November 2022

INTELLIGENT AUTO-ACTIVATION OF CLOSED CAPTIONS DURING ONLINE MEETINGS AND CALLS FOR PRODUCTIVE DISCUSSION

Anupam Mukherjee

Vibhor Jain

Rajarathinam Chidambaram

Recommended Citation

Mukherjee, Anupam; Jain, Vibhor; and Chidambaram, Rajarathinam, "INTELLIGENT AUTO-ACTIVATION OF CLOSED CAPTIONS DURING ONLINE MEETINGS AND CALLS FOR PRODUCTIVE DISCUSSION", Technical Disclosure Commons, (November 04, 2022)
https://www.tdcommons.org/dpubs_series/5432

INTELLIGENT AUTO-ACTIVATION OF CLOSED CAPTIONS DURING ONLINE MEETINGS AND CALLS FOR PRODUCTIVE DISCUSSION

AUTHORS:
Anupam Mukherjee
Vibhor Jain
Rajarathinam Chidambaram

ABSTRACT

During a meeting or a call via a collaboration service (which brings together different capabilities such as video conferencing, online meetings, screen sharing, webinars, Web conferencing, and calling), a closed captions feature enables engaging and productive conversations for hard of hearing users and users with different levels of language proficiency. However, such a feature is by default turned off in collaboration meeting and calling services. Additionally, not all of the participants will be aware of such a feature. Techniques are presented herein that offer a unique cognitive algorithm for the intelligent automatic activation of a live captions feature for struggling participants. The algorithm may, in real time, examine various heuristic, audio, and video patterns (such as a participant's geolocation, their home language, their accent, their fluency, their facial expression and body language, and optionally the context and intent of their speech) to identify the need for closed captions for the participants. Current meeting and calling solutions do not offer a mechanism to automatically detect the need for live captions. Employing the presented techniques may help promote the use of a closed captions feature and, in turn, enable much more engaging and productive conversations during a meeting or a call.

DETAILED DESCRIPTION

As an initial matter, it will be helpful to confirm the meaning of an element of nomenclature. The discussion below makes reference to an online communication and collaboration service. Such a service, which for simplicity of exposition may be referred to herein as a collaboration service, brings together different capabilities such as video conferencing, online meetings, screen sharing, webinars, Web conferencing, and calling.

Within such a collaboration service, a live captions feature facilitates engaging and productive conversations for users with different levels of language proficiency during a one-to-one, three-way, or n-way meeting or call. Additionally, such a feature makes the experience during a meeting or call more accessible and inclusive for users who may be hard of hearing.

However, such a closed captions feature is typically turned off by default in both the meeting and the calling services of a collaboration service (if the feature is enabled at an organization level, at a location level, or at a user level). In a meeting there may be mix of participants from, for example, the United States, the European Union, India, China, Japan, etc. Similarly, any person from any geographic region may make a call to another person in another region or country. Due to different levels of language proficiency, or insecurity with a non-native language, the communication frequently becomes unproductive. Sometimes a speaker's accent, their use of vocabulary, their intonation, and their voice modulation also make the conversation difficult to understand.

Not all of the participants will be aware of a closed captions feature while joining a call or meeting. Hence, most of the time many participants do not use such a feature in spite of facing difficulty in understanding the speech of other participants or the context itself and thus remain less interactive during the discussion. Frequently, a participant will request that the speaker repeat a phrase, or even hesitantly express their inability to understand the conversation, which in turn affects the momentum of the discussion. Thus, the original purpose of a live captions feature is defeated.

Currently, there is no intelligent mechanism available in any meeting or calling solution that can automatically turn on a live captions feature and translate according to the home language settings of a participant (or recommend a live captions feature to the participant) just by sensing the participant's inability to understand the conversation and, optionally, their hesitation to express the limitation.

To address the challenge that was described above, techniques are presented herein that may intelligently turn on closed captions for a struggling participant provided that the closed captions feature is enabled by an administrator at an organization level, a location level, or at a user level. Aspects of the presented techniques employ a unique algorithm comprising heuristic techniques in tandem with audio and/or video intelligence to detect the need for a closed captions feature in the most cognitive fashion. According to further aspects of the presented techniques, settings may be exposed at the user level to turn the above-described algorithm on or off.

