INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# A Study on Visualization of Voiceprint Feature

*Jian Zhang*$^{1,2}$, *Liang He*$^{1,2,3,\dagger}$, *Xiaochen Guo*$^{1,2}$, *Jing Ma*$^{1,2}$

[1]School of Information Science and Engineering, Xinjiang University, Urumqi 830017, China
[2]Xinjiang Key Laboratory of Signal Detection and Processing, Urumqi 830017, China
[3]Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

zhangjian@stu.xju.edu.cn, heliang@mail.tsinghua.edu.cn

## Abstract

Despite the remarkable success of convolutional neural networks (CNNs) in voiceprint recognition, we still lack a comprehensive understanding of the specific features extracted by these models. To address this issue, we adopt an attribution approach in this paper to explain the voiceprint identification model and visualize the relevant features. Using five attribution methods, we successfully identify the features extracted by the ECAPA-TDNN model and confirm the reliability of our attribution techniques.We also explore two distinct methods for visualizing voiceprint features, with one approach aimed at interpreting features in unknown speech and the other focused on known speech. Through the attribution method, we are able to more precisely capture voiceprint features within speech data without significantly impacting the performance of the voiceprint recognition model. It would help us to do a more detailed study of the voiceprint features in the future.

**Index Terms**: visualization, voiceprint features, attribution

## 1. Introduction

With the remarkable success of deep learning [1, 2, 3] comes an increasing concern about its black-box nature. The lack of interpretability in machine learning models can negatively impact the level of trust in deep learning systems. This is especially true for applications where machines make predictions or decisions related to critical aspects of human life, health, safety, and property. The interpretability of deep learning models [4, 5] has become a crucial factor in determining whether users can trust these models. By providing insights into the reasoning behind the model's predictions and decisions, interpretability can improve the transparency of the system, enhancing the user's confidence and trust. As a result, researchers have been exploring various approaches to improve the interpretability of deep learning models, making them more accessible and understandable to users.

The remarkable success of deep convolutional neural networks (CNNs) in various fields is closely tied to the concept of interpretable learning. By utilizing various visualization tools, it is possible to interpret the decisions made by these models and continuously optimize them. While much of the research in voiceprint recognition models has focused on improving the accuracy [1, 6, 7] of predictions and studying speech representations [8, 9, 10, 11], these studies have greatly facilitated the extraction of voiceprint features by deep CNNs and the interpretation of voiceprint recognition model predictions using visualization tools. However, current research on visualization tools is more prevalent in the field of computer vision, such as the use of imputation methods to trace model predictions back to the input image and visualize the imputation results. This enables the identification of which pixels in the input image are influential in determining the output results, providing intuitive insights into the model's decision-making process. Despite the lack of attention to visualization tools in voiceprint recognition, they offer great potential for improving the interpretability and trustworthiness of these models.

Currently, the mainstream attribution methods used for deep neural networks include LIME [12], Integrated Gradients [13], DeepLift [14], DeconvNet [5], Guided Backpropagation [15], CAM [16, 17], SHAP [18, 19] and other methods. In [13], The authors of this paper identify two basic axioms of attribution: Sensitivity and Implementation Invariance, and design a new attribution method called Integrated Gradients based on these axioms. Many attribution methods before this paper did not satisfy these two axioms, and Integrated Gradients is simple to use and does not require modifying the structure of the original model.In [12], The authors use interpretable models such as linear models and decision trees to locally approximate the prediction of the target black box model. This method detects changes that occur in the output of the black box model by slightly perturbing the input and trains an interpretable model at the point of interest based on this change. The SHAP [18] model is a more versatile approach to model interpretability, which can be used for both global and local interpretations of the relationship between predicted values and certain features in a single sample. SHAP constructs an additive explanatory model in which all features are considered as "contributors." For each prediction sample, the model generates a prediction value, and the SHAP value is the value assigned to each feature in that sample. In [20], The authors use three variants of CAM (Grad-CAM [21], Score-CAM [22], and Layer-CAM [23]) to visualize the voiceprint features, making this paper the only one to explain the features extracted by the voiceprint recognition model. The main idea is to determine the position of the voiceprint features on the spectrogram by masking the effect of the part of the input features on the output results. Their experiments on ResNet34SE [24] demonstrate the effectiveness of the three methods. In this paper, we first experimented the two implementation paths on images as a way to verify the feasibility of the method, and finally on ECAPA-TDNN. The contributions of this paper are as follows:

