Automatic Video Dubbing at AppTek

Mattia Di Gangi, Nick Rossenbach, Alejandro Pérez, Parnia Bahar
Eugen Beck, Patrick Wilken, Evgeny Matusov
AppTek GmbH
Aachen, Germany
mdigangi@apptek.com

Abstract

Automatic Video Dubbing is the process of automatically revoicing a video with a new script to make it accessible to a new audience. In this paper, we describe AppTek Dubbing, a product that will be available in Q3 2022 to automatically dub a video into a target language. We plan multiple releases of the product with incremental features, as well as the possibility to allow human intervention for increased quality.

1 Introduction

Video dubbing is the activity of revoicing a video while offering a viewing experience equivalent to the original video. The revoicing usually comes with a new script, and it should reproduce the original emotions, coherent with the body language, and be lip synchronized. Öktem et al. (2019) and Federico et al. (2020) introduced two automatic dubbing systems as a cascade of automatic speech recognition (ASR), machine translation (MT) and Text-to-Speech (TTS), enhanced with a prosodic alignment (PA) component to transfer prosody through the pipeline. In this project, we aim to build an AD system in two phases: (1) voice-over; (2) full dubbing, and enhance it with human-in-the-loop capabilities for a higher quality. The product will be released in the form of REST APIs and a web interface in Q3 of the current year. The pricing will follow a pay-per-use scheme, with possible variations according to requested quality control or if the script to dub is provided by the user for higher quality dubbing.

2 Current Features

Our current system is designed as an enhanced pipeline of ASR, MT and TTS. Our ASR system includes speaker diarization (the task of detecting "who speaks when") so that consecutive segments from the same speaker can be translated as coherent units, and each speaker is assigned a unique voice. Our MT system is a Transformer-based encoder-decoder, augmented with metadata features for style adaptation (Matusov et al., 2020) and output length control (Lakew et al., 2019). The translations are performed from and to subtitle files to preserve the timestamps and use them as boundaries for the synthesized voices. Additionally, we use speaker-adaptive TTS to reproduce the voice features of the original actor for the given segment in the new language. Finally, the background sound, obtained via source separation, is merged with the synthesized voices for the final audio and video rendering. This system can already translate video contents and dub the output videos in a voice-over style.

3 Voice-over

Voice-over is a simpler solution than dubbing, where the original voice's volume is lowered down, and the new voice is rendered with a natural volume over it, usually with a delay of some frames. Our system is already capable of performing voice-over for some language pairs[1] but some aspects can be improved:

Diarization: speaker diarization can be improved in the cases when the audio quality is low, or one speaker speaks for less than one second.

[1] see demo at https://www.apptek.com/post/automatic-dubbing-for-user-generated-content

Prosody Alignment: we plan to add prosody alignment for transferring the pauses from the source to the target speech, but also the emphasis applied to sentences and single words.

MT Output Length: although in voice-over we have time constraints less strict than in dubbing, some translations do not fit the allocated space, and it is important to have a fine-grained control over the MT output length.

4 Emotional Voice-over

The main limitation of the current system is the synthesized voice speaking with a "flat" tone, which does not match the emotions expressed in the original video. Our research effort for achieving emotional speech is aimed to release the feature in 2023 and will affect the whole pipeline:

Emotion Detection: emotions need to be detected from the source audio and matched with the recognized text, in order to annotate the latter with emotions tags.

Emotion-aware MT: Expand AppTek's MT systems to support emotions as part of their metadata. Additional research effort will focus on letting the MT system annotate the output text with emotions at a word level, to be used from our TTS system.

Emotion-aware TTS: develop TTS systems that can generate emotional speech for different emotions. Such a task can be challenging given the low data availability, particularly for languages other than English.

5 Full Dubbing

A fully-fledged AD system improves the voice-over approach by fully synchronizing audio and video time. Lip-syncing is a strict requirement that can be achieved using orthogonal technologies:

Isometric translations: improve the methods to generate translations under length constraints.

Lips motion: modify the lips' movement in the video to match the synthesized speech, building over the work described in (Furukawa et al., 2016).

6 Language Support

Our initial release will include English-to-Arabic and English-to-Spanish. In the following two years we plan to expand it to English to many European languages, including French, German, Italian, Polish and Ukrainian, plus Russian and Chinese. The reverse directions will also be rolled out soon after.

7 Human in the Loop

An AD system can make errors in multiple points of its pipeline, and the earlier the errors occur, the more harmful they can be for the final result. For this reason, we plan to let users adding manual transcripts or the final scripts to obtain a higher-quality video at the cost of more manual work, using our internal tool for easy editing parallel data.

8 Conclusion

AppTek Dubbing is an ambitious pioneering project that combines MT with other technologies to provide a high-quality and localized translated video, with the goal of making dubbing accessible beyond the movie industry. Intermediate product releases will support simpler re-voicing modes and a human-in-the-loop approach to allow the users to trade-off costs with quality.

References

Federico, M., R. Enyedi, R. Barra-Chicote, R. Giri, U. Isik, A. Krishnaswamy and H. Sawaf. 2020. From Speech-to-Speech Translation to Automatic Dubbing. *Proceedings of the 17th International Conference on Spoken Language Translation*, pp. 257–264.

Furukawa, S., T. Kato, P. Savkin and S. Morishima. 2016 Video Reshuffling: Automatic Video Dubbing without Prior Knowledge. *ACM SIGGRAPH 2016 Posters* pp. 1–2.

Lakew, S. M., M. A. Di Gangi, and M. Federico. 2019. Controlling the Output Length of Neural Machine Translation. *16th International Workshop on Spoken Language Translation.*

Matoušek, J. and J. Vít. 2012. Improving Automatic Dubbing with Subtitle Timing Optimisation Using Video Cut Detection. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2385–2388.

Matusov, E., P. Wilken, and C. Herold. 2020. Flexible Customization of a Single Neural Machine Translation System with Multi-dimensional Metadata Inputs. *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pp. 204–216

Öktem, A., M. Farrús, and A. Bonafonte. 2019. Prosodic Phrase Alignment for Machine Dubbing. *Proceedings of Interspeech 2019*, pp. 4215–4219.