May 2023

INCREASING COMPREHENSION THROUGH PLAYBACK OF TRANSLATED SPEECH

Bhramara Tirupati
Meta Platforms Technologies, LLC

INCREASING COMPREHENSION THROUGH PLAYBACK OF TRANSLATED SPEECH

Inventor:
Bhramara Tirupati

FIELD OF THE INVENTION

[0001] The present disclosure generally relates to speech translation, and specifically relates to increasing comprehension through playback of translated speech.

BACKGROUND

[0002] Listening comprehension of a non-native language is a key challenge in several use cases that involve traveling for instance or day to day activities that require comprehension for non-native English speakers. Even when equipped with good reading comprehension, listening comprehension for non-native speakers remains a big challenge due to varying accents. A conventional solution to the problem involves mobile applications which can provide a speech to text type of translation, but often do not provide real time translation in day-to-day situations.

DETAILED DESCRIPTION

[0003] Users who are non-native speakers of a language may have trouble understanding the language when spoken with an accent. An accent refers to a way of pronouncing a language that is distinctive to a particular area or country, or background. For example, a native English speaker located in the United States may have acquired one of a variety of accents, such as a Boston accent, or a Southern accent. An accent may have features such as the stress, pitch, and

Atty. Docket No.: 31718-54782

intonation on consonants or vowels. To illustrate different vowel pronunciations, the word "lot" pronounced by a person with an American accent may sound like "laht" (e.g., lɑt), while the word "lot" pronounced with an English accent may sound like "lawt" (e.g., lɒt). A non-native English speaker may not be able to discern the content of the speech due to these differences, and the speed at which words are spoken. However, the non-native speaker is more likely able to discern the content of the speech once hearing the language spoken in a voice with similar speech patterns as themselves. Thus, an alternative solution includes capturing speech audio from a sound source, modifying the captured speech audio to have speech patterns that match that of the user, and playing back the modified speech audio content to the user to facilitate comprehension. Such a design once tested has applicability in social situations for international visitors as well as immigrant populations in a foreign country. Another use case for this technology would be improving reading and listening comprehension in children or assisting special needs children and adults with an assistive technology. In such cases, the playback audio could be the voice of a caretaker, parent or a medical professional as appropriate for the situation.

[0004] An audio system that is configured to translate captured speech audio signals and modify captured speech audio based on the characteristics of a user's voice, is disclosed herein. The audio system may be implemented in wearable devices which includes, and is not limited to, head-mounted devices such as artificial reality headsets. In some embodiments, the audio system can translate from one language to another.

[0005] The audio system may include a transducer array, a sensor array, and an audio controller. Some embodiments of the audio system may have more or fewer components than described here. The audio system captures speech audio from a sound source, modifies the

captured speech audio to have speech patterns that match that of the user, and presents the audio content to the user using one or more transducers of the audio system. The audio system generates one or more acoustic transfer functions for a user and may use the one or more acoustic transfer functions to generate audio content for the user. The audio controller may include a speech translation module, and a data storage. Similarly, other embodiments of the audio controller may have more or fewer components than described. In some embodiments, the audio system may use machine learning models to perform functionalities described herein. Example machine learning models include regression models, support vector machines, naïve bayes, decision trees, k nearest neighbors, random forest, boosting algorithms, k-means, and hierarchical clustering. The machine learning models may also include neural networks, such as perceptrons, multi-layer perceptrons, convolutional neural networks (CNNs), recurrent neural networks (RNNs), sequence-to-sequence models, generative adversarial networks, automatic speech recognition (ASR) models, or transformers.

[0006] The sensor array of the audio system detects sounds within the local area of the headset. The sensor array includes a plurality of acoustic sensors. An acoustic sensor captures sounds emitted from one or more sound sources in the local area (e.g., a room). Each acoustic sensor is configured to detect sound and convert the detected sound into an electronic format (analog or digital). The acoustic sensors may be acoustic wave sensors, microphones, sound transducers, or similar sensors that are suitable for detecting sounds. The data store stores data for use by the audio system. Data in the data store may include sounds recorded in the local area of the audio system (e.g., speech from a sound source), speech profiles associated with certain speech and/or audio characteristics, audio content, head-related transfer functions (HRTFs), transfer functions for one or more sensors, array transfer functions (ATFs) for one or more of the

acoustic sensors, sound source locations, virtual model of local area, direction of arrival estimates, sound filters, and other data relevant for use by the audio system, or any combination thereof.

[0007] The audio controller of the audio system processes information from the sensor array that describes sounds detected by the sensor array. The audio controller may comprise a processor and a computer-readable storage medium. The audio controller may include a speech translation module. The speech translation module may be configured to translate the captured speech audio into a target language. In other embodiments, the speech translation module may modify the captured/translated speech audio based on the speech patterns of the user's voice. In some embodiments, the translation functionality of the audio system may be user activated (e.g., wake up word, depressing a button on the wearable device). In other embodiments, the audio system may automatically process detected speech audio above a threshold amplitude.

[0008] The audio system may capture and analyze recordings of the user's voice to create a speech profile for the user. A speech profile may be associated with one or more determined speech parameters, the speech parameters describing characteristics of a recording of speech audio, such as the spoken language or dialect, stress, pitch, and intonation on consonants or vowels. A speech profile can be associated with English spoken with a type of American accent, such as a Boston accent or a Southern accent. In some embodiments, the user may select, from a list of voices, their preferred playback voice, each associated with a corresponding speech profile. In some embodiments, the user's own voice may be selected as a playback voice.