- In this paper, we successfully use five attribution methods applied to the voice recognition model and make a verification of the reliability of the five attribution methods.
- We used two ways to analyze the voiceprint features extracted by the voiceprint recognition model, one for known speech

This work was supported by the National Key R&D Program of China under Grant No. 2022ZD0115801. Experimental data available: https://github.com/xju-sword/vis-voiceprint

[†] Corresponding author.

10.21437/Interspeech.2023-1286

(speech in the training set) and the other for unknown speech (speech not in training).

# 2. Related work

## 2.1. Neural Network Interpret

In [25], the authors provide an exhaustive classification of the interpretability of neural networks. The interpretability is classified into two major categories: passive [26] and active [27, 28]. Passive: Post hoc explain trained neural networks. Active: Actively change the network architecture or training process for better interpretab.

It has been noted that the captum toolkit includes a large number of attribution algorithms. These algorithms [29] have been tried for the visualization of voiceprint features, and five of them have been successfully applied in this study. However, it is important to note that visualizing acoustic features is different from visualizing features in computer vision. In computer vision[30], we can easily observe whether important features are visualized or not, whereas in the case of acoustic features, it is difficult to determine if the features are actually visualized. Therefore, to validate the visualization of acoustic features, another experiment was designed using the attribution algorithm backpropagation for acoustic features. The results of this experiment demonstrated that the visualization of acoustic features achieved in this study is reliable.

## 2.2. ECAPA-TDNN

ECAPA-TDNN, proposed in [1], has been widely regarded as the state-of-the-art system in voiceprint recognition.

Table 1: *The proposed ECAPA-TDNN architecture, which mainly integrates three modules: SE-Res2Block, Multi-layer feature aggregation and summation(MFA), Attentive statistic pooling(ASP).*

| Layer name | Output size | Structure |
|---|---|---|
| Input | - | C = 512, 80 × T |
| conv1d | 512 × T | C, k=5, p=2 |
| SE-Res2Block-1 | 512 × T | C, k=3, d=2 |
| SE-Res2Block-2 | 512 × T | C, k=3, d=3 |
| SE-Res2Block-3 | 512 × T | C, k=3, d=4 |
| MFA | 1536 × T | 3 × C, k=1, d=1 |
| ASP | 3072 × 1 | - |
| FC | 192 × 1 | - |
| AAM-Softmax | 1211 × 1 | - |

In deep learning models for voiceprint recognition, the final classification layer FC is usually used to compute the loss during training and to make predictions during testing. The embedding layer, which outputs a fixed-length vector that represents the input speech signal, is used as the input to the final classification layer. The purpose of the final classification layer is to map the embedding vector to a specific speaker ID or speaker-independent class label.

# 3. Methodology

## 3.1. Attribution algorithm

### 3.1.1. Integrated Gradients

In [13], the authors assume that there is a function $F : R^n \in [0,1]$ representing a deep network, and an input $x = (x_1, ..., x_n) \in R^n$. The authors define a tensor $x' = (1, ..., 1) \in R^n$ ( When the input is a picture) as the baseline. An attribution of the prediction at input $x$ relative to a baseline input $x'$ is a vector $A_F(x, x^{'}) = (a_1, ..., a_n) \in R^n$. $a_i$ is the contribution of $x_i$ to the prediction $F(x)$. Here, $\frac{\partial F(x)}{\partial x_i}$ is the gradient of $F(x)$ along the $i^{th}$ dimension.

$$a_i = (x_i - x_i^{'}) \times \int_{\alpha=0}^{1} \frac{\partial F(x^{'} + \alpha \times (x - x^{'}))}{\partial x_i} d\alpha \quad (1)$$

