August 14, 2018

Self-managed Speech Therapy

Dimitri Kanevsky

Sagar Savla

Thad Starner

Self-managed speech therapy

ABSTRACT

Speech defects are typically addressed by having the patient or learner undergo several sessions with speech therapists, who apply specialized therapeutic tools. Speech therapies tend to be expensive, require the scheduling of appointments, and do not lend themselves easily to self-paced self-improvement.

This disclosure presents techniques that automatically provide speech-improvement feedback, thereby enabling self-managed speech therapy. Given a speech utterance by a user, the techniques cause display of a sequence of images of speech-organ positions, e.g., tongue, lips, throat muscles, etc., that correspond to the actual utterance as well as a targeted, ideal utterance. Further phonetic feedback is provided to the user using visual, tactile, spectrogram, or other modes, such that a speaker who is hard of learning can work towards a target pronunciation. The techniques also apply to foreign language learning.

KEYWORDS

- Speech therapy
- Oral-cavity sensor
- Phonetic feedback
- Stabilized auditory images
- Spectrogram
- Tactile feedback
- Automatic speech recognizer
- Phoneme

BACKGROUND

Speech defects are typically addressed by having the patient or learner undergo several sessions with speech therapists, who apply specialized therapeutic tools. Speech therapies tend to be expensive, require the scheduling of appointments, and do not lend themselves easily to self-improvement at a self-paced rate.

Satisfactory speech production depends on the coordinated action of several factors, e.g., tongue position, pressure, air flow etc. For example, different combinations of these physical factors can lead to the same pronunciation. Speech patients or learners benefit from continuous feedback regarding whether something was spoken correctly or not, and if not, what should be changed, e.g., in terms of speech-organ positions, to improve pronunciation.

Further, there is a category of people with hearing difficulties, e.g., deaf or hard-of-hearing individuals, for whom speech therapies are especially difficult. Audio feedback, a typical tool used in speech therapies, is not of much help to such individuals, since they are unable to hear the feedback. This effect is sometimes manifested when such individuals speak louder than necessary, since they lack an understanding of the volume level of their own utterances.

Some speech therapies introduce sensors into the learner’s mouth to obtain a real-time reading of speech-organ positions during speech production. Such readings are then used to create learning targets for the user. However, such therapy is expensive, time consuming (e.g., may require visits to a dentist to make casts of the mouth), and inconvenient.

