# Filters Know How You Feel: Explaining Intermediate Speech Emotion Classification Representations

Anubhav Anand, Shubham Negi and Narendra N
Wipro Limited, India
E-mail: anubhav.anand1@wipro.com, shubham.negi1@wipro.com, narendra.n05@wipro.com

***Abstract*—Emotion recognition from speech is gaining popularity amongst the research community. Speech Emotion Recognition (SER) systems have applicability in variety of application scenarios like health-care systems, monitoring systems and automatic driving systems to name a few. However, interpreting the results of the SER system and providing human understandable reasoning is a topic very few have touched upon. We propose a SincNet based emotion recognition engine which makes use of the interpretable filters of the first layer to explain the reasoning behind the model inference. We use the IEMOCAP dataset and compare our results of emotion recognition with the state of the art algorithms. We also propose an explainability technique to provide understanding of the model as well as the inferences. To the best of our knowledge, the proposed scheme is novel and achieves good performance for emotion recognition using speech.**

**Index Terms**: speech emotion recognition, explainable AI, SincNet

## I. Introduction

With the ever growing use of smart automated machines built using artificial intelligence, in day to day life, it is becoming all the more important for the research community to work on all types of data users can offer. There has been plenty of research on structured data and research on visual data has also grown exponentially in the past few years. Now is the time of incorporating speech into AI. There has been good amount of interest of the community towards this area. Significant progress has been made in terms of understanding speech signals and training the machine to make decisions according to that.

Recognizing emotions from speech signals is an active area of research and has gained a lot of attention in recent times. Emotion is a conscious mental reaction subjectively experienced as strong feeling, usually directed towards a specific object and typically accompanied by physiological and behavioral changes in the body. Understanding emotions is vital in designing systems that take decisions on the behalf of users. Speech Emotion Recognition (SER) systems try to make machines understand the emotions of the speaker and the affective state she is in. It is important to make Human-Machine interaction more accessible. SER systems have a vast number of applications in daily life. SER can also become an essential part of health-care systems, monitoring systems and automatic driving systems.

Early work in emotion recognition concentrated on identifying prosodic features like pitch and formants (spectral peak) to name a few [1], [2], [3]. Other works considered spectral features like Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Cepstral Coefficients (LPCC) [4], [5] to aid in classification. These features were fed to classifiers like HMM [6], GMM [7], SVM [8] followed by more sophisticated neural networks [9], [10] to identify the emotions. Recent advances in deep learning techniques allow us to get a representation from the raw audio waveform for emotion classification. [11] provides a good overview of state of the art techniques for emotion recognition. Deep learning techniques, though state of the art miss out on providing interpretability of the results. SincNet [12] proposed by Bengio et. al provides an architecture to interpret the filters for a speaker recognition task.

Explainable AI (XAI) is an area focusing on providing interpretability to black box deep learning models. A plethora of work has been done to provide interpret ability to models involving text, image and tabular data as input. [13] provides a good overview of the current trends in XAI.

Our work involves combining the power of SincNet and XAI to provide a human centric interpretation of the black box model. We take up the task of emotion recognition using the SincNet architecture and follow it up to provide interpretation of the inferences provided by the SincNet model. We see an improvement over the state of the art in the classification accuracy for the four emotion classification task on the IEMOCAP [14] dataset. We also try to explain the model in the global as well as local context with global explainability interpreting what the model says as a whole and local interpretability explaining the output of a given input sample.

Section II provides an overview of SincNet whereas Section III provides an overview of model explainability. Section IV presents the ideas behind our work on classification and interpretation. We also discuss the dataset and the implementation details in Section IV. Section V focuses on the conclusion and ideas for future work.