### 3.1.2. SHAP

Define a simpler model as any interpretable approximation of the original model. Let $f$ be the original prediction model to be interpreted and g be the interpreted model. Here we focus on the local approach used to interpret the prediction $f(x)$ based on a single input $x$. As proposed in LIME: the explanatory model typically uses a simplified input $x'$ that is mapped to the original input via the mapping function $x = h_x(x')$, and when $x' \approx z'$, the local approach attempts to ensure that $g(z') \approx f(h_x(z'))$

$$g(z^{'}) = \phi_0 + \sum_{i=1}^{M} \phi_i z_i^{'} \quad (2)$$

where $z' \in [0,1]^M$ and $M$ is the number of simplified input features. The explanatory model calculates a contribution $\phi_i$ for each feature $x_i$, i.e., $\phi_i$ is the feature-attributed Shapley value for feature $i$.The sum of the contributions of all feature attributes approximates the output of the original model $f(x)$ .

## 3.2. Realization path

### 3.2.1. Traversal Method

In this study, the speech data is first converted into Mel-frequency cepstral coefficients (MFCCs), which are used as input $x = (x_1, ..., x_n)$ to the voiceprint recognition system $F(x)$. To establish a baseline, $x'$ is defined to be of the same type and size as $x$, but with all elements set to 1.

$$e = F(x) \quad (3)$$

$$A_{e_i}(x, x^{'}) = (a_{1i}, a_{2i}, a_{3i}, ..., a_{ni}) \quad (4)$$

$$A_F = \frac{1}{N} \times \sum_{i=1}^{N} A_{e_i}(x, x^{'}) \quad (5)$$

After passing the input $x$ through the voice recognition system $F(x)$ , a speaker embedding of output length $N$ is obtained. Each element in the speaker embedding $e = (e_1, ..., e_N)$ is generated based on the input $x$, making it reasonable to use each element($e_i$) in $e$ as the results for the system prediction. $A_{e_i}$ denotes the attribution of the input to the $e_i$ in the speaker embedding. Finally, $A_F$ can be understood as attributing input $x$ to the whole embedding. This is shown in Fig. 1.

### 3.2.2. Classification Method

Recently, researchers have applied attribution algorithms to classification models. In other words, the model's output is a two-dimensional tensor corresponding to the probability of the category predicted by the model, with the maximum probability as the target of the attribution algorithm. However, researchers usually take the speaker's embedding as the model's output in voiceprint recognition. And the classification layer is generally placed in the loss. So borrowing from previous researchers, we migrate the classification layer in the loss to the last layer of the model. $G(x)$ is our new model after adding the classification layer, and $A_G$ is the attribution. This is shown in Fig. 2.

$$G(x) = FC_{AAM-Softmax}(F(x)) \tag{6}$$

$$A_G(x, x^{'}) = (a_1, a_2, a_3, ..., a_n) \tag{7}$$

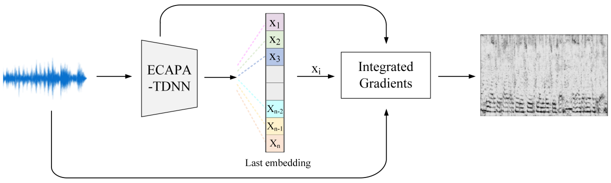

Figure 1: *The input is involved in training or not: the speech goes through ECAPA-TDNN, the last layer of embedding is output, and $x_i$ denotes the i-th value in the embedding.*

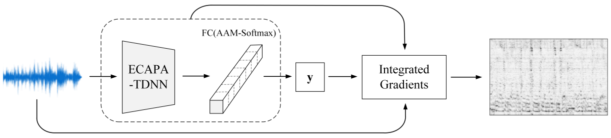

Figure 2: *Input needs to be involved in training, y is the result of the model prediction.*