DESCRIPTION

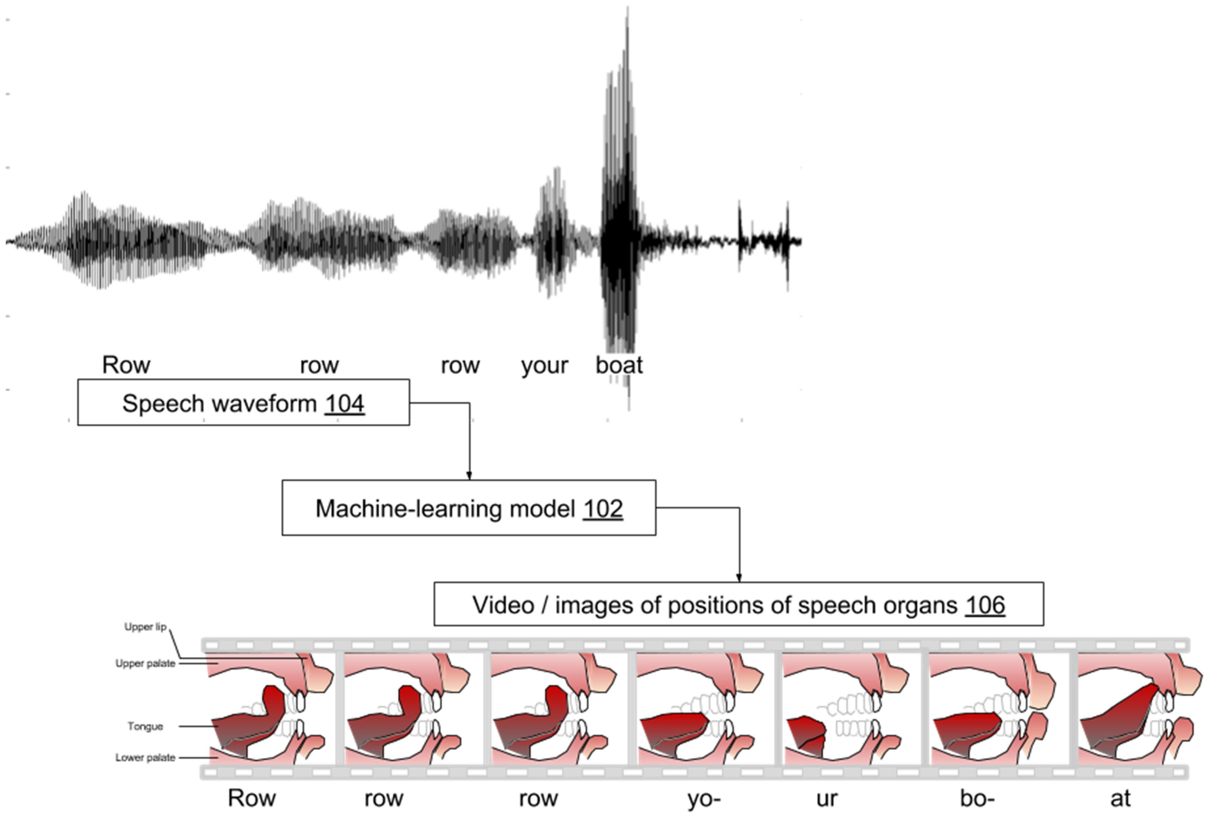

Fig. 1: Mapping speech to a video of movements of speech-production organs

Fig. 1 illustrates an example of mapping speech to a video of movements of speech-production organs, per techniques of this disclosure. A trained machine-learning model (102) receives as input a speech waveform (104) and produces a sequence of images or video of the movements of speech-production organs (106) that would result in the input waveform. In the example shown in Fig. 1, the speech waveform corresponds to a phrase “row, row, row your boat,” and the video shows the sectional positions of the tongue, lips, upper and lower palates as the speech waveform progresses in time. In a similar manner, videos of other speech-production organs, e.g., throat muscles, teeth, oral cavity shape, etc. are generated.

The generated videos could be sectional (as in Fig. 1), or could be taken from other views of the oral cavity, e.g., plan, elevation, etc. The videos can show additional detail, for example, a plan video of tongue movements could graphically display the points (or regions) of contact of the tongue with the upper (or lower) palate, teeth, etc., along with the pressure that the tongue exerts on the palates, teeth, etc. The generated video can be indicative of particular types of speech impairments, or of no impairments at all.

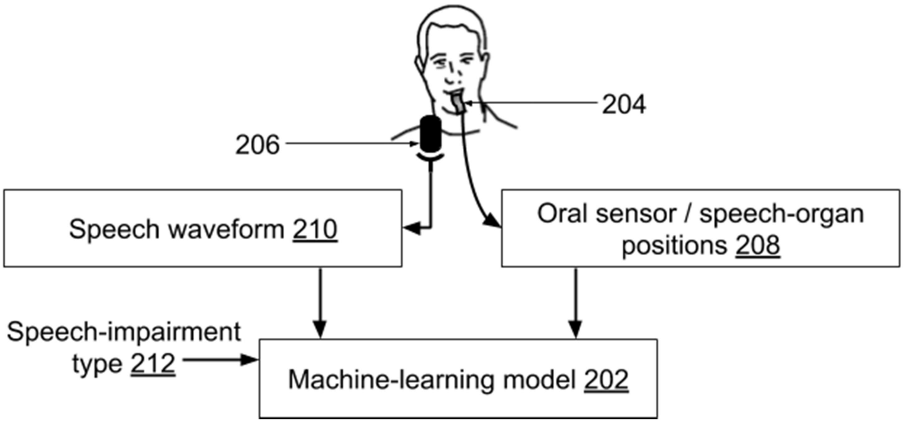
204
206
Speech waveform 210
Oral sensor / speech-organ positions 208
Speech-impairment type 212
Machine-learning model 202

Fig. 2: Training of the machine-learning model

Training of the machine learning model that translates a speech waveform into a video of speech-production organs is illustrated in Fig. 2. A human speaker that voluntarily contributes data for training, who may or may not have a speech impairment, is fitted with oral-cavity sensors (204) that monitor real-time positions of speech-production organs such as tongue, teeth, palates, throat muscles, jaw-opening angle, mouth-cavity shape, lips, etc. Oral-cavity sensors can include sensors that use micro-radar, imaging, electromagnetic, haptic, proximity-sensing, and other technologies.