As noted previously, aspects of the techniques presented herein encompass an algorithm that may employ heuristic observations as well as artificial intelligence (AI) technologies to process both audio and (if a camera is on) video streams to identify a need for a closed captions feature for one or more participants. Such an algorithm may run centrally in the meeting and calling servers. Based on a computed “Need-Caption Classification" result, the algorithm may instruct the specific endpoint of a struggling participant to turn on (or show a recommendation for) live captions. Alternatively, under a lite version of aspects of the presented techniques the calling or meeting clients (i.e., endpoints) may be equipped with the algorithm which may process the video and audio inputs locally using lightweight AI models and act according to a computed "Need-Caption Classification" result.

In the following detailed discussion of the algorithm that was introduced above, it is important to note that the terms "meeting" and "calling" will be used interchangeably. Additionally, a live captions feature needs to be in a turned-off mode for the invocation of the AI-based algorithm.

According to aspects of the techniques presented herein, the above-referenced algorithm may track a participant’s information along with the audio and (if video is turned on) video stream of a meeting, at a participant level, to extract multiple attributes (such as a participant’s geolocation, their home language, their accent, their fluency, their facial expression and body language, and optionally the context and intent of their speech) and then analyze those attributes in real time to identify the need for closed captions for the participant. Among other things, the algorithm may examine various heuristic, audio, and video patterns to arrive at a conclusion.

Figure 1, below, depicts elements of an exemplary solution architecture that is possible according to aspects of the techniques presented herein.

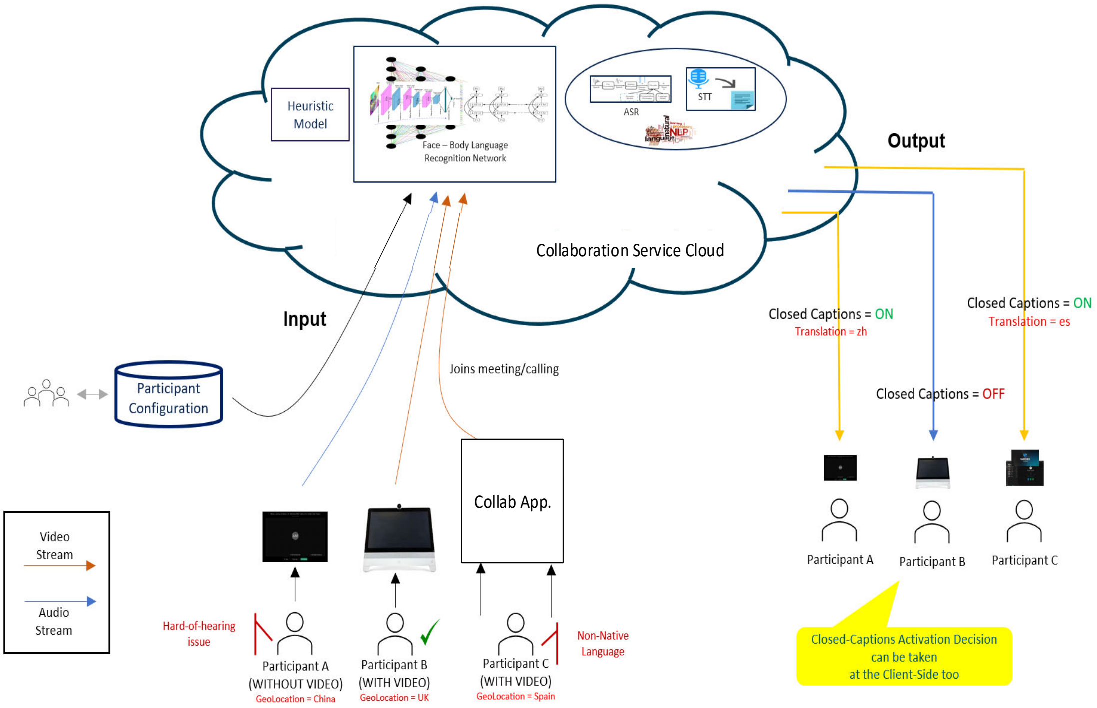

Figure 1: Exemplary Solution Architecture

Figure 2, below, illustrates elements of an exemplary control flow that is possible according to aspects of the techniques presented herein.