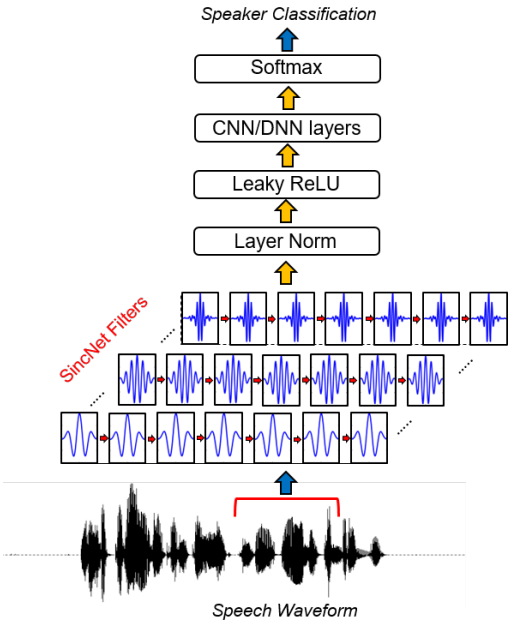



Fig. 1: SincNet Architecture

## II. SincNet

Convolutional Neural Networks (CNN) based filtering is becoming popular in tasks involving speech recognition. CNNs learn low-level speech representations from waveforms, potentially capturing important characteristics such as pitch and formants. However, there are two problems with this. One, the design of neural networks become crucial for getting a good representation. And second, the dimensionality of input waveform is very high. SincNet allows us to replace the first convolution layer with meaningful filters. This offers a very compact and efficient way to derive a customized filter bank specifically tuned for the desired application. Fig 1 captures the SincNet architecture proposed in [12]. The first layer of a standard CNN performs a set of time-domain convolutions between the input waveform and some Finite Impulse Response (FIR) filters

$$y[n] = x[n] * h[n] = \sum_{l=0}^{L-1} x[l].h[n-l] \quad (1)$$

where $L$ is the length of the filter. SincNet however performs convolution with a pre-defined parametric function say $g$

$$y[n] = x[n] * g[n, \theta] \quad (2)$$

A reasonable choice of $g$ would be a filter bank of rectangular band pass filters. A band-pass filter is designed as a difference of two low pass filters with different cut-off frequencies say $f_1$ and $f_2$. The time domain representation of such a filter will yield the difference of two sinc functions

$$g[n, f1, f2] = 2f_2 sinc(2\pi f_2 n) - 2f_1 sinc(2\pi f_1 n) \quad (3)$$

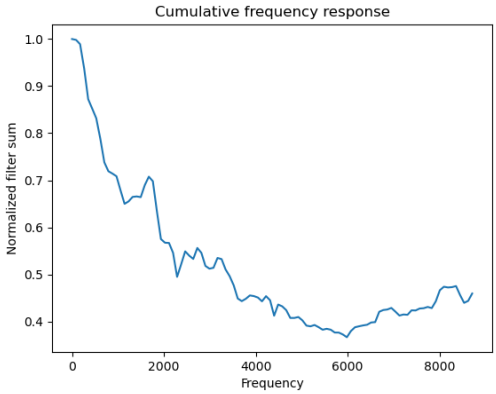



Fig. 2: Cumulative Frequency Response

The function $g$ is fully differentiable and hence the cut-off frequencies can be jointly optimized through gradient descent methods. These result in providing an interpretable filter response to the input speech.

## III. Model Explainability

Model explainability is gaining traction in the research community as AI penetrates more and more into our lives. For example, in the current context AI is being used to diagnose Covid-19 through CT-scan reports. Once the AI engine identifies a patient as positive from the scan, need of the hour is to provide an explanation to the patient as to why the engine interpreted it to be positive. Large amount of work in this regard has been done in the area of image classification with various techniques available to provide an explanation. However, very little work has been seen in the context of explaining speech models. Model explainability as a whole can be divided into two categories

