Waveforms and Spectrograms: Enhancing Acoustic Scene Classification Using Multimodal Feature Fusion

***Dennis Fedorishin**, Nishant Sankaran, Deen Dayal Mohan, Justas Birgiolas, Philip Schneider, Srirangaraj Setlur, Venu Govindaraju*

Introduction

- Acoustic scene classification (ASC) has seen tremendous progress due to advances in CNNs and other signal processing techniques.
- While Mel-spectrograms are the most commonly used audio representation, *we explore the fusion of multiple representations of audio signals*: the raw waveform and Mel-spectrogram.

Methods

We design an end-to-end fusion model based on two CNN feature extractors and a unified classification layer.

- The waveform and spectrogram latent representations l_w and l_s from branches F_s and F_w shown in the figure are fused together for the classification layers F_c.
- The classification ĉ of an audio sample's waveform and spectrogram x_w and x_s is defined as:

$$\hat{c}_{(x_w, x_s)} = \underset{c \in \mathbb{C}}{\operatorname{argmax}}\, F_c(F_w(x_w) + F_s(x_s))$$

Along with the fusion model, we utilize two sub-networks to investigate interactions and dynamics between modalities:

Spectrogram Sub-Network $F_c(F_s(x_s))$

- The spectrogram sub-network is trained only with Mel-spectrograms, omitting the waveform branch.

Waveform Sub-Network $F_c(F_w(x_w))$

- The waveform sub-network is trained only with waveforms, omitting the spectrogram branch.

