Explaining the Predictions of the MOSNet Classifier via the LIME Framework

Ada Lamba*
Department of Computer Science and Engineering
The Ohio State University
lamba.39@osu.edu

Abstract

Deep neural networks are used widely in speech assessment to predict the quality of audio signals. For example, a recent model, MOSNet, predicts the mean opinion score (MOS) of audio signals. However, many networks use objective metrics which correlate poorly with human perceptual scores [1]. To better understand this poor correlation, one needs to understand what aspects of an audio signal networks are focusing on when making their decisions. In this work, we apply a recent explanation framework, LIME, to better understand MOSNet's decision-making process [2]. Our results show that MOSNet focuses on lack of noise and frequency, but the results are non-intuitive and difficult to understand.

1 Introduction

Speech assessment is primarily concerned with two metrics: *quality* and *intelligibility*. Quality describes how well the signal is perceived (e.g., a noisy signal typically has a low quality score) and can be highly subjective. Intelligibility describes how well the speech in a signal is understood and is objective (i.e., the listener either understood what words were spoken in a recording, or they did not). Approaches for measuring quality or intelligibility can be either *subjective* or *objective*. Subjective measures utilize human listeners to provide a quality or intelligibility score while objective measures produce a score by either comparing the signal to a "clean" reference signal or utilizing signal properties.

Subjective measures are the gold standard but are expensive as they require time and money to employ the human listeners. As a result, many recent approaches instead use objective measures [3, 4, 5, 6]. While objective measures are typically cheaper than subjective measures, the resulting scores do not correlate highly with human-provided scores.

In recent years, deep neural networks (DNNs) have become the standard approach for machine learning algorithms that predict the quality score of a signal and many use objective metrics to avoid the cost associated with human subjects. However, these networks are not perfect and often make mistakes or exhibit low correlation with human predictions. To better understand *why* these networks are inaccurate or poorly correlated with human perceptive scores, researchers need to understand how the network came to its decision. This is, broadly, the field of network *explainability*.

Neural networks are not, technically, black-boxes. They are well-defined mathematically and, given a specific input, one can calculate what the resulting output from the network would be. However, as the networks grow in complexity, the number of nodes and connections makes it intractable to reason on a broad scale about how an arbitrary input influences the network's output. This is where explainability comes in. Explainability approaches aim to determine which parts of the network input are particularly relevant in the network's decision.

*CSE 5539 - Khalili AU23

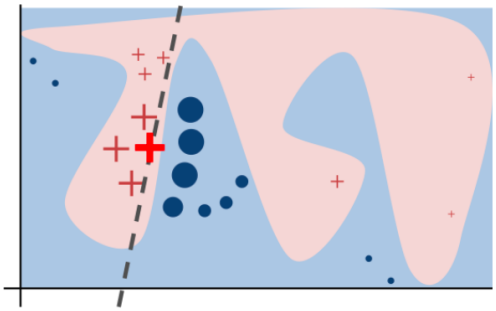


Figure 1: A toy example presented in the original lime publication [2] which illustrates how LIME generates its explanations. In this example, the red/blue boundary is the decision boundary of the target model and the red bold cross is the target data sample to explain. LIME learns the dotted line which locally approximates the red/blue decision boundary around the target red-bold sample.

model	**BLSTM**	**CNN**	**CNN-BLSTM**
input layer	input (***N X 257 mag spectrogram***)		
conv. layer		$\left\{\begin{matrix} conv3-(channels)/1 \\ conv3-(channels)/1 \\ conv3-(channels)/3 \end{matrix}\right\} X4$ $channels$ = [16, 32, 64, 128]	
recurrent layer	BLSTM-128		BLSTM-128
FC layer	FC-64, ReLU, dropout	FC-64, ReLU, dropout	FC-128, ReLU, dropout
	FC-1 (***frame-wise scores***)		
output layer	average pool (***utterance score***)		

Figure 2: The MOSNet architecture from [1]. This work uses the CNN-BLSTM configuration in the rightmost column. In the convolutional layer row, 3 corresponds to the kernel size, and the denominator (1 or 3) refers to the stride length. Each convolutional layer uses the ReLU activation function and N denotes the number of frames.

