JOINT ACOUSTIC AND SUPERVISED INFERENCE FOR SOUND EVENT DETECTION

Technical Report

Sangwook Park, Ashwin Bellur, Sandeep Kothinti, Masoumeh Heidari Kapourchali, Mounya Elhilali

Johns Hopkins University, Dept. of Electrical and Computer Engineering, Baltimore, USA,
{spark190, abellur1,skothin1,mheidar1,mounya}@jh.edu

ABSTRACT

This is a technical report about a sound event detection system for the task 4 of DCASE2020. The purpose of a sound event detection is to find event class label as well as its time boundaries. To achieve this purpose, we considered several methods such signal enhancement and event boundary detection, and built five systems by integrating these methods with supervised system trained by using Mean Teacher model. In particular, we estimate event boundaries of weakly labeled data by performing a event boundary detection. Then, we used the estimated strong label in training the supervised system. In addition, we adopt a fusion method by calculating weighted averaging posterior over the five outputs from each individual system. In experiments with validation set, we found that a final result of our system shows an improvement about 11 % in class averaging f-score compared to a baseline performance.

Index Terms— Separation, enhancement, salience detection, predictive coding, data augmentation, posterior fusion

1. INTRODUCTION

This report describes a Sound Event Detection (SED) system which is submitted to the task 4 in the Detection and Classification of Acoustic Scenes and Events (DCASE) 2020 challenge. From the task description, 10-domestic sound events are considered as the target events such Alarm/bell/ringing, Blender, Cat, Dishes, Dog, Electric shaver/tooth brush, Frying, Running water, Speech, and Vacuum cleaner. And, the system is designed to recognize not only event class but also time boundaries given that multiple events can be present in an audio recordings. To develop a SED system, we can use three types of dataset such strong labeled including both event class and its time boundaries, weakly labeled having event label only, and unlabeled dataset which has no information about the events.

Our approach is composed of five subsystems designed by integrating sound enhancement, event boundary detection, and supervised system (Fig. 1). In the first part of our work, we trained a supervised system based on Mean Teacher model [1] with three types of dataset and estimated event labels by performing a salience detection method on weakly labeled data. In terms of pre-processing, we applied two methods for speech separation and background noise suppression to audio recordings. After performing these preprocessing, the supervised systems were applied for posterior calculation. Note that estimate labels for all classes were used in training version I while the estimate for Dishes class was excluded in training version II due to ( ). In system 03, a speech separation method outputs two audio clips for speech and non-speech classes. The posteriors for these audio clips are individually calculated in

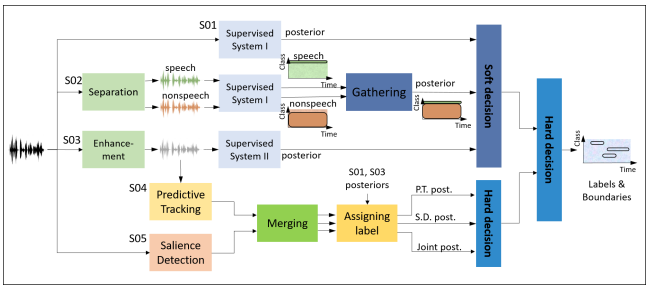

Figure 1: Overall System Architecture

the following supervised system. Then, the output is constructed by gathering a speech posterior in the output of speech audio and non-speech posteriors in the output of non-speech audio. In parallel, the salience detection used in event label estimation for weakly labeled data and a predictive tracking are performed to find event interval in an audio recording. Then, we assigned target event class on each interval based on the output of supervised systems. The final result of our approach is determined by performing soft and hard decision with the outputs of the five systems in fusion.

2. METHODS

2.1. DESED enhancement

A distributed CNN with attentional mechanisms, first proposed for music source separation [2], was trained in this instance to enhance the DESED foreground. This system is denoted as Enhancement in the system architecture in Fig. 1. A mixture audio consisting of DESED and FUSS training set was created similar to the source separation baseline. A distributed CNN system was trained using this data to attend to either FUSS or DESED foreground. The system consists of two map memories and two anchor memories [2], representing the DESED foreground and FUSS auditory scenes respectively. Given a mixture input, the output of the system when *attending* to DESED foreground was used as the enhanced DESED waveform on which supervised event detection was performed.

