TIME-FREQUENCY AND GEOMETRIC ANALYSIS OF TASK-DEPENDENT LEARNING IN RAW WAVEFORM BASED ACOUSTIC MODELS

Devansh Gupta and Vinayak Abrol

Infosys Centre for AI, IIIT Delhi, India

ABSTRACT

End-to-end raw-waveform modelling with learnable feature extraction front-ends has shown promising results in various speech/audio tasks. Despite its varied success, there have not been many attempts to understand how spectral/temporal feature integration from raw inputs helps recognize task-dependent information. Towards this aim, this work presents data-dependent and data-independent methods for understanding the modelling behavior of acoustic models. The first method employs time-frequency analysis to visualize input-specific response spectra as a function of short-time front-end block processing. The second method employs geometric properties of layer-wise weights to quantify the impact of architectural choices on signal propagation and trainability of the model. We demonstrate potential of the proposed methods with help of case studies on speech classification, speaker identification, and spoofing classification tasks.

Index Terms— Spectral visualization, mutual coherence, acoustic modelling, raw-waveform models

1. INTRODUCTION

Deep neural networks have proved to be successful in a variety of acoustic modelling tasks such as audio captioning [1], acoustic scene classification [2, 3], speech recognition [4] and speaker recognition [5]. While most existing systems are typically fed with spectral features, [6, 7, 8], it is now widely accepted that such features have limited spectral information constrained by choice of filter-bank type or time-frequency (TF) resolution. Several works [4, 3, 5, 9, 10] have tried to address this issue with a waveform based acoustic model by directly modelling the raw waveforms using 1D convolutions and making the spectral feature extraction learnable. While different studies claim different performance gains in different training and architectural setups, there have not been many attempts to understand the kind of information these networks have extracted in a generalized setting. Understanding what information is modelled as a whole by a DNN is an active field of research in computer vision. In particular, it has been shown that gradient-based methods helps visualizing the influence of each pixel in the input image on the prediction score via a relevance map [11, 12, 13]. Similarly, [14] also proposed a gradient-based spectral relevance map visualization for raw-waveform models. However, recently it has been shown that gradient-based methods are performing (partial) signal recovery and thus these relevance maps are unrelated to the network decisions in practice [15]. Towards understanding the modelling behaviour of raw-waveform models, we explore two alternative methods.

1. The first method is *data-dependent*, and uses example specific short-time response spectra as a TF visualization tool for analysis of acoustic models. The motivation for this method lies in the fact that in raw-waveform models, the first layer acts as an information bottleneck for the spectral information propagating through the network. This makes it critical to analyze the information captured, which depends on kernel design, i.e., the inherent block-processing as segmental (processing a signal of about 1-3 pitch period duration), sub-segmental, or multi-resolution. This spectral visualization approach enables us to understand feature integration and task-dependent modelling in acoustic models with different front-end processing pipelines. Further, we demonstrate that acoustic models are implicitly biased towards learning a function with more power at low frequencies.
2. The second method is *data-independent*, and employs tools from linear algebra to compute geometric properties of layer-wise weights. A weight matrix is considered as a learned dictionary whose geometric properties such as upper/lower bound and mutual coherence properties based on the angle between learned vectors [16] enables us to quantify the signal propagation, information flow and trainability, which can be linked to generalization error of the acoustic model.

We demonstrate the potential of the proposed methods with the help of case studies on speech classification and speaker identification tasks by analysing the training behaviour of raw-waveform acoustic models with different processing front-ends. We also considered the spoofing classification task to make our work more exhaustive. Here we extend our study to analyse deep SOTA pre-trained models based on ResNet/InceptionNet models [17]. We demonstrate how such deep models flexibly integrate information from raw-waveforms and how changes in geometric properties correlate with the architectural choices, e.g., skip connection in the case of ResNet. All the experiments are reproducible and implementation available on GitHub* with additional animated plots for better understanding.