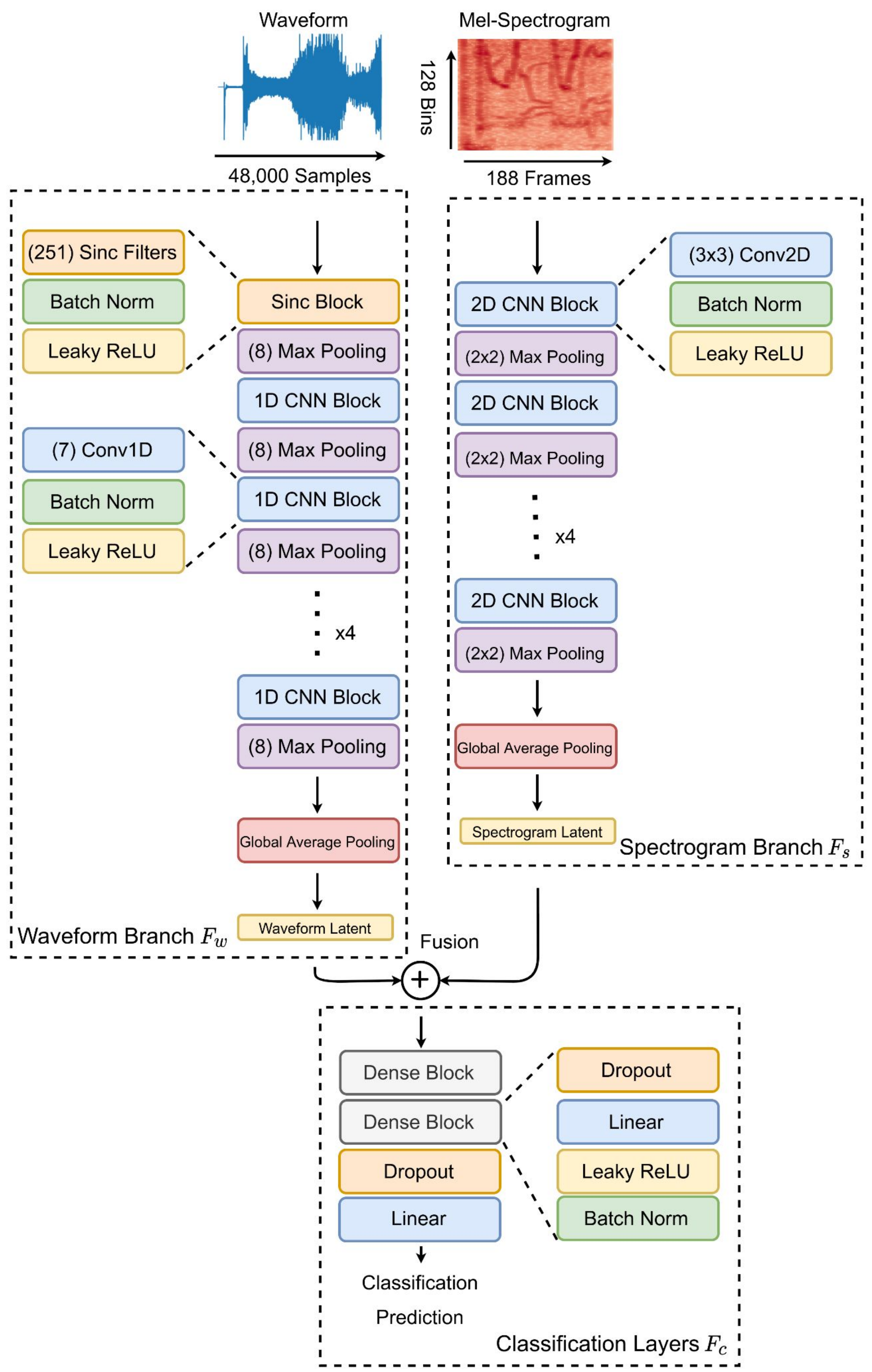

Illustration of fusion model.

All experiments were conducted on the DCASE 2021 Challenge Task 1B Audio-Visual Scene dataset, using only the audio modality. Classification is performed on one-second intervals according to the challenge guidelines.

Results

We experimentally determine that fusing features learned from waveform and Mel-spectrogram representations of audio improve ASC performance beyond a single modality.

- 5.7% increase in accuracy over DCASE 2021 Challenge Task 1B audio network baseline.
- 4.3% increase in accuracy over independent sub-networks, showing that complementary features are learned.

Table 3: Model performance compared to challenge baseline.

Model	Accuracy %	Log Loss	# Params
Audio baseline [20]	65.1	1.048	-
Waveform sub-network	64.79	1.045	1.0M
Spectrogram sub-network	66.46	1.072	1.1M
Fusion Model	**70.78**	**0.915**	1.4M

Our proposed fusion method outperforms various other fusion paradigms, showing latent vector fusion performs strongly.

Table 4: Fusion method comparisons.

Model	Accuracy %	Log Loss	# Params
Wavegram-Logmel-CNN	68.35	1.063	80.2M
Decision fusion	68.65	0.955	2.0M
Decision ensemble	70.47	**0.845**	2.0M
Proposed late fusion	**70.78**	0.915	1.4M

Ablation Study Insights

When training the fusion model end-to-end, each sub-network learns disparate features that when fused together, improve ASC performance.

- Large performance drops when removing each branch.

Table 6: Feature branch removal ablation study.

Model	Accuracy %	Log Loss
Spectrogram sub-network	66.46	1.072
Fusion spectrogram branch only	51.33	1.720
Waveform sub-network	64.79	1.045
Fusion waveform branch only	31.51	2.500

Certain classes have the lowest loss within one branch of the fusion model, lower than the fusion model with both branches.

- A stronger fusion method can fully exploit modality complementary to further improve ASC performance.

Table 8: Class-Wise losses of the fusion model.

Class-Wise loss	Fusion	Fusion spec. branch only	Fusion wave. branch only
Airport	**0.901**	1.226	3.441
Shopping Mall	**0.944**	0.995	1.612
Metro Station	**1.053**	2.030	1.827
Street Pedestrian	1.104	**1.069**	2.638
Public Square	1.321	1.384	**0.663**
Street Traffic	**0.424**	0.843	3.038
Tram	**0.899**	2.106	4.182
Bus	0.747	4.905	**0.723**
Metro	**1.145**	2.825	4.324
Park	0.535	**0.443**	2.913

Conclusion

We present a novel ASC model that fuses complementary features of the raw waveform and Mel-spectrogram representations of audio.

- Our proposed **fusion design outperforms** various **other experimentally tested methods.**
- Each sub-network learns **disparate but complementary features**, improving overall ASC performance.
- We achieved **1st place** against the DCASE 2021 Challenge Task 1B audio-only submissions for validation accuracy and **2nd place** against validation loss.

This work was supported by the Center for Identification Technology Research (CITeR) and the National Science Foundation (NSF) under grant 1822190.

References

- H. Purwins, B. Li, T. Virtanen, J. Schluter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 2 ,pp. 206–219, 2019.
- S. Wang, A. Mesaros, T. Heittola, and T. Virtanen, "A curated dataset of urban scenes for audio-visual scene analysis" 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.
- Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D.Plumbley, "PANNS: Large-scale pretrained audio neural networks for audio pattern recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp.2880–2894, 2020.

University at Buffalo The State University of New York

DCASE2021 Workshop
Workshop on Detection and Classification of Acoustic Scenes and Events

Center for Unified Biometrics and Sensors
School of Engineering and Applied Sciences