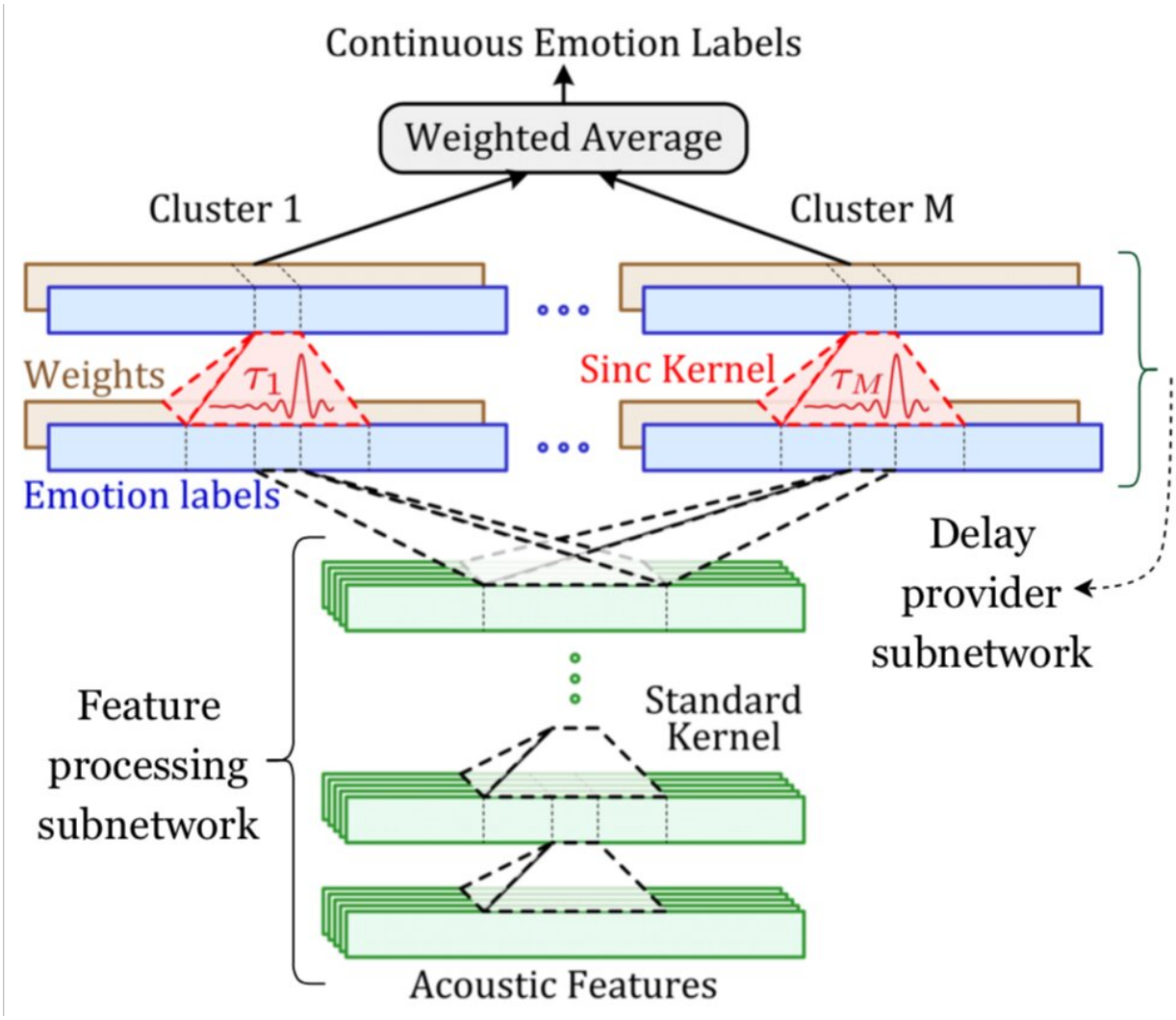
A convolutional network to align and predict emotion annotations

July 22 2019, by Ingrid Fadelli



A system diagram of the MDS Network. Credit: Khorram, McInnis & Provost.

Machine learning models that can recognize and predict human emotions have become increasingly popular over the past few years. In order for most of these techniques to perform well, however, the data used to train them is first annotated by human subjects. Moreover, emotions continuously change over time, which makes the annotation of videos or voice recordings particularly challenging, often resulting in discrepancies between labels and recordings.

To address this limitation, researchers at the University of Michigan have recently developed a new convolutional neural network that can simultaneously align and predict emotion annotations in an end-to-end fashion. They presented their technique, called a multi-delay sync (MDS) network, in a paper published in *IEEE Transactions on Affective Computing*.