2. METHODS FOR ANALYSIS

In this section, we briefly describe two tools to study the impact of different feature processing front-ends via TF and geometric analysis of layer-wise weight matrices in raw-waveform models.

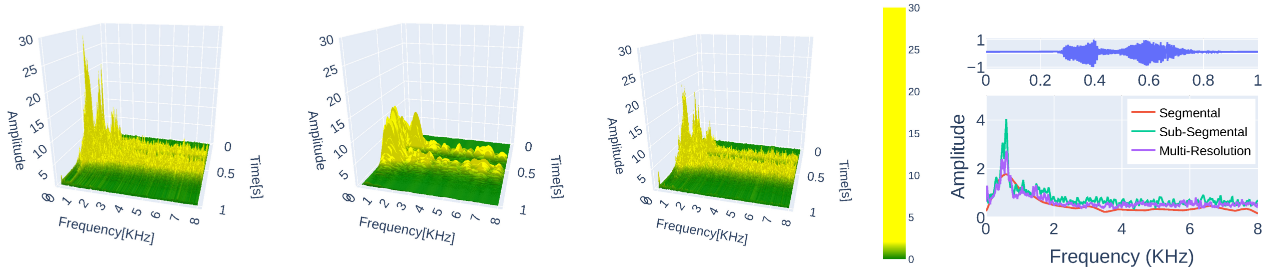
2.1. Time-Frequency Analysis

The TF spectral visualization method used in this study is adapted from [4] based on which we define two generalized input specific response spectra for a raw-waveform model, namely, Short-time response spectra (STRS) and Cumulative response spectra (CRS).

Corresponding Author: devansh19160@iiitd.ac.in

*Code: `https://github.com/Cross-Caps/SWAM`

978-1-6654-0540-9/22/$31.00 ©2022 IEEE



(a) Segmental (b) Sub-Segmental (c) Multi-Resolution (d) Cumulative

Fig. 1. STRS and CRS of an example input [*Backward* class] for model trained on speech classification task

2.1.1. *Short Time Response Spectra*

Given a set of first layer filters defined as $\{f_u \in \mathbb{R}^{kW_u}, kW_u > 0 \mid 1 \leq u \leq n_f\}$, the 3D STRS map is defined as

$$\mathrm{STRS}(i,j) = \left|\sum_{k=1}^{n_f} F_k[i] \cdot R_k[j]\right|, \quad F_u[k] = \sum_{n=0}^{n=N_f-1} f_u^p[n] \cdot e^{-i\frac{2\pi}{N_f}kn}$$

$$R_u[k] = \sum_{i=0}^{i=kW_u-1} f_u[i] \cdot \mathbf{x}\left[k+i-\frac{kW_u-1}{2}\right], \tag{1}$$

Here, f_u^p, kW and k denotes an appropriately padded filter, the kernel width and stride, respectively. If filters f_u forms a Fourier basis then STRS reduces to the Fourier magnitude spectrum of the input. Note that (1) can be easily adapted to other transforms such as Constant-Q transform with logarithmically scaled basis filters.

Cumulative Response Spectra: STRS map is summarized to average frequency information across temporal dimension as

$$\mathrm{CRS}[i] = \frac{1}{N_t}\sum_{j=0}^{N_t-1} \mathrm{STRS}(i,j) \tag{2}$$

Later we demonstrate how CRS can be used to evaluate the discrimination ability of the first layer filters in a raw-waveform model.

2.2. Analysis using Geometric Properties

To quantify the quality of a model, we propose to analyze them using geometric properties of their layer-wise weights as dictionaries. The first set of properties is based on Eigenvalue distribution that indicates how the filters are distributed on the unit ball in different directions. The highest and lowest SVs or the upper (A) and lower (B) bounds depict the number of filters that are pointing in the direction most and least densely represented [16]. The second set consists of coherence properties (such as mutual coherence μ) that quantify the correlation between filters [18]. Ideally, mutual coherence should be small, and the upper and lower bound should be equal and as small as possible, which is not true in practice as even two correlated filters can make μ close to one even if all other atoms are well spread. Similar to [18, 19] we will also use Wigner-Ville (WV) distribution to study the geometric properties of first layer filters in the TF plane and their link to the generalization capability of the model.