As the speaker speaks, a microphone (206) picks up the speaker's speech waveform (210), and the oral-cavity sensors pick up positions of speech-production organs (208). Both speech waveform and speech-organ positions are fed as training data to machine-learning model

(202). Using this input data, the model is trained to associate speech waveforms with positions of speech-production organs. With permission of the speaker, the speech impairment type for the particular speaker (212), if any, is fed as a parameter to the machine-learning model.

In this manner, once trained, the machine-learning model can receive as input a speech waveform and generate a video of speech-organ positions. In a similar manner, the machine-learning model can accept as input strings of phonemes, spectrogram data, stabilized auditory images, speech features extracted from an automatic speech recognizer, or other representations of speech signals, and produce as output the most probable sequence of positions of speech organs. It is noted that in operation, the machine-learning model generates the video of speech-organ positions without use of oral-cavity sensors, thereby freeing the user of the burden and inconvenience of hosting oral-cavity sensors for the purpose of speech therapy.

The machine-learning model can be implemented as a multi-layer neural network, e.g., a long short-term memory (LSTM) neural network. Other types of models, e.g., recurrent neural networks, convolutional neural networks, and techniques such as support vector machines, random forests, boosted decision trees, etc., can also be used to implement the model.

An important category of speech learners includes learners who are deaf or hard of hearing. For such individuals, audio feedback relating to their speech or pronunciation is difficult or impossible. This disclosure provides techniques that provide visual or tactile feedback relating to or assessing a learner’s speech.