2.2. Separating speech and non-speech

The speech class is the most frequently encountered class in this event detection task. Also, the presence of speech overlapping presents challenges to the supervised system in detecting and classifying the overlapping events. Therefore, another distributed CNN with attention [2] was trained using the synthetic data for the pur-

pose of separating speech class from the non-speech classes of the challenge. In this case, the two anchor memories and the two map memories, represent the speech class and the non-speech classes. Given a waveform, the system generates two outputs, one on attending to the speech class only and the other other on attending to the non-speech class alone. The supervised system then estimates posteriors on each of this waveforms. Only the speech posteriors are retained from the speech waveform while the posteriors for the remaining nine classes are retained from the non-speech waveform (indicated as *gathering* in the system architecture).

2.3. Salient Event Detection

Salience is an acoustic driven attribute of audio objects that makes them stand out from the other objects in the scene. In this work, salience is used to detect object boundaries with the assumption that salient onsets correspond to event onsets in the scene. Salience detection is unsupervised and consists of a feature extraction stage and a boundary detection stage. These stages are described below.

Feature extraction: A variational auto-encoder(VAE) [3] trained to separate the input feature stream into different modulation frequencies is used as a feature extractor. By separating the features based on modulations, the VAE is expected to separate objects in the embedding space. For this work, we used 3 non-overlapping bands with band-pass frequencies [0, 2.4Hz], [2.4Hz, 6Hz] and [6Hz, 12Hz]. The embedding space is trained with gaussian priors with unit variance and means restricted to the modulation frequencies mentioned above. Embeddings corresponding to the means of the posterior gaussians are used for the boundary detection stage.

Boundary detection: For detecting the boundaries from the embeddings, we used a similar method to that used in [4]. Briefly, a derivative following by smoothing in time and averaging across units is used as measure of temporal salience of the input audio. Peaks in the salience are considered event onsets and the corresponding offsets are computed by thresholding the short-term energy. In contrast to the previous work, we chose the band that had the best performance for the classes in a given scene which were determined by the supervised system posteriors.

2.4. Predictive tracking

Linear predictive coding (LPC) [5] is used to detect event boundaries by tracking acoustic features. A set of fifteen temporal and spectral acoustic features introduced in [6] is extracted from the DESED dataset audio signals producing fifteen timeseries. An LPC is applied on each timeseries for online predictive tracking. The coefficients of LPCs are optimized by minimizing the squared prediction error using a normalized stochastic gradient descent approach. A failure in accurate prediction (i.e. prediction error exceeding the threshold) is considered to be a change in the acoustic events. The threshold is fixed based on the best results from the DESED synthetic dataset. The prediction error for each individual feature is used as a score to detect onset of the events. The scores among features are integrated using a weighted average where the weights are learned by applying a logistic regression algorithm on the synthetic dataset. A change in data (feature values) distribution indicates the offset of events.

2.5. Post-processing for unsupervised methods

The post-processing for unsupervised methods is composed of Merging and Assigning event class. In merging, joint intervals are additionally extracted by applying AND rule in between outputs of salience detection and predictive tracking. And, the joint interval is going to the next block with the outputs of the unsupervised methods. Because these outputs have no information about event class, class-wise posteriors obtained by supervised systems are used to assign event classes to each detected interval. First, correlations between event-present probabilities within the interval and posteriors for each class are calculated. After normalization by the duration, event classes are assigned to the interval if a normalized correlation is larger than 0.5. Note that an estimate interval could be have multiple class labels if the correlations in those classes are larger than 0.5.

2.6. Fusion

A final result is produced by integrating the outputs of each individual system. In the first stage, supervised and unsupervised methods are separately integrated by calculateing weighted averaging posterior of the outputs. For supervised methods, class-wise f-scores of each system are defined as the weights and the weight is normalized by summation in each class [7]. Thus, we denote this integration as soft-decision in Fig 1. A supervised posterior is calculated by using this soft weight. On the other hand, unsupervised posterior is obtained by gathering posterior for a class from a system that shows the best f-score in the class. This integration is denoted as hard-decision in the figure. Similarly, extra hard decision is performed again in between the supervised and unsupervised posteriors. Once a final posterior is calculated, we applied a threshold (0.3) to posterior for each class for detecting event intervals.

3. EXPERIMENTS

3.1. Database

To develop and evaluate our approach, we used the database for sound event detection in DCASE2020. Basically, a supervised system is trained with not only strong labeled data (synthetic data) but also weakly labeled and unlabeled data (real data) which are subsets of the DESED [8, 9]. In the enhancement system, the FUSS [10, 11, 12] and DESED mixture was used to train the system, whereas the synthetic data was used to train the system that separated speech and non-speech. For developing predictive tracking system, synthetic training set is used to optimize threshold.