3. EXPERIMENTS

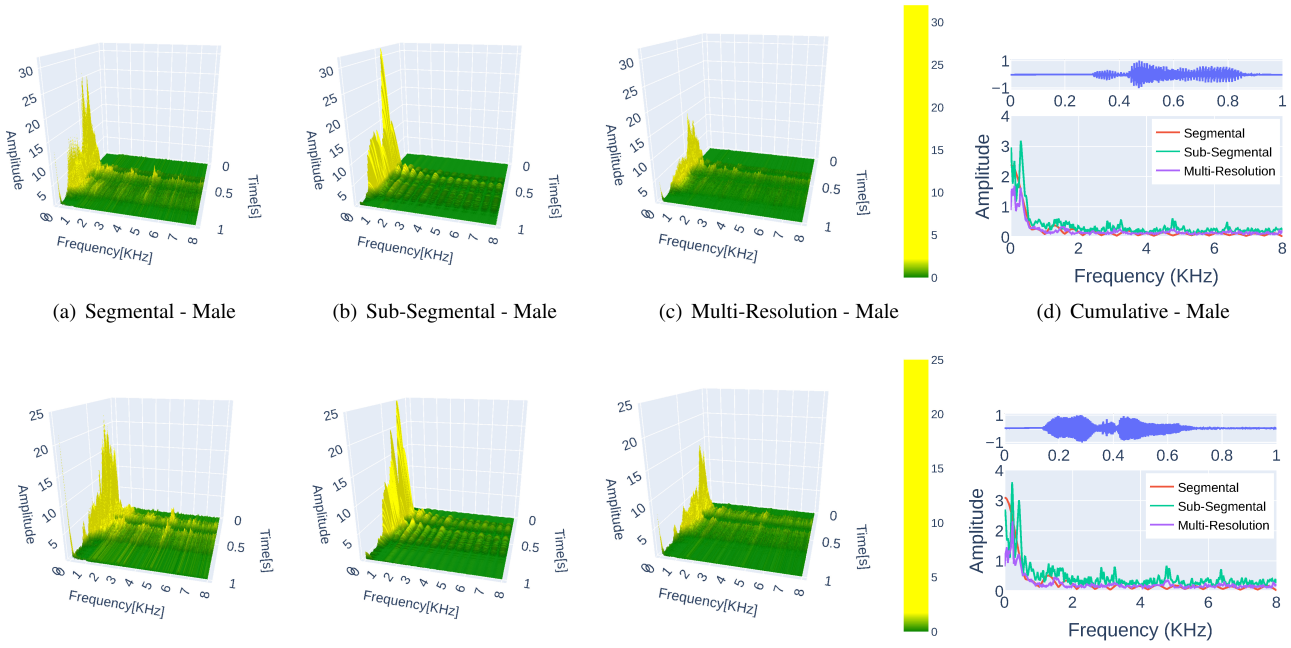
We perform our experiments on two tasks, namely speech classification and speaker identification using Google's speech commands [20] and LibriSpeech[train-other(100hrs, 251 speakers)] [21] dataset, respectively with 60:20:20-train/val/test splits. All models are trained with Adam optimizer, an initial learning rate of 10^{-3} for 100 epochs on a workstation utilizing a single K80 GPU. Extended results shown for spoofing task uses ASVSpoof19 [22] test set.