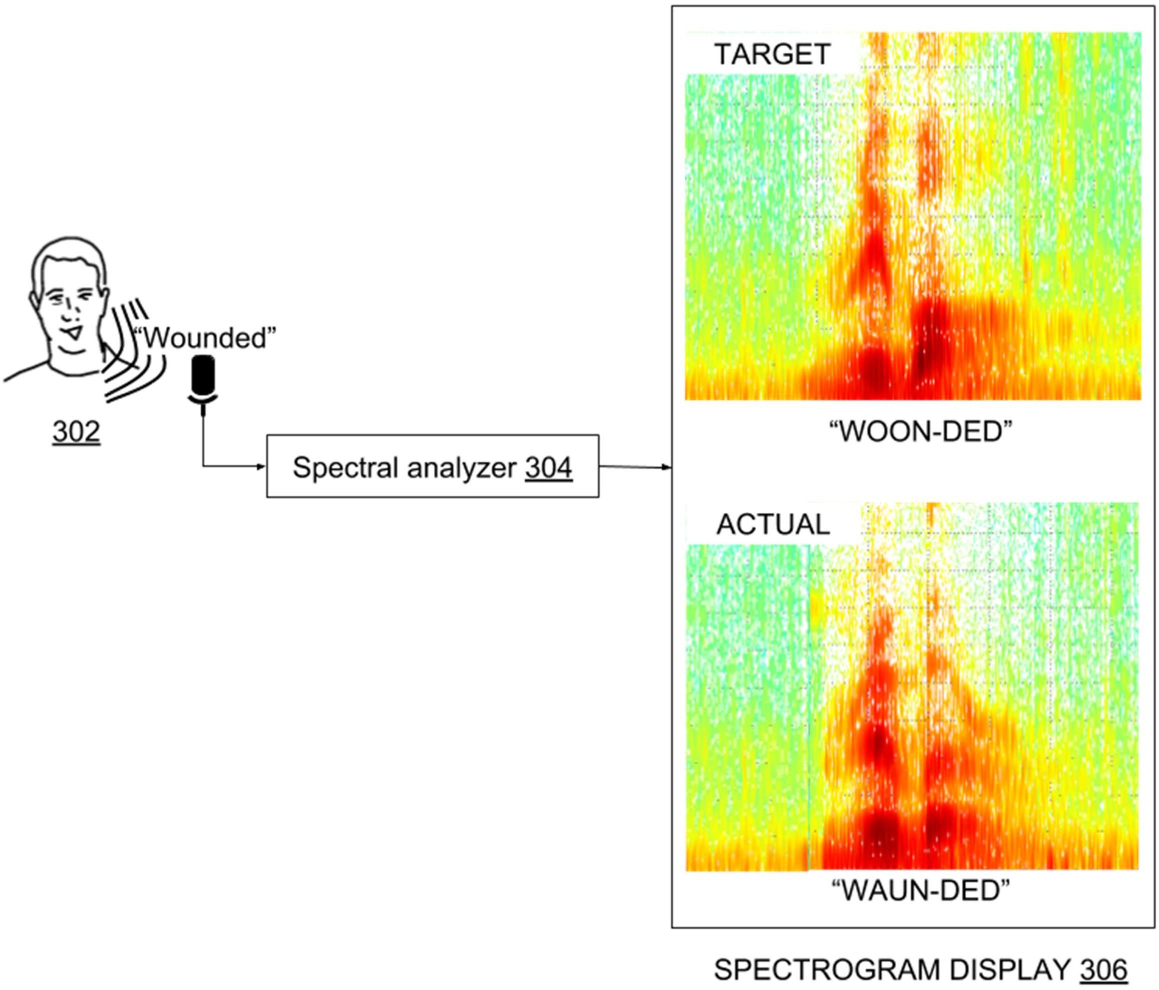

Fig. 3: Visual pronunciation feedback using spectrograms

Fig. 3 illustrates an example of visual pronunciation feedback, per techniques of this disclosure. A learner (302), who is hard of hearing, attempts to pronounce the English word "wounded," and mispronounces it as "WAUN-ded," whereas the correct pronunciation is "WOON-ded." A spectral analyzer (304) captures and analyzes the speech and displays a spectrogram of both the actual utterance ("WAUN-ded") and the correct (target) pronunciation ("WOON-ded"). In this manner, visual feedback is presented to the learner, who, despite being unable to accept audio feedback, nevertheless perceives a visual difference between actual and target pronunciations. The speech learner can then modify their speech until the spectrogram of the utterance matches the target. The spectral analyzer can be portably incorporated into consumer devices, e.g., as a mobile device application.

Alternative to spectrogram display, stabilized auditory images can be used. Stabilized auditory images (SAI) are a real-time, two-dimensional, visual display of sound that exhibits distinct patterns for different phonemes and environmental sounds.

