June 05, 2019

Context Based Speech Enhancement Techniques for Voice and Media Applications

Snehitha Singaraju

Context-Based Speech-Enhancement Techniques for Voice and Media Applications

Abstract:

Smart devices continue to proliferate, and they provide a variety of functions to assist end users. Voice-interactive (voice-activated) and audio-interactive devices continue to increase in popularity, availability, and functionality. Voice-interactive and audio-interactive devices may improve by monitoring and determining context when distinguishing speech from background noise and distinguishing speech intended for the voice-interactive and audio-interactive device and speech nearby but irrelevant to the device. A machine-learning model can process detected keywords, contextual cues, historical cues, and other information in determining relevant speech from background noise.

Keywords:

Machine learning, ML, learning, training model, neural network, artificial intelligence, AI, speech, audio signal, voice, conversation, utterance, sound, speak, keyword, attribute, query, text, word, word string, dialect, query, content, context, topic, intent, aim, purpose, objective, goal, target, environment, surrounding, situation, background noise, party, concert, enhance, increase, intensify, magnify, strengthen, upgrade, improve, privacy, confidential, security, anonymity, seclude.

Background:

Human beings interact and communicate with one another using audio, visual, contextual, historical, and other cues. For example, a child may clearly identify the voice of his or her parents among many voices in a crowded room because of experience with the parents' voices. A teacher may focus on one student among many asking a question because the student is also raising his or her hand. Likewise, the teacher may focus in on the conversation of one group of students and tune out other conversations of nearby students while moving about the classroom. In such circumstances, human beings focus on certain information and exclude other information based on the variety of cues they receive. Human beings often make these focus determinations automatically or autonomically as part of their cognitive function.

Unlike human beings, computing devices do not possess this cognitive functionality. Computing devices may need additional assistance to determine where to focus, what to focus on, or where and what to ignore. Consider the classroom example above. An audio receiver, such as a microphone of a desktop computing assistant, receives sound indiscriminately in the surrounding area. The desktop computing assistant cannot initially distinguish between a speaker attempting to interact with the desktop computing assistant and other nearby speakers holding an unrelated conversation or general background noise. Without more information, the desktop computing assistant may consider all the received audio as being potentially relevant, which is unlikely to produce the result desired by the speaker that is attempting to interact with the desktop computing assistant.

Description:

Smart devices provide a variety of functions to assist end users. Voice-interactive (voice-activated) and audio-interactive functions for these devices continue to increase in popularity, availability, and functionality. Voice-interactive and audio-interactive devices may be improved by monitoring and determining context when distinguishing speech from background noise and distinguishing speech intended for the voice-interactive and audio-interactive device and speech nearby but irrelevant to the device. A machine-learning model can process detected keywords, contextual cues, historical cues, and other information in determining relevant speech from background noise.

Consider the kitchen depicted in Figure 1. A man stands near the kitchen sink. The kitchen contains a smart thermostat, a countertop or desktop computing assistant, appliances (some may be "smart" or Internet-connected others may be of a more conventional variety), and other kitchen items. In the absence of other sounds, the man can communicate and interact with the desktop computing assistant relatively smoothly and easily. For example, the man can instruct the assistant to make a phone call to his mother asking for assistance in preparing a favorite recipe. During the conversation with his mother, other non-voiced sounds, like a beeping microwave, running water at the sink, or a mixing device, may complicate the operations of the assistant. Nevertheless, mechanical sounds can be repetitive and of a different frequency composition from human vocal speech, which the computing assistant may distinguish. The computing assistant, or a speech-recognition module of the assistant, can filter out or ignore repetitive sounds that are more likely to be made by a machine than by the man.

Figure 1

The introduction of additional human voices presents a challenge to the computing assistant. Consider the same kitchen in Figure 1 but with the addition of two more people, as shown in Figure 2.