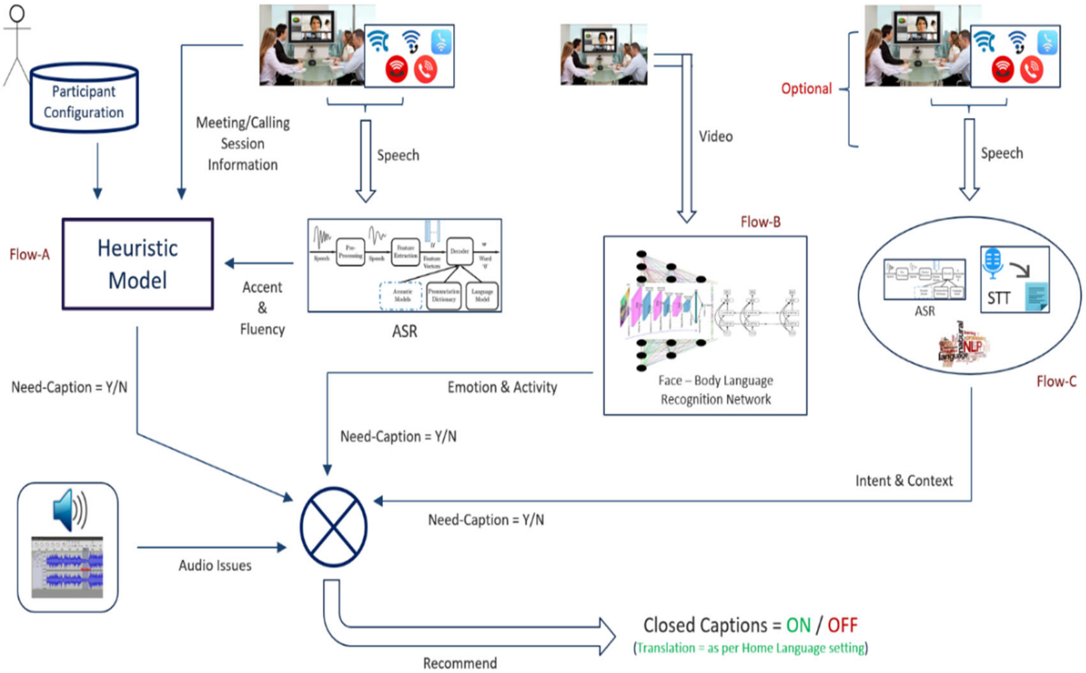

Figure 2: Exemplary Control Flow

As depicted in Figure 2, above, aspects of the techniques presented herein may employ multiple approaches to detect the need for closed captions during a meeting.

A first approach encompasses using configuration heuristics and, from the audio stream, a participant's accent and fluency in a non-native language. Since the collaboration meeting service or calling solution is aware of a participant’s country of origin (even state and county information as well), if there is a mix of participants from different geographic regions or countries in a meeting or calling session the algorithm may use such a heuristic pattern as one of the triggers in its decision tree graph.

The algorithm may also keep track of the native or home language settings of the different participants. Consequently, during a meeting if there are participants from different geographic regions having various home language settings which are mostly different from the official language that is used in the meeting, the algorithm may incline towards turning on (or recommending) closed captions.

During this phase, the algorithm may use pre-trained automatic speech recognition (ASR) systems to assess, from the conversation (i.e., an audio stream) using speaker diarization, the accent and the fluency of the participants in an official meeting language before turning on (or recommending) closed captions. Additionally, the algorithm may only check the fluency for those participants who have joined the meeting from a geographic region having different home or native language settings or who have a non-native accent in the official meeting language. If the fluency score of a participant is below a threshold value, the algorithm may turn on an activation flag for closed captions (or a recommendation of the same).

A second approach encompasses using a video stream to track the facial expression and the body language of a participant. Using readily available face and activity recognition software (or a homegrown long-term recurrent convolutional network (LRCN), popularly known as a convolutional neural network (CNN)-long short-term memory (LSTM) model, typed pre-trained models) the algorithm may process a video stream of the participants provided that the video is turned on (which happens mostly in collaboration meetings, during collaboration calling most of the calls are audio only calls).

AI models may be used by the algorithm to extract the emotion (i.e., a facial expression) and the activity (i.e., the body language) of the participants in a meeting from the captured video stream and then classify those attributes for the recognition of the right intent. If the facial expression of a participant indicates that he or she does not understand the conversation (or the discussion), and hence is confused, the models are able to recognize it and classify accordingly. Additionally, the models are also capable of tracking hesitation from the facial expression or the body language of a participant. Similarly, if a participant leans towards the microphone to hear the sentences clearly, the models are able to detect that body language as well and classify it as "inability to understand the conversation."

Consequently, if the video stream analyzer models identify that a participant is confused or not able to clearly grasp or hear the context (and hesitating, as well) from the facial expression and the body language of the participant, that may serve as a major trigger in the algorithm for the activation of closed captions (or the recommendation of the same). The algorithm may turn on an activation flag accordingly.

