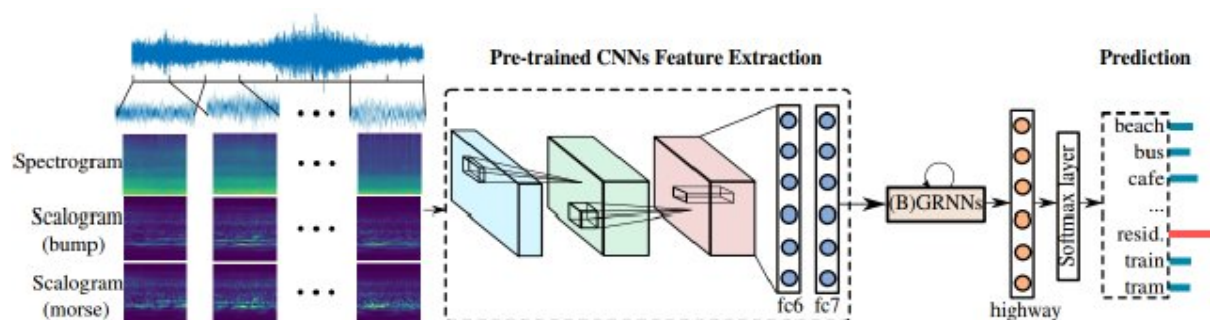# Researchers develop more comprehensive acoustic scene analysis method

April 10 2018

Predictions of sounds were achieved by an improved method developed by an international team of researchers. Credit: *IEEE/CAA Journal of Automatica Sinica*

Researchers have demonstrated an improved method for audio analysis machines to process our noisy world. Their approach hinges on the combination of scalograms and spectrograms—the visual representations of audio—as well as convolutional neural networks (CNNs), the learning tool machines use to better analyze visual images. In this case, the visual images are used to analyze audio to better identify and classify sound.

The team published their results in the journal *IEEE/CAA Journal of Automatica Sinica* (*JAS*), a joint publication of the IEEE and the Chinese Association of Automation.

"Machines have made great progress in the analysis of speech and music, but general sound analysis has been lagging a big behind—usually, mostly isolated sound 'events' such as gun shots and the like have been targeted in the past," said Björn Schuller, a professor and chair of Embedded Intelligence for Health Care and Wellbeing at the University of Augsburg in Germany, who led the research. "Real-world audio is usually a highly blended mix of different sound sources—each of which have different states and traits."

Schuller points to the sound of a car as an example. It's not a singular audio event; rather different parts of the car's parts, its tires interacting with the road, the car's brand and speed all provide their own unique signatures.

"At the same time, there may be music or speech in the car," said Schuller, who is also an associate professor of Machine Learning at Imperial College London, and a visiting professor in the School of Computer Science and Technology at the Harbin Institute of Technology in China. "Once computers can understand all parts of this 'acoustic scene', they will be considerably better at decomposing it into each part and attribute each part as described."

Spectrograms provide a visual representation of audio scenes, but they have a fixed time-frequency resolution, that is the time at which frequencies change. Scalograms, on the other hand, offer a more detailed visual representation of acoustic scenes than spectrograms, for instance, acoustic scenes like the music or the speech or other sounds in the car now can be better represented.

"There are usually multiple sounds happening in one scene so... there should be multiple frequencies and they change with time," said Zhao Ren, an author on the paper and a Ph.D. candidate at the University of Augsburg who works with Schuller. "Fortunately, scalograms could solve

this problem exactly since it incorporates multiple scales."

"Scalograms can be employed to help spectrograms in extracting features for acoustic scene classification," Ren said, and both spectrograms and scalograms need to be able to learn to continue improving.

"Further, pre-trained neural networks build a bridge between [the] image and audio processing."

The pre-trained neural networks the authors used are Convolutional Neural Networks (CNNs). CNNs are inspired by how neurons work in animals' visual cortex and the artificial neural networks can be used to successfully process visual imagery. Such networks are crucial in machine learning, and in this case, helping improve the scalograms.

CNNs receive some training before they're applied to a scene, but they mostly learn from exposure. By learning sounds from a combination of different frequencies and scales, the algorithm can better predict the sources and, eventually, predict the result of an unusual noise, such as a car engine malfunction.

"The ultimate goal is machine hearing/listening in a holistic fashion... across speech, music, and sound just like a human being would," Schuller said, noting that this would combine with the already advanced work in speech analysis to provide a richer and deeper understanding, "to then be able to get 'the whole picture' in the audio."