

SPEECH-TO-TEXT GENERATION USING VIDEO-SPEECH MATCHING FROM A PRIMARY SPEAKER

BACKGROUND

Field of the Technology

[001] The present disclosure relates generally to speech-to-text generation employing pattern recognition, speech separation, video-speech mapping, lip reading, and natural language processing principles.

Brief description of the Prior Art

[002] Digital assistant systems have become quite pervasive in today's hi-tech world. They may be in the form of software running on mobile cell phones or browser-based website applications, or they may be in the form of in-home devices that look like speakers that provide users with a way to ask questions and receive information through voice inputs. For example, interested readers may review US Patent 9,548,050, titled "Intelligent Automated Assistant," and US Patent 6,584,439, titled "Method and Apparatus for Controlling Voice Controlled Device." These digital assistant systems primarily rely on a vocal input from a user and speech-to-text generation technologies to receive and translate the acoustic vocal input into machine-readable text that can then be intelligently interpreted. The speech-to-text generation for these digital assistant systems is, however, often inaccurate.

[003] Speech-to-text generation, also known as speech recognition or speech processing, is the process of converting an acoustic signal, captured by a microphone or a telephone, to a set of words. Currently available speech-to-text generator systems may be in the form of software running on mobile cell phones or browser-based website applications that provide users with a way to speak into a microphone or telephone and receive a visual output of text representative of the input speech on some form of display screen. For example, interested readers may review US Patent No. 6,173,259, titled "Speech to text conversion", and US Patent No. 6,064,957, titled "Improving speech recognition through text-based linguistic post-processing".

[004] In recent years, there have been many improvements to speech-to-text generators to improve transcription accuracy. In particular, US Patent Application No. 12/628,514, issued as (hereinafter “Kim”) describes a combined lip reading and voice recognition multimodal interface system that allows a driver to speak to a digital assistant and receive directions from the digital assistant while driving a car. This disclosure requires a vocal input as well as a lip video image input from a user to provide a more accurate representation of what the user was trying to say. Unfortunately, Kim is compatible with only one user to provide the vocal and lip video image input; in other words, Kim’s voice recognition or speech-to-text generator system would not work properly if more than one user was speaking at the same time.

[005] The problem of separating speech samples from multiple users in a single acoustic signal was addressed in US Patent No. 7,664,643 (hereinafter “Gopinath”), which discloses a system and method for speech separation and multi-talker speech recognition. Although Gopinath discloses a system for speech separation of a signal that contains the simultaneous speech of multiple speakers, Gopinath does not disclose a system or method for improving the speech recognition or speech-to-text generation for one of the separated speech samples. Further Gopinath does not disclose a system or method for identifying a primary speaker for a digital assistant system when multiple people are speaking simultaneously nor does he use a video image input to associate a speech sample with a speaker.

[006] Person identification using video-speech matching was disclosed in U.S. Patent Application 10/076,194, published as US 2003/0154084 (hereinafter “Li”). Li essentially disclosed a system that segmented video and audio features from video data to then correlate the segmented video and audio features. Li, however, failed to disclose a system for identifying a primary speaker and his or her vocal input from video data. U.S. Patent Application No. 14/036,728, published as US 2015/0088515 (hereinafter “Beaumont”), however, discloses a method for identifying a primary speaker from a visual sensor and one or more microphones in an information handling device when multiple speakers are speaking concurrently. Both Li and Beaumont, however, are not concerned with using a segmented video image of a speaker to determine the content of a segmented speech sample, but rather they are concerned with using the segmented video image data to determine the identity of the speaker providing the segmented

speech sample.

[007] In summary, no method currently exists that generates text from speech using video-speech matching from a primary speaker. What is needed is a new and improved method for generating text from speech from a primary speaker when multiple speakers are speaking simultaneously

SUMMARY

[008] Disclosed herein is a method for improving the accuracy of speech-to-text generation by using video-speech matching from a primary speaker. In one embodiment, the method comprises receiving a video image input and receiving an audio input. The method further provides segmenting the video image input into a plurality of video features and segmenting the audio input into a plurality of audio features. The method further provides matching the plurality of video features to the plurality of audio features, selecting a primary speaker from one of the matched video and audio features, and generating text from the matched video feature and audio feature of the primary speaker.

