INVESTIGATING WAVEFORM AND SPECTROGRAM FEATURE FUSION FOR ACOUSTIC SCENE CLASSIFICATION

Technical Report

Dennis Fedorishin[1], *Nishant Sankaran*[1], *Deen Mohan*[1], *Justas Birgiolas*[2,3]
Philip Schneider[2], *Srirangaraj Setlur*[1], *Venu Govindaraju*[1]

[1] University at Buffalo, Center for Unified Biometrics and Sensors, USA,
{dcfedori, n6, dmohan, setlur, govind}@buffalo.edu
[2]ACV Auctions, USA, {jbirgiolas, pschneider}@acvauctions.com
[3]Ronin Institute, USA.

ABSTRACT

This technical report presents our submitted system for the DCASE 2021 Challenge Task1B: Audio-Visual Scene Classification. Focusing on the audio modality only, we investigate the use of two common feature representations within the audio understanding domain, the raw waveform and Mel-spectrogram, and measure their degree of complementarity when using both representations for fusion. We introduce a new model paradigm for acoustic scene classification by fusing features learned from Mel-spectrograms and the raw waveform from separate feature extraction branches. Our experimental results show that our proposed fusion model has a 4.5% increase in validation accuracy and a reduction of .14 in validation loss over the Task 1B baseline audio-only sub-network. We further show that learned features of raw waveforms and Mel-spectrograms are indeed complementary to each other and that there is a consistent classification performance improvement over models trained on Mel-spectrograms alone.

Index Terms— Audio classification, Acoustic scene classification, Feature fusion, Multi-modal features.

1. INTRODUCTION

Mel-spectrograms are the de-facto audio feature representation and they have been widely used throughout the entire history of audio understanding. Mel-spectrograms are created by calculating the short-time fourier transform (STFT) of an audio signal, then passing the STFT frequency responses through band-pass filters spaced on the Mel(logarithmic)-scale and often further passed through a logarithmic compression to replicate the human's non-linear perception of signal pitch and loudness, respectively.

With the advent of deep neural networks, many methods have been introduced that perform audio understanding tasks such as acoustic scene classification (ASC) and sound event detection that use Mel-spectrogram representations of audio as the input to a convolutional neural network [1, 2]. Researchers also explored the use of other feature representations such as the gammatone and Constant-Q (CQT) spectrogram, and Mel Frequency Cepstrum Coefficients (MFCC) [3, 4]. [5] found that fusing these representations together allows for a network to learn complementary features, creating a stronger model for ASC.

In parallel, other researchers utilize the raw waveform as the input into neural networks, bypassing the need for hand crafted features [6, 7]. Waveform-based networks are able to be trained end-to-end, while networks that utilize spectrograms need to create these hand crafted features that may often be sub-optimal for the given task. Regardless, many state of the art methods in ASC, speaker recognition, sound event detection, and other tasks still utilize spectrogram representations [8, 9]. Further, [10] introduced a fully learnable variation of spectrograms, where they can be trained end-to-end to automatically find optimal parameters.

As a result, there is still no clear distinction as to the best feature representation that performs strongly across various audio understanding tasks. Works such as [7, 11] have begun to bridge together methods using both waveform and spectrogram representations in a fusion setting. Although a performance improvement is exhibited, these methods do not deeply explore the degree of complementarity and effects of fusing these features together.

In this report, we investigate waveform and Mel-spectrogram feature fusion and propose a new acoustic scene classification model that learns complementary features from both modalities. We evaluate our proposed model using the DCASE 2021 Challenge Task 1B dataset to prove the effectiveness and complementarity of waveform and Mel-spectrogram feature fusion.

2. DATASET

2.1. DCASE 2021 Task 1B: Audio-Visual Scene Classification

Task 1B is based on the TAU Audio-Visual Urban Scenes 2021 dataset, a dataset containing synchronized audio and video recordings from 12 European cities in 10 different scenes. Audio is recorded using a Soundman OKM II Klassik/studio A3 microphone paired with a Zoom F8 audio recorder, sampled at $48kHz$ at a 24-bit resolution. Video is recorded using a GoPro Hero5 Session. The dataset contains 12,292 samples of each modality spread across the 10 scenes. The provided train/validation split consists of 8,646 samples in the training set and 3,645 samples in the validation set. [12]

2.2. Data Preprocessing

For Task 1B, we input the raw waveform and its generated Mel-spectrogram into their respective feature extractors. According to the Task 1B rules, we split the development dataset samples into 1 second audio files to perform classification at the 1 second level.