3.1. Models

In the first two tasks, we consider three models with segmental, sub-segmental, and multiresolution 3-layer CNN front-ends for feature extraction followed by 2 dense layers for classification [23]. Segmental model's CNN layers uses kernel width [300, 7, 3], stride [10, 1, 1] and filters [80, 80, 100]. The sub-segmental model differs only in the first layer kernel width of 30. Similarly, the multiresolution model's CNN layers use 10 filters each of kernel width [10, 30, 50, 70, 90, 150, 300, 500] and stride 10 in the first layer, 20 kernels each of width [3, 6, 9, 12] and stride 1 in the second layer, and 25 kernels each of width [3, 5, 7, 9] and stride 1 in the last convolutional layer. Each CNN layer is followed by an ELU activation, batch-normalization, and max-pooling. CNN features are fed to 2 dense layers with ReLU activation and dropout regularization. The dimension of dense layers is 1024 and the number of classes, respectively. Our trained models achieved test accuracies of 99.81, 99.85, and 99.82% for the speech classification and 88.14, 88.58 and 91.24% in the speaker identification task on the sub-segmental, segmental, and multiresolution models, respectively. For the spoofing classification task, we used pretrained SOTA models based on ResNet/InceptionNet architecture from [17]. ResNet and InceptionNet models are based on sub-segmental and multiresolution kernels, respectively. We performed the analysis on the ASVSpoof19 test set.

3.2. Time Frequency Analysis

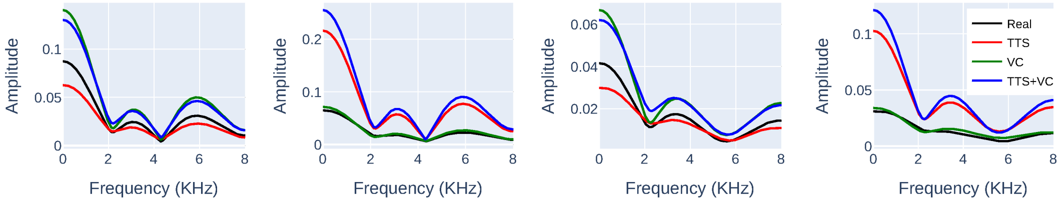
Figs. 1 and 2 shows the original waveform and corresponding STRS and CRS maps obtained from models trained on speech and speaker tasks. We observe that the choice of block-processing in terms of segmental, sub-segmental and multi-resolution front-ends impacts the TF behaviour of the model. This behaviour is task-dependent, the information of interest is localized in both time and frequency, and the results are consistent for different speakers/genders. These visualizations show that the band around 1kHz [i.e., first two formants] is important for capturing speech characteristics while the band around



(a) Segmental - Male (b) Sub-Segmental - Male (c) Multi-Resolution - Male (d) Cumulative - Male

(e) Segmental - Female (f) Sub-Segmental - Female (g) Multi-Resolution - Female (h) Cumulative - Female

Fig. 2. STRS and CRS of example inputs for model trained on speaker identification task



(a) Male - InceptionNet (b) Female - InceptionNet (c) Male - ResNet (d) Female - ResNet

Fig. 3. CRS of example inputs for pre-trained spoofing model

Table 1. Spoofing Classification using CRS features

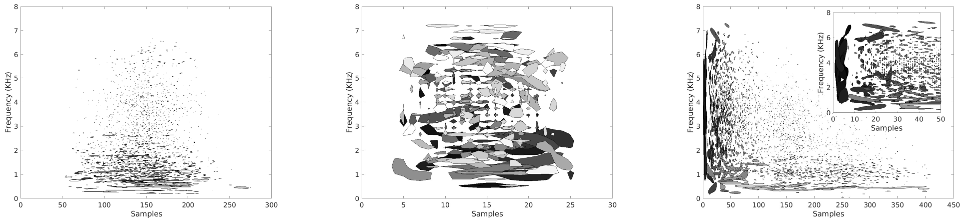
Model Class	ResNet	InceptionNet
TTS	0.90 ± 0.015	0.89 ± 0.017
TTS + VC	0.91 ± 0.02	0.90 ± 0.01
VC	0.67 ± 0.056	0.62 ± 0.043

500Hz [i.e., pitch and first formant] is important for speaker task- These observations reveal that acoustic models are indeed implicitly biased towards a function with more power at low frequencies during the training. Secondly, our quantitative analysis revealed that there are two high-frequency regions that are emphasized: between 2-3.5kHz and 4-5kHz, which is consistent with existing studies that found mid/high frequencies to be speaker discriminative [24, 25].