One such explanation method is LIME [2], published in 2016. LIME explains the decision for a specific input by learning an already-interpretable model that locally approximates the DNN for that input. In this work, we implement LIME explanations to better understand how the MOSNet [1] deep neural network for speech quality assessment makes its predictions.

2 Background

In this section, we briefly recap the design of the LIME framework and the MOSNet deep neural network [2, 1].

LIME provides a framework for explaining the predictions of any regressor or classifier by locally approximating the target network. An illustration of this is shown in Figure 1. Intuitively, LIME learns a linear model which is approximately similar to the target model for a given data instance by perturbing the target data to create a neighborhood that the new model fits to. The idea is then that this linear model is already interpretable and more intuitive to understand than the target model.

MOSNet is a deep neural network developed by Lo et al. in 2019 [1]. MOSNet takes the magnitude spectrogram of an audio file as input and predicts its MOS score. MOSNet is comprised of 12 convolutional layers and a recurrent bidirectional long short-term memory layer followed by dropout and averaging layers. The architecture can be seen in the rightmost column of Figure 2.

3 Methods

This section describes the details of implementing LIME for MOSNet. Broadly, the work falls into three categories: (1) modifying the input data set to fit the differing input formats of MOSNet and LIME; (2) initializing LIME and updating MOSNet to have a compatible prediction function; and (3) displaying the explanations. The code associated with this project is publicly available on GitHub for a limited time[2].

3.1 Data Preprocessing

The first, and most laborious, task was to modify the input data to fit LIME and MOSNet. For this work, we used the Voice Conversion Challenge 2018 (VCC 2018) data set which was used to train and evaluate MOSNet [7]. VCC 2018 is comprised of professional English speakers in a noise-free environment from the device and production speech (DAPS) data set[8]. There are 12 speakers in the VCC 2018 data set: 6 male and 6 female chosen from DAPS.

Given the VCC 2018 data files are of the .wav format and MOSNet takes magnitude spectrograms as input, a Fourier transform is first performed to calculate the spectrograms. LIME, on the other hand, takes images as input so for each calculated magnitude spectrogram, we plot the spectrogram and save the plot as a .png image. These images then are read in as input to LIME.

3.2 Initialization

The LIME framework has several versions for explaining text, images, and tabular data. In this work, we use the LIME Image Explainer. An image explainer was used due to ease of converting data formats between it and MOSNet and understanding the explanations. The explanations produced are attention maps of the original spectrogram image.

The LIME Image Explainer only supports classifiers. Since MOSNet is a regressor, we converted MOSNet's output (a real-valued number between 1 and 5) to a classification by truncating the prediction to be an integer in the range [1, 5] and then using a one-hot vector to represent the class probabilities for each class 1-5.

