LARGE-SCALE WEAKLY LABELLED SEMI-SUPERVISED CQT BASED SOUND EVENT DETECTION IN DOMESTIC ENVIRONMENTS

Technical Report

Rojin Raj, Shefali Waldekar, Goutam Saha

Electrical and Electronics Communication Department
Indian Institute of Technology Kharagpur
West Bengal,India
rojinraj@iitkgp.ac.in,{shefaliw,gsaha}@ece.iitkgp.ernet.in

ABSTRACT

This paper proposes a constant quality transform based input feature for baseline architecture to learn the start and end time of sound events (strong labels) in an audio recording given just the list of sound events existing in the audio without time information (weak labels). This is achieved using constant quality transform coefficients as input feature for convolutional recurrent neural network. The proposed method is a contribution to the challenge of detection and classification of acoustic scenes and events (DCASE 2018, Task 4) and evaluated on a publicly available dataset from youtube with 10 sound event classes. The method achieves the best error rate of 1.48 and F-score of 14.55 %.Based on the results obtained using a CPU based system there is a decrease of 7.5 % in case of error rate and increase of 11.5 % in case of F-score as compared to baseline results.

Index Terms— constant quality transform,sound event detection, weak labels, deep neural network, CNN, GRU

1. INTRODUCTION

Understanding the sounds of everyday life has received great attention in recent years due to its practical applications such as the hearing impaired, smart cars and smart appliances. Among others, Sound Event Detection (SED) [1] is a particularly challenging task because it predicts not only possible descriptive words of environment sound but also their start and end times. Most SED systems are based on hard annotated data where both event classes and their time-stamps are present. However, it is time consuming and expensive to construct a large dataset with such labels. This motivates the community to explore weakly labelled dataset. Weak labels only need to determine whether an event in the recording is present or absent. This greatly reduces the resource needed for collecting such dataset.

2. THE PROPOSED SYSTEM

2.1. Input feature

The input feature used in the proposed system is constant quality transform coefficients [2, 3].Figure 2 shows the structure of the proposed sytem. Constant-Q transform (CQT) here refers to a technique that transforms a time-domain signal into the time frequency domain so that the center frequencies of the frequency bins are geometrically spaced and their Q-factors are all equal. In effect, this means that the frequency resolution is better for low frequencies and the time resolution is better for high frequencies.

From auditory perspective, the frequency resolution of the peripheral hearing system of humans is approximately constant-Q over a wide range from 20kHz down to approximately 500Hz, below which the Q-values get progressively smaller. From perceptual audio coding, we know that the shortest trans-form window lengths have to be of the order 3ms in order to retain high quality, whereas higher frequency resolution is required to carry out coding at low frequencies. All this is in sharp contrast with the conventional discrete Fourier transform (DFT) which has linearly spaced frequency bins and there-fore cannot satisfy the varying time and frequency resolution requirements over the wide range of audible frequencies. CQT has not widely replaced the DFT due to the computational intensity.

The CQT transform $X^{CQ}(k,n)$ of a discrete time-domain signal x(n) is defined by

$$X^{CQ}(k,n) = \sum_{j=n-\lfloor N_k/2 \rfloor}^{n+\lfloor N_k/2 \rfloor} x(j) a_k^*(j-n+N_k/2) \qquad (1)$$

where k = 1, 2,. . . ,k indexes the frequency bins of the CQT, $a_k^*(n)$ denotes the complex conjugate of $a_k(n)$. The basis functions $a_k(n)$ are complex-valued waveforms, also called time-frequency atoms, and are defined by

$$a_k(n) = \frac{1}{N_k} w\frac{n}{N_k} exp\left[-2\pi\, n\frac{f_k}{f_s}\right] \qquad (2)$$

where f_k is the centre frequency of bin k, f_s denotes the sampling rate, and $w(t)$, is a continuous window function (for example Hann or Blackman window), sampled at points determined by t. The window function is zero outside the range $t \,\epsilon\, [0, 1]$.The window lengths $N_k \,\epsilon\, R$ in (1),(2) are real-valued and inversely proportional to f_k in order to have the same Q-factor for all bins k.