BRIEF DESCRIPTION OF THE DRAWINGS

[009] The following description and associated figures teach illustrative embodiments of the disclosure. For the purpose of teaching inventive principles, some conventional aspects of the embodiments can be simplified or omitted. The following claims specify the scope of the disclosure. Note that some aspects of the embodiments cannot fall within the scope of the disclosure as specified by the claims. Thus, those skilled in the art will appreciate variations from the claimed embodiments that fall within the scope of the disclosure. Those skilled in the art will appreciate that the features described below can be combined in various ways to form multiple variations of the disclosure. As a result, the disclosure is not limited to the specific examples described below, but only by the claims and their equivalents.

[0010] Figure 1 is a block diagram illustrating the general functional components of a method of speech-to-text generation using video-speech matching from a primary speaker.

[0011] Figure 2 is a flow diagram illustrating a method of speech-to-text generation using video-speech matching from a primary speaker.

DETAILED DESCRIPTION OF THE EXEMPLARY EMBODIMENTS

[0012] Embodiments of the present disclosure will now be described in detail with reference to the drawings, which are provided as illustrative examples of the disclosure so as to enable those skilled in the art to practice the disclosure. Notably, the figures and examples below are not meant to limit the scope of the present disclosure to a single embodiment, but other embodiments are possible by way of interchange of some or all of the described or illustrated elements.

[0013] Moreover, where certain elements of the present disclosure can be partially or fully implemented using known components, only those portions of such known components that are necessary for an understanding of the present disclosure will be described, and detailed descriptions of other portions of such known components will be omitted so as not to obscure the disclosure. In the present specification, an embodiment showing a singular component should not be considered limiting; rather, the disclosure is intended to encompass other embodiments including a plurality of the same component or nested stages, and vice-versa, unless explicitly stated otherwise herein. Moreover, applicants do not intend for any term in the specification or claims to be ascribed an uncommon or special meaning unless explicitly set forth as such. Further, the present disclosure encompasses present and future known equivalents to the known components referred to herein by way of illustration.

[0014] In general, the present application is designed to improve speech-to-text generation accuracy using video-speech matching. The present application also facilitates a user's ability to speak into a microphone or telephone and have their recorded speech accurately transcribed when other people are present and speaking concurrently into the same microphone. The present application further uses video and audio segmentation processing to determine a primary speaker speaking into a microphone. The present application also uses lip reading software to more accurately generate text representative of recorded speech. The hardware used in this disclosure are video and audio recording devices, and can be deployed in almost any location. There is no requirement for additional hardware from the user to interact with the video and audio recording devices and have their speech translated into text.

[0015] The present disclosure is motivated by the need to provide a more accurate textual representation or transcription of a user's speech input in a noisy environment. The resulting method is a completely new disclosure that does not exist anywhere, or in any form. Some of the key software programs include pattern recognition, speech separation, video-speech mapping, lip reading, and natural language processing. The new method requires receiving a video image input, receiving an audio input, segmenting the video image input into a plurality of video features, segmenting the audio input into a plurality of audio features, matching the plurality of video features to the plurality of audio features, selecting a primary speaker from one of the matched video and audio features, and generating text from the matched video feature and audio feature of the primary speaker.

[0016] The new proposed method improves upon every one of these existing methods in a variety of different ways. The new method uses concurrent video and audio recordings of the speech of more than one user to first segment the video recordings into video features and segment the audio recordings into audio features, each feature representative of a human speaker. The new method then matches the video and audio features of the human speakers to associate a certain speech with an image, such as a user's face or lip movements. The new method further processes the video and audio feature matches to determine a primary speaker. The new method last uses the matched video and audio features of the primary speaker to then generate a text corresponding to the primary speaker's speech input.

[0017] As described above, the key components required to successfully create the method of the disclosure include, but are not limited to, pattern recognition, speech separation, video-speech matching, lip reading, and natural language processing.

