Technical Disclosure Commons

Defensive Publications Series

December 14, 2018

AUTOMATIC DUBBING OF VIDEOS WITH MULTIPLE SPEAKERS

Sandro Feuz

Mohammadamin Barekatain

AUTOMATIC DUBBING OF VIDEOS WITH MULTIPLE SPEAKERS

Online availability of video content with sound is growing exponentially. Typically videos uploaded in content hosting and serving platforms have associated sound in one specific language. Hence, those videos may not be fully enjoyed by users who do not speak that language. Automatic captioning based on the original sound and textual translations of those captions may be supported by existing systems. However, textual translation sterilizes most of the contextual information (such as tonality, nuance, dialect, accent, age, gender etc.) contained in the speaker's voice, and fails to capture emotions in the speech. Also, the overall viewing experience is generally degraded as the user has to read along while watching the video. Moreover, textual translations do not effectively cater to some segments of users, for example, users with a visual impairment or users with reading disabilities (e.g., dyslexia).

An additional challenge is presented when the original audio-visual clip contains multiple speakers speaking in a language that a viewer does not understand. Existing textual translation systems do not support the use-case of multiple speakers. Existing speech synthesizers may improve user experience somewhat, because the viewer does not have to read along, but listen to the synthesized speech. However, even existing speech synthesizers remove emotive components from the audio that differentiate one speaker from another. Finally, just like existing textual translation systems, existing speech synthesizers too lose additional contextual information available in a multi-speaker audio-visual clip.

An end-to-end automatic dubbing system is proposed here that addresses many of the shortcomings described above by automatically replacing the spoken audio portion in a user-generated audio-visual clip with a translated audio stream. The automatic dubbing system proposed here synthesizes a translated version of the audio-visual clip with different voices

mirroring different speakers. The emotions and sentiments in different speakers' voices in the source video are represented as authentically as possible, along with the content of the speeches being translated from one or more source languages to one or more destination languages selected by a user, or suggested by an intelligent system.

At the core of the automatic dubbing system proposed here is an end-to-end machine learning-based dubbing model. An artificial intelligence network, such as a deep neural network, capable of running the model, receives an audio-visual clip which can be separated into a video stream, and one or more audio streams that are to be dubbed. Figure 1 depicts a flow diagram of a method showing how the artificial intelligence network, generates automatically dubbed audio-visual clips. One or more operations of Figure 1 may be performed by an automatic dubbing system operatively coupled to an online content hosting and serving platform.

The method starts at step 101, where the automatic dubbing system receives an audio-visual clip once the clip is uploaded by a user to the content hosting and serving platform.

Then, at step 102, it is determined whether the audio stream of the audio-visual clip should be converted (i.e., the audio stream is identified as a candidate for translation) into a dubbed audio stream. This can be performed by a classification system that inspects the audio-visual clip to make that determination. Classification can be performed automatically based on the content of the audio-visual clip. For example, the classification system may determine that a music video is not a good candidate for translation because the appeal of the original music video may be lost by the process of machine translation. Classification may also be performed by user action. Note that, the creator of the original audio-visual clip may have the option of enabling or disabling the dubbing feature. For example, a creator/distributor of a music video may not want fans to listen to a dubbed version of the music video, and hence, may choose to disable the

dubbing capability while uploading the music video. On the other hand, some creators of audio-visual content may prefer that their content is available in a wide variety of languages so that a larger audience can access and enjoy the content.

In step 103, in response to determining that the audio stream should be translated (determination is performed by the classification system or by the user, as described above), the machine-learning-based dubbing model is invoked for that specific language. Note that the model can support a number of languages, and when a specific language is selected, the corresponding model parameters are invoked. Alternatively, there may be a separate model for each different language or for languages of each linguistic group. The machine-learning based dubbing model takes both the video stream and the audio streams of the audio-visual clip that need to be dubbed. The model can also receive meta-data associated with the audio-visual clip (e.g., textual annotation, location data, etc.) as input.

In step 104, in case of an audio-visual clip with multiple speakers, as a pre-processing step, or alternatively, as part of the model, a set of speakers appearing in the audio-visual clip is extracted. This is performed using known face embedding and/or body embedding techniques. The extracted speakers are associated with various parameters, such as features describing the appearance (i.e. features of the face and the body), timestamps (when a speaker appears), bounding boxes (where a particular speaker appears in a frame).

In step 105, “voice embedding” techniques are used to associate dubbed audio streams to match the corresponding speakers. The model receives audio streams separated from the video stream, metadata of the audio-visual clip, as well as a list of detected speakers alongside their extracted features.

The model may have a number of functional blocks dedicated to specific subtasks to achieve the overall voice embedding function. For instance, one of the subtasks may be extracting audio masks for each speaker. Another subtask may be extracting the content of the speech given the mask and the input audio streams. Another subtask may be translating the audio, based on the content. This subtask may or may not utilize a textual representation of the content. Another subtask involves replicating the correct intonation in the translated audio (using techniques akin to text-to-speech (TTS) models).

As discussed above, the output of the machine-learning-based model performing voice-embedding is a raw audio stream which is the dubbed version of the original audio stream. The output of the model depends on the efficacy of the training method and the availability and comprehensiveness of the training data. The whole model can be trained jointly as one big model or some functional blocks performing specific subtasks can be trained independently. Professionally dubbed content (such as content from movies) may be used as training data. Also, manually dubbed content from the same multi-lingual speaker may be used as training data.

