June 2022

# REMOVING POTENTIAL DISTURBANCES FROM AUDIO BASED ON CONTEXT

Ramprasad Sedouram

Puneetha Pai B P

# REMOVING POTENTIAL DISTURBANCES FROM AUDIO BASED ON CONTEXT

## ABSTRACT

A computing device (e.g., a smartphone, a laptop computer, a tablet computer, a smartwatch, etc.) may monitor a media presentation and adjust the media presentation based on a detected real-life context. The computing device may use microphones and/or other sensors (e.g., position sensor, accelerometer sensor, pressure sensor, temperature sensor, force sensor, vibration sensor, piezo sensor, humidity sensor, etc.) to detect the context. The computing device may evaluate the media presentation and identify media elements. Based on the identified media elements and the context, the computing device may mute or censor potentially distracting, dangerous, or undesirable sounds or images for the context. For example, sounds and noises from media presentations may be mistaken for real-life sounds, which may be distracting in certain environments. When driving a car, context-related noises, such as honk sounds, crash or brake/engine sounds, traffic sounds, and yelling, may distract the driver resulting in unsafe driving. The computing device may improve safety by muting these distracting noises from the media presentation.

## DESCRIPTION

FIG. 1 below is a conceptual diagram illustrating a system 100 that includes a computing device 102 and a computing system 122. In accordance with various techniques described in this publication, computing device 102 may use a machine learning model(s) 104 to mute or remove distracting sounds from an audio presentation based on context.

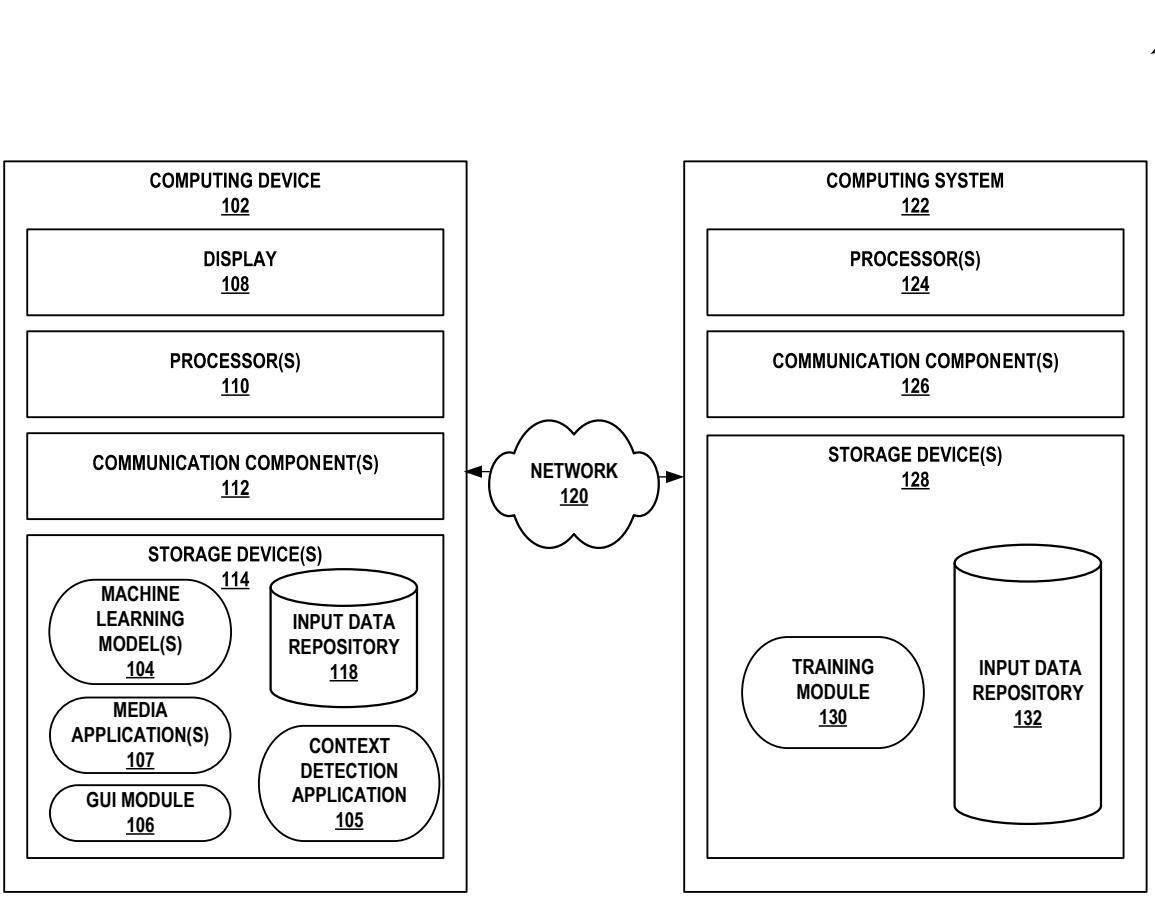


**FIG. 1**

As shown in FIG. 1, computing device 102 may include a display 108, one or more processors 110, one or more communication components 112 ("COMM components 112"), and one or more storage devices 114. Storage devices 114 may include machine learning model(s) 104, a graphical user interface module 106 ("GUI module 106"), and an input data repository 118. Computing device 102 may be any mobile or non-mobile computing device, such as a cellular phone, a smartphone, a desktop computer, a laptop computer, a tablet computer, a portable gaming device, a portable media player, an e-book reader, a watch (including a so-called smartwatch), a gaming controller, and/or the like.

