March 2023

Automated Conversion of Impaired Speech in Communication Applications

Gang Feng

Xia Zhang

Pedro Moreno Mengibar

Fadi Biadsy

Liyang Jiang

See next page for additional authors

Inventor(s)
Gang Feng, Xia Zhang, Pedro Moreno Mengibar, Fadi Biadsy, Liyang Jiang, Oleg Rybakov, Yuexin Wu, and Joseph Chen

Automated Conversion of Impaired Speech in Communication Applications

ABSTRACT

Voice communication can be difficult for those with impaired or accented speech. When such users communicate with others via applications on their devices, listeners often find it difficult to understand them. This disclosure describes techniques that dynamically process impaired or accented speech and convert it to synthesized canonical speech with permission. Generation of the synthesized speech is performed with low latency as a user speaks, enabling the parties to engage in smooth communication that is unaffected by the speaker's speech impairment. The listeners receive clear, fluent speech automatically generated by suitably trained models. In addition, users can personalize the operation based on their specific speech impairments. The techniques can be integrated within any messaging, conferencing, or phone calling/ dialer application on any device and can make the applications more accessible to users with impaired speech and enhance the user experience.

KEYWORDS

- Speech impairment
- Accented speech
- Speech conversion
- Synthesized speech
- Canonical speech
- Dialer
- Automated Speech Recognition (ASR)
- Speech-To-Text (STT)

BACKGROUND

Temporary or chronic speech impairment can result from a variety of causes, including speech-related disorders and ailments such as dysarthria, amyotrophic lateral sclerosis (ALS), Parkinson’s disease, stroke, down syndrome, aphasia, etc. Voice communication with others can be difficult for those whose speech is impaired. When users with impaired speech communicate with others via applications on their devices, listeners often find it difficult, if not impossible, to understand what the speaker is saying. Listeners cannot overcome such difficulties with accessibility features, such as captions, because automated speech recognition (ASR) models cannot transcribe impaired speech accurately. Users can encounter similar challenges in understanding speech when listening to a person who speaks with an unfamiliar accent.

Some approaches (e.g., [1]) employ single end-to-end neural models for direct conversion of any speech into fluent speech that matches a canonical speech pattern that is likely to be understood by nearly everyone. Such models can be trained at scale and perform the conversion for inference in a streamable manner with very low latencies, e.g., less than 200ms per utterance. Although such technology typically requires deploying the models via a server, the conversion can be performed locally on the device in a fully streamable fashion with a short smart dynamic buffer based on the context.

Such voice conversion techniques have been utilized to help those with speech impairments communicate with others as well as with speech-enabled applications (e.g., voice-based virtual assistants) and devices (e.g., smart speakers). The technology can convert the original impaired speech into fluent synthesized speech that can be understood by the listener. Instead of relying on an ASR engine to convert speech by choosing words from the ASR vocabulary, the text transcript and audio of the synthesized canonical speech is generated in parallel, which enables the model to stay loyal to the speaker's original speech. Such an approach can also be effective for noise removal, speech separation, and accent conversion.

DESCRIPTION

This disclosure describes techniques to integrate technologies for processing and converting impaired speech into applications for voice messaging and calls. With user permission, the user's original speech with impairment can be converted on-the-fly or on-device to canonical synthesized speech via suitably trained speech processing and conversion models. The converted speech is relayed to the other parties with whom the user is conversing on the call. Since speech generation is performed with low latency as the user speaks, the parties can engage in smooth communication, unaffected by the speaker's speech impairment because the listening parties receive clear, fluent speech instead of the original impaired speech.