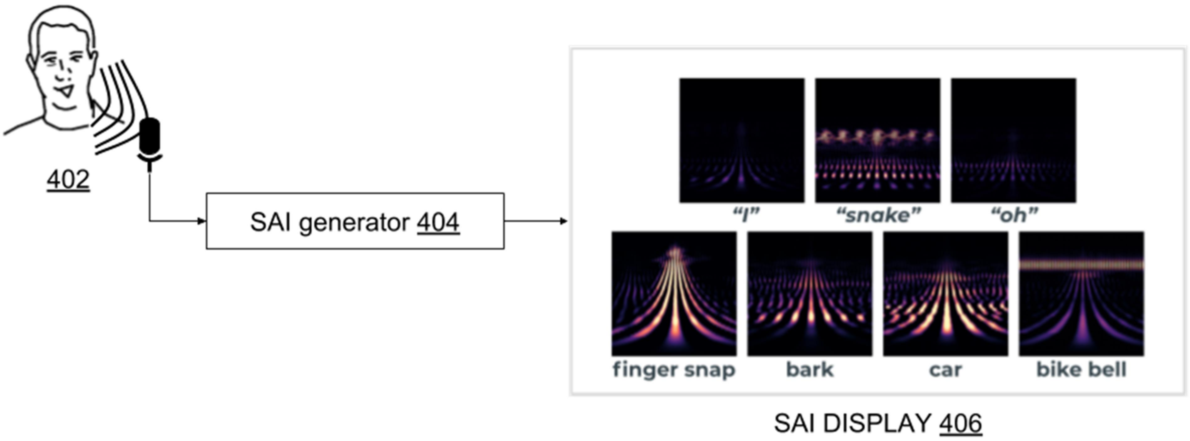

Fig. 4: Visual pronunciation feedback using stabilized auditory images

Fig. 4 illustrates an example of visual pronunciation feedback using stabilized auditory images. As is seen in the example of Fig. 4, distinct phonemes ("I", "snake", "oh") or environmental sounds (finger snap, bark, car, bike bell) generate distinct SAI patterns (406). The phonemes of the speech of a user (402), who is hard of hearing, are transformed into a sequence of stabilized auditory images using SAI generator (404). The user associates phonemes with distinct SAI patterns.

SAI patterns for correctly-pronounced versus incorrectly-pronounced phonemes are shown to the user. Despite being hard of hearing, the user thereby learns to pronounce correctly using the visual display afforded by stabilized auditory imagery. Further, the brightness of the

SAI pattern can be made proportional to the speaker's volume in relation to background noise. The user thereby learns to adjust the volume of speech despite being hard of hearing.

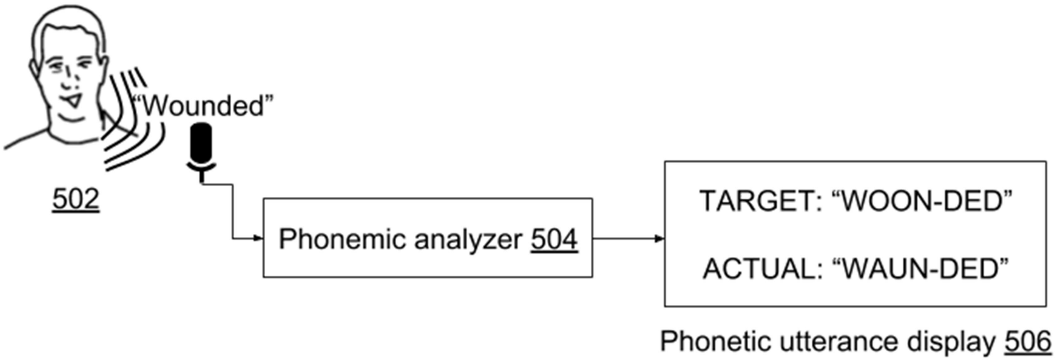

Fig. 5: Phonetic pronunciation feedback

Fig. 5 illustrates an example of another technique to provide visual pronunciation feedback to speech learners who are hard of hearing. A speech learner (502), who is hard of hearing, utters a word ("Wounded"), and a phonemic analyzer (504) displays the phonetic utterance (506) of both the actual speech ("WAUN-ded") as well as the target pronunciation ("WOON-ded"). In this manner, visual feedback is presented to the learner, who, despite being unable to accept audio feedback, nevertheless perceives a phonetic difference between actual and target pronunciations. The learner then modifies their speech until the phonetics of the utterance matches target. The phonemic analyzer can be portably incorporated into consumer devices, e.g., as a mobile device application.

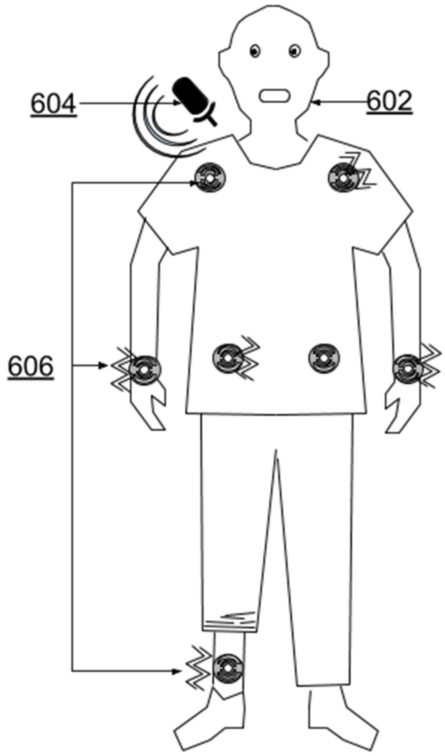

Fig. 6: Tactile pronunciation feedback

Fig. 6 illustrates an example of a technique to provide tactile pronunciation feedback to speech learners who are hard of hearing. A speech learner (602), who is hard of hearing, wears sensors or actuators (606) over their body. The sensor-actuators can be embedded in consumer devices, e.g., smart-watches, smart-clothing, etc. They can also be embedded, e.g., in vests, within the oral cavity, etc.

