Introduction

Many audio signals can be captured from inside the occluded ear such as speech [1], biosignals [2] and other events (listed in Table 1) studied in the present research. The classification of nonverbal human audio events could have many applications such as health monitoring or silent interfaces for human-machine interactions.

Methodology

To classify the events, different features, feature structure and machine learning algorithms were compared.

Feature Used:
- Mel-Frequency Cepstral Coefficient (MFCC) with delta and acceleration + Zero Crossing Rate (ZCR).
- Auditory-Inspired Amplitude Modulation Features (AAMF).

Feature Structures:
- Framed: Information in one frame of audio (50 ms).
- Contextual: Information of one audio frame with the N previous and following frames concatenated.
- Concatenated: All the sample’s frame concatenated in one vector.

Classifiers Used:
- Gaussian Mixture Models (GMM)
- Support Vector Machines (SVM)
- Multilayer Perceptron (MLP)

Database:
- 25 participants
- 11 classes
- Samples of 400 ms
- Total of 3027 samples

Industrial plants are the main environment where the intra-aural device is currently used. To train our algorithm to be resistant to this kind of noise, a noisy dataset was created by adding plant noise from the NOISEX-29 database to our existing samples at 10 dB SNR.

Results

The results were obtained by doing a 10-fold cross validation over our database and the accuracy was calculated by averaging the score of each of the ten folds.

- The GMM classifier has the best accuracy compared to the other classifiers. The best results were obtained using the contextual framing with N = 4.
- When looking at the confusion matrices, two groupings can be clearly seen where a lot of confusion happens between similar classes:
  - MF, ce and of, which are all clicking sounds coming from the mouth.
  - Sn and cl, which are all clicking sounds created when closing the eyes.
- As seen in Table 2, the addition of the AAMF features increases the accuracy when testing on the noisy dataset.

Conclusions & Future work

Classification of nonverbal events was achieved with 75.5% accuracy across 11 classes with the clean database and 72.8% when testing on the noisy database. This score was achieved by adding the AAMF feature to the MFCC, which suggest that the nature of nonverbal and the way we extract their features events should be investigated more closely.

Other machine learning techniques were tested after writing the conference paper. A better accuracy of 77.1% on clean dataset was achieved by using a bag-of-features approach with the MFCC features and contextual framing. The focus is now on trying to find features that represent better our nonverbal human produced sounds.

References