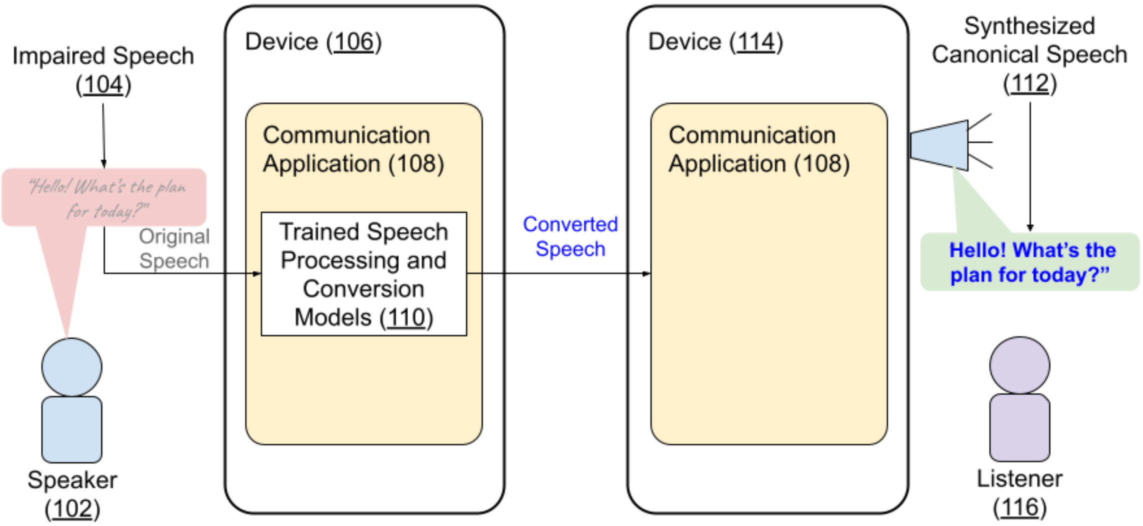
Impaired Speech (104)
"Hello! What's the plan for today?"
Original Speech
Speaker (102)
Device (106)
Communication Application (108)
Trained Speech Processing and Conversion Models (110)
Converted Speech
Device (114)
Communication Application (108)
Synthesized Canonical Speech (112)
Hello! What's the plan for today?"
Listener (116)

Fig. 1: Relaying synthesized canonical speech generated by processing impaired speech

Fig. 1 shows an example of operational implementation of the techniques described in this disclosure. A user (102) with impaired speech (104) is using a communication application

(108) on a device (106) to call another user (116) using the same application on another device (114). With the speaker's permission, the original impaired speech is analyzed via suitably trained speech processing models (110) and converted to synthesized speech in a canonical form (112). The listener receives and hears the converted fluent speech which can be understood without difficulty.

With user permission, the models can be executed locally on the device in a streaming fashion to ensure negligible latency and provide a lag-free user experience (UX) for the communication. Alternatively, or in addition, the models can be located externally (e.g., on a server, in the cloud, etc.) with user permission, if such a configuration is suitable depending on context.

The techniques described herein can be utilized to handle any type of temporary or permanent speech impairment as well as any kind of accented speech. Users can personalize the operation based on their specific speech impairments. For instance, a user can provide a sample set of utterances to build a speech processing and conversion model personalized to the user's speech. In the case of accented speech, the models can be pre-trained for a given accent rather than needing to be personalized to each user with a similar accent. The operation can be easily internationalized to handle any spoken language for which processing and conversion models are available.

With user permission, the techniques described in this disclosure can be integrated within any messaging, conferencing, or phone calling (dialer) application on any device. Integration within the application enables users to derive the benefits of obtaining legible speech without being burdened with specific interactive steps when using the applications to communicate. Implementation of the techniques can make calling and voice-based interaction more accessible to users with impaired speech and enhance the UX for everyone when communicating with such users.

Further to the descriptions above, 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 speech and/or accent, 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.

CONCLUSION

This disclosure describes techniques that dynamically process impaired or accented speech and convert it to synthesized canonical speech with permission. Generation of the synthesized speech is performed with low latency as a user speaks, enabling the parties to engage in smooth communication that is unaffected by the speaker’s speech impairment. The listeners receive clear, fluent speech automatically generated by suitably trained models. In addition, users can personalize the operation based on their specific speech impairments. The techniques can be integrated within any messaging, conferencing, or phone calling/ dialer application on any device and can make the applications more accessible to users with impaired speech and enhance the user experience.

REFERENCES

1. Biadsy, Fadi, Ron J. Weiss, Pedro J. Moreno, Dimitri Kanevsky, and Ye Jia. "Parrotron: An end-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation." *arXiv preprint arXiv:1904.04169* (2019).
2. Biadsy, Fadi, Youzheng Chen, Xia Zhang, Oleg Rybakov, Andrew Rosenberg, and Pedro J. Moreno. "A Scalable Model Specialization Framework for Training and Inference using Submodels and its Application to Speech Model Personalization." *arXiv preprint arXiv:2203.12559* (2022).