Supplementary material to "Multiview Embeddings for Soundscape Classification"

Dhanunjaya Varma Devalraju, Padmanabhan Rajan.
School of Computing and Electrical Engineering, Indian Institute of Technology, Mandi
s18023@students.iitmandi.ac.in, padman@iitmandi.ac.in

I. Other bases for NAP

Other techniques such as kernel PCA and dictionary learning can also be used to learn the basis. We have performed an initial analysis of these techniques, but they did not result in tangible benefits. Table I gives the results obtained with various techniques used to learn the NAP basis. Note that these results are for the setup described in [1], but the use of NAP to suppress part of the background (or foreground) is the same.

TABLE I: DCASE 2017 Results, when PCA, FDDL (Fisher Discriminant Dictionary Learning) and KPCA (kernel PCA) are used to learn the class-specific NAP basis for background suppression followed by attention and SVM for classification.

Basis	Accuracy (%)
PCA	75.06
FDDL	75.43
KPCA	75.86

II. Supervised acoustic scene classification network

To determine the effect of the MvLDAN framework, we have removed it from the proposed method and combined the information in the embeddings obtained from RPCA followed by NAP. This is done via a multi-input deep neural network, and early/late fusion. This is shown in Figure 1 and the results are in Table II. It can be seen that the MvLDAN brings about a considerable improvement over the information obtained from only RPCA and NAP. The representations after NAP have considerable classwise overlap (there may be many basis components common to various classes). Seeing the foreground-suppressed and background-suppressed embeddings as multiple views of the given audio recording, and using the class information, reduces the overlap among classes in the embedding space.

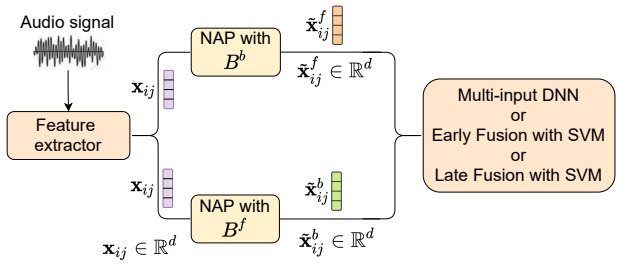

Fig. 1: Illustration of supervised acoustic scene classification using separated background and foreground signals. This method does not use MvLDAN.

TABLE II: Results using various supervised fusion methods to combine the background and foreground suppressed embeddings for DCASE 2017 dataset.

Fusion system	*Accuracy (%)*
Multi-input deep neural network	68.52
Early fusion + SVM	52.41
Late fusion + SVM	73.21

III. Synthetic data generation

We have attempted to evaluate the performance of RPCA for foreground-background separation by utilising the Scraper soundscape simulator. The procedure we followed is described below.

1) Synthetic audio samples of varying complexity levels are generated using Scaper [2]. Complexity levels range from low (meaning foreground dominates) to high (background dominates). The source background and foreground events used to compile each audio sample are saved for future use (required for step 3).
2) RPCA is performed on the audio samples to obtain the corresponding background and foreground signals.
3) SI-SNR improvement (SI-SNRi) [3] metric is computed between the background obtained from RPCA and the source background. The same is also computed between the foreground obtained from RPCA and the source foreground.
4) For comparison, we compare the separation performance of RPCA with that of the TDCN++ method described in [3] using the same metric. The results are shown in Figure 2.

We generated synthetic data using Scaper [2] and by using the sample background and foreground sound events that comes with Scaper. We used combinations of one background and two foreground events with the below given specifications to generate 30 audio samples of length 10 sec each. We used "street" as the background and "human voice" and "siren" as foreground sound events. The duration of the foreground events is chosen randomly from truncated normal distribution with mean 3 sec and standard deviation 1 sec. Further, these foreground events are added anywhere between 0 and 9 seconds, chosen uniformly. The Scaper settings while generating the data [1] as above is described in Table III.

The RPCA algorithm separates a mixture into a single foreground and a single background. The SI-SNRi is computed between the foreground returned by RPCA and the time-domain combination of the source foregrounds of Scraper. While computing the SI-SNRi for TDCNN++, the same procedure is also done to the (multiple) foreground(s) returned by the TDCN++ algorithm.

The plots below indicate that the separation of RPCA degrades gracefully as the complexity of the soundscape increases. Also, it is to be noted that the TDCNN++ is a source separation algorithm, whereas RPCA only separates the foreground and the background. Thus the above procedure has made some approximations by combining multiple foregrounds.

TABLE III: Reference dB and SNR of foreground events compared to background (in LUFS) used to create audio samples of varying complexities namely High, Medium and Low.

	Low	*Medium*	*High*
Ref dB	-20	-5	10
SNR	10	-5	-20

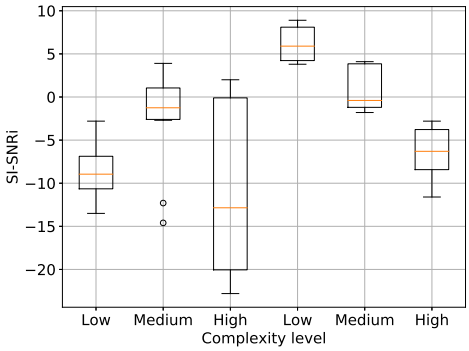

Fig. 2: Illustration of SI-SNR improvement corresponding to audio samples of varying complexities (High, Medium and Low). The first three boxplots correspond to TDCN++ and the last three correspond to RPCA.

References

[1] D. V. Devalraju, H. Muralikrishna, P. Rajan, and A. D. Dileep, "Attention-driven projections for soundscape classification," *Proc. Interspeech 2020*, pp. 1206–1210, 2020.

[2] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 344–348.

[3] S. Wisdom, H. Erdogan, D. P. 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 *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 186–190.

[1] The data is available at https://drive.google.com/drive/folders/11pcD9yTHsg5F7jNrI3t5uqjy_FCFFAKf?usp=sharing