3.2. Experiment setting

Supervised system II differes from I in two aspects. Firstly, synthetic data used for the strong label based loss is augmented by weakly labeled data with a single class. The weakly labels were converted to strong labels by computing the boundaries using salience as described in section 2.3. Secondly, the exponential softmax attention is replaced with a linear softmax.

A distributed CNN with attention architecture same as the one detailed in [2] was employed for the system enhancing DESED foreground (section 2.1) and as well as for the system separating speech and non-speech (section 2.2).

VAE used in the salience based event detection was trained using weakly labeled and unlabeled data of the real data. A 128 di-

Table I. Performance evaluation with DCASE validation set: class-wise f-score [%]

	DCASE	Supervised methods			Unsupervised methods		
	Baseline	S01	S02	S03	S04	S05	S04&S05
Alarm_bell_ringing	37.64	**45.72**	43.70	44.47	52.22	40.77	**56.94**
Blender	30.37	38.69	**39.13**	37.80	**44.44**	39.36	41.76
Cat	43.26	40.45	**44.20**	40.20	**40.70**	28.30	37.52
Dishes	24.78	**23.76**	14.32	21.06	**33.07**	16.03	19.73
Dog	20.64	**30.77**	21.47	25.97	**38.59**	28.60	29.56
Electric_shaver_toothbrush	35.56	**49.30**	46.62	51.20	27.21	**40.00**	28.40
Frying	24.05	**37.93**	19.78	23.36	4.83	**23.64**	3.79
Running_water	33.18	33.26	29.68	**33.80**	34.54	31.98	**35.23**
Speech	47.89	52.70	**55.32**	54.54	41.15	34.07	**42.11**
Vacuum_cleaner	46.45	51.28	**62.14**	56.65	35.11	**56.67**	30.04
Avg.	**34.38**	**40.39**	**37.64**	**38.91**	**35.19**	**33.94**	**32.51**

Table II. Performance evaluation with DCASE validation set : class-wise f-score [%]

	DCASE Baseline	Supervised fusion	Unsupervised fusion	Final fusion
Alarm_bell_ringing	37.64	46.11	**56.94**	56.94
Blender	30.37	37.97	**44.44**	44.44
Cat	43.26	**41.91**	40.70	41.91
Dishes	24.78	25.08	**33.07**	33.07
Dog	20.64	28.78	**38.59**	38.59
Electric_shaver_toothbrush	35.56	**47.68**	40.00	47.68
Frying	24.05	**32.07**	23.64	32.07
Running_water	33.18	**38.14**	35.23	38.14
Speech	47.89	**56.50**	42.11	56.50
Vacuum_cleaner	46.45	**64.45**	56.67	64.45
Avg.	**34.38**	**41.87**	**41.14**	**45.38**

mensional biomimetic spectrogram extracted at 100Hz is used as the input feature to the VAE. Encoder of the VAE is constructed using 3 pairs of 2D Convolution-Maxpooling layers with convolution kernel sizes 16x16 and ReLu activation followed by a linear projection layer with output size 360. 180 of these embeddings are used as means and remaining 180 are used as log-variance of the latent posterior gaussian. The decoder uses the similar architecture with Maxpooling operations replaced by upsampling. Stochastic gradient descent based training with reparametrization trick [3] was used to train the VAE with a batch size of 1.

In evaluation of our approach, we performed cross-validation test by random selection for 20% data in the DCASE valdiation set. And, we summarized as the mean and standard deviation over the 20-times repetition.

3.3. Results

Using the DCASE validation set, our approach was tested and the results are summarized on Table I for individual systems and Table II for fusion. We found an improvement of class averaging f-score in supervised methods compared to DCASE baseline. By comparing the results between baseline and S01, we found the effectiveness of using estimated labels for weakly labeled data in training. In S02, separated audio clips for speech and nonspeech are tested in the supervised system used in S01. In this result, we found an improvement in Speech and Vacuum cleaner classes although the f-score in Frying class is decreased. Similarly, enhanced audio recordings by performing a background noise reduction can help to improve the f-scores. In the result of S03, almost f-score are improved compared to baseline except Cat and Dishes classes.

On the other hand, we found improvements in several classes by performing unsupervised methods. In between unsupervised methods, the predictive tracking (S04) can significantly improve the f-scores for Blender and Dog classes while a f-score for Frying is dropped down. And, the salience detection (S05) can help to improve f-scores in Electric shaver toothbrush and Vacuum cleaner. In particular, a joint result (S04&S05) between two unsupervised methods shows the best f-scores in Alarm bell ringing and Running water classes among the unsupervised methods. Because a duration in detected interval by each method can be reduced by applying the AND rule in between two unsupervised methods. It can help to find short time repeating event like alarm or discrete sound like speech due to silent syllable.