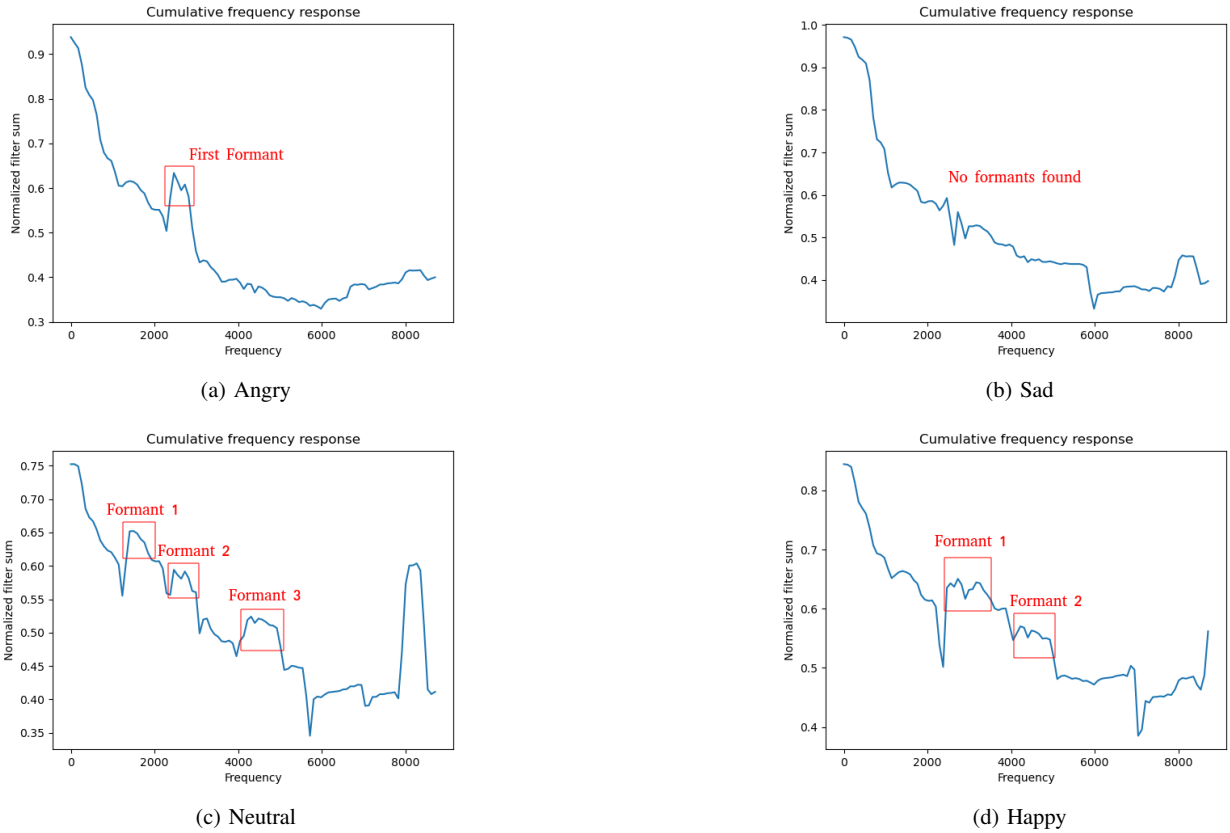
- Global Explanation: This involves explaining what the model has learnt as a whole with respect to the training data.
- Local Explanation: Interpret the output of a particular test input and provide the explanation in a human understandable form.

## IV. What are the filters learning?

SincNet provides interpretable filters as the first layer of the model. We use the SincNet architecture to solve the emotion recognition problem.

### A. Dataset

We use the IEMOCAP dataset provide by University of Southern California for our experiments. The IEMOCAP dataset has 12 hours of audio-visual data. It includes video, speech, text transcriptions and motion capture of the face. It consists of sessions where hired actors perform improvisations or read scripted dialogues, specifically selected to suggest emotional expressions of the speaker. The dataset is annotated by multiple annotators. There are 10 classes in the database




(a) Angry (b) Sad

(c) Neutral (d) Happy

Fig. 3: Cumulative Frequency Response for individual emotions

such as anger, happiness, neutral etc. Although the data has video and text too, we focus only on speech signals. This is,

- To remain consistent with prior research [15], [16]
- In daily life use-cases of SER, video and text data will not be easily available.