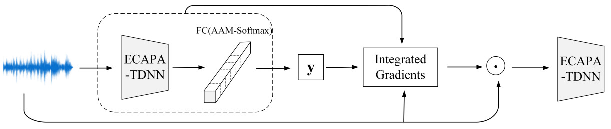

Figure 3: *Verify the reliability of the attribution algorithm.*

### 3.3. Reliability

#### *3.3.1. Realization of ideas*

The verification process described in the paper involves using the attribution method to identify important features in the input speech and forming attribution values on the original speech data points. These attribution values are then fed into the model for training, and the final result is compared to the result without the attribution method. If the final result is similar, the attribution algorithm is deemed reliable, otherwise, it is deemed unreliable. The paper explains that overwriting a part of the data to continue training does not affect the accuracy of the final voice recognition, which suggests that the data fed into the model still contains the voice features. This process aims to verify the reliability of the attribution algorithm when applied to the voiceprint recognition model.

#### *3.3.2. Make ready*

The dataset used in the study, which is voxceleb1. This dataset contains speech recordings of 1251 celebrities, extracted from short video clips on YouTube. We chose this dataset because it is publicly available and contains a large number of speakers, making it suitable for training a voiceprint recognition model. We trained an ECAPA-TDNN model on this dataset and saved its parameters as model_baseline. We also created a model_classification for voiceprint recognition by extracting the FC layer from AAM-Softmax and connecting it to the embedding of the last layer. The purpose of this study was to verify the reliability of attribution methods when applied to the voiceprint recognition model.

#### *3.3.3. Validation*

The process described here involves using the Integrated Gradients algorithm with the input speech data, prection (prediction result), model_classification, and model_baseline as inputs to obtain the attribution value of important features. This attribution value is then multiplied with the initial input speech to obtain the resulting input of ECAPA-TDNN. Additionally, to improve the efficiency of training the voiceprint recognition model, all the voice data of voxceleb1 is processed by the attribution algorithm and saved, so that during validation experiments, only the processed data needs to be imported to ECAPA-TDNN. This is shown in Fig. 3

## 4. Experiments

The reliability of the attribution method is critical to the accuracy and validity of the subsequent analysis of the voiceprint features. The experiments in the study demonstrated that the attribution method used is reliable and produces results that are generally consistent with the original ECAPA-TDNN model training. The use of masks and different thresholds in the validation experiments allowed for the filtering out of smaller attribution values, which slightly reduced the accuracy of the results, but the effect on the visualization was negligible. In fact, the trade-off between accuracy and visualization efficiency is worthwhile(compare Figure 5 and Figure 6), as it allows for the identification of important voiceprint features more efficiently. Overall, the reliability of the attribution method is important in forensic identification and the study supports the usefulness of the method in identifying important voiceprint features.

Table 2: *Comparing the effect of adding attribution methods with and without attribution methods and changing the threshold value of mask on model performance.*

| **Model** | **Attribution algorithm** | **EER(%)** |
|---|---|---|
| ECAPA-TDNN | None | 3.48 |
| | Integrated Gradients | 4.21 |
| | Integrated Gradients+mask(0.1) | 4.29 |
| | GradientsShap | 4.55 |

To return to the topic, our main work is to explain the voiceprint recognition model. In order to understand the voiceprint features extracted by the voiceprint recognition model, we successfully applied five attribution algorithms and two pathways for the visualization of the voiceprint features. This is shown in Figure 4, Figure 5, and Figure 6.

Based on the analysis of Fig. 5, we can draw several conclusions: Firstly, The voiceprint recognition system still relies on non-speaker information in the final extracted high-level features, indicating that other factors beyond individual voiceprint information influence the system's classification of speakers. Secondly, The darker color in the figure indicates

