research community.

The accompanying video shows our vision of what Version 2 would be, if both problems were solved, again omitting the time for human intervention. The synthesizer has been trained using only around 100 sentences to illustrate the quality currently achievable with a limited amount of speech material.

5. Conclusions

Unbabel Talk was developed with the purpose of enabling voice instant messaging between people that speak different languages. This can be achieved through a pipeline of ASR, MT and TTS, with in-the-loop human intervention. Some major challenges are still being addressed such as the delivery time of translations, which can take more than expected for a real time conversation when human quality must be ensured. Our vision for the future encompasses much greater challenges such as adapting the TTS voice to the characteristics of the user, in near real-time, and in a crosslingual scenario.

6. Acknowledgements

The authors would like to thank Unbabel's communities and Labs Team for all the feedback provided, with a special mention of Ricardo Araújo that edited the video, and Christine Maroti for the voice-over. This work has been supported by national funds through FCT with reference UID/CEC/50021/2019, and by PT2020 funds, under the project "Unbabel Scribe: AI-Powered Video Transcription and Subtitle" with the contract number: 038510.

7. References

[1] C. Stokel-Walker, "Voice messaging conversational gain or pain?" https://www.theguardian.com/technology/2018/dec/02/five-reasons-why-voice-messaging-is-the-next-big-thing, Dec. 2018, in The Guardian. 11 Apr. 2019.

[2] D. Lee, "Instagram is bringing voice messaging to your dms," https://www.theverge.com/2018/12/10/18134675/instagram-voice-messaging-direct-message-dm, Dec. 2018, in The Verge.11 Apr. 2019.

[3] F. Kepler, J. Trnous, M. Treviso, M. Vera, and A. F. T. Martins, "OpenKiwi: An open source framework for quality estimation," https://arxiv.org/abs/1902.08646, 2019, arXiv:1902.08646.

[4] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[5] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," *CoRR*, vol. abs/1712.05884, 2017.

[6] D. Sundermann, H. Ney, and H. Hoge, "Vtln-based cross-language voice conversion," in *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*, Nov. 2003, pp. 676–681.

[7] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. Lopez-Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *CoRR*, vol. abs/1806.04558, 2018.