[0018] Pattern recognition generally refers to a type of machine learning that focuses on the recognition of patterns and regularities in data. Pattern recognition is also the assignment of a label or category to a given input value. Pattern recognition systems are first trained by inputting multiple training patterns and forming categories based on the received input training patterns. After training, the pattern recognition system takes input data and compares it to categories and, based on the comparison, identifies the data as belonging to one of the categories learned from

training. In the present disclosure, pattern recognition is used to classify video image data into categories (or “video features”) of speakers. For instance, pattern recognition is used to identify how many speakers are in a given image, to train the pattern recognition system the visual aspects of each identified speaker (such as the speaker’s face), and to classify successive video image data into a plurality of classes of speakers.

[0019] Speech separation generally refers to a process of receiving a single acoustic signal containing human speech from two or more people and modelling the signal in such a way that the various frequencies contained in the acoustic signal can be separated or filtered to render separate acoustic signals representative of one person’s speech. Single channel speech separation has been attempted using Gaussian mixture models as well as using the modeling techniques discussed in “Gopinath” above. In the present disclosure, speech separation is used to separate a received acoustic signal containing the speech of two or more people into two or more acoustic signals, each containing the speech of a different person.

[0020] Video-speech matching is a process that matches video data with corresponding audio data. Certain applications of video-speech matching, such as face-speech matching, use low-level audio and visual features to associate faces with speech. For instance, lip movement detection and analysis in video data can be used to attempt to match the video data with audio data, as demonstrated in “Li” referenced above. In the present disclosure, video-speech matching is used to associate the faces or lip movements of speakers with the separated speech samples of each identified speaker.

[0021] Lip reading is the visual detection and analysis of the movement of a speaker’s lips to determine what the speaker is saying. Some digital assistant systems integrate lip reading systems with audio-based voice recognition systems to process verbal commands from a user, as demonstrated in “Kim” discussed above. In the present disclosure, lip reading software is used, along with natural language processing from a received audio input, to determine the content of a primary speaker’s speech.

[0022] Natural language processing is utilized to record incoming speech from users and then

convert the incoming speech into text. Natural language processing is currently used by several major technology corporations for a variety of reasons. One major commercial application is for converting speech into text to be displayed on some form of screen. This facilitates word processing for users what cannot type or use a keyboard efficiently. These common applications are generally accomplished through the use of microphones and digital assistant software that is embedded into mobile cell phones and into pieces of hardware that look like speakers. Instances of natural language processors include IBM's Watson, the Google Cloud Natural API, Window's Cortana, Apple's Siri, Intel's RealSense, and Dragon's Speech Recognition Software, among others. In the present disclosure, natural language processing is used to generate text from a primary speaker's recorded speech.

[0023] As a first example employing the method of speech-to-text generation using video-speech matching from a primary speaker, Figure 1 is presented. Figure 1 is a block diagram illustrating the general functional components of a method of speech-to-text generation using video-speech matching from a primary speaker. In the non-limiting embodiment, the components of the method 100 include a speech-to-text generator system 101, an audio capture device 102, a video capture device 103, an audio segmentation module 104, a video segmentation module 105, an audio and video matching module 106, a primary speaker selection module 107, and a text generator module 108.

[0024] The speech-to-text generator system 101 may be one or more computing devices configured to perform the speech-to-text generation method. The speech-to-text generator system 101 may include a plurality of processing systems that can be implemented within a single processing device or can also be distributed across multiple processing devices or sub-systems that cooperate in executing program instructions, such as cloud computing systems. The audio capture device 102 may be in the form of a microphone embedded on a computer monitor or a telephone. The audio capture device 102 is preferably configured to capture human speech. The audio capture device 102 may also comprise an array of microphones. The video capture device 103 may be a visual sensor, e.g. a camera, an IR sensor, or even an acoustic sensor utilized to form image data. For ease of illustration and description, the audio capture device 102 and the video capture device 103 are described separately in this description; however, this is only by way of example.