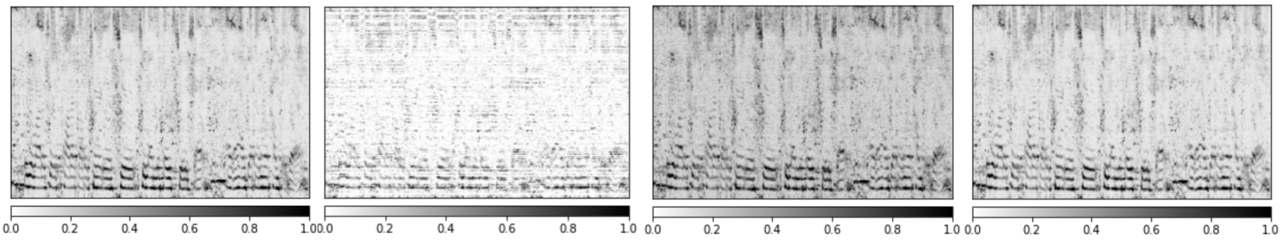

Figure 4: *Visualization of voiceprint features for four attribution methods other than the attribution algorithm Integrated Gradients.*

the feature's robust association with the speaker, suggesting that these features are more reliable for identifying speakers. Thirdly, The Traversal Method, which traverses all the elements in the embedding, produces a darker color than the Classification Method, indicating that it enhances the voiceprint features and leads to more reliable results. Overall, these findings suggest that the voiceprint recognition system relies on more than just individual voiceprint information when classifying speakers.

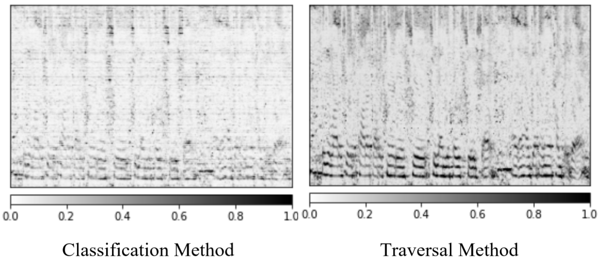

Figure 5: *Comparison of two ways to visualize voiceprint features. input: MFCCs.*

Based on the validation experiments, it was found that setting the mask threshold to 0.1 (filtering out attribution values less than 0.1) resulted in the training of the voiceprint recognition model being one percentage point smaller than the EER without adding mask. Although this result is acceptable for the visualization of voiceprint features, it is essential to keep in mind that setting the mask threshold may result in loss of some important information.Setting the threshold of the mask helps to visualize the voiceprint features with different levels of importance, which can reduce the time required to find the voiceprint features of a particular person during the identification of voiceprint patterns. However, it is important to carefully select the threshold value to ensure that the resulting feature captures the essential characteristics of the voiceprint accurately. In summary, setting the mask threshold can aid in visualizing voiceprint features, but it is necessary to carefully balance between information loss and reducing the time required for feature identification. As shown in Fig. 6

As shown in Figure 4, based on the experiments conducted with the four additional attribution algorithms, it has been observed that the shapley value sampling algorithm performs the worst in terms of sound pattern visualization. The visualized voiceprint features are less compared to the other four algorithms, as can be seen from the figure. Additionally, it has been observed that the human voiceprint features are mainly concentrated in the fundamental frequency, specifically in the first and second resonance peaks. Finally, it is noted that the visualized voiceprint patterns generated by any of the algorithms contain non-voiceprint pattern information. This indicates that the current voiceprint recognition systems do not effectively separate the voiceprint information from non-voiceprint information.

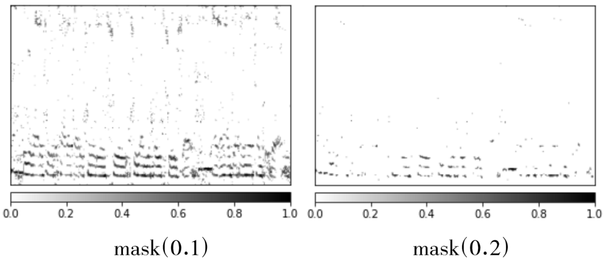

Figure 6: *Features of different importance are extracted according to the threshold value.*

## 5. Conclusions

The use of attribution methods in voiceprint recognition can help professionals quickly identify important voiceprint features that contribute to identifying the speaker's identity. This can improve the efficiency and accuracy of the identification process. Additionally, by comparing the performance of different voiceprint recognition systems using attribution methods, professionals can choose a system that better suits their needs and is more reliable in identifying voiceprints. Overall, the use of attribution methods can greatly enhance the capabilities of professionals in forensic identification, making the process faster and more accurate.