Figure 2

Here, another man and a woman enter the kitchen. If the second man and the woman are engaged in conversation, the assistant may have difficulty distinguishing their background conversation from the conversation of the man relevant to the phone call. Some computing

assistant speech-recognition systems can be trained to favor selection of voices that interact with the computing assistant frequently through time and experience. If the second man and the woman are not familiar to the computing assistant, it can likely filter out their conversation as part of the background noise. Nevertheless, in the case of the two men and the woman being siblings, all the individuals in the room may be known to the computing assistant. The computing assistant may struggle to select or isolate a particular voice relevant to the phone call, which may result in unnecessary noise being included on the call or relevant vocal sound being filtered out and excluded. Further, the challenge is exacerbated by each of the siblings periodically engaging in the phone call conversation with their mother and periodically engaging in conversation with one another. Even this simple scenario can be difficult for a traditional computing assistant to effectively-manage human interactions.

An enhanced speech-recognition system can monitor the conversation among the mother on the phone call and the siblings in the room to determine a voice on which to focus. For example, the man, during his conversation with his mother, says something like “Eliza just walked in” and the system reduces background sound filtering. The speech-recognition system may have access to a contact list and recognizes that Eliza is the man’s sister, the mother’s daughter, and responds by including Eliza’s voice in the transmission to the mother. Similarly, the man says “your grandson Jimmy just walked in” and, also possibly from a contact list or the content of the words of the man, the speech-recognition system recognizes the intent of the grandmother to engage in conversation with the grandson and adjust its speech-recognition operations to account for either a child’s voice or both the child’s and the adult’s voices. Likewise, the mother says “Go and get my grandson” or “Put Eliza on”, and the speech-recognition assistant adjusts its operations to account for any number of speakers.

Gatherings of many people present another challenge. Consider the meeting of several professionals, as illustrated in Figure 3. The professionals may be gathering at a job fair, networking event, or after-work social event. Tom, the gray-colored man in the center holding a computing tablet, has chosen to attend the event to meet two colleagues, Sally and Katie. Tom intends to introduce another member of this team, Jane, using a video-conferencing application. On a traditional computing device, Tom may be required to engage a user interface to shift a mode of the speech-recognition system. In such a crowded and voice-heavy location, the computing device may rely on Tom selecting either a private mode, which can filter out all but familiar voices or preserve the loudest voice under the assumption that Tom's voice will be louder because he is closer to the device, or a group mode, which may include all voices or even all received sounds. Nevertheless, Tom is surrounded by multiple conversations, which may render the transmission too cluttered to be useful to Jane, the other member of his team.