"Emotion varies continuously in time; it ebbs and flows in our conversations" Emily Mower Provost, one of the researchers who carried out the study, told TechXplore. "In engineering, we often use continuous descriptions of emotion to measure how emotion varies. Our goal then becomes to predict these continuous measures from speech. But there is a catch. One of the biggest challenges in working with continuous descriptions of emotion is that it requires that we have labels that continuously vary in time. This is done by teams of human annotators. However, people aren't machines."

As Mower Provost goes on to explain, human annotators can sometimes be more attuned to particular emotional cues (e.g., laughter), but miss the meaning behind other cues (e.g., an exasperated sigh). In addition to this, humans can take some time to process a recording, and thus, their reactions to emotional cues is sometimes delayed. As a result, continuous emotion labels can present a lot of variation and are sometimes misaligned with speech in the data.

In their study, Mower Provost and her colleagues directly addressed these challenges, focusing on two continuous measures of emotion: positivity (valence) and energy (activation/arousal). They introduced the multi-delay sync network, a new method to handle misalignment between speech and continuous annotations that reacts differently to different types of acoustic cues.

"Time-continuous dimensional descriptions of emotions (e.g., arousal, valence) provide detailed information about both short-time changes and long-term trends in emotion expression," Soheil Khorram, another researcher involved in the study, told TechXplore. "The main goal of our study was to develop an automatic emotion recognition system that is able to estimate the time-continuous dimensional emotions from speech signals. This system could have a number of real-world applications across different fields including human-computer interaction, e-learning, marketing, healthcare, entertainment and law."

The convolutional network developed by Mower Provost, Khorram and their colleagues has two key components, one for emotion prediction and one for alignment. The emotion prediction component is a common convolutional architecture trained to identify the relationship between acoustic features and emotion labels.

The alignment component, on the other hand, is the new layer introduced by the researchers (i.e. the delayed sync layer), which applies a learnable time-shift to an acoustic signal. The researchers compensated for the variation in delays by incorporating several of these layers.

"An important challenge in developing automatic systems for predicting time-continuous emotion labels from speech is that these labels are generally not synchronized with the input speech," Khorram explained. "This is mainly due to delays caused by reaction-time, which is inherent in human evaluations. In contrast with other approaches, our

convolutional neural network is able to simultaneously align and predict labels in an end-to-end manner. Multi-delay sync network leverages traditional signal processing concepts (i.e. sync filtering) in modern deep learning architectures to deal with the reaction delay problem."

The researchers evaluated their technique in a series of experiments using two publicly available datasets, namely the RECOLA and the SEWA datasets. They found that compensating for annotators' reaction delays while training their emotion recognition model led to significant improvements in the model's emotion recognition accuracy.

They also observed that the reaction delays of annotators when defining continuous emotion labels do not typically exceed 7.5 seconds. Finally, their findings suggest that parts of speech that include laughter generally require smaller delay components compared to those marked by other emotional cues. In other words, it is often easier for annotators to define emotion labels in segments of speech that include laughter.

"Emotion is everywhere and it is central to our communication," Mower Provost said. "We are building robust and generalizable emotion recognition systems so that people can easily access and use this information. Part of this goal is accomplished by creating algorithms that can effectively use large external data sources, both labeled and not, and by effectively modeling the natural dynamics that are a part of how we emotionally communicate. The other part is accomplished by making sense of all of the complexity that is inherent in the labels themselves."

Although Mower Provost, Khorram and their colleagues applied their technique to emotion recognition tasks, it could also be used to enhance other machine learning applications in which inputs and outputs are not perfectly aligned. In their future work, the researchers plan to continue investigating ways in which emotion labels produced by human annotators can be efficiently integrated into data.

"We used a sync filter to approximate the Dirac delta function and compensate for the delays. However, other functions, such as Gaussian and triangular, can also be employed instead of the sync kernel," Khorram said. "Our future work will explore the effect of using different types of kernels that can approximate the Dirac delta function. Additionally, in this paper we focused on the speech modality to predict continuous emotion annotations, while the proposed multi-delay sync network is a reasonable modeling technique for other input modalities too. Another future plan is to evaluate the performance of the proposed network over other physiological and behavioral modalities such as: video, body language and EEG."

More information: Soheil Khorram et al. Jointly aligning and predicting continuous emotion annotations. arXiv:1907.03050 [cs.LG]. arxiv.org/abs/1907.03050

Soheil Khorram et al. Jointly Aligning and Predicting Continuous Emotion Annotations, *IEEE Transactions on Affective Computing* (2019). DOI: 10.1109/TAFFC.2019.2917047
ieeexplore.ieee.org/document/8716568