[0025] The audio segmentation module 104 and the video segmentation module 105 are one or more computing devices that are configured to perform audio and video segmentation processing, respectively. The audio segmentation module 104, more specifically, comprises one or more processors and memory to receive audio data from the audio capture device 102 and segment the received audio data into audio features, the audio features comprising audio data from different speakers. In a preferred embodiment, the segmented audio data is an audio recording of a speaker's speech. The video segmentation module 105, more specifically, comprises one or more processors and memory to receive video data from the video capture device 103 and segment the received video data into video features, the video features comprising video data from different speakers. In a preferred embodiment, the segmented video data is a video recording of a speaker's face, lip movements, mouth movements, body movements, among other visual aspects.

[0026] The audio and video matching module 106 is one or more computing device configured to match the segmented audio features from the audio segmentation module 104 with the video features of the video segmentation module 105. The audio and video matching module 106 may assign confidence scores to various combinations of audio and video features and then "match" the combined audio and video features with the highest confidence scores. It should be understood that other techniques may be used to draw similarities between the plurality of audio and video features. For instance, the plurality of audio features can be first processed by a computing device using speech-to-text generator software to generate a text representative of the plurality of audio features. The plurality of video features, including the plurality of speakers' mouth movements, can be then processed by a computing device using lip reading software to generate a text representative of the spoken utterances contained in the plurality of video features. The texts generated from the plurality of audio and video features can then be matched for similarity to associate the plurality of audio features with the plurality of video features.

[0027] The primary speaker selection module 107 is one or more computing devices configured to select a "match" of the plurality of audio and video features that is most likely the primary speaker. For example, a primary speaker may be a person who comprises the largest number of pixels in a given video feature, and is therefore the person closest to the video capture device 103. Alternatively, the primary speaker may be the speaker whose spoken decibel level is the highest

among other audio features. Further, the primary speaker may be one who is using certain a vocabulary typically used in a retail environment, such as “I would like to buy” or “May I please have”. It should be understood that the determination of a primary speaker is not limited to the above examples, and the operator of the method of speech-to-text generation is free to determine what characteristics define a primary speaker. In a preferred embodiment, there is only one primary speaker.

[0028] Last, the text generator module 108 is one or more computing devices configured to generate text from the audio and video features of the determined primary speaker. In operation, the speech to text generator system 101 receives concurrent audio and video data from the audio capture device 102 and the video capture device 103, segments the audio data into a plurality of audio features using the audio segmentation module 104, and segments the video data into a plurality of video features using the video segmentation module 105. The audio and video matching module 106 then matches the pluralities of audio features with the plurality of video features. The primary speaker selection module 107 then determines a primary speaker from one of the plurality of audio and video feature matches. Last, the text generator module 108 generates text representative of the matched audio and video feature of the primary speaker. For example, the text generator module 108 may use the video feature of the primary speaker to perform a lip-reading analysis and generate text representative of the movement of the primary speaker’s lips. Further, the text generator module 108 may perform natural language processing on the audio feature of the primary speaker to generate text representative of the primary speaker’s recorded speech. The text generator module 108 may be configured to use only the primary speaker’s audio feature, video feature, or both, to generate a text representative of the primary speaker’s speech. In a preferred embodiment, the text generator module 108 uses both the audio and video feature of the primary speaker to determine the most accurate text representative of the primary speaker’s speech.

[0029] It should be understood that the general functional components of a method of speech-to-text generation using video-speech matching from a primary speaker, as presented in Figure 1, are connected by a variety of communication links, wired or wireless. These links are, however, omitted for clarity.

[0030] To further illustrate the method of the disclosure, Figure 2 is presented. Figure 2 is a flow diagram 200 illustrating a method of speech-to-text generation using video-speech matching from a primary speaker. The method provides first receiving a video input (201). The video input can be of any format, examples of which include but are not limited to: MP3, MP4, AVI, MKV, MOV etc. The method then provides receiving an audio input (202), preferably through a microphone. The method then, at the video segmentation module 105, segments the received video input into a plurality of video features (203). The method then, at the audio segmentation module 104, segments the received audio input into a plurality of audio features (204). The method further provides, at the audio and video matching module 106, matching the plurality of video features to the plurality of audio features (205). The method then provides, at the primary speaker selection module 107, selecting a primary speaker from one of the matched video and audio features (206). The method then concludes, at the text generator module 108, by generating text from the matched video feature and audio feature of the primary speaker (207).