### *B. Implementation*

We restrict our analysis to 4 emotion classes viz. angry, happy, sad and, neutral, to compare our results with the state of the art algorithms. IEMOCAP official release provides us with dialogue level classification. There are approximately 10,000 dialogues in the whole dataset with respective classes. Due to the imbalance in the data points of these 4 classes, we combine signals with happy and excited into one class to make it balanced. The modified 4 classes contain approximately 7,000 utterances. We also consider only the improvised speech conversations to be consistent with the state of the art. This provides us a dataset of 2943 utterances. The duration of the utterances range from as small as 0.5 seconds to as large as 37 seconds. Each utterance is given one of the 4 labels.
We use a sliding window length of 200ms chunks with 50% overlap over the input speech. We consider each chunk to have the same label as the overall utterance. The first layer performs convolutions based on the sinc filters learnt during training. We use 80 such filters of length 251 followed by 2 layers of CNN with 60 filters each of length 5. This is followed by three fully connected layers of length 2048. All layers use leaky-Relu as the activation function. The model is trained using RMSprop optimizer with a learning rate of 0.001. We divided the dataset in the ratio 80:20 where 80 percent was used for training and 20 for validation. A sentence level classification was computed by averaging the predicted probabilities over all the chunks of the input speech and choosing the one with the maximum posterior after averaging. We also performed a 3-fold validation for the dataset. We achieved an overall accuracy of 77.19%. Table I summarizes the comparison of our implementation with the state of the art.
We see that the SincNet architecture produces better overall accuracy as compared to the existing state of the art techniques. What is also important to note is that we were able to achieve better accuracy by not including any other modalities apart from speech. Fig 2 shows the cumulative frequency response of the filters. This is obtained by summing up the response of all the learnt filters in the first layer of the model.

| Method | Overall Accuracy |
|---|---|
| Lee [17] (Bi-LSTM) | 62.8 |
| Satt[18] (CNN + LSTM) | 68.8 |
| Ramet [19] (Attn. Bi-LSTM) | 68.8 |
| Zhang [20] (Attn. CNN) | 70.4 |
| Yenigalla [15] (CNN) | 71.3 |
| MHA + PE + MTL [16] | 76.4 |
| **SincNet (Ours)** | **77.19** |

TABLE I: Comparison with state of the art

### C. Interpretation

*1) Global Explainability:* The cumulative frequency response can be interpreted as the response of the filter over all the four emotions. We try to see if we can interpret the frequency response with respect to individual emotions. To do this, we consider the samples in the training dataset over individual emotions. We then pass these samples to the learnt model. For every sample we feed as input, we try to understand the importance of every filter in the first layer. We have 80 filters in the first layer and we mask each filter one by one and see the effect it has on the final classification. If the classification accuracy varies by more than 10% of the baseline (With all filters present), we consider the filter to be important for the input. For example, consider a speech signal labeled as **angry**. We pass the speech through the model to get a classification as **angry** with a confidence score of 0.99. We now start removing the filters in the first layer one at a time and observe the classification output. If we see a particular filter affecting the classification output by more than 10% i.e if classification confidence drops below 0.9 in the case of this example, we term the filter to be important. We maintain a list of all the important filters and plot the cumulative frequency response of the important filters over all the training examples. We then average the cumulative response over all examples to visualize the responses over individual emotions.

We make the following observations from the plots in Fig.3.

- We see identifiable formants in the neutral speech with first formant around 1KHz followed by second around 2KHz
- Angry emotion is quantified by a single predominant formant around 2 - 2.5KHz
- Sad is characterized by no particular formant
- We see a set of frequencies from 2KHz to 5KHz which appear to be important for Happy emotion

*2) Local Explainability:* We now turn our attention to explaining the inference of a particular speech input. We repeat the same experiment of masking the filters and checking their response with respect to classification accuracy. A comparison of the filters selected as important by the masking technique are compared with the most frequently selected filters of the emotions. Automatic identification of the emotion based on the selected filters helps us in providing an explanation of the inference. We consider the average cumulative frequency response of the filters for each emotion based on the training data. We then identify the most important filters for a particular input speech by masking technique. We then compute the Mean Squared Error (MSE) between the average frequency response of all emotions with respect to the cumulative frequency response of the input speech. We then try to check if the minimum MSE obtained corresponds to classification output which would then provide a good base for explainability of the result. Table II provides a quantitative analysis of the results we obtained for local explainability over validation dataset for each emotion. We get an overall accuracy of **57.3%** over all emotions.