Table 1: Detailed overview of proposed model design.

Spectrogram Branch F_s	
Input shape	[128,188]
2D CNN kernel size	3×3
2D CNN stride	1
Filter responses	32, 64, 128, 256
Max pooling size	2×2
Global average pooling output l_s	$\langle 1024 \rangle$
Waveform Branch F_w	
Input shape	[48000]
Sinc kernel size	251
Sinc stride	1
1D CNN kernel size	7
1D CNN stride	1
Filter responses	32, 64, 128, 256
Max pooling size	6
Global average pooling output l_w	$\langle 1024 \rangle$
Classification Layers F_c	
Input shape $(l_w + l_s)$	$\langle 1024 \rangle$
Dropout p	0.3
Dense layer outputs	$512, 256, 10$
Output classes	10

This brings the training dataset to 86,460 samples and the validation dataset to 36,450 samples. Audio files are sampled at $48kHz$ and therefore have a sample length of $[48000]$. In addition, the audio waveforms are scaled to the range $[0, 1]$. Mel-spectrograms are generated using 128 frequency bins, a hop length of 256 samples, and a Hann window size of 2048 samples, creating a final size of $[128 \times 188]$. The Mel-spectrograms are also passed through a logarithmic compression and then normalized at an instance level using Z-Score normalization such that each sample has a mean of 0 and unit standard deviation.

3. PROPOSED METHOD

To investigate and understand the complementarity between learning features from Mel-spectrograms and raw waveforms, we designed a fusion model based on two CNN feature extractors, and a unified classification layer. Figure 1 illustrates the design of our model. The spectrogram branch, F_s, is comprised of repeating 2D CNN blocks followed by a max pooling operation. The CNN blocks contain a convolution layer using a kernel size of 3×3, followed by a batch normalization and a Leaky ReLU nonlinear activation.

The waveform branch, F_w, is of a similar structure, however the two-dimensional convolutional layers are replaced with one-dimensional convolutions with a kernel size of 7. In addition, the first convolutional layer in the waveform branch are parameterized to Sinc functions, as described in [13].

As both branches are working with different-sized input data, the feature responses from each branch vary in size. We utilize global average pooling (GAP) layers to condense both waveform and spectrogram features into a vector of 1024 units, denoted by l_w and l_s, respectively.

Feature fusion of both feature extraction branches is accomplished at the latent representation level, where features for both the waveform and spectrogram branch are extracted independently, then fused together into a unified representation. Fusion is accomplished using elementwise-summation such that the final latent representation $(l_w + l_s)$ has the same shape as its constituents.

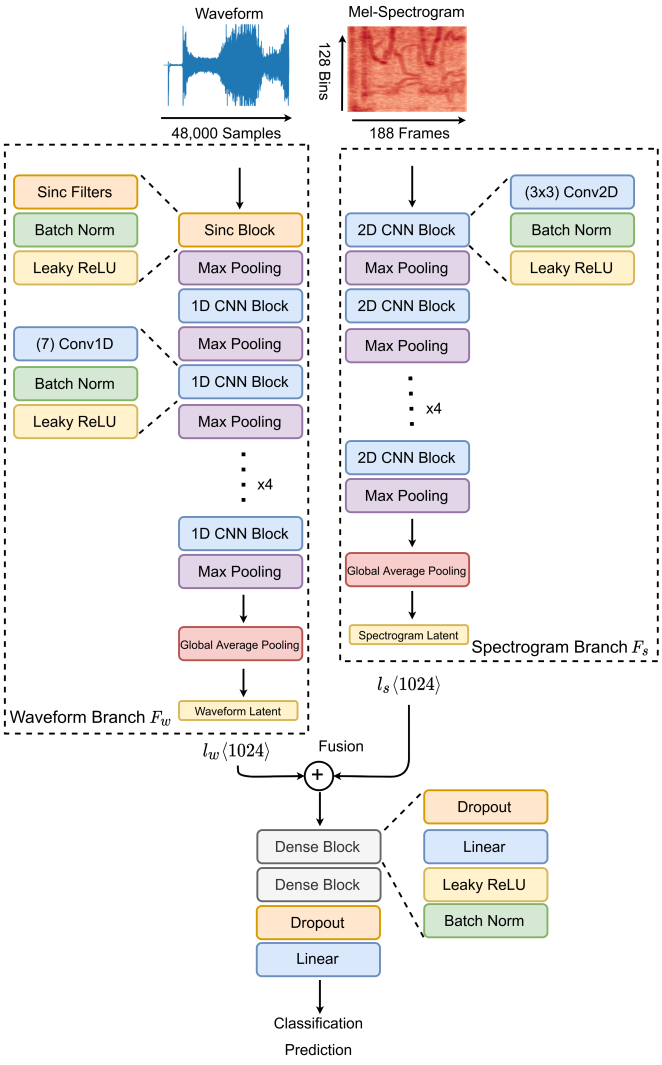

Figure 1: Illustration of the proposed fusion model.