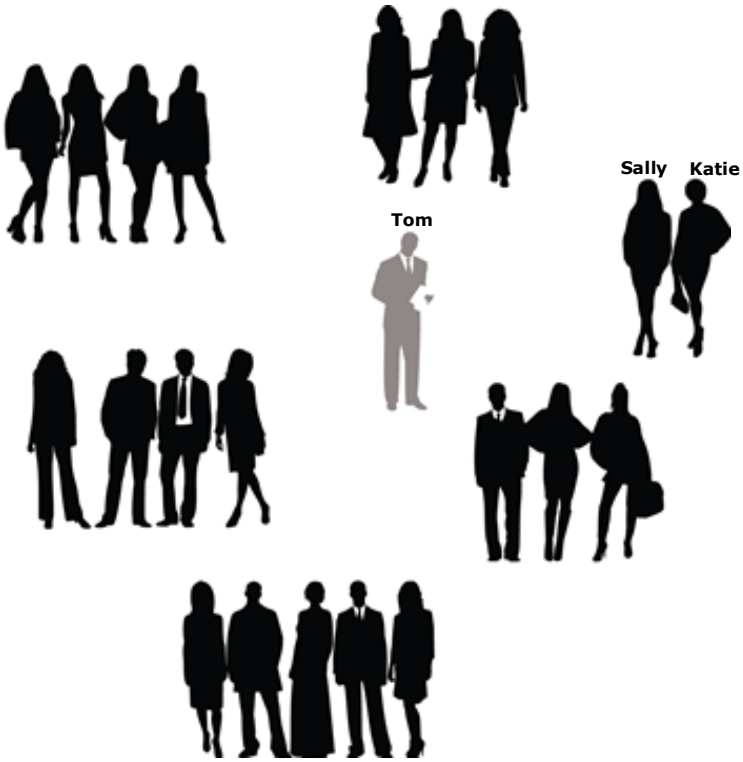

Figure 3

Tom has a tablet computing device, however, that includes an enhanced speech-recognition system. The speech-recognition system analyzes cues in Tom's behavior and speech, which have been learned or assembled using a machine-learning model to estimate or determine Tom's intent based on the context of the conversation. For example, while engaging with conversation with Jane, Tom sees Sally and Katie. During the conversation, Tom may say something, such as "Sally and Katie just arrived. I'll walk over to them now." As shown in Figure 4, Tom moves towards Sally and Katie. Hearing and interpreting this cue, the speech-recognition system may monitor other systems of the computing tablet to further confirm that the man intends to approach or is approaching Sally and Katie, such as monitoring movement sensed by an accelerometer of the tablet computing device. The system may also query a calendar or scheduling application to confirm an appointment. In some circumstances, the speech-recognition system deduces from historical interactions that the man always meets to discuss business matters with Sally and Katie at this same time or place. Having further confirmed that the system correctly interpreted Tom's intent, the speech-recognition system changes its method of operation to better account for the addition of Sally's and Katie's voices to the conversation.

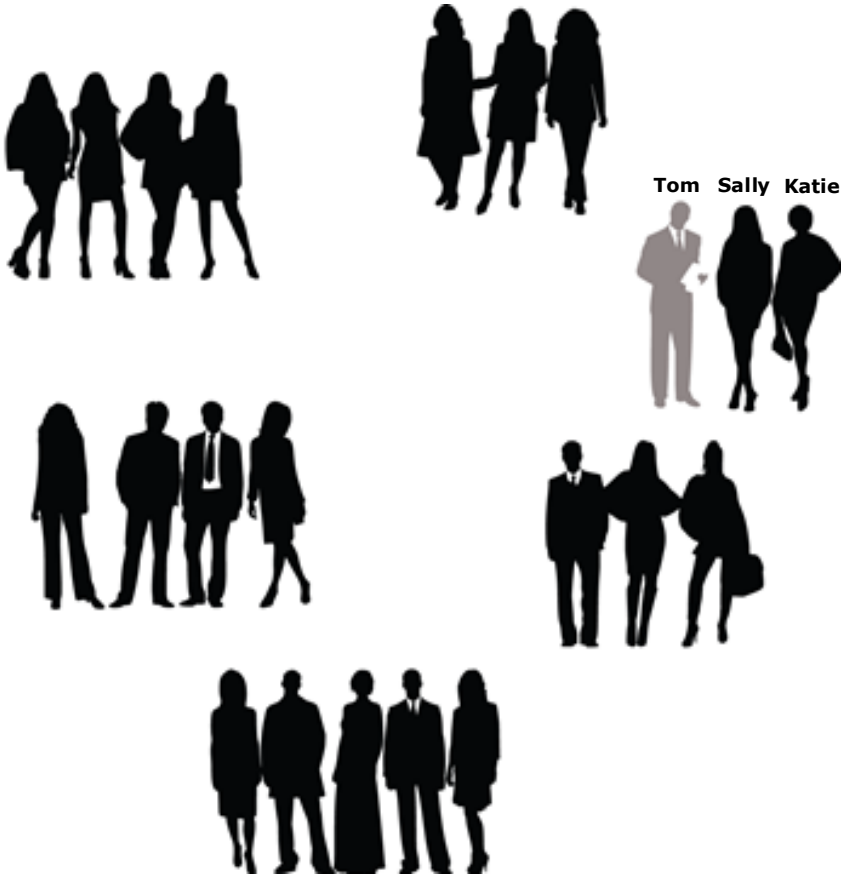

Figure 4

As the conversation among Sally, Katie, Jane, and Tom progresses, the speech-recognition system continues to monitor the conversation for cues that suggest the manner of operation needs to respond to changed or changing circumstances. In Figure 5, Tom, Sally, and Katie, approach a television screen on which a keynote speaker can be seen and heard delivering a message. Such a display may be off to one side of the social gathering at which they have all met. During their conversation with Jane, Tom says something, such as "Jane, you need to hear this" and the speech-recognition system adjusts its methods of operation to no longer filter out background conversations, which may include the sounds coming from the television or a broadcast system.