Table III. Performance evaluation in cross-validation test: class-wise f-score [%]

	Supervised fusion		Unsupervised fusion		Final fusion	
	Mean	Std.	Mean	Std.	Mean	Std.
Alarm_bell_ringing	45.37	6.61	**57.06**	8.79	57.06	8.79
Blender	40.49	8.32	**48.39**	6.48	48.39	6.48
Cat	**42.52**	9.34	40.70	11.15	42.52	9.34
Dishes	25.75	4.62	**34.00**	4.39	34.00	4.39
Dog	26.92	4.73	**36.89**	6.28	36.89	6.28
Electric_shaver_toothbrush	**47.88**	13.77	40.76	12.59	47.88	13.77
Frying	**29.97**	6.18	25.10	8.17	29.97	6.18
Running_water	**40.36**	6.73	35.99	9.51	40.36	6.73
Speech	**56.13**	2.97	41.35	2.66	56.13	2.97
Vacuum_cleaner	**64.32**	8.98	61.37	9.73	64.32	8.98
Avg.	41.97	2.99	42.16	3.51	45.75	3.13

As shown in Table I, each system has pros and cons depending on event class. To compensate drawbacks of each system, we considered a fusion method among the individual systems and summarize the results in Table II. In fusion among the supervised methods, we found further improvement in five classes such Alarm bell ringing, Dishes, Running water, Speech, and Vacuum cleaner. And, the best cases in each class are gathered by performing hard decision among the unsupervised methods. As shown in Table II, our supervised fusion system seems to help for finding long time events while unsupervised fusion system can help to detect short time events. With these results, the final result is derived in the second hard-decision. In our approach, the class averaging f-score is improved about 11% compared to baseline performance.

Due to the risk of using hard decision, we additionally performed a cross-validation test. Table III shows mean and standard deviation of f-score over the repetition. The result shows same trend with the numbers in Table II, and three class averaging f-scores in Table II is bounded within each 95% confidence interval calculated by mean and standard deviation in Table III (95% confidence interval for; supervised fusion: 41.97 ± 1.31, unsupervised fusion: 42.16 ± 1.54, and final fusion: 45.75 ± 1.37).

4. SUMMARY

In this report. we describe our approach for sound event detection which is submitted in task 4 of DCASE2020 challenge. For this challenge, we built a fusion system composed of five individual subsystems. And, each subsystem was designed by integrating several methods such speech separation, background noise removal, salience detection, and predictive tracking. In particular, we applied salience detection to weakly labeled data for getting information about event class as well as its time boundaries. Then, we used the estimated labels in training a supervised system. The effectiveness of this method was demonstrated in experiments. For combining five outputs from each subsystem, we calculated weighted averaging posterior with soft and hard weight. The final result was derived by applying threshold (0.5) to a posterior for each class. According to the results, it is shown that the fusion method can help to improve the performance.

5. REFERENCES

[1] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, 2017, pp. 1195–1204.

[2] A. Bellur and M. Elhilali, "Bio-mimetic attentional feedback in music source separation," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 8718–8722.

[3] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[4] S. Kothinti, K. Imoto, D. Chakrabarty, G. Sell, S. Watanabe, and M. Elhilali, "Joint Acoustic and Class Inference for Weakly Supervised Sound Event Detection," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 36–40. [Online]. Available: https://ieeexplore.ieee.org/document/8682772/

[5] D. O'Shaughnessy, "Linear predictive coding," *IEEE potentials*, vol. 7, no. 1, pp. 29–32, 1988.

[6] N. Huang and M. Elhilali, "Auditory salience using natural soundscapes," *The Journal of the Acoustical Society of America*, vol. 141, no. 3, pp. 2163–2176, 2017.

[7] S. Park, S. Mun, Y. Lee, and H. Ko, "Score fusion of classification systems for acoustic scene classification," DCASE2016 Challenge, Tech. Rep., September 2016.

[8] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, New York City, United States, October 2019. [Online]. Available: https://hal.inria.fr/hal-02160855

[9] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," in *ICASSP 2020 - 45th International Conference on Acoustics, Speech, and Signal Processing*, Barcelona, Spain, 2020. [Online]. Available: https://hal.inria.fr/hal-02355573

[10] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50k: an open dataset of human-labeled sound events," in *arXiv*, 2020.

[11] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 411–412.

[12] S. Wisdom, H. Erdogan, D. P. W. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman, and J. R. Hershey, "What's all the fuss about free universal sound separation data?" in *in preparation*, 2020.