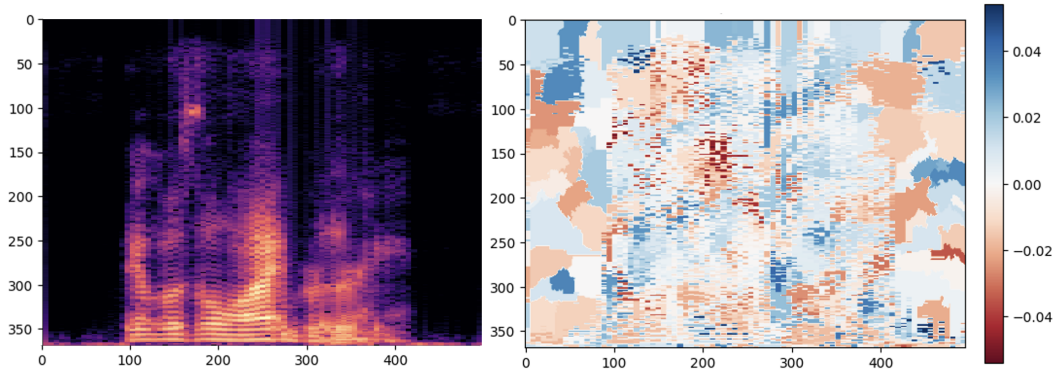
3.3 Generating Explanations

To call LIME's Image Explainer for a specific input, the target model's (i.e., MOSNet's) prediction function must be provided so that LIME can get predictions on the neighborhood of the given input. However, MOSNet's original prediction function can't be used since it expects a magnitude spectrogram but LIME only has an image. To resolve this, we wrote two wrapper functions around MOSNet's prediction functions. First, we stored the the names of the image files that will be given to LIME to explain as an attribute in MOSNet model. Then, we wrote a new prediction function which can look up the original image file name, look up the corresponding audio file, calculate the magnitude spectrogram, and pass that along to the original MOSNet prediction function. During explanation, there is a hyper-parameter which sets the size of the neighborhood around the target data. For this work, we used 100 samples as the neighborhood size. A neighborhood of 100 is somewhat small and was due to computational constraints.

4 Results

Figure 3 shows an example of LIME's output for MOSNet. This data sample has a true MOS score of 2 and MOSNet prediction of 2. The original input magnitude spectrogram is shown in Figure 3(a) and the corresponding explanation for why this sample was classified as having a MOS score of 2 is shown in Figure 3(b). We can see that the strong harmonic pattern at low frequencies about halfway through the signal (approximately pixels (200, 325)) has a strong positive association with the predicted MOS score while the minimal high frequency noise at the end of the signal (approximately pixels (425, 75)) has a strong negative impact. Overall, this explanation is difficult to draw conclusions from.

[2] `https://github.com/abarach/mosnet-lime`



(a) Original input image of a magnitude spectrogram

(b) Associated LIME explanation of MOSNet's prediction for the image in subfigure (a). The true MOS value was 2 and this explanation was generated for MOSNet's prediction of 2. The red regions indicate evidence against the prediction and bluer regions are evidence for.

Figure 3: Example of the input image and corresponding output of the LIME explanations for MOS predictions.

One possible reason for the difficulty in reading the explanation in Figure 3(b) could be that it is not a contrastive explanation. A contrastive explanation is one in which we ask why something is the case as opposed to an alternative. Contrastive explanations thus require reasoning about the differences between the target and the alternative instead of reasoning about the target in general. Studies have shown that humans use contrastive explanation more often than not [9]. For example, if presented with a picture of a raven, humans will (often implicitly) ask what makes the picture a raven *as opposed to some other type of bird*. An explanation of why the picture is of a raven requires reasoning about the whole class of birds (e.g., there is a beak, wings) instead of nuanced differences (e.g., it's a raven due to its black, curved beak and eye placement). This mismatch between explanation methods could cause difficulty in understanding automated explanation methods. While there has been work in developing contrastive explanation methods for deep neural networks in recent years, more research is warranted [10, 11].

Other work in explainability has approximated a contrastive explanation by computing a heatmap for the highest probability class, such as in Figure 3(b), as well as the second highest probability class. Then, they subtract the heatmap for the second most probable class from that of the most probable class. They then argue that this difference is an approximate answer to the question "Why did the network predict the most probable class over the second most probable class?" [10]. This experiment was not performed in this work due to the one-hot vector used in subsection 3.2. The one-hot vector approach results in class probabilities being 0 for every class except the one predicted.

5 Related Work

DNN explainability has been gathering popularity in recent years. In 2021, the National Institute of Standards and Technology (NIST) defined four principles of artificial intelligence (AI) (explanation, meaningfulness, accuracy, and knowledge) of which three require the ability to interpret a network's decisions [12]. Ras et al. later published a detailed survey of explanation methods in AI, grouping them into visualization methods, distillation methods, or intrinsic methods [13]. Using this system, LIME is a (local approximation) distillation method.

AudioLIME was proposed as an extension to the LIME framework. LIME natively supports matrix data, text, and images. AudioLIME extends LIME to be able to explain regression models for audio signals [14].

Layer-wise relevance propagation (LRP) is a visualization method that is compatible with any network that supports backpropagation [15]. In the same manner as backpropagation, LRP assigns a relevance score to each node and distributes this relevance backward through the network until the input layer is

reached. Then, one can plot the input layer's relevance scores as a heatmap showing the importance of each input feature to the ultimate output. LRP was originally developed for vision, but has been translated and applied to audio applications in a work called AudioMNIST [16]. Another visualization method, GradCam, creates a localization map from the gradients flowing into the final convolutional layer of a network [17].