The classification layers, F_c, take $(l_w + l_s)$ as input and perform the final classification using two repeating dense blocks, as shown in Figure 1. We use dropout layers with $p = 0.3$, followed by linear layers, a Leaky ReLU activation, and batch normalization. The classification $\hat{c}$ of the set of classes $c \in \mathbb{C}$ of an audio sample with its raw waveform x_w and Mel-spectrogram x_s can be described as:

$$\hat{c}_{(x_w, x_s)} = \underset{c \in \mathbb{C}}{\operatorname{argmax}} F_c(F_w(x_w) + F_s(x_s)) \quad (1)$$

Table 1 describes in detail the configuration of both feature extraction branches and the final classification layers. For our experiments, we utilize three variations of the described model to investigate modality complementarity:

Spectrogram sub-network: The spectrogram branch in Figure 1 is used independently with the classification layers, without the waveform branch. In this model, training is conducted only using Mel-spectrograms, omitting $F_w(x_w)$ from (1).

Table 2: Waveform sub-network data augmentation strategies.

Method	Accuracy %	Log Loss
No augmentation	61.32	1.149
Mixup ($\alpha = 0.2$)	63.00	1.080
Time masking	61.76	1.161
Pitch shifting	59.25	1.164
Time stretching	61.32	1.206
Time shifting	62.60	1.172
Random gaussian noise	58.04	1.182
Mixup + Time shifting	**64.19**	**1.066**

Waveform sub-network: The waveform branch in Figure 1 is used independently with the classification layers, without the spectrogram branch. In this model, training is conducted only using raw waveforms, omitting $F_s(x_s)$ from (1).

Fusion model: Both the spectrogram and waveform branch are trained end-to-end with their respective inputs. The latent representations of each branch are fused together for classification.

4. TRAINING CONFIGURATION

All models are trained using the SGD optimizer paired with the one-cycle learning rate scheduler and learning rate range test described in section 4.1. Training batch size is set to 128 and the models are trained for 50 epochs. The models were trained on an RTX 6000 GPU with the most complex model taking 1.5 hours to fully train.

4.1. One-cycle Learning Rate Scheduler

The one-cycle learning rate policy [14] was introduced as an extension beyond cyclical learning rate schedulers, where an initial learning rate is annealed to a large maximum value, then annealed back to a value much lower than the initial learning rate, over the entire training procedure. [14] showed that this procedure leads to faster training times, in addition using large learning rates for a portion of the training procedure acts as a form of regularization.

[15] introduced the learning rate range test, a method of programatically finding a near-optimal maximum learning rate for the one-cycle scheduler. We use a modified version of the learning rate range test described in [15]. Learning rate values $\lambda_1, \lambda_2, ..., \lambda_n$ are sampled over a uniform space and used for a single forward pass within the model using a batch of training samples. The learning rate that produces the lowest loss value, λ_t, is selected as the optimal maximum learning rate, divided by a factor of 10. Due to the stochasticity of data within batches, we run this operation m times and take the median learning rate to remove any possible outliers. The algorithm can be described as:

$$\lambda_{max} = \frac{median(\lambda_{t_1}, \lambda_{t_2}, ..., \lambda_{t_m})}{10} \quad (2)$$

We set $m = 7$ and sample $n = 50$ learning rates over the log space of $[-7, 2]$. We experimented with various learning rate schedulers and found that the one-cycle scheduler paired with (2) reduced the number of epochs needed to achieve convergence.

4.2. Data augmentation

We conducted a search to find the optimal data augmentation strategies to improve the classification performance of the proposed models. We tested various augmentations on the raw waveform: mixup (50% application chance), time masking, pitch shifting, time stretching, time shifting, and adding random gaussian noise. As shown in Table 2, time shifting, time masking and mixup are beneficial to model performance. However, when combining both mixup and time stretching together, we get a further improvement in classification accuracy beyond any other combination. For consistency, we utilize time shifting and mixup augmentations on the spectrogram as well. In the fusion setting, time shifting is applied independently to the waveform and spectrogram, such that both modalities may be shifted by varying degrees. Further research should be conducted in studying the effects of using independent data augmentations for each modality.

