Mimicking the brain to realize 'human-like' virtual assistants

February 3 2022

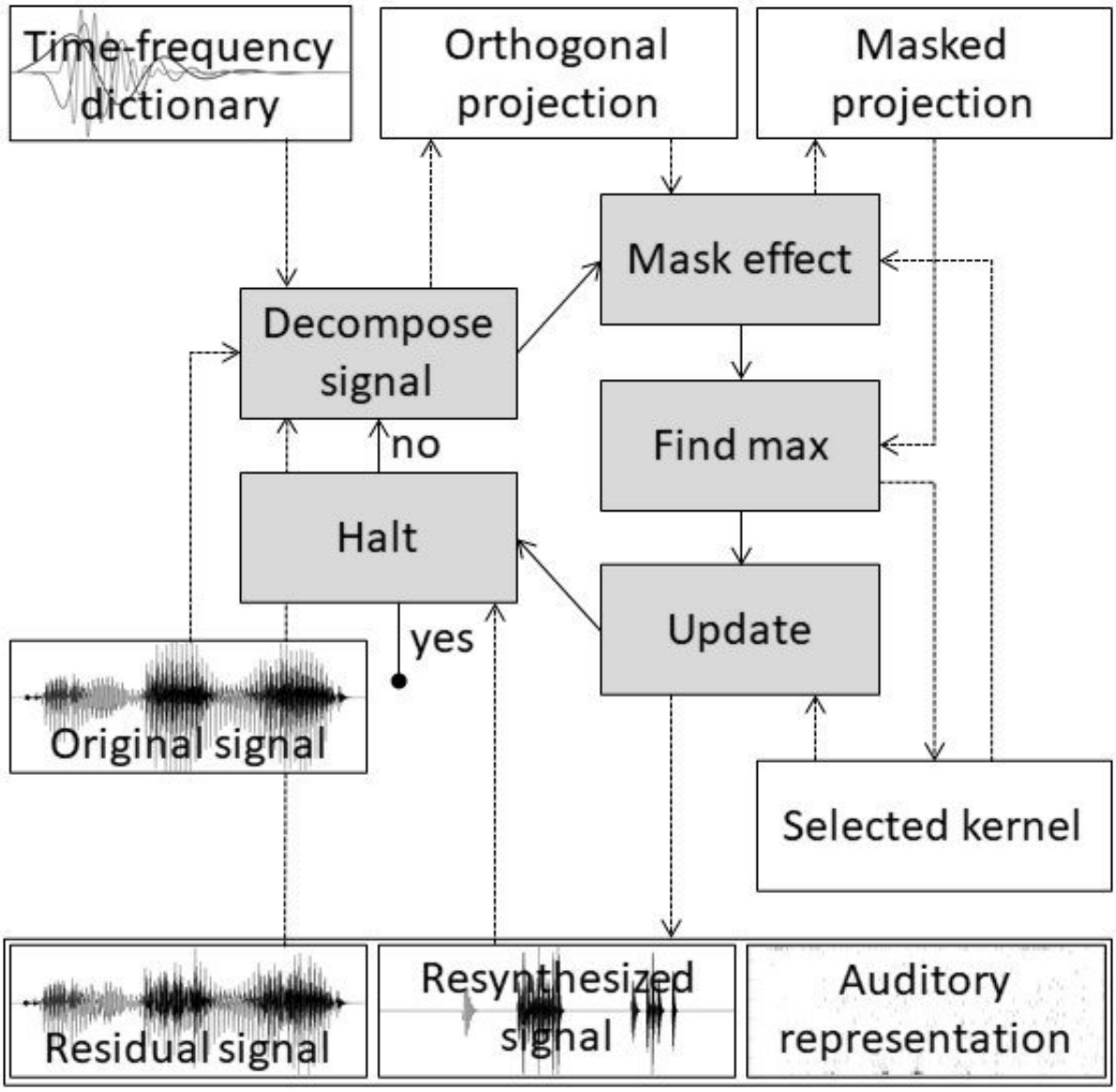

Figure 1. A representation of the algorithm used to mimic human speech. Caption: Processing pipeline of perceptual matching pursuit algorithm used to

derive auditory sparse representations from speech signals. The five main processing steps are illustrated by gray blocks and solid arrows. The first step is to decompose signal, second is to apply mask effect, third is to find max, fourth is to update, and last is to halt. Information about selected kernel found after find-max step is used to create auditory sparse representation, resynthesized signal, and residual signal. Credit: Masashi Unoki from JAIST

Speech is more than just a form of communication. A person's voice conveys emotions and personality and is a unique trait we can recognize. Our use of speech as a primary means of communication is a key reason for the development of voice assistants in smart devices and technology. Typically, virtual assistants analyze speech and respond to queries by converting the received speech signals into a model they can understand and process to generate a valid response. However, they often have difficulty capturing and incorporating the complexities of human speech and end up sounding very unnatural.

Now, in a study published in the journal *IEEE Access*, Professor Masashi Unoki from Japan Advanced Institute of Science and Technology (JAIST), and Dung Kim Tran, a doctoral course student at JAIST, have developed a system that can capture the information in speech signals similarly to how humans perceive speech.

"In humans, the auditory periphery converts the information contained in input speech signals into neural activity patterns (NAPs) that the brain can identify. To emulate this function, we used a matching pursuit algorithm to obtain sparse representations of speech signals, or signal representations with the minimum possible significant coefficients," explains Prof. Unoki. "We then used psychoacoustic principles, such as the equivalent rectangular bandwidth scale, gammachirp function, and masking effects to ensure that the auditory sparse representations are

similar to that of the NAPs."

To test the effectiveness of their model in understanding voice commands and generating an understandable and natural response, the duo performed experiments to compare the signal reconstruction quality and the perceptual structures of the auditory representations against conventional methods. "The effectiveness of an auditory representation can be evaluated in terms of three aspects: the quality of the resynthesized speech signals, the number of non-zero elements, and the ability to represent perceptual structures of speech signals," says Prof. Unoki.

To evaluate the quality of the resynthesized speech signals, the duo reconstructed 630 speech samples spoken by different speakers. The resynthesized signals were then rated using PEMO-Q and PESQ scores—objective measures for sound quality. They found the resynthesized signals to be comparable to the original signals. Additionally, they made auditory representations of certain phrases spoken by 6 speakers.

The duo also tested the model on its ability to capture voice structures accurately by using a pattern-matching experiment to determine if the auditory representations of the phrases could be matched to spoken utterances or queries made by the same speakers.

"Our results showed that the auditory sparse representations produced by our method can achieve high quality resynthesized signals with only 1,066 coefficients per second. Furthermore, the proposed method also provides the highest matching accuracy in a pattern matching experiment," says Prof. Unoki.

From smartphones to smart televisions and even smart cars, the role of voice assistants is becoming more and more indispensable in our daily

lives. The quality and the continued usage of these services will rely on their ability to understand our accents and our pronunciation and respond in a way we find natural. The model developed in this study could go a long way in imparting human-like qualities to our voice assistants, making our interactions not only more convenient but also psychologically satisfying.

More information: Dung Kim Tran et al, Matching Pursuit and Sparse Coding for Auditory Representation, *IEEE Access* (2021). DOI: 10.1109/ACCESS.2021.3135011

Provided by Japan Advanced Institute of Science and Technology