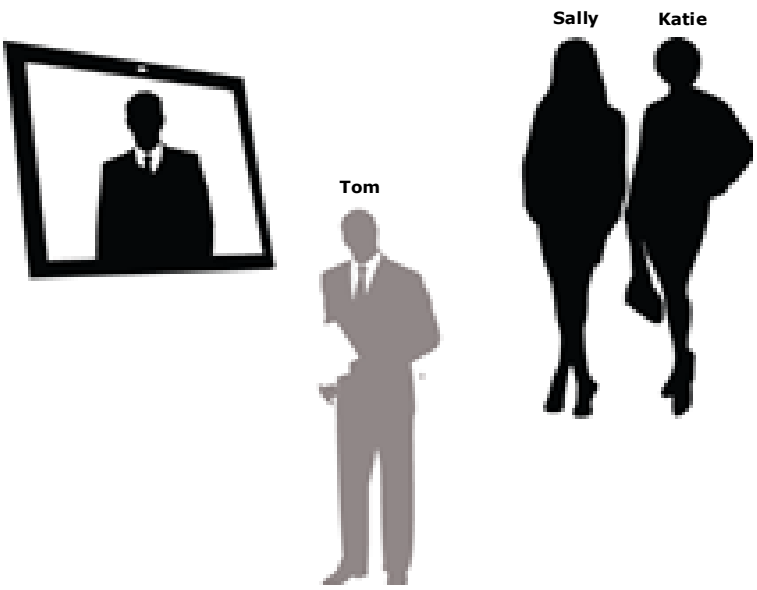

Figure 5

Ultimately, the user is in control of the actions taken by the speech-recognition system. As part of the learning, calibration, or control mechanisms of the system, the user may be presented with information and make decisions based on the information. For example, in the initial setup of the assistant or the video-conferencing application of Figures 1 and 3, the user can be presented with several different prompts, such as "Would you like me to monitor the conversation and adjust my speech recognition operations accordingly?" or "It appears that you are talking with multiple people. During the conversation, this video-conferencing application can adjust speech-recognition operations to better collect speech from multiple speakers? Would you like these operations performed automatically?". Likewise, after using the computing assistant or the video-conferencing application, the user can determine that he or she may have preferred the speech-recognition system assist in transmitting speech and vocal sound but forgot to enable it. The user can direct the machine-learning model to prompt, suggest, or automatically engage the speech-recognition system during future use. In this manner, the user ultimately retains control over the

collection, analysis, storage, and sharing of information collected by the speech-recognition system and its components on the computing device.

Additionally, a user may be provided with controls allowing the user to make an election as to both if and when systems, programs, or features described herein may enable collection of user information (*e.g.*, information about a user's social network, social actions or activities, profession, a user's preferences, or a user's current location), and if the user is sent content or communications from a server. In addition, certain data may be treated in one or more ways before it is stored or used, so that personally identifiable information is removed. For example, a user's identity may be treated so that no personally identifiable information can be determined for the user, or a user's geographic location may be generalized where location information is obtained (such as to a city, ZIP code, or state level), so that a particular location of a user cannot be determined. Thus, the user may have control over what information is collected about the user, how that information is used, and what information is provided to the user. Here, the user may restrict access, or grant access to, contact lists or other personal information in directing the machine-learning model to direct the speech-recognition operations.

Smart devices with voice-activated, voice-interactive, audio-interactive, or speech-recognition systems may be improved by monitoring and determining context when distinguishing speech from background noise and distinguishing speech intended for the voice-interactive and audio-interactive device and speech nearby but irrelevant to the voice-interactive and audio-interactive device. A machine-learning model can process detected keywords, contextual cues, historical cues, and other information in determining relevant speech from background noise.

References:

[1] Kristjansson, Trausti, and Matthew I. Lloyd. Geotagged and weighted environmental audio for enhanced speech recognition accuracy. US Patent 8175872, filed September 30, 2011, and issued May 8, 2012.