BodyScope: A Wearable Acoustic Sensor for Activity Recognition

Summary

The paper describes a wearable acoustic sensor BodyScope which uses the sounds produced in user's mouth and throat and can differentiate between 12 different activities like eating, drinking, speaking, laughing etc. The device is a modified Bluetooth headset embedded with a unidirectional microphone and a chest piece that amplifies the sounds produced. In the testing phase, the device can differentiate between eating, drinking, speaking and laughing with an accuracy of 71.5%.

The paper points out that even when the user is still they get a sound which is the sound of the blood pumping through the carotid artery. Although it could be used for health purposes but the sound is lost in the sound of other activities. They are able to differentiate between normal and deep breaths and as such can identify if the user is involved in some sort of physical activity. While eating as well, they are able to differentiate between the sounds of eating a cookie and bread. Most importantly they are able to clear identify the sound of swallowing of the food.
In the case of drinking, the sound of gulp reaches as high as 1500Hz and can be easily differentiated from other sounds. They also observed that gulping sound is stronger than the swallowing sound. They also differentiate between a hot and cold drink by using the sound of sipping.
Next, they tried to differentiate between whispering and speaking sound. In the case of speaking, there are clear harmonics in the sound of humans which can be used to differentiate between eating and drinking. But in the case of whispering the sound harmonics disappear along with a decrease in sound intensity.

In the case of laughing, sighing and coughing they were able to identify laughing dues to high intensity but chuckle was often misinterpreted with a deep breath as the power distribution of the two were similar. Furthermore, sighing and coughing were also pretty difficult to identify as the sound pattern of sighing is similar to that of drinking the hot liquid. On the other hand, spectrum of sighing was often confused with that of drinking the cold drink.

In order to capture all the information, the sampling rate was kept at 22050Hz thus covering up to 11025Hz which is double of what was actually required. Three different domain features were used for machine learning classification: time, frequency, and cepstral.

Apart from this, they used 3 different algorithms namely: SVM, Naive Bayes, and 5-nearest neighbor. The last two were computationally less expensive as compared to SVM but the accuracy was not as good as SVM. These can be used for real-time classification. For SVM, Radial Basic Function(RBF) was used as the kernel and one-against-one strategy was used for classification. Naive Bayes classifier employed a Gaussian Mixture Model and Nearest neighbor employed Euclidian distance without weight as its classifier.

The participants were asked to sit relaxed and then asked to perform certain activities as they would in their natural environments. The classifiers were trained with the data collected based on the feature mentioned earlier for each of the classification. The training and testing employed two protocols namely: Leave-one-participant-out cross-validation and Leave-one-sample-per-participant-out cross validation. As apparent by the name Leave-one-participant-out trains the classifier in a user-independent manner resulting in better performance in the case of Leave-one-sample-per-participant-out. Leave-one-participant-out showed an accuracy of approximately 50% for all three machine learning techniques whereas Leave-one-sample-per-participant-out's accuracy was close to 75%.

Critique

In my opinion, the paper is very well crafted and describes the whole procedure of repeating the steps of the experiment really very well so that anyone can recreate the experiment and check the results or can improve upon it. They also describe the various feature used to differentiate between various activities based on their sounds with graphs which makes it pretty easy to visualize and understand the differences.

They experimented with the various locations of the device and realized the wearing it around the Larynx can interfere with certain activities like speaking, drinking and adapted their device to prevent this from happening. They actually took care of the feasibility of wearing such a device and modified it so that it is comfortable for the participants giving up on the accuracy of the results by doing so.

Also, the confusion matrix provided by them consisted of actual numbers instead of just the colors to indicate the level of confusion. This provided with a much clearer idea of the level of confusion between the activities as opposed to the one which just contains various shades of grey to convey this information.

But having said this there were some shortcomings as well:

1. The accuracy presented by them was from the Leave-one-sample-per-participant-out cross validation (approx. 75%) which was way higher than that of the Leave-one-participant-out cross validation (approx. 50%). The Leave-one-sample-per-participant-out seems to be overfitting in my opinion.
2. They start off with classifying 12 activities initially but the final testing and results are for classifying just 4 simple activities.