Comparison of Baseline System with Perceptual Linear Predictive Feature Using Neural Network for Sound Event Detection in Real Life Audio

Khizer Feroze

Institute of Space Technology Islamabad
Pakistan
khixr999@gmail.com

Dr. Abdur Rahman Maud

Institute of Space Technology Islamabad
Pakistan
armaud@gmail.com

ABSTRACT

For sound event detection of polyphonic sounds, we compare the performance of perceptual linear predictive (PLP) feature with Mel frequency cepstral coefficients (MFCC) using neural network classifier. The results are further compared with the performance of the baseline system given by DCASE 2017 (task 3). Our results show that using PLP based classifier, individual error rate (ER) for each event is improved compared to the baseline system. For car event, ER is improved by 10%, for large vehicle event 23%, for people walking event 26% and some improvements are also observed in other events.

Index Terms— Sound event detection, perceptual linear predictive, neural networks, DCASE

1. INTRODUCTION

This paper deals with a detection of audio scenes. In which the most important techniques of digital signal processing and machine learning are used. Where the sound event detection (SED) has many real life applications like in a scene recognition for robotics [1], automatic surveillance of acoustic activities [2] and in video retrieval using audio [3]. Normally the most of work which was previously done by different research groups are on monophonic sounds. Monophonic sound means there is no overlapping in different class of sounds. So it is quite easy to get good performance on monophonic sounds. Where as in task 3 dataset contains polyphonic sounds where all dataset contains overlapping sounds. Due to interaural time difference human can easily recognized the polyphonic sounds in real time. Concept of time difference of arrival (TDOA) is used by Adavanne [4] to make a system look like human auditory system, he use both spatial and harmonic features with long short term memory recurrent neural network classifier and he got top position in DCASE 2016. The gated neural network is used with Mel Frequency Cepstral Coefficient (MFCC) and preprocessed output data in such a way that if the output event is smaller than 0.1 sec then neglect it and if there are some constant gaps of 0.1 second then ignore it, and they also got better performance as compare to last year baseline [5]. MFCC is also use by Toan H. Vu [6] but they use this feature with BIRRN classifier with 50 hidden units. Deep neural network and GMM is used with MFCC by Qiuqiang Kong [7] and he concluded that DNN gives better result than GMM.

2. DATASET

It is based on a problem given by research group of Tempere University named as DCASE 2017(Task 3). DCASE announced four challenges in which all ones are of classification or detection of different audio scenes. Labeling is a major issue while making any classifier model for overlapping intervals. This task contains six type of classes. On the other side to get the good performance it is also an issue that there is no control over overlapping of sounds, there may be no overlapping events or at some place three classes or six classes are overlapped and the person who is recoding and labeling that sounds labeled that sounds on his own guess. Whereas each class is further divided into different classes like "car engine running" or when "car passing by" and also when "car is running" that all type of sounds are labeled as a single class called "car" other classes are also labeled like car. Sounds (dataset) are recorded in different streets on different atmospheres in Finland and each sound file is 3 to 5 minutes long. The two type of datasets are given by DCASE one is named as development and the other one called evaluation dataset. Development dataset is further divided into four folders. Each folder contain training and test file which is to be initially use to get the good performance. After it this all development dataset is for training and the evaluation dataset is for testing.

3. TASK EVALUATION

Task 3 is evaluated on segment-base error rate (SBER). One second of frame length is selected for this task to evaluate performance. Formula to calculate performance is:-

$$\text{Error rate} = \frac{\sum S(k) + \sum D(k) + \sum I(k)}{\sum N(k)}$$

Where,
$S(k) = \min(FN(k), FP(k))$
$D(k) = \max(0, FN(k) - FP(k))$
$I(k) = \max(0, FP(k) - FN(k))$

N=Number of sound events in a ground truth
F-score is also use to evaluate the performance. Its formula is:-

$$\text{F} - \text{Score} = \frac{2.Precision.Recall}{Precision + Recall}$$

4. PROPOSED ALGORITHM

We have used Perceptual Linear Predictive (PLP) as a feature. PLP is firstly computes by H. Hermansky [8]. It is widely used for speaker recognition. PLP outputs two type of things, that are spectral and cepstral coefficients. We have used only spectral coefficients of it. Two audio recorder sensors are used to record sounds. We take both recorded sounds and find there features separately and make them a single vector after combining them. Figure 1 shows Feature Exration.

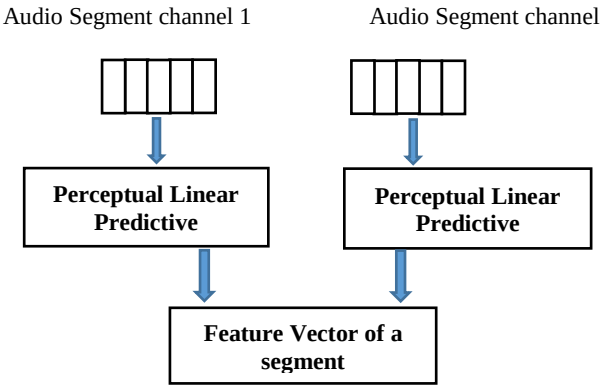

Figure 1. Flow diagram of Feature Extraction of each segment using both channels

We have not use multilabing for this task. We separatly label each class by using annotations given by DCASE. Mainly people use neural network as a classifier, as we discussed in introduction section. We also used BINN where we took 40 hidden states. We break each training folds recordings into 0.1 sec frame and further divide each frame into 25ms sub frame and extract features of each frame where we take hop size of 10ms. Labeling is also done for every 0.1 sec frame by rounding off the give annotation data. Both labels and features are given to that three classifiers. For test folders we applied the same approach. Figure 2 shows Training process.