A third approach encompasses an optional extension that uses an audio stream to verify the intent and the context of the speech of a participant. Many existing solutions employ AI or non-AI based approaches to analyze the context of a discussion from the audio stream of a meeting. Using one of those existing solutions, the algorithm may track the context of the overall conversation from the audio stream.

The algorithm may monitor each participant's speech using available speech-to-text (STT) application programming interfaces (APIs), intent analysis, and context analysis software (or homegrown Bidirectional Encoder Representations from Transformers (BERT)-based natural language processing (NLP) models) to identify a participant's intent and context.

If the algorithm detects any verbal statement that expresses confusion (such as "sorry, I am unable to understand what you are discussing" or a similar type of statement), that indicates an inability due to a lack of proficiency in the official meeting language (such as "sorry, I cannot speak fluently" or a similar type of statement), that is full of grammatical mistakes (using any of the available grammar checking tools), of that contains a totally out of context statement (which may be in response to a question that was asked by the other participant(s) or in general be based on the context of the overall discussion) it may turn on the activation flag in favor of closed captions (or the recommendation of the same).

It is important to note that the algorithm may also simultaneously track both the audio volume and the audio quality for all of the participants. If the audio quality degrades due to jitter, packet loss, or latency (such as audio breakup) or if the volume is low at a participant's side (which may be tracked from the respective endpoint settings), the algorithm may ignore the findings from the audio and video stream and continue the monitoring.

The three above-described flows may be referred to as Flow-A (the first approach), Flow-B (the second approach), and Flow-C (the third approach encompassing an optional extension). Based on the activation flags that are received from Flow-A, Flow-B, and Flow-C, the algorithm may either recommend that the user turn on closed captions or continue the monitoring. At that moment, a user may or may not follow the suggestion. Whatever action the user does take, that may terminate the algorithm for the current session. However, user interface (UI) settings may be exposed in a client application to turn the algorithm on again in the same session.

According to aspects of the techniques presented herein, the above-described algorithm may be further augmented by introducing the intelligence of a translation suggestion as the conclusive step which may operate as an extension to Flow-A. As explained above in connection with Flow-A, a collaboration meeting or calling solution is aware of a participant’s country of origin (even state and county information as well) and can also map the corresponding home language for each participant. If an activation flag is turned on in Flow-A, the algorithm may suggest the translation of a transcription (i.e., closed captions according to the appropriate home language mapping of the participant).

It is important to note that, according to aspects of the techniques presented herein, the STT engine that was described and illustrated in the above narrative needs to be localized to consider country- and region-specific languages, accents, and dialects. Otherwise, an STT capability may yield an inaccurate transcription which, in turn, may defeat the purpose of a closed captions feature.

The application of the techniques presented herein offers a number of benefits, several of which will be briefly described below.

First, the presented techniques offer a unique cognitive algorithm for the intelligent automatic activation of a live captions feature which otherwise is one of the least used features in any meeting solution. Most of the participants with different levels of language proficiency or hard of hearing users do not use such a feature even if they face difficulty in understanding the speech of other participants or the context itself. Thus, the discussion becomes unproductive, and the original purpose of a live captions feature is defeated. Aspects of the presented techniques may help to promote the use of a closed captions feature and, in turn, enable much more engaging and productive conversations for users during a one-to-one, three-way, or n-way meeting or call

Second, the algorithm (that was mentioned above and which was previously described in the above narrative) examines, in real time, various heuristic, audio, and video patterns (such as a participant's geolocation, their home language, their accent, their fluency, their facial expression and body language, and optionally the context and intent of their speech) to identify the need for closed captions for the participant. Such a capability is quite unique in nature. No other meeting or calling solution uses such a mechanism to automatically detect the need for live captions. Rather, such solutions only expose a manual option to turn on closed captions.

In summary, techniques have been presented herein that offer a unique cognitive algorithm for the intelligent automatic activation of a live captions feature for struggling participants. The algorithm may, in real time, examine various heuristic, audio, and video patterns (such as a participant's geolocation, their home language, their accent, their fluency, their facial expression and body language, and optionally the context and intent of their speech) to identify the need for closed captions for the participants. Current meeting and calling solutions do not offer a mechanism to automatically detect the need for live captions. Employing the presented techniques may help promote the use of a closed captions feature and, in turn, enable much more engaging and productive conversations during a meeting or a call.