[0009] The audio system may recognize the captured speech audio as the target language chosen by the user. The wearable device plays back the captured speech in the user's preferred playback voice. In other embodiments, the captured speech audio is translated into English and a

corresponding text transcription may be displayed to the user on the display elements of the wearable device or on an application on the user's mobile device, in addition to being played back to the user in a voice with similar speech patterns of the user in real time.

[0010] The audio system may recognize the captured speech audio as a language different from the target language chosen by the user. The user may select, from a list of languages, a target language to translate recorded speech audio into. For example, if English is chosen as a target language, captured speech audio in a different language (e.g., Japanese) is translated into English and played back to the user in a voice with similar speech patterns as the user in real time. The audio controller may implement one or more machine-learned models (e.g., ASR models) to predict the speech profile of a captured speech audio by using extracted speech parameters of the captured speech audio recording and modify the predicted speech profile of the captured speech audio to the speech profile of the user. In some embodiments, the speech translation module is configured to convert the captured speech audio into one or more representations of the captured speech audio for input to one or more machine-learned models. The one or more machine-learned models may receive, as input, one or more representations of the captured speech audio, and outputs a speech profile associated with the determined characteristics of the one or more representations. An example representation of the captured speech audio includes a spectrogram, which is a visual representation of the amplitude and frequencies of the audio signal over time. In other embodiments, the speech translation module may be configured to convert captured speech audio into Mel-Frequency Cepstral Coefficients (MFCCs), a representation of short-term spectrum of sounds.

[0011] The one or more machine-learned models may be configured to learn a mapping between the predicted speech profile of the captured speech audio (e, g, the sound source) and

the speech profile of the user. The machine-learned models may be configured to learn the conversion of words and phrases between speech profiles using determined speech parameters. The machine-learned models may also be configured to modify the speech pattern of the captured speech audio to resemble the speech pattern of the user's voice. For example, the machine-learned models may slow down the speed of the captured speech audio to match the speed at which the user speaks. The modified captured speech audio content is presented to the user through the one or more transducers of the audio system.

[0012] Embodiments of the invention may include or be implemented in conjunction with an artificial reality system. Artificial reality is a form of reality that has been adjusted in some manner before presentation to a user, which may include, e.g., a virtual reality (VR), an augmented reality (AR), a mixed reality (MR), a hybrid reality, or some combination and/or derivatives thereof. Artificial reality content may include completely generated content or generated content combined with captured (e.g., real-world) content. The artificial reality content may include video, audio, haptic feedback, or some combination thereof, any of which may be presented in a single channel or in multiple channels (such as stereo video that produces a three-dimensional effect to the viewer). Additionally, in some embodiments, artificial reality may also be associated with applications, products, accessories, services, or some combination thereof, that are used to create content in an artificial reality and/or are otherwise used in an artificial reality. The artificial reality system that provides the artificial reality content may be implemented on various platforms, including a wearable device (e.g., headset) connected to a host computer system, a standalone wearable device (e.g., headset), a mobile device or computing system, or any other hardware platform capable of providing artificial reality content

to one or more viewers.

Additional Configuration Information

[0013] The foregoing description of the embodiments has been presented for illustration; it is not intended to be exhaustive or to limit the patent rights to the precise forms disclosed. Persons skilled in the relevant art can appreciate that many modifications and variations are possible considering the above disclosure.

[0014] Some portions of this description describe the embodiments in terms of algorithms and symbolic representations of operations on information. These algorithmic descriptions and representations are commonly used by those skilled in the data processing arts to convey the substance of their work effectively to others skilled in the art. These operations, while described functionally, computationally, or logically, are understood to be implemented by computer programs or equivalent electrical circuits, microcode, or the like. Furthermore, it has also proven convenient at times, to refer to these arrangements of operations as modules, without loss of generality. The described operations and their associated modules may be embodied in software, firmware, hardware, or any combinations thereof.

[0015] Any of the steps, operations, or processes described herein may be performed or implemented with one or more hardware or software modules, alone or in combination with other devices. In one embodiment, a software module is implemented with a computer program product comprising a computer-readable medium containing computer program code, which can be executed by a computer processor for performing any or all the steps, operations, or processes described.

[0016] Embodiments may also relate to an apparatus for performing the operations herein. This apparatus may be specially constructed for the required purposes, and/or it may comprise a

general-purpose computing device selectively activated or reconfigured by a computer program stored in the computer. Such a computer program may be stored in a non-transitory, tangible computer readable storage medium, or any type of media suitable for storing electronic instructions, which may be coupled to a computer system bus. Furthermore, any computing systems referred to in the specification may include a single processor or may be architectures employing multiple processor designs for increased computing capability.

[0017] Embodiments may also relate to a product that is produced by a computing process described herein. Such a product may comprise information resulting from a computing process, where the information is stored on a non-transitory, tangible computer readable storage medium and may include any embodiment of a computer program product or other data combination described herein.

[0018] Finally, the language used in the specification has been principally selected for readability and instructional purposes, and it may not have been selected to delineate or circumscribe the patent rights. It is therefore intended that the scope of the patent rights be limited not by this detailed description, but rather by any claims that issue on an application based hereon. Accordingly, the disclosure of the embodiments is intended to be illustrative, but not limiting, of the scope of the patent rights, which is set forth in the following claims.

What is claimed is:

1. A method comprising:

capturing speech audio signals from a sound source in a local area of an audio system;

determining a speech profile of the sound source using the speech audio signals;

generating translated audio signals using the captured speech, the speech profile of the sound source, and a speech profile of a user of the audio system, the translated audio signals having the speech profile of the user; and

presenting the translated speech signals as audio content to the user.