Table 3: Kernel parameterization performance.

Parameterization	Accuracy %	Log Loss
Unparameterized (normal)	62.20	1.068
Complex-Gabor [10]	42.10	2.006
Sinc [13]	**64.82**	**1.067**

Table 4: Model performance compared to challenge baseline.

Model	Accuracy %	Log Loss	# Params
Audio baseline [16]	65.1	1.048	-
Wave. sub-network	61.79	1.051	1.0M
Spec. sub-network	66.29	1.046	1.1M
Wave. + Spec. fusion	**69.58**	**0.907**	1.4M

5. EXPERIMENTAL RESULTS

5.1. Waveform Kernel Parameterizations

[13] introduced the use of parameterized Sinc filters for speech recognition. Further, [10] introduced the use of learnable complex-valued Gabor filters to extract audio features, similar to Sinc filters. These methods have shown to outperform normal one-dimensional kernels as they are less prone to overfitting because they are constrained to their respective functions.

Table 3 shows model performance when replacing the first convolutional layer of the waveform branch with a parameterized kernel, instead of an unparameterized, fully learnable kernel. We test real-valued Sinc filters [13] and complex-valued Gabor filters [10] and compare model performance across these setups. As shown, using parameterized Sinc filters outperform both the conventional unparameterized filters in addition to the Gabor filters. Using Sinc filters also allows us to reduce model complexity and use the filter's interpretability to further investigate what is being learned within the waveform branch.

5.2. Waveform and Spectrogram Feature Fusion

Table 4 demonstrates classification performance of the Task 1B baseline, compared to the three different model variations proposed. The waveform sub-network is not able to outperform the baseline in terms of accuracy and loss, however the spectrogram sub-network performs marginally better than the baseline in both accuracy and loss. The fusion model outperforms both the baseline and models trained on single modalities, specifically a 4.5% improvement in accuracy and a reduction of 0.14 in loss over the baseline. Furthermore, we see that the fusion model outperforms the spectrogram

Table 5: Feature fusion methods experiment.

Fusion Method	Accuracy %	Log Loss	# Params
Element-wise sum	69.58	**0.907**	1.4M
Concatenation	**70.85**	0.924	1.9M
MFB [17]	70.13	0.943	7.6M

Table 6: Parameterization ablation study.

Model	Accuracy %	Log Loss	# Params
Fusion model	**69.58**	**0.907**	1.4M
Large spec. sub-network	66.48	1.043	4.2M
Large wave. sub-network	63.44	1.041	3.9M

sub-network by 3.3% accuracy and a .14 reduction in loss. This improvement shows that there are features being learned within the raw waveform that are complementary to features being learned from the Mel-spectrogram, resulting in a more discriminative classification model.

5.3. Multi-Modal Feature Fusion Methods

Most approaches to multi-modal feature fusion utilize simple linear methods, such as element-wise summation and concatenation of vectors and feature maps. A more advanced operation, bilinear pooling, has been shown to capture more dependencies between vectors being fused. Multimodal Factorized Bilinear Pooling [17] has been used within the visual question answering domain and has shown to capture more expressive features than linear methods while being less computationally expensive than conventional bilinear pooling.

We experiment using these fusion methods to see whether we can fuse features in a more expressive fashion. Table 1 and Figure 1 depict the design for element-wise summation fusion. For concatentation, latent vectors l_w and l_s are combined to a final size of 2048 units. This new vector is passed into the classification layers, with the dense layers outputting $1024, 512, 10$ units, respectively. For MFB fusion, we set $k = 3$ and $o = 1024$, as described in [17]. The MFB fusion model has the same design as Figure 1, but the element-wise summation operation is replaced with MFB.

Table 5 shows the performance of our fusion model when utilizing element-wise sum, concatenation, and MFB. All methods perform similarly, however element-wise summation produces the lowest validation loss model. Fusion by concatenating latent vectors results in the highest accuracy model. We select element-wise fusion as it produced the lowest loss in addition to it being the least computationally expensive operation.

6. ABLATION STUDIES

Although we examine a classification performance improvement when fusing waveform and spectrogram features, it is important to validate that the improvement is coming from complementary features being extracted from both modalities. As we are performing late-stage feature fusion, using two separate feature extractors inherently increases the size of the model. It may be the case that the feature extraction branches themselves are underparameterized, and when adding more parameters the model performs better solely due to the increase in parameterization and not the second modality.