A microphone (604) picks up speech of the user and transmits it to sensor-actuators 606. The sensor-actuators vibrate in a manner corresponding to the user's speech. The sensor-actuators can also vibrate in a manner corresponding to a target pronunciation. In this manner, the user receives tactile feedback as to their pronunciation. Despite being unable to accept audio feedback, the user nevertheless perceives a tactile difference between actual and target pronunciations. The learner then modifies their speech until the tactile feel of the utterance matches that of the target.

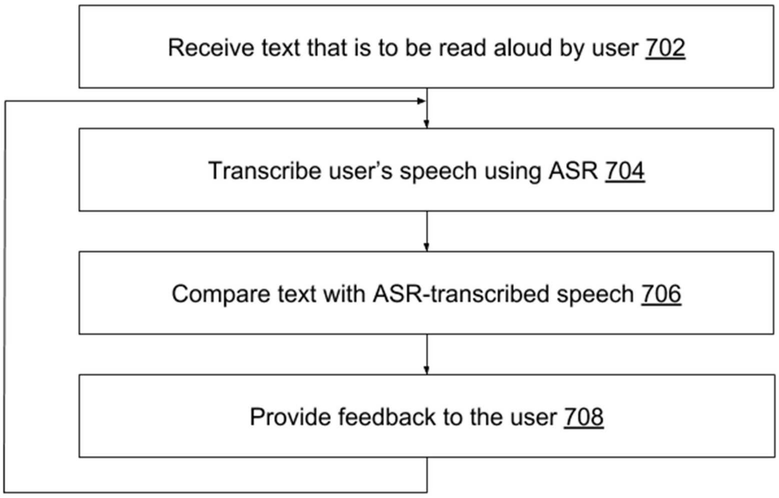
Receive text that is to be read aloud by user 702
Transcribe user's speech using ASR 704
Compare text with ASR-transcribed speech 706
Provide feedback to the user 708

Fig. 7: Providing pronunciation feedback to speech learners using automatic speech recognizers

Fig. 7 illustrates an example of feedback to speech learners using automatic speech recognizers (ASR). A user reads a written text aloud. The written text and resulting speech is received (702) and transcribed by an ASR (704). The written text and ASR-transcription are compared (706) to determine differences. Differences between written text and ASR-transcriptions that are mispronunciations, e.g., diphthong where there is no diphthong, are provided as feedback (708) to the user. Further, or as an alternative, correctly pronounced words are juxtaposed with mispronounced words to provide phonemic guidance, e.g., pronounce "e" in "gallaudet" as the "e" in "test."

The feedback to the user can be in the form of tactile feedback, e.g., using cochlear implants, hearing aids, on-body sensor-actuators, oral-cavity sensor-actuators, etc. The feedback can also be visual, e.g., phonemic (e.g., display reads "say O instead of AU"), in the form of spectrograms or stabilized auditory imagery, in the form of a video of movements of speech organs, etc. Visual feedback can be provided using augmented reality displays, head-up displays, smart glasses, or other wearable/portable devices that can be configured to respond to correctly-pronounced or incorrectly pronounced words in different ways. The process of Fig. 7 can be continuous and is implemented within the control of the speech learner, so that the speech learner continuously improves their speech at their own pace.