To demonstrate the usefulness of our visualizations for large scale tasks, we extend our study to deep pretrained ResNet & InceptionNet based models for the spoofing task. Due to space constraints, we only show the CRS maps from these models for both real and spoofed examples (generated via Text-To-Speech (TTS), Voice Changer (VC), and both) in Fig. 3. Consistent with the earlier results, spectral visualization also reveals important cues for understanding deep models. For instance, observe how the CRS maps are different for real and fake inputs, which shows the discriminative ability of the feature processing front-end and the information bottleneck resulting from chosen block processing. To quantify this behaviour, we trained a SVM Classifier with a polynomial kernel of degree 3 using a CRS map computed from respective models as input features to classify real/fake audio clips on the ASVSspoof19 test set. These results are reported in Table 1 which demonstrate that except for the VC case, the first layer itself is capturing important features for discrimination. This shows the effectiveness of raw-waveform models in capturing task-dependent information using learned filters compared to conventional Fourier/Mel-scale filters.

3.3. Geometric Analysis using Dictionary Metrics

Similar to TF visualization, analysis of geometric properties of network weights provides insights into learning behaviour with the advantage of being data-independent and applicable to all layers. Few of such properties (averaged over 10 trials) for CNN



(a) Segmental

(b) Sub-Segmental

(c) Multi-Resolution

Fig. 4. WV contour plots (at threshold of 0.7) of learned filters for speech classification task. The grey scale represents the total energy.

Table 2. Geometric properties for speech classification model

Layer	Sub-Segmental			Segmental			Multi-Resolution		
Metric	I	II	III	I	II	III	I	II	III
A	0.15	0	0	0	0	0	0	0	0
B	6.46	5.09	4.42	2.51	4.82	4.09	4.08	4.30	4.60
β_{avg}	61.5	74.4	78.1	79.3	76.2	78.6	72.2	76.5	79.2
μ	0.62	0.63	0.57	0.29	0.54	0.45	0.69	0.67	0.45
μ_{avg}	0.47	0.26	0.20	0.18	0.23	0.19	0.30	0.22	0.18

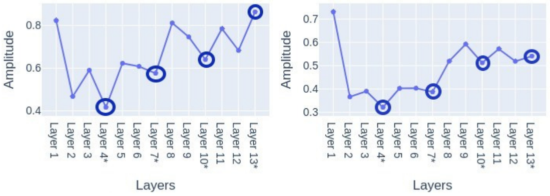
Table 3. Geometric properties for speaker identification model

Layer	Sub-Segmental			Segmental			Multi-Resolution		
Metric	I	II	III	I	II	III	I	II	III
A	0.54	0	0	0	0	0	0	0	0
B	14.86	7.36	6.45	8.11	6.82	6.80	5.54	5.20	4.64
β_{avg}	51.9	66.0	73.7	65.7	68.3	72.9	64.4	73.4	77.8
μ	0.96	0.88	0.84	0.85	0.84	0.69	0.65	0.78	0.64
μ_{avg}	0.59	0.39	0.27	0.40	0.36	0.29	0.42	0.28	0.20