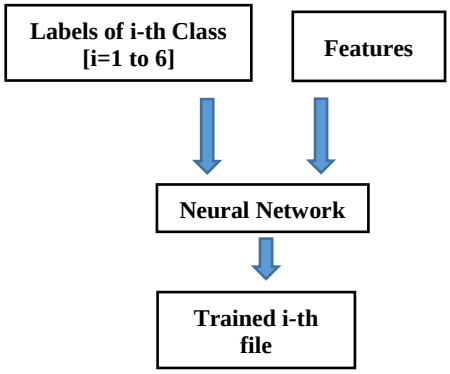

Figure 2. Flow diagram of Training extracted features using neural network

We first dilate the output results (which we obtained after neural network classification) where we take dilation mask of 0.3 seconds. Purpose of dilation is to close zeros if they lie within 0.3 seconds. We further erode the results which we get after dilation and we take erosion mask of 0.2 seconds. We have done both erosion and dilation because it is not possible that any class happen only for just 0.2 or 0.3 second. Figure 3 shows Testing.

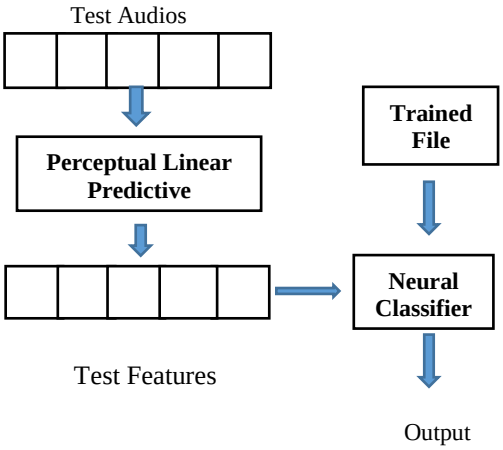

Figure 3. Testing flow diagram

5. CROSS VALIDATION RESULTS

In baseline MFCC log energies are used as a feature with multilayer perceptron neural network classifier. We proposed that using PLP spectra is better than MFCC. We did their comparisons by using bi-neural network. Results prove that both error rate and F-Score improved by using PLP spectra features. Overall ER 0.84 and F-score 35 % obtained by using MFCC where overall ER 0.76 and F-Score 47% obtained by using PLP. They were not improved as such due to a false positives and false negatives occurring on other events at same segment but it might be improved by using both baseline and our system in parallel or use PLP with baseline classifier. Whereas individual ER is concerned it gets improved by using PLP with Neural network. Results of cross validation are listed in Table 1.

Events	Baseline		Neural Network (Using MFCC)		Neural Network (Using PLP)	
	F-Score	Error rate	F-Score	Error rate	F-Score	Error rate
Children	0	1.35	4	1.01	0	1.21
Car	74.1	0.57	55	0.51	63	0.47
Large vehicle	50.8	0.9	27	0.80	38	0.67
People walking	55.6	0.84	10	0.93	48	0.58
People speaking	18.5	1.25	0	0.99	10	0.92
Brakes squeaking	4.1	0.98	0	1	7	0.95

Table 1. Cross Validation results of each folds

6. CONCLUSION

The two evaluation methods were given by DCASE which were F-score and error rate. F-Score is calculated by using TP, FP and FN. Error rate is also compute through TP, FP and FN but the major difference in error rate and F-score is N (number of events occur in a segment). Due to independency of F-Score with N it will not always happen that if error rate is getting improved than F-score will also get improved and it will also not always happen that by improving individual error rate overall error rate also get improved because error rate is dependent on N and it also effected by other event's FP and FN occurring on that segment. As far as our algorithm is concerned, by using PLP it is concluded that error rate gets reduced and giving good performance as compare to MFCC log energies. If we compare our system we concluded that individual error rate can be improved by using PLP feature and neural network as classifier. So if we use proposed feature with a baseline system it may give more improved results.

7. REFERENCES

[1] A. Harma, M. F. McKinney, and J. Skowronek. 2005 . "Au-tomatic surveillance of the acoustic activity in our living environment." *IEEE International Conference on Mul-timedia and Expo (ICME).*

[2] Hermansky, Hynek. n.d. "Perceptual linear predictive (PLP) analysis of speech."

[3] M. Zöhrer, F. Pernkopf. 2016. "ated recurrent networks applied to acoustic scene classification and acoustic event detection ." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* . IEEE. 1304 - 1314.

[4] Qiuqiang Kong, Iwnoa Sobieraj, Wenwu Wang, Mark Plumbley. 2016. "Deep neural network baseline for DCASE challenge ."

[5] Radha, N. 2016. "Video retrieval using speech and text in video." *Inventive Computation Technologies (ICICT), International Conference on.* Coimbatore, India: IEEE.

[6] S.Chu, S.Narayanan, C.C.J.Kuo, and M.J.Mataric,. 2006. "Where am I? Scene recognition for mobile robots using audio features." *Int. Conf. Multimedia and Expo (ICME).* IEEE. p. 885.

[7] Sharath Adavanne, Giambattista Parascandolo, Pasi Pertila, Toni Heittola, Tuomas Virtanen. 2017. "Sound event detection in multichannel audio using spatial and harmonic features." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* New Orleans, LA, USA: IEEE Conference Publications. 771,775.

[8] Toan H. Vu, Jia-Ching Wang. n.d. "Acoustic scene and event recognition using recurrent neural networks."