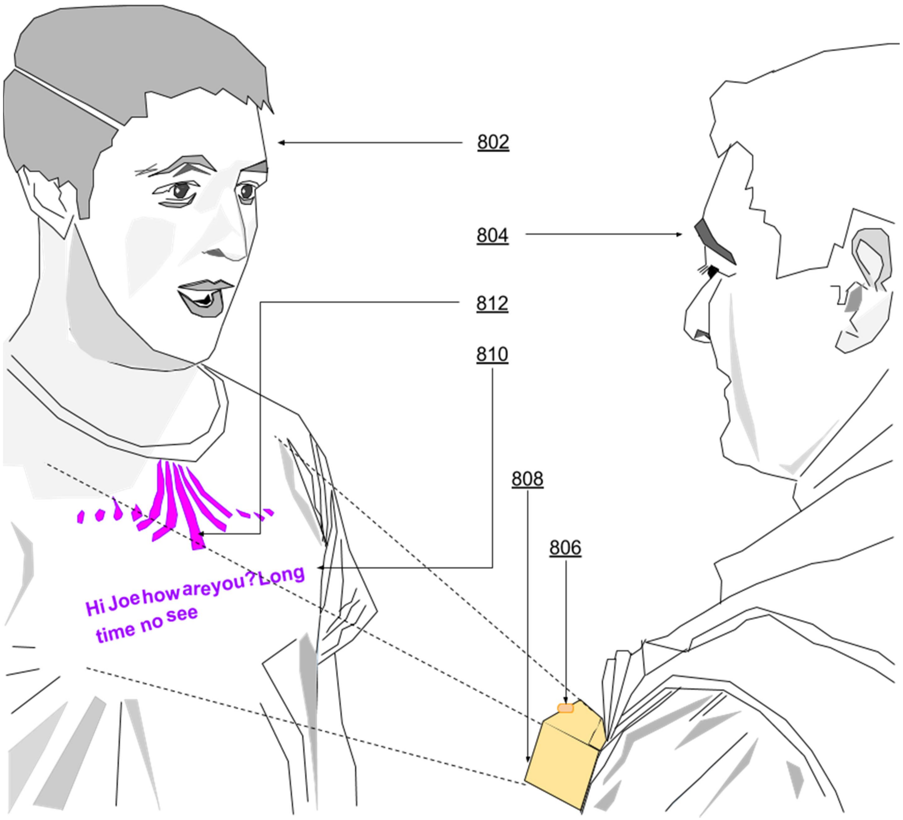

Fig. 8: Augmented reality for conversation feedback

Fig. 8 illustrates an example of the use of augmented reality for conversation feedback and assistance, per techniques of this disclosure. Two participants are in face-to-face conversation. A first participant (802) is conventionally abled, and the other participant (804) is hard of hearing. The participant who is hard of hearing wears a microphone (806), a portable projector (808), and processing equipment capable of transcribing received speech signals into written transcriptions.

With consent from the participants, as the conventionally-abled participant speaks, the microphone picks up the speech signal. The projector projects (810) onto the body (or torso) of the conventionally-abled speaker a transcription of their speech signal. The projector also projects other information that can assist the hard-of-hearing participant, e.g., a sequence of stabilized auditory images (812) corresponding to the speech signal. In a similar manner, the augmented reality components can include lipreading, with virtual information such as transcribed text, phonetic utterances, visualization of tongue positions, etc. being conveyed to the individual who is hard of hearing.

In this manner, the techniques of this disclosure use visual and/or tactile modes of feedback to enable a user to self-correct speech impairments. The techniques are also applicable to learners of foreign languages or accents. Types of visual feedback include auto-generated videos of speech-organ positions, sound visualizations, stabilized auditory imagery, spectrograms, etc. that map to spoken utterances of the user. Types of tactile feedback include sensor-actuators situated on the body, or within the oral cavity of the user, and that map to spoken utterances of the user. Automatic speech recognizers and phonetic translators are used to provide feedback relating to correctly pronounced versus incorrectly pronounced words.

CONCLUSION

This disclosure presents techniques that automatically provide speech-improvement feedback, thereby enabling self-managed speech therapy. Given a speech utterance by a user, the techniques cause display of a sequence of images of speech-organ positions, e.g., tongue, lips, throat muscles, etc., that correspond to the actual utterance as well as a targeted, ideal utterance.

Further phonetic feedback is provided to the user using visual, tactile, spectrogram, or other modes, such that a speaker who is hard of learning can work towards a target pronunciation. The techniques also apply to foreign language learning.