EC 601
Project 3
Team Product Review
Yifan Wang
Instructor: Osama Alshaykh

Multi-Speaker Identification for Streaming Services Transcription

Product Mission

The product we will work on is Streaming Multi-speaker Automatic Real-time Transcription (SMART). This product is using multi-speaker detection and speech to text transcription to achieve the automatic transcription for movies and videos and songs on any streaming services. Currently for music and movie transcription, the streaming services such as Spotify and Netflix sometimes offer the subtitles, captions or lyrics for the contents based on if the uploaders or editors uploaded the caption or not, so it is limited since what they are doing is to manually input the lyrics or subtitle files and then matching the timeframe in order to let the viewers and listeners to look at the transcript. But by using SMART, the transcription should be able to automatically generated for any content on the streaming network and thus for any content and any scenes or part of music, users can always see who are the speakers or singers and what is the caption. So, this product is for the steaming service providers to give their users the full uninterrupted experience of viewing their contents with the correct transcripts.

Current Situation

Nowadays streaming services like Netflix have higher and higher demand of closed captions and subtitles and lyrics in order to fulfill the worldwide users' need, so the requirements and standards predictably becoming higher too. Localization has become imperative to ensure that viewers are not only able to understand the content but are able to read the text naturally in their own language. A considerable amount of research has also gone into the timing and accuracy of the text to ensure readability. But the solution so far is still trying to let hosts and uploaders manually upload the caption files, and there are some companies provide high quality transcriptions of audios and videos for the contents to reach wider audiences, which is not only time-consuming no matter how fast can they transcribe and translate, but also complicated the procedure and increased the cost for transcription. For each and every content, they need to go through these extra steps and went through several pairs of hands before the caption can be ready. Transcript makes any content more accessible to users, and letting a person more a team to simply listen to each episode and write down every word spoken by the actors or hosts and guests is not the most efficient solution for it, by using multi-speaker identification, speech-to-text transcription and natural language machine translate processing, these goals can be achieve easier and can be finished in real-time without any latency between streaming and captioning.

MVP User Stories

Our product has three basic user stories. First, using multi-speaker identification to separate and identify the speakers in a certain segment from either movies/videos or songs/podcasts. By using the multi-speaker detection, it will first get the mixed sound signal with different people's voice intervals, and then based on the vocal-print feature detection, it will separate different speech signals from the mixed file and then compare it with the database to figure out who are the speakers, then use the speaker tracking to get the full interval of each speaker's speech in each scene. Second, using speech-to-text automatic speech recognition based on deep learning neural network algorithms to generate the transcription of each speaker's speech interval with speaker identity in order to provide the captions for the viewers/listeners in real time regardless of the language the speakers are using. The separated audio signals from stream will first input into the product and then by going through the natural language processing and analyzing, the caption will be generated and then will go through the speech to text functionality to get the draft of the transcript, then it will be evaluated by going through the deep learning virtual machine to be polished and compiled for the broadcast and streaming services to use. Third, based on the transcript generated, using the natural language processing and machine translation to automatically converting one natural language into another in order to achieve the multi-language multi-media multi-speaker transcription and provide users from different language background to get the content and view the accurate captions/subtitles/lyrics in real time. So the minimum viable product would be the combination of the above features including multi-speaker detection and tracking from a multi-media content, speech-to-text automatic transcription and multi-language translation of the captions in order to satisfy the early stage users to get the complete idea of SMART.