layers in speech and speaker models are reported in Table 2 and 3, respectively. Geometric properties are typically related e.g., $\mu_{mse}^2 = \sum \lambda_i^2/(\sum \lambda)^2; \mu = \cos \beta_{min}$; but $\mu_{avg} \neq \beta_{avg}$ (see [16] for more details). We observe that layer-wise geometric properties are more consistent for models with multi-resolution kernels, which correlates well with better performance than segmental and sub-segmental models. For instance, the average mutual coherence for the multi-resolution model is the lowest, which is expected due to its capability to learn more independent features. To further demonstrate the link between the TF behaviour of filters and the generalization capability of a model, we analyze the learned filters using the Wigner-Ville (WV) distribution [18]. In Fig. 4, the filters can be grouped based on TF spread as: many narrow-band contours in the lower frequency band (like sinusoidal components), several compact wide-band contours, and a few with fragmented WV distribution. Importantly, we observe how information localisation in the TF plane changes for different kernels, where the multi-resolution front-end simultaneously captures the localized relationships in time and frequency. This explains the reason for a higher difference between the upper and lower bound of our learned models [16]. Model's performance improves if the learned filters are evenly distributed in both lower and higher bands covering the whole TF plane. This concept can be interpreted in training networks on orthogonal manifolds, which have shown to be theoretically and empirically the optimal choice in existing studies. Interestingly, geometric properties also reveal the impact of architectural choice for a model. The layer-wise properties of InceptionNet for spoofing task is reported in Table 4. Each layer(II-V) is a multi-resolution block consisting of multiple dilated CNN layers. Observe that for such a complex model value of various metrics is higher. Here, while the model deviates from ideal desired properties, it compensates by achieving better performance due to high expressivity with depth. We observed similar trends for the ResNet model which uses sub-segmental kernels. However, due to space constraints we focus on the impact of skip connections. Fig. 5 shows the layer-wise coherence properties of the ResNet model. Observe that μ and μ_{avg} increase with depth, but the skip connection inhibits the increase, hence the resultant features are less dependent. This can be interpreted in terms of conditioning of the network's Jacobian for better signal propagation [26, 27].

Table 4. Geometric properties of pre-trained net for spoofing task

Metrics	Layer I	Layer II	Layer III	Layer IV	Layer V
A	0.3457	0.0	0.0	0.0	0.0
B	4.5024	3.0179	3.7923	8.8337	12.5639
β_{avg}	41.2059	72.0779	70.5769	65.8626	62.1708
μ	0.8925	0.4402	0.5149	0.7679	0.7788
μ_{avg}	0.7434	0.307	0.3313	0.4056	0.462



(a) μ

(b) μ_{avg}

Fig. 5. Layerwise properties of pre-trained ResNet for spoofing task. Layers with residual connection are marked with circle.

4. CONCLUSION

In this paper, we presented both data-dependent/independent tools for analyzing the spectral and geometric information captured by raw waveform-based acoustic models. With case studies on three different tasks, we demonstrated how deep models flexibly integrate information from raw waveforms using TF visualizations and how changes in geometric properties correlate with the architectural choices and generalization error. In particular, multi-resolution models are more effective in learning task-dependent uncorrelated features and capturing hierarchical TF relationships from the raw waveform. In the future, we will develop a theoretical explanation of observed behaviours and extend our study to more complex speech tasks such as TTS, speech recognition, and information retrieval.

5. REFERENCES

[1] K. Chen, Y. Wu, Z. Wang, X. Zhang, F. Nian, S. Li, and X. Shao, “Audio captioning based on transformer and pre-trained CNN,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, Nov. 2020, pp. 21–25.

[2] K. Koutini, H. Eghbal-zadeh, M. Dorfer, and G. Widmer, “The Receptive Field as a Regularizer in Deep Convolutional Neural Networks for Acoustic Scene Classification,” in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, A Coruña, Spain, 2019.

[3] V. Abrol and P. Sharma, “Learning hierarchy aware embedding from raw audio for acoustic scene classification,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1964–1973, 2020.