It is interesting to note that while the attribution methods were effective in explaining the voiceprint recognition model, they did not completely decouple the content of the speech from the voiceprint features. This suggests that there may be non-voiceprint features present in the visualizations obtained through attribution. The research highlights the importance of evaluating voiceprint recognition models beyond just their accuracy, particularly in fields such as voiceprint identification and criminal investigation. Moving forward, we plan to improve the generalization ability of the model by focusing on stable learning and decoupling content information from voiceprint features. This will enhance the model's ability to accurately extract voiceprint features and improve its reliability in practical applications.

# 6. References

[1] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification,” in *Proc. Interspeech 2020*, 2020, pp. 3830–3834.

[2] Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, and K. Yu, “Deep feature for text-dependent speaker verification,” *Speech Communication*, vol. 73, pp. 1–13, 2015.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[4] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: visualising image classification models and saliency maps.” ICLR, 2014, pp. 1–8.

[5] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, 2014, pp. 818–833.

[6] K. Okabe, T. Koshinaka, and K. Shinoda, “Attentive Statistics Pooling for Deep Speaker Embedding,” in *Proc. Interspeech 2018*, 2018, pp. 2252–2256.

[7] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[8] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, “Deep neural networks for small footprint text-dependent speaker verification,” in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[9] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, “Deep neural network embeddings for text-independent speaker verification.” in *Interspeech*, vol. 2017, 2017, pp. 999–1003.

[10] N. Chen, Y. Qian, and K. Yu, “Multi-task learning for text-dependent speaker verification,” in *Sixteenth annual conference of the international speech communication association*, 2015.

[11] Z. Chen, S. Chen, Y. Wu, Y. Qian, C. Wang, S. Liu, Y. Qian, and M. Zeng, “Large-scale self-supervised speech representation learning for automatic speaker verification,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6147–6151.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin, “” why should i trust you?” explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[13] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[14] A. Shrikumar, P. Greenside, and A. Kundaje, “Learning important features through propagating activation differences,” in *International conference on machine learning*. PMLR, 2017, pp. 3145–3153.

[15] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for simplicity: The all convolutional net,” in *ICLR (workshop track)*, 2015. [Online]. Available: http://lmb.informatik.uni-freiburg.de/Publications/2015/DB15a

[16] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[17] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks,” in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.

[18] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in neural information processing systems*, vol. 30, 2017.

[19] M. Ancona, C. Oztireli, and M. Gross, “Explaining deep neural networks with a polynomial time algorithm for shapley value approximation,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 272–281.

[20] P. Li, L. Li, A. Hamdulla, and D. Wang, “Reliable Visualization for Deep Speaker Recognition,” in *Proc. Interspeech 2022*, 2022, pp. 331–335.

[21] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[22] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, “Score-cam: Score-weighted visual explanations for convolutional neural networks,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 24–25.

[23] P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, “Layercam: Exploring hierarchical class activation maps for localization,” *IEEE Transactions on Image Processing*, vol. 30, pp. 5875–5888, 2021.

[24] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, “Utterance-level aggregation for speaker recognition in the wild,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5791–5795.

[25] Y. Zhang, P. Tiňo, A. Leonardis, and K. Tang, “A survey on neural network interpretability,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2021.

[26] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas *et al.*, “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav),” in *International conference on machine learning*. PMLR, 2018, pp. 2668–2677.

[27] G. Plumb, M. Al-Shedivat, Á. A. Cabrera, A. Perer, E. Xing, and A. Talwalkar, “Regularizing black-box models for improved interpretability,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 10 526–10 536, 2020.

[28] M. Wojtas and K. Chen, “Feature importance ranking for deep learning,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 5105–5114, 2020.

[29] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, “Sanity checks for saliency maps,” *Advances in neural information processing systems*, vol. 31, 2018.

[30] X. Zhang, P. Cui, R. Xu, L. Zhou, Y. He, and Z. Shen, “Deep stable learning for out-of-distribution generalization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 5372–5382.