Explanation association is an intrinsic explanation method for sentiment analysis that uses joint training with a secondary task of learning rationale extraction [18]. A generator learns candidates for the rationale explanation, an encoder uses these candidates during sentiment analysis prediction, and then the explanation for the prediction is the association of the model prediction with a critical set of words from the candidate rationale explanation.

6 Conclusion

In this work we applied the LIME framework to explain the predictions of the MOSNet audio quality assessment model. The results show that MOSNet is, in fact, focusing on recognizable speech patterns within the signal but was ultimately inconclusive. Further work is needed to accurately portray MOSNet's decision-making process.

6.1 Limitations

This work has several limitations. First, and most notably, is the use of the LIME Image Explainer. Converting from a magnitude spectrogram to an image of that spectrogram is a lossy process. That is, one cannot recover the magnitude spectrogram from an image. This information loss likely reduces the effectiveness of the resulting explanation. A possible solution would be to use the Lime Tabular Explainer or the AudioLIME module [14].

The other limitation of this work is the conversion of MOSNet's predictions from a regression to a classification. As mentioned in section 4, the one-hot vector zeros out any class that was not ultimately predicted. This prevents approximate contrastive explanations from comparing the heatmaps for the most probable and second most probable classes. Using a LIME module intended for regression models would thus be an improvement.

6.2 Future Work

As hinted in the previous section, future work for this project is to use the AudioLIME module to yield more accurate explanations. Another interesting direction would be to compare the results from several different approaches (e.g., Contrastive Explanation Methods, LRP, GradCAM, etc.) to get a better understanding of MOSNet's predictions.

References

[1] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "MOSNet: Deep learning based objective assessment for voice conversion," in *Proc. Interspeech 2019*, 2019.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?" Explaining the Predictions of Any Classifier," in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[3] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6493–6497.

[4] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, 2001, pp. 749–752 vol.2.

[5] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual objective listening quality assessment (polqa), the third generation itu-t standard

for end-to-end speech quality measurement part i—temporal alignment," *journal of the audio engineering society*, vol. 61, no. 6, pp. 366–384, 2013.

[6] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 4214–4217.

[7] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. H. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," *ArXiv*, vol. abs/1804.04262, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:4796554

[8] G. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech? – a dataset, insights, and challenges," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1006–1010, 2015.

[9] P. Lipton, "Contrastive explanation," *Royal Institute of Philosophy Supplements*, vol. 27, p. 247–266, 1990.

[10] A. Feghahati, C. R. Shelton, M. J. Pazzani, and K. Tang, "Cdeepex: Contrastive deep explanations," in *ECAI 2020*. IOS Press, 2020, pp. 1143–1151.

[11] M. Prabhushankar, G. Kwon, D. Temel, and G. AlRegib, "Contrastive explanations in neural networks," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 3289–3293.

[12] P. J. Phillips, C. A. Hahn, P. C. Fontana, A. N. Yates, K. Greene, D. A. Broniatowski, and M. A. Przybocki, "Four principles of explainable artificial intelligence," National Institute of Standards and Technology, Gaithersburg, MD, Interagency or Internal Report 8312, September 2021.

[13] G. Ras, N. Xie, M. van Gerven, and D. Doran, "Explainable deep learning: A field guide for the uninitiated," *Journal of Artificial Intelligence Research*, vol. 73, pp. 329–396, January 2022.

[14] V. Haunschmid, E. Manilow, and G. Widmer, "audiolime: Listenable explanations using source separation," 2020.

[15] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[16] S. Becker, M. Ackermann, S. Lapuschkin, K.-R. Müller, and W. Samek, "Interpreting and explaining deep neural networks for classification of audio signals," *CoRR*, vol. abs/1807.03418, 2018.

[17] 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 *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[18] T. Lei, R. Barzilay, and T. Jaakkola, "Rationalizing neural predictions," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 107–117. [Online]. Available: https://aclanthology.org/D16-1011