To test this hypothesis, we expand the waveform and spectrogram sub-networks such that their total number of parameters exceed the fusion model. For both sub-networks, we double each of the CNN block filter responses, increase latent vectors from 1024 units to 2048 units, and double the classification layer responses. Table 6 shows these trained expanded sub-networks in comparison to the original fusion model. Even with the increase in model size, both of the sub-networks were unable to surpass the performance of the fusion model, showing that the performance improvement in the fusion model is from the added modality.

Table 7: Feature branch training ablation study.

Model	Accuracy %	Log Loss
Spec. sub-network	66.29	1.046
Fusion spec. branch only	51.33	1.72
Wave. sub-network	61.79	1.051
Fusion wave. branch only	31.51	2.50

To further understand the differences of each sub-network's performance when trained alone or in a fusion setting, we compare each sub-network to their equivalent sub-network trained in the fusion setting. Examining the performance drop when removing each feature extraction branch in the fusion model may give clues into how the branches train alone versus in the fusion setting.

The trained waveform and spectrogram sub-networks depicted in Table 4 are compared to the fusion model's respective sub-network. As shown in Table 7, the sub-networks that are trained in the fusion setting have a substantial performance loss when removing the opposite sub-network, far below the performance of the respective sub-network that is trained independently. We infer that when trained end-to-end, each of the sub-networks in the fusion model learn to focus on different, specific features that overall improve classification performance.

7. SUBMITTED SYSTEMS

We submitted two systems to the Task 1B challenge. Both models are trained using the described training procedure on the provided development training dataset.

Fusion model: Our main submission is the waveform and spectrogram feature fusion model depicted in Figure 1 and Table 1. This model has 1,351,562 parameters and achieved 69.58% accuracy and 0.907 loss on the development validation dataset.

Expanded Fusion model: The second submission is an expanded version of the fusion model depicted in Table 1. We added an additional CNN block to the end of each feature branch, with an output response of 512 filters. The latent vectors are also expanded from 1024 to 2048 units. This model has 5,422,730 parameters and achieved 68.56% accuracy and 0.990 loss on the development validation dataset.

8. CONCLUSION

In this technical report, we describe our submitted systems to the 2021 DCASE Challenge Task 1B. We investigate feature fusion of two common audio representations, the raw waveform and Mel-spectrogram, and show that there are complementary features being learned that improve ASC performance. Our proposed fusion model utilizes these features to outperform the Task 1B audio baseline by 4.5% accuracy and .14 validation loss.

9. REFERENCES

[1] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 131–135.

[2] 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

[3] L. Pham, H. Phan, T. Nguyen, R. Palaniappan, A. Mertins, and I. McLoughlin, "Robust acoustic scene classification using a multi-spectrogram encoder-decoder framework," *Digital Signal Processing*, vol. 110, p. 102943, 2021.

[4] Y. Su, K. Zhang, J. Wang, D. Zhou, and K. Madani, "Performance analysis of multiple aggregated acoustic features for environment sound classification," *Applied Acoustics*, vol. 158, p. 107050, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X19302701

[5] H. Wang, Y. Zou, and D. Chong, "Acoustic scene classification with spectrogram processing strategies," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 210–214.

[6] T. Kim, J. Lee, and J. Nam, "Sample-level cnn architectures for music auto-tagging using raw waveforms," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 366–370.

[7] B. Zhu, C. Wang, F. Liu, J. Lei, Z. Huang, Y. Peng, and F. Li, "Learning environmental sounds with multi-scale convolutional neural network," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[8] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Conformer-based sound event detection with semi-supervised learning and data augmentation," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 100–104.

[9] L. Sarı, K. Singh, J. Zhou, L. Torresani, N. Singhal, and Y. Saraf, "A multi-view approach to audio-visual speaker verification," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6194–6198.

[10] N. Zeghidour, O. Teboul, F. d. C. Quitry, and M. Tagliasacchi, "Leaf: A learnable frontend for audio classification," *arXiv preprint arXiv:2101.08596*, 2021.

[11] 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.

[12] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen, "A curated dataset of urban scenes for audio-visual scene analysis," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, accepted. [Online]. Available: https://arxiv.org/abs/2011.00030

[13] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[14] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay," 2018.

[15] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006. International Society for Optics and Photonics, 2019, p. 1100612.

[16] S. Wang, T. Heittola, A. Mesaros, and T. Virtanen, "Audio-visual scene classification: analysis of dcase 2021 challenge submissions," 2021.

[17] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," 2017.