As shown in FIG. 1, computing system 122 may include one or more processors 124, one or more communication components 126, and one or more storage devices 128. Storage devices

128 may include training module 130 and an input data repository 132. Computing system 122 may be any suitable remote computing system, such as one or more desktop computers, laptop computers, mainframes, servers, cloud computing systems, virtual machines, etc., capable of sending and receiving information via a network 120. Computing system 122 may be a cloud computing system that provides one or more services via network 120. For example, computing system 122 may be a distributed computing system.

Display 108 of computing device 102 may be implemented using various display hardware. Display 108 may function as an input device using a presence-sensitive input component, such as a presence-sensitive screen or touch-sensitive screen, that receives tactile input from a user of computing device 102. The presence-sensitive input component may determine a contact location (e.g., an (x,y) coordinate) of the presence-sensitive input component at which the object was detected.

Processors 110 and processors 124 may implement functionality and/or execute instructions associated with computing device 102 and computing system 122, respectively. Examples of processors 110 and processors 124 may include one or more of an application specific integrated circuit (ASIC), a field programmable gate array (FPGA), an application processor, a display controller, an auxiliary processor, a central processing unit (CPU), a graphics processing unit (GPU), one or more sensor hubs, and any other hardware configure to function as a processor, a processing unit, or a processing device. Machine learning model(s) 104, and GUI module 106 may be operable by processors 110 to perform various actions, operations, or functions of computing device 102. Training module 130 may be operable by processor 124 to perform various actions, operations, or functions of computing system 122.

COMM components 112 and COMM components 126 may receive and transmit various types of information, such as input data stored in input data repository 118 and input data repository 132, over network 120. Network 120 may include a wide-area network such as the Internet, a local-area network (LAN), a personal area network (PAN) (e.g., Bluetooth®), an enterprise network, a wireless network, a cellular network, a telephony network, a Metropolitan area network (e.g., WIFI, WAN, WiMAX, etc.), one or more other types of networks, or a combination of two or more different types of networks (e.g., a combination of a cellular network and the Internet).

COMM components 112 and COMM components 126 may include wireless communication devices capable of transmitting and/or receiving communication signals using network 108, such as a cellular radio, a 3G radio, a 4G radio, a 5G radio, a Bluetooth® radio (or any other PAN radio), an NFC radio, or a WIFI radio (or any other WLAN radio). Additionally, or alternatively, COMM components 112 and COMM components 126 may include wired communication devices capable of transmitting and/or receiving communication signals via a direct link over a wired communication medium (e.g., a universal serial bus ("USB") cable).

Storage devices 114 and storage devices 128 may include one or more computer-readable storage media. For example, storage devices 114 and storage devices 128 may be configured for long-term, as well as short-term storage of information, such as, e.g., instructions, data, or other information used by computing device 102 and computing system 122, respectively. Storage devices 114 and storage devices 128 may include non-volatile storage elements. Examples of such non-volatile storage elements include magnetic hard discs, optical discs, solid-state discs, and/or the like. In other examples, in place of, or in addition to the non-volatile storage elements, storage devices 114 and storage devices 128 may include one or more so-called "temporary"

memory devices, meaning that a primary purpose of these devices may not be long-term data storage. For example, the storage devices may comprise volatile memory devices, meaning that the devices may not maintain stored contents when the devices are not receiving power. Examples of volatile memory devices include random access memories (RAM), dynamic random-access memories (DRAM), static random-access memories (SRAM), and/or the like. A user of computing device 102 may provide user input via display 108 to navigate with respect to GUIs of computing device 102.

Various aspects of the techniques described in this publication enable computing device 102 to use machine learning model(s) 104 to adjust a media presentation based on context. Context detection application 105 may use microphones or other sensors at the computing device 102 to detect the context using machine learning model(s) 104. Media applications 107 may use machine learning model(s) 104 to search for potentially distracting audio elements based on the user's context. For example, certain sounds and noises from media presentations may be mistaken for real-life sounds, which may be distracting in specific environments, such as driving in a car.

Computing system 122 may train the machine learning model(s) 104. Training module 130 may train the machine learning model(s) 104 on computing system 122 based on training data in the input data repository 132.

Training module 130 may train machine learning model(s) 104 by optimizing an objective function. The objective function may represent a loss function that compares (e.g., determines a difference between) output data generated by the model from the training data and labels (e.g., ground-truth labels) associated with the training data. For example, the loss function may evaluate a sum or mean of squared differences between the output data and the labels.