A key ability of the proposed dubbing system is preserving the “sentiment” in the original speaker’s voice, and maintaining consistency of the synthesized voice throughout the audio stream during voice embedding. Training data for voice embedding may be gathered for multiple languages, where a set of audio clips is collected, and labeled into fixed “sentiment” categories (e.g., sounding angry, sounding happy, sounding excited, etc.). The model is trained based on the labeled dataset such that the data points with the same sentiment cluster closely. A speech decoder can be “seeded” with the correct sentiment to produce a more authentic dubbed audio.

As shown in step 106, once the dubbed version of the audio stream is combined with the original video steam, a modified audio-visual clip is generated. The modified audio-visual clip is stored on the server that serves the viewer if the viewer requests the modified audio-visual clip, i.e. the dubbed version of the original audio-visual clip.

In summary, the proposed system enables users to select automatic dubbing the same way it is currently possible to generate automatic textual subtitles. In addition to content hosting platforms, this system can be used in messaging applications where videos are shared, allowing a receiving user to consume the video in their preferred language. The system can be extrapolated to other entertainment products, such as live-action movies, animations, television programs etc.

ABSTRACT

A machine-learning model that automatically converts audio streams from an audio-visual content from a source language to a destination language is described. In response to determining that an audio stream should be translated, a machine-learning-based dubbing model is invoked for a specific destination language. In case of multiple speakers, voice embedding techniques are used to match dubbed audio streams to the corresponding speakers. The sentiment in the original speaker's voice is preserved by training the model with targeted data set in the destination language.

Keywords: Machine Learning, Automatic Dubbing, Sentiment Detection, Audio, Audio-visual

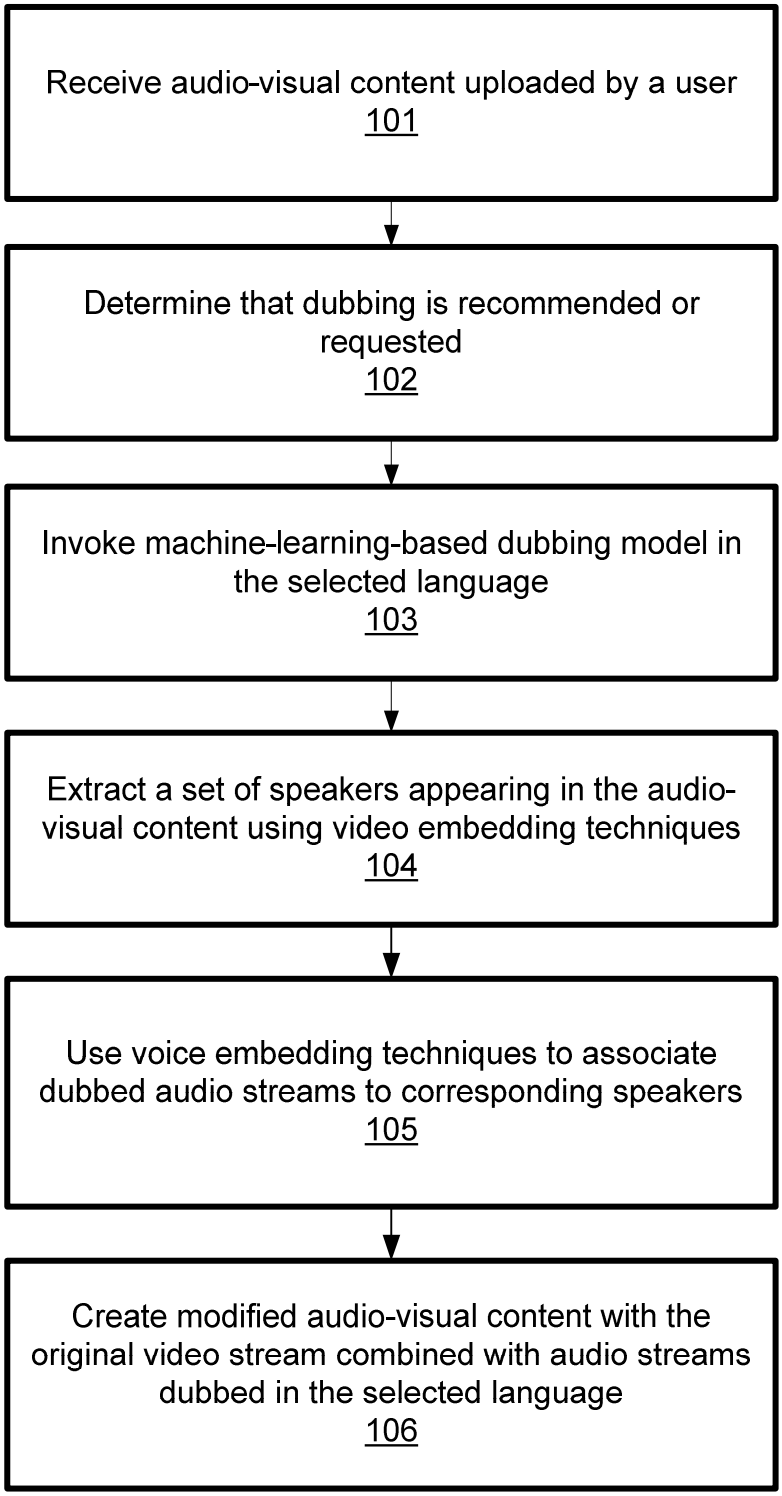
Receive audio-visual content uploaded by a user
101
Determine that dubbing is recommended or requested
102
Invoke machine-learning-based dubbing model in the selected language
103
Extract a set of speakers appearing in the audio-visual content using video embedding techniques
104
Use voice embedding techniques to associate dubbed audio streams to corresponding speakers
105
Create modified audio-visual content with the original video stream combined with audio streams dubbed in the selected language
106

FIG. 1