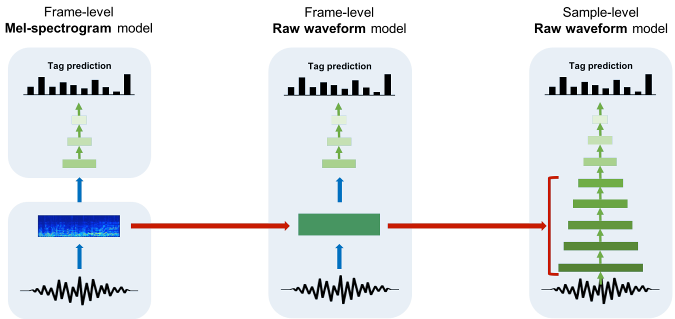
COMBINING MULTI-SCALE FEATURES USING SAMPLE-LEVEL DEEP CONVOLUTIONAL NEURAL NETWORKS FOR WEAKLY SUPERVISED SOUND EVENT DETECTION

Jongpil Lee[1], Jiyoung Park[1], Sangeun Kum[1], Youngho Jeong[2], Juhan Nam[1]
[1] Music and Audio Computing Lab, Graduate School of Culture Technology, KAIST
[2] Realistic AV Research Group, ETRI, Korea
{richter, jypark527, keums, juhannam}@kaist.ac.kr, yhcheong@etri.re.kr

Overview

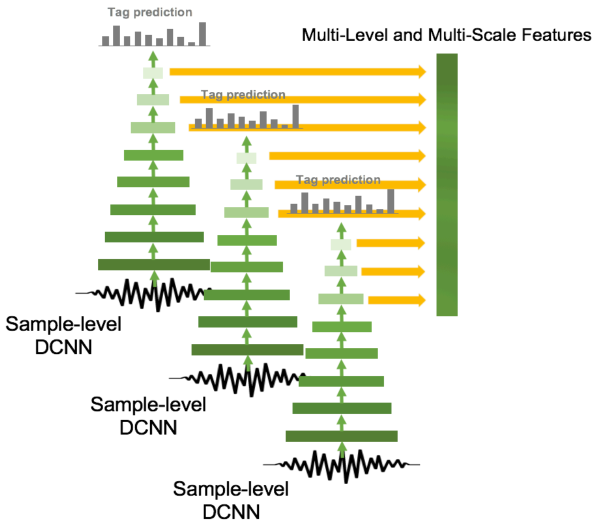
Our method submitted to large-scale weakly supervised sound event detection for smart cars in the DCASE Challenge 2017 Task 4. It is based on two deep neural network methods suggested for music auto-tagging. One is training sample-level Deep Convolutional Neural Networks (DCNN) using raw waveforms as a feature extractor. The other is aggregating features on multi-scaled models of the DCNNs and making final predictions from them. With this approach, we achieved the best results, 47.3% in F-score on subtask A (audio tagging) and 0.75 in error rate on subtask B (sound event detection) in the evaluation. These results show that the waveform-based models can be comparable to spectrogram-based models when compared to other DCASE Task 4 submissions.

Sample-level Deep Convolutional Neural Networks



Combination of Multi-Scale Features

Event sounds have different timbre patterns in terms of feature hierarchy and time-scales. The sample-level DCNNs take different input sizes to capture both local and global characteristics of the sounds.

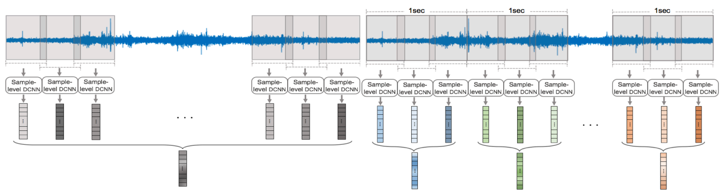


Submissions

- **SDCNN**: Sample-level DCNN that takes 893ms of audio as input. This is one of the models used as a feature extractor for the rest submissions.
- **MLMS5**: Multi-level and Multi-scale features extracted from models taking 372ms, 557ms, 627ms, 743ms and 893ms as input.
- **MLMS3**: Multi-level and Multi-scale features extracted from models taking 1486ms, 2678ms and 3543ms as input.
- **MLMS8**: Multi-level and Multi-scale features extracted from models taking 372ms, 557ms, 627ms, 743ms, 893ms, 1486ms, 2678ms and 3543ms as input.

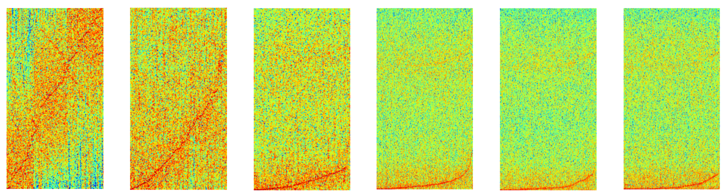
Feature Aggregation and Final Classification

- Subtask A
The features of all segments are averaged into a single feature vector for each model.
- Subtask B
Segment-level features are averaged every second.



Lastly, the final prediction is performed using a fully-connected neural network for each subtask.

Filter Visualization



Layer 1 Layer 2 Layer 3 Layer 4 Layer 5 Layer 6

Spectrum of the filters in the sample-level convolution layers which are sorted by the frequency at the peak magnitude. The x-axis represents the index of the filters and the y-axis represents the frequency. We can observe that they are sensitive to more log-scaled in frequency as the layer goes up.

Results

- Instance-based results for subtask A

	Development set			Evaluation set		
	F-score	Prec.	Rec.	**F-score**	Prec.	Rec.
SDCNN	37.8%	26.7%	64.8%	40.3%	31.3%	56.7%
MLMS5	44.3%	38.8%	51.7%	47.3%	48.0%	46.6%
MLMS3	42.2%	39.0%	45.9%	47.2%	49.6%	45.0%
MLMS8	43.8%	39.2%	49.5%	47.1%	48.5%	45.9%

- Instance-based results for subtask B

	Development set		Evaluation set	
	ER	F-score	**ER**	F-score
SDCNN	0.88	28.1%	0.82	39.4%
MLMS5	0.86	30.7%	0.78	42.6%
MLMS3	0.86	31.2%	0.78	44.2%
MLMS8	0.84	34.2%	0.75	47.1%

Discussion

- The feature aggregation and final classification stage improve performance compared to the direct result of SDCNN.
- Class-wise performance indicates that audio clips with different tags are optimal in different time scales.

Reference

1. Jongpil Lee, Jiyoung Park, Keunhyoung Luke Kim, and Juhan Nam. Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms. Sound and Music Computing Conference (SMC), 2017.