Training module 130 may train machine learning model(s) 104 using supervised learning techniques.

Computing device 102 may download machine learning model(s) 104 (or, if computing device 102 already locally stores a version of machine learning model(s) 104, update machine learning model(s) 104 by downloading a more recent version of machine learning model(s) 104) from the computing system 122.

Computing system 122 may use deep learning to produce machine learning model(s) 104, which may include a machine learning model to determine a context and a machine learning model to produce audio signal classifiers or labels.

Computing system 122 may train a machine learning model to detect certain sounds in a media presentation. Such a machine learning model may be a sound event detection (SED) network, such as a deep learning network that uses a convolutional neural network (CNN) layer operating on spectrograms of an audio presentation to detect certain sounds. Computing system 122 may train such a machine learning model using training data, including labeled audio snippets. Computing device 102 may use the machine learning model to detect specific sounds in the media presentation.

Computing system 122 may train a machine learning model to produce a transcript of an audio presentation. Such a machine learning model may be an automatic speech recognition (ASR) network or speech-to-text network, such as a convolutional neural network (CNN) and/or recurrent neural network (RRN) that uses spectrograms of an audio presentation to produce a transcript. Computing system 122 may train such a machine learning model using training data, including previously transcribed audio. Computing device 102 may use the machine learning model to determine certain words or phrases in a media presentation to modify.

Computing system 122 may train a machine learning model to remove specific sounds from a media presentation. Such a machine learning model may be a deep neural network trained using training data, including baseline audio and the baseline audio with added sounds. Computing device 102 may use the machine learning model to remove specific sounds without removing speech or other sounds from the media presentation.

Computing system 122 may train a machine learning model to determine the context. Such a machine learning model may be a deep neural network trained using labeled sensor data. Computing device 102 may use the machine learning model to determine the context. The computing device 102 may also use predetermined rules to help determine context.

Computing device 102 may replace sensitive non-verbal audio snippets from the audio content presented when context-specific distraction sounds occur. For example, computing device 102 may determine the context based on the user's activity, time of the day, location, and other related factors. For example, computing device 102 may determine the context as "driving" based on detected car sounds and/or the speed at which the device is moving. Computing device 102 may determine the speed at which the device is moving using an accelerometer system or by using a global navigation satellite service (GNSS), such as the global positioning system (GPS) or the global navigation satellite system (GLONASS).

The computer device 102 may determine what sounds have to be filtered/altered/tweaked from the audio content to ensure that the user is not distracted in view of the context. For example, the computing device 102 may modify the audio content to remove distracting sounds (e.g., road traffic sounds, voice assistant invocation catchphrases, system sound alerts, loud accident-like sounds, sounds that clash with users' phobias, etc.) from the audio content. The

computing device 102 may also ensure that the content presented remains acceptably understandable by not removing too much of the non-distracting audio of the media presentation.

In the case of pre-recorded audio files, the computing device 102 or computing system 122 may pre-process the pre-recorded audio files to identify the start and end times of audio snippets to be modified. In the case of a live audio stream, the computing device 102 may cache a few seconds of audio to do this processing.

In an exemplary case, a driver drives his car while listening to his favorite podcast. The podcast includes a short advertisement about the best engine oil accompanied by racers racing with car drifts, revs, and other traffic sounds and honks. The computing device 102 may determine that' considering the user's context, the user should not be disturbed by these sounds as they may be confused with real-life traffic. The computing device 102 may selectively replace these distracting sounds with alternatives like background music or remove them. The computing device 102 may retain the voice-over, so the user still gets the context of the advertisement.

In another exemplary case, a cook is in the kitchen cutting vegetables and listening to her favorite podcast via smart speakers. The cook had previously indicated to computer device 102 that she was afraid of reptiles. The computer device 102 may also determine that a user has a phobia of reptiles based on previously detected user behavior and/or monitored indications like heart rate.

The computing device 102, such as a smartwatch, may understand that she is chopping vegetables via movement and/or audio sensors. The computing device 102 may also determine that the podcast is about to discuss snakes. Computing device 102 may determine this by pre-processing the podcast audio to transcribe it and check for sensitive words based on the context.

Computing device 102 may selectively skip the section of the podcast including the discussion of reptiles to avoid any accident while the cook is working with a sharp knife.

It is noted that the techniques of this disclosure may be combined with any other suitable technique or combination of techniques. As one example, the techniques of this disclosure may be combined with the techniques described in U.S. Patent Application Publication No. 2020/0389718. In another example, the techniques of this disclosure may be combined with the techniques described in U.S. Patent Application Publication No. 2015/0195641. In yet another example, the techniques of this disclosure may be combined with the techniques described in U.S. Patent Application Publication No. 2021/0279634. In yet another example, the techniques of this disclosure may be combined with the techniques described in U.S. Patent Application Publication No. 2022/0076689.