In the CQT considered here, the centre frequencies f_k obey

$$f_k = f_1 2^{\frac{k-1}{B}} \qquad (3)$$

where f_1 is the centre frequency of the lowest-frequency bin, and B determines the number of bins per octave. In practice, B is the

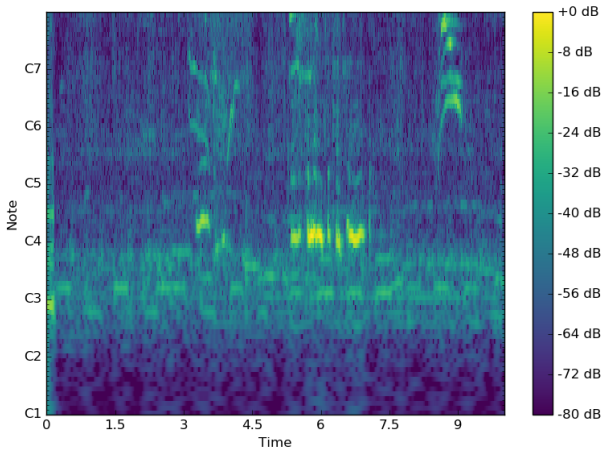

Figure 1: A sample CQT spectrum.

most important parameter of choice when using the CQT, because it determines the time-frequency resolution trade-off of the CQT. Fig 1 shows a sample CQT spectrum image.

2.2. Neural network architecture

The baseline system is used with minor modifications[4, 5]. The system is based on two convolutional recurrent neural network (CRNN)[6, 7] using 84 CQT coefficients magnitudes as features. 10 seconds audio files are divided in 431 frames. Using these features, we train a first CRNN with three convolution layers (84 filters (3x3), max pooling (4) along the frequency axis and 30 % dropout), one recurrent layer (84 Gated Recurrent Units GRU with 30 % dropout on the input), a dense layer (10 units sigmoid activation) and global average pooling across frames. The system is trained for 100 epochs (early stopping after 15 epochs patience) on weak labels (1578 clips, 20 % is used for validation). This model is trained at clip level (file containing the event or not), inputs are 431 frames long (10 sec audio file) for a single output frame. This first model is used to predict labels of unlabelled files (unlabel in domain, 14412 clips). A second model based on the same architecture (3 convolutional layers and 1 recurrent layer) is trained on predictions of the first model (unlabel in domain, 14412 clips; the weak files, 1578 clips are used to validate the model). The main difference with the first pass model is that the output is the dense layer in order to be able to predict event at frame level. Inputs are 431 frames long, each of them labelled identically following clip labels. The model outputs a decision for each frame. Pre-processing (median filtering) is used to obtain events onset and offset for each file. The baseline system includes evaluations of results using event-based F-score [8] as metric.

3. EVALUATION RESULTS

Evaluation is done based on the F-score (F) and error rate (ER)[8] and results are generated using a CPU based system (Intel Xeon 2.1GHz, Gallium 0.4 LLVM, 8GB RAM).The results using CQT features and log mel energy features as input is shown in Table 1 . (Please note the results are based on old dataset).There is a decrease of 3.8 % in error rate as compared to baseline(as given in

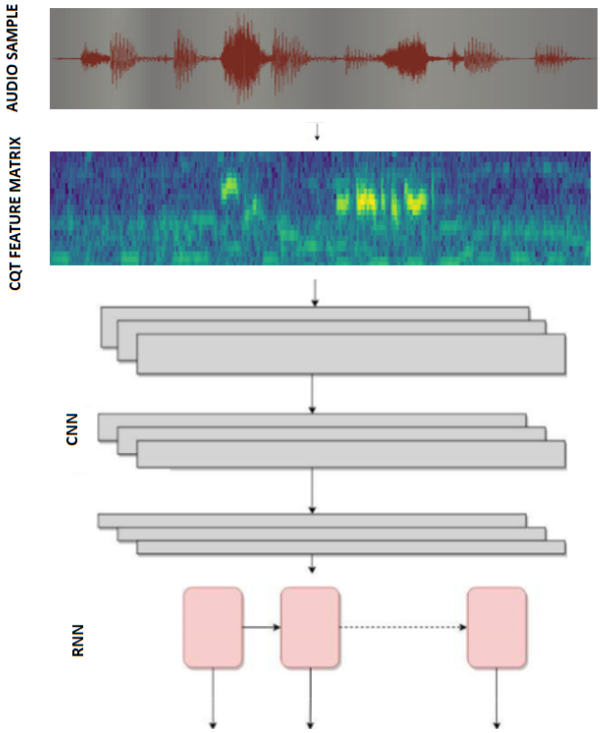

Figure 2: Baseline system

DCASE website) and there is a decrease of 7.5 % as compared to baseline(obtained by running in our system).Figure 3 shows a comparison of the baseline system with the proposed sustem in terms of F-score and error rate.The proposed system gives better results for the blender, dog, electric shaver, frying and running water event classes with regard to F-score. Considering the error rate for the event classes, blender, dishes, dog, electric shaver, frying and vaccum cleaner ,the proposed system qualify as a bettter system as compared to the baseline system.

The macro average results for F-score and ER (error rate) are 14.55 % and 1.48 respectively which is better than the baseline results (F-score-14.68 % (according to DCASE website),13.04 %(our system) , ER-1.54 (according to DCASE website),1.6 (our system)).