[4] D. Palaz, M. Magimai-Doss, and R. Collobert, “End-to-end acoustic modeling using convolutional neural networks for hmm-based automatic speech recognition,” *Speech Communication*, vol. 108, pp. 15–32, 2019.

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

[6] A. Mohamed, G. E. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[7] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn, “Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4277–4280.

[8] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. G.-Dominguez, “Deep neural networks for small footprint text-dependent speaker verification,” in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 4052–4056.

[9] T. Kim, J. Lee, and J. Nam, “Sample-level cnn architectures for music auto-tagging using raw waveforms,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 366–370.

[10] D. Tang, P. Kuppens, L. Geurts, and T. van Waterschoot, “End-to-end speech emotion recognition using a novel context-stacking dilated convolution neural network,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 18, May 2021.

[11] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, “SmoothGrad: removing noise by adding noise,” in *ICML workshop on visualization for deep learning*, 2017, pp. 1–10.

[12] J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for simplicity: The all convolutional net,” in *International Conference on Learning Representations (ICLR)*, 2015, pp. 1–14.

[13] Mukund Sundararajan, Ankur Taly, and Qiqi Yan, “Axiomatic attribution for deep networks,” in *International Conference on Machine Learning (ICML)*, 2017, p. 3319–3328.

[14] H. Muckenhirn, V. Abrol, M. Magimai-Doss, and S. Marcel, “Understanding and visualizing raw waveform-based cnns,” in *Interspeech 2019*. 2019, pp. 2345–2349, ISCA.

[15] Weili Nie, Yang Zhang, and Ankit Patel, “A theoretical explanation for perplexing behaviors of backpropagation-based visualizations,” in *International Conference on Machine Learning (ICML)*, 2018, pp. 3806–3815.

[16] K. Skretting and K. Engan, “Learned dictionaries for sparse image representation: properties and results,” in *Wavelets and Sparsity XIV*. International Society for Optics and Photonics, 2011, vol. 8138, pp. 404 – 417, SPIE.

[17] Guang Hua, Andrew Beng Jin Teoh, and Haijian Zhang, “Towards end-to-end synthetic speech detection,” *IEEE Signal Processing Letters*, vol. 28, pp. 1265–1269, 2021.

[18] V. Abrol, P. Sharma, and A.K. Sao, “Greedy double sparse dictionary learning for sparse representation of speech signals,” *Speech Communication*, vol. 85, pp. 71–82, 2016.

[19] Samer A. Abdallah and Mark D. Plumbley, “If the independent components of natural images are edges, what are the independent components of natural sounds?,” in *International Workshop on Independent Component Analysis and Blind Separation (ICA)*, September 2001, pp. 534–539.

[20] P. Warden, “Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition,” *ArXiv e-prints*, Apr. 2018.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[22] Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Hector Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee, “Asvspoof 2019: Future horizons in spoofed and fake audio detection,” 2019.

[23] V. Abrol, S.P. Dubagunta, and M. Magimai-Doss, “Understanding raw waveform based CNN through low-rank spectro-temporal decoupling,” Tech. Rep., Idiap, October 2019.

[24] Tomi Kinnunen, *Spectral features for automatic text-independent speaker recognition*, Ph.D. thesis, Department of Computer Science, University of Joensuu, 2003.

[25] Laura Fernández Gallardo, Michael Wagner, and Sebastian Möller, “Spectral sub-band analysis of speaker verification employing narrowband and wideband speech,” in *Odyssey: The Speaker and Language Recognition Workshop*, January 2014, pp. 81–87.

[26] Wojciech Tarnowski, Piotr Warchol, Stanislaw Jastrzkebski, Jacek Tabor, and Maciej A. Nowak, “Dynamical isometry is achieved in residual networks in a universal way for any activation function,” in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, April 2019, pp. 2221–2230.

[27] M. Murray, V. Abrol, and J. Tanner, “Activation function design for deep networks: linearity and effective initialisation,” *Applied and Computational Harmonic Analysis*, 2022.