Cognitive neuroscience could pave the way for emotionally intelligent robots

28 April 2021

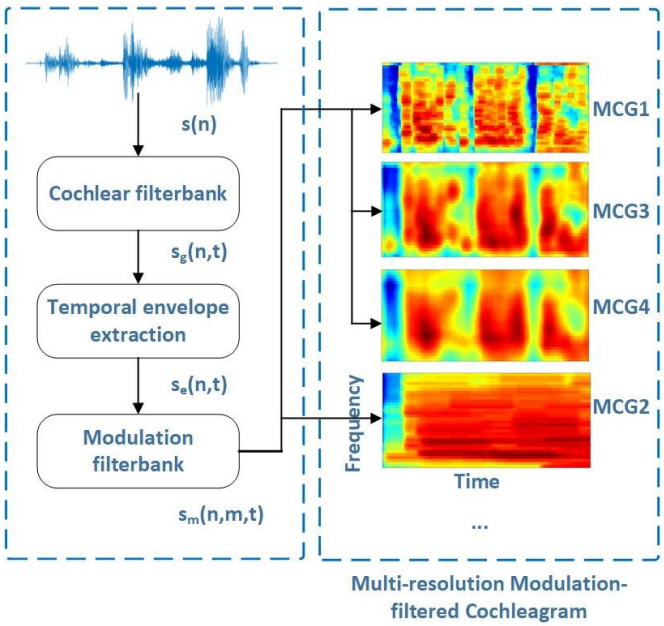

Figure 1. Extraction of multi-resolution modulation-filtered cochleagram (MMCG) features. The left panel shows the process of extracting temporal modulation cues from the auditory front end, while the right panel shows modulation-filtered cochleagram (MCG1–MCG4) at four different resolutions. Credit: Japan Advanced Institute of Science and Technology

Human beings have the ability to recognize emotions in others. Although perfectly capable of communicating with humans through speech, robots and virtual agents are only good at processing logical instructions, which greatly restricts human-robot interaction (HRI). Consequently, a great deal of research in HRI is about emotion recognition from speech. But first, how do we describe emotions?

Categorical emotions such as happiness, sadness and anger are well understood by us but can be hard for robots to register. Researchers have focused on "dimensional emotions," which constitute a gradual emotional transition in natural speech. "Continuous dimensional emotion can help a robot capture the time dynamics of a speaker's emotional state and accordingly adjust its manner of interaction and content in real time," explains Prof. Masashi Unoki from Japan Advanced Institute of Science and Technology (JAIST), who works on speech recognition and processing.

Studies have shown that an auditory perception model simulating the working of a human ear can generate what are called "temporal modulation cues" that faithfully capture the time dynamics of dimensional emotions. Neural networks can then be employed to extract features from these cues that reflect these time dynamics. However, due to the complexity and variety of auditory perception models, feature extraction turns out to be pretty challenging.

In a new study published in *Neural Networks*, Prof. Unoki and his colleagues, including Zhichao Peng, from Tianjin University, China (who led the study), Jianwu Dang from Pengcheng Laboratory, China, and Prof. Masato Akagi from JAIST, have now taken inspiration from a recent finding in cognitive neuroscience suggesting that our brain forms multiple representations of natural sounds with different degrees of spectral (i.e., frequency) and temporal resolutions through a combined analysis of spectral-temporal modulations.

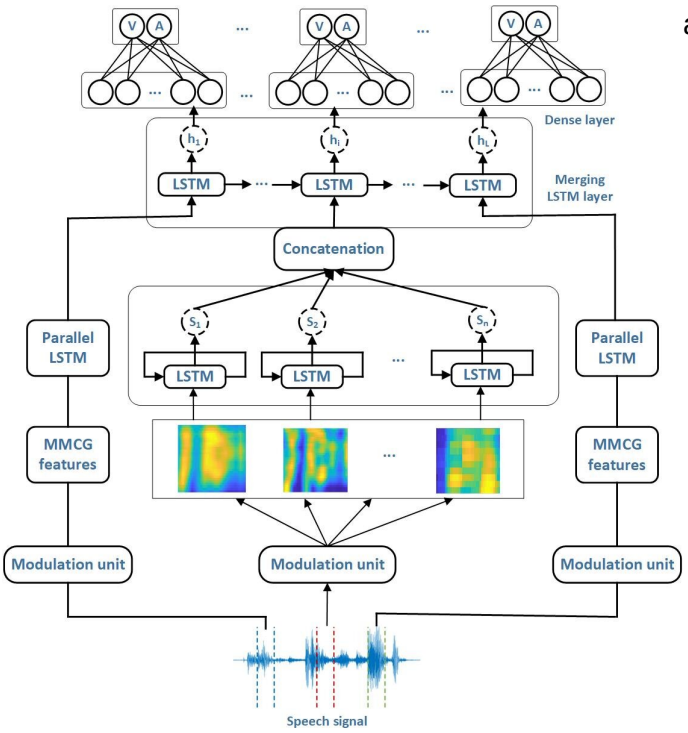

Figure 2. Parallel LSTM network architecture for dimensional emotion recognition. A parallel LSTM network takes in MMCG features with different resolutions and yields outputs that are concatenated together and then sent to a merging LSTM layer and a dense layer to yield the valence (V) and arousal (A) sequences. Credit: Japan Advanced Institute of Science and Technology

Accordingly, the researchers have proposed a novel feature called multi-resolution modulation-filtered cochleagram (MMCG), which combines four modulation-filtered cochleagrams (time-frequency representations of the input sound) at different resolutions to obtain the temporal and contextual modulation cues. To account for the diversity of the cochleagrams, researchers designed a parallel neural network architecture called "long short-term memory" (LSTM), which modeled the time variations of multi-resolution signals from the cochleagrams and carried out extensive experiments on two datasets of spontaneous speech.

The results were encouraging. The researchers found that MMCG showed a significantly better emotion recognition performance than traditional acoustic-based features and other auditory-based features for both the datasets. Furthermore, the parallel LSTM network demonstrated a superior prediction of dimensional emotions than that with a plain LSTM-based approach.

Prof. Unoki is thrilled and contemplates improving upon the MMCG feature in future research. "Our next goal is to analyze the robustness of environmental noise sources and investigate our feature for other tasks, such as categorical emotion recognition, speech separation, and voice activity detection,"he concludes.

More information: Zhichao Peng et al. Multi-resolution modulation-filtered cochleagram feature for LSTM-based dimensional emotion recognition from speech, *Neural Networks* (2021). DOI: 10.1016/j.neunet.2021.03.027

Provided by Japan Advanced Institute of Science and Technology