4. CONCLUSION

The results point to the fact that when compared to log Mel features CQT features performs better with the baseline system. The disadvantage with the CQT is high computational intensity as compared to the the mel features.

5. REFERENCES

[1] S. Adavanne, G. Parascandolo, P. Pertil{"a, T. Heittola, and T. Virtanen, "Sound event detection in multichannel audio using spatial and harmonic features," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*. Tampere University of Technology. Department of Signal Processing, 2016, pp. 6–10.

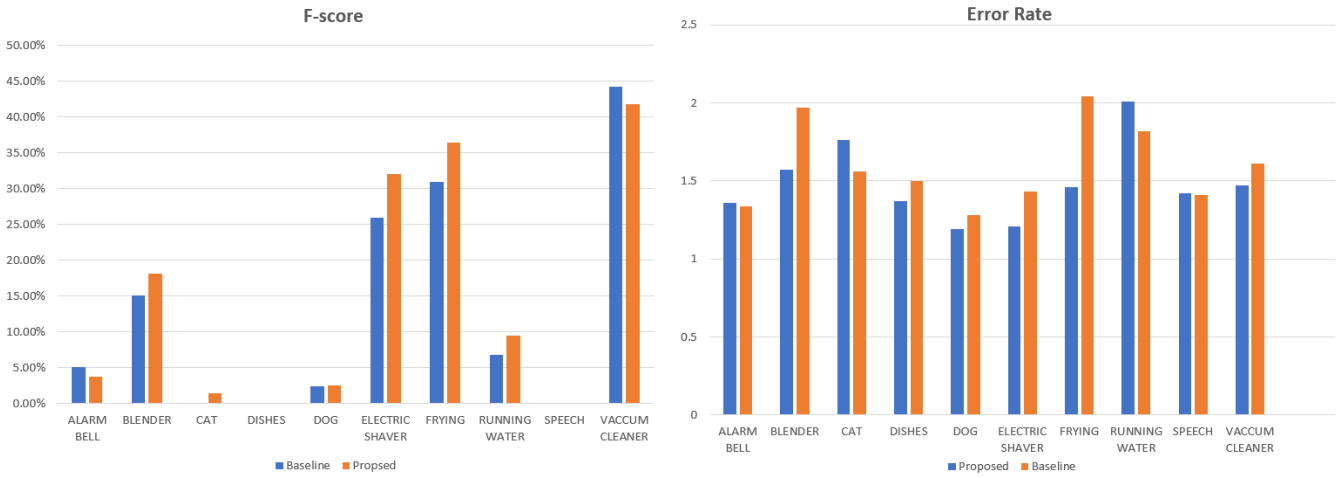

Figure 3: Results comparison.

Table 1: Event based comparison of baseline and proposed system

	PROPOSED		**BASELINE**	
EVENT LABEL	F-SCORE	ERROR RATE	F-SCORE	ERROR RATE
ALARM BELL	3.80 %	1.36	5.10 %	1.34
BLENDER	18.2 %	1.57	15.1 %	1.97
CAT	1.40 %	1.76	0.00 %	1.56
DISHES	0.00 %	1.37	0.00 %	1.50
DOG	2.60 %	1.19	2.40 %	1.28
ELECTRIC SHAVER	32.0 %	1.21	25.90 %	1.43
FRYING	36.4 %	1.46	31.0 %	2.04
RUNNING WATER	9.50 %	2.01	6.80 %	1.82
SPEECH	0.00 %	1.42	0.00 %	1.41
VACCUM CLEANER	41.8 %	1.47	44.2 %	1.61

[2] C. Schörkhuber and A. Klapuri, "Constant-q transform toolbox for music processing," in *7th Sound and Music Computing Conference, Barcelona, Spain*, 2010, pp. 3–64.

[3] J. C. Brown and M. S. Puckette, "An efficient algorithm for the calculation of a constant q transform," *The Journal of the Acoustical Society of America*, vol. 92, no. 5, pp. 2698–2701, 1992. [Online]. Available: https://doi.org/10.1121/1.404385

[4] http://dcase.community/challenge2018.

[5] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. Parag Shah, "Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments," July 2018, submitted to DCASE2018 Workshop. [Online]. Available: https://hal.inria.fr/hal-01850270

[6] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," *CoRR*, vol. abs/1710.00343, 2017. [Online]. Available: http://arxiv.org/abs/1710.00343

[7] K. Choi, G. Fazekas, M. B. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," *CoRR*, vol. abs/1609.04243, 2016. [Online]. Available: http://arxiv.org/abs/1609.04243

[8] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, 2016.