[0031] Persons skilled in the art would understand that the steps of flowchart 200 are exemplary. Additional steps may be added, sequence of some of the steps may be altered, and/or some steps may be omitted based on particular embodiments.

[0032] There are numerous advantages to the method discusses in this disclosure, the overarching advantage being the ability to generate, from one or more computing devices, an accurate transcription of a primary speaker's speech or utterance when more than one person is speaking at or around the one or more computing devices. In a preferred embodiment, the primary speaker's speech can be processed in real-time or as a recording contained in an audio and video file. The method allows speech to text generators to work properly and accurately in noisy environments where one primary speaker is verbally requesting something.

[0033] One application of the method of the disclosure would be to produce text on a word processor from a received speech input in a noisy environment. Using both an audio recording and a concurrent video image to accurately isolate and transcribe speech will effectively improve the accuracy of the transcription. Another application of the method of the disclosure would be to generate text from speech in a retail sales environment, office environment, or other location that

requires services through a verbal exchange (ex. shopping mall directions). For instance, a real-time system employing the method of the disclosure may be used to facilitate a verbal exchange between two people where one person is hearing-impaired and cannot hear what the other person is trying to say. Some specific non-limiting illustrative use cases are described below.

[0034] Use case 1: An elderly man is not very skilled at typing and would like to write a letter to his grandson. The elderly person will open a word processor and speak into a microphone. The word processor is communicatively coupled to a speech-to-text generator system that uses both a camera and a microphone to produce an accurate transcription of the elderly man's speech. As the elderly man speaks into the microphone and towards the camera, his wife approaches from behind and starts asking him something. The speech-to-text generator system identifies the two speakers, segments the old man and his wife's respective audio and video channels, matches the audio and video channels for the old man and his wife, determines that the old man is the primary speaker, and generates text representative of the old man's speech that is immediately displayed on the word processor in front of the old man.

[0035] Use case 2: A girl wishes to place a fast food order at a digital kiosk. The digital kiosk comprises a camera, a microphone, and one or more computing devices that perform the speech to text generation method of the disclosure. As the girl verbally places her order, a loud man is talking on his cell phone in line behind her. The speech-to-text generator system contained in the digital kiosk identifies the two speakers, segments the girl's and the loud man's respective audio and video channels, matches the audio and video channels for the girl and the loud man, determines that the girl is the primary speaker, and generates text representative of the girl's speech. The digital kiosk then takes the generated text and applies an artificial intelligence program to process and respond to the generated text.

[0036] The disclosure has been described in terms of particular embodiments. Other embodiments are within the scope of the following claims. For example, the steps of the disclosure can be performed in a different order and still achieve desirable results.

[0037] Insofar as the description above and the accompanying drawings disclose any additional

subject matter that is not within the scope of the claims below, the portions of the disclosure are not dedicated to the public and the right to file one or more applications to claim such additional embodiments is reserved.

[0038] It should be recognized the scope of this disclosure is much broader than presented by the claims. Applicants reserve the right to present different claims (that may include broader claims) in an application that claims the benefit of priority from this application.

CLAIMS

1. A method of generating text from speech to identify content of the speech received from a primary speaker, the method comprising:

receiving a video input;

receiving an audio input;

segmenting the video input into a plurality of video features;

segmenting the audio input into a plurality of audio features;

matching the plurality of video features to the plurality of audio features;

selecting a primary speaker from one of the matched video and audio features; and

generating text from the matched video feature and audio feature of the primary speaker,

wherein one or more of the above steps are performed using a computer processor that accesses a memory.

2. The method of claim 1, wherein the video input is received through a camera.

3. The method of claim 1, wherein the video input is in an mp4 format.

4. The method of claim 1, wherein the audio input is at least one human's voice.

5. The method of claim 1, wherein the audio input is received through a microphone.

6. The method of claim 1, wherein the video features are human lip movements.

7. The method of claim 1, wherein the audio features are audio samples of different speakers.

8. The method of claim 1, wherein the primary speaker is the speaker physically closest to an audio and video input device.

9. The method of claim 1, wherein the primary speaker is the speaker with the loudest audio feature.

ABSTRACT

A method of generating text from speech using video-speech matching from a primary user is disclosed herein. The method requires first receiving a video and audio input. The video and audio inputs are then segmented into a plurality of video and audio features, respectively. The plurality of video and audio features are then matched according to their similarities. A primary speaker is then determined from one of the matched video and audio features. The primary speaker's matched video and audio features are then used to generate a text representative of the primary speaker's speech.

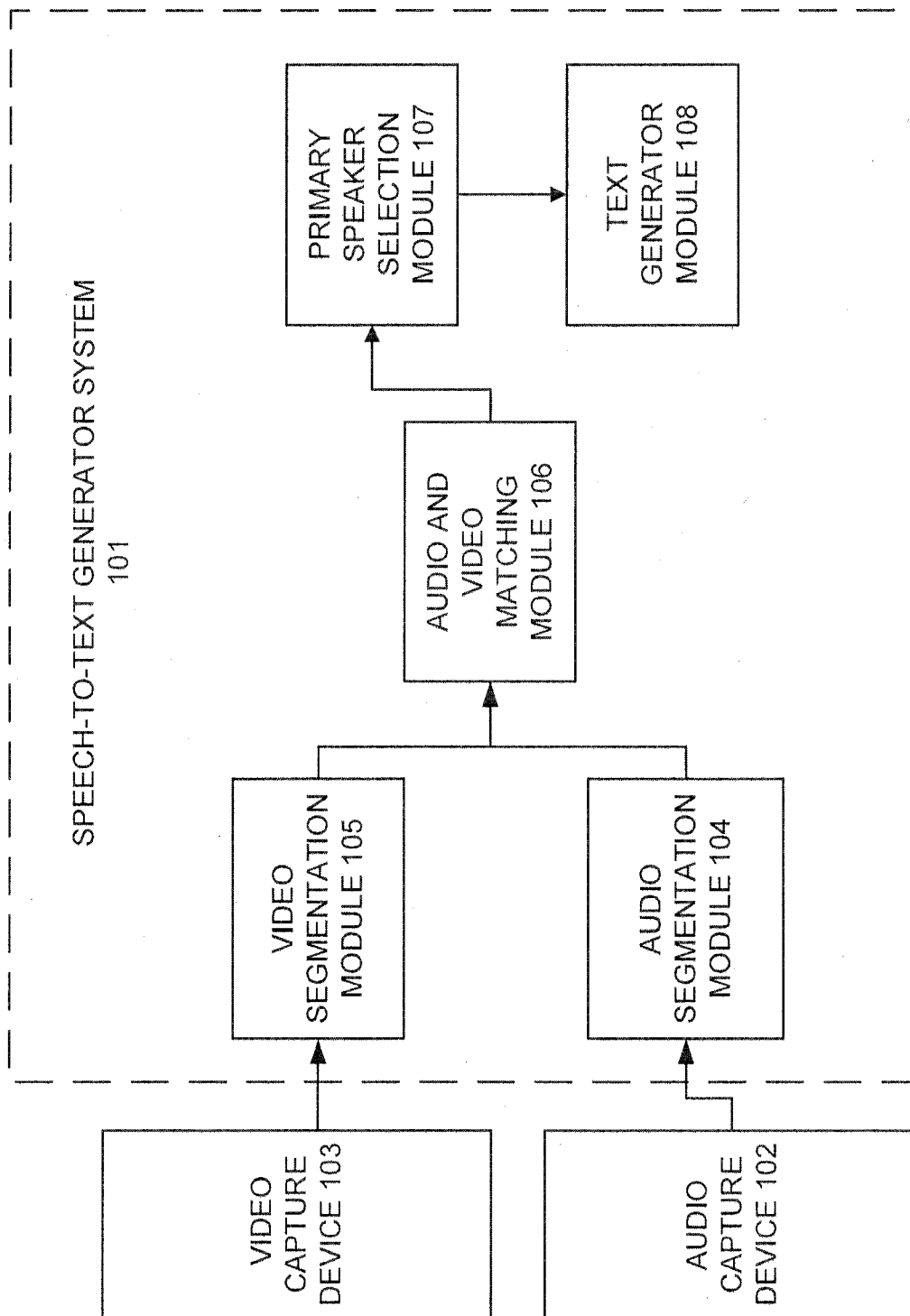


FIGURE 1

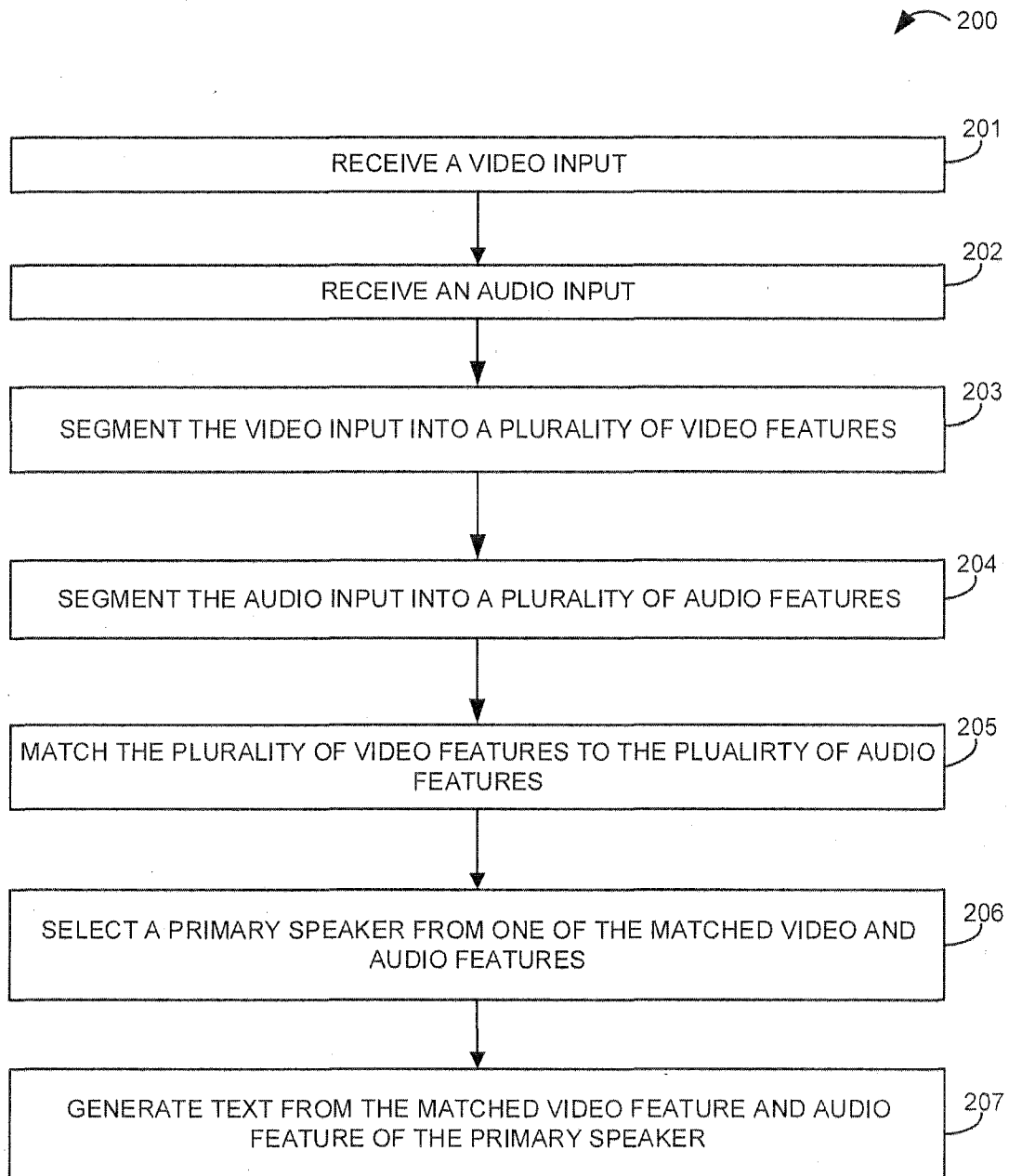


FIGURE 2