| Emotion | Accuracy |
|---|---|
| Happy | 71.83% |
| Angry | 70.83% |
| Neutral | 47.05% |
| Sad | 40.29% |

TABLE II: Accuracies for Explainability

We consider an example from the validation set of a neutral emotion which has been classified correctly by the model. Fig. 4 gives the frequency response of the filters identified as important. We can see that we are able to identify 2 formants around 1KHz and 2KHz which is similar to the plots we obtained over the training data for a neutral emotion. Similar plot was obtained for an sad emotion which is seen in Fig. 5 where we do not see any predominant formant. A sample explanation for sad emotion would be "*The model has inferred that the emotion is sad because of the absence of any predominant formant in the frequency response*".

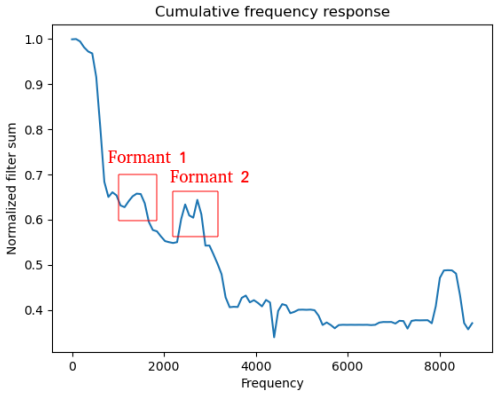



Fig. 4: Frequency response for local explanation for neutral emotion

## V. Conclusion & Future Work

We propose an explainable SincNet based model for emotion recognition from speech signals. We find that we are able to achieve an overall accuracy better than the state of the art without using any other modalities apart from speech on the IEMOCAP dataset for a four emotion classification problem. There is scope to look at recurrent architecture which would

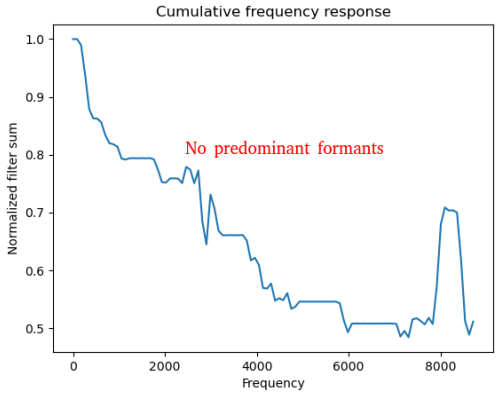



Fig. 5: Frequency response for local explanation for sad emotion

also involve identifying relationships between patches of the samples in the input speech to improve the performance. We also propose means to provide explainability for the model in both global and local sense. We propose a method to identify the cumulative frequency response over individual emotions to describe the model. We also propose to use those responses as references to provide explanation for inferences obtained from the model. There is a scope to improve the explainability of the results by considering other masking techniques. In future, we would like to extend the work to more emotion classes as well as other speech processing tasks.

## REFERENCES

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2008.

[2] J.-C. Lin, C.-H. Wu, and W.-L. Wei, "Error weighted semi-coupled hidden markov model for audio-visual emotion recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 142–156, 2011.

[3] K. S. Rao, S. G. Koolagudi, and R. R. Vempada, "Emotion recognition from speech using global and local prosodic features," *International journal of speech technology*, vol. 16, no. 2, pp. 143–160, 2013.

[4] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International journal of speech technology*, vol. 15, no. 2, pp. 99–117, 2012.

[5] S. Kuchibhotla, H. D. Vankayalapati, R. Vaddi, and K. R. Anne, "A comparative analysis of classifiers in emotion recognition through acoustic features," *International Journal of Speech Technology*, vol. 17, no. 4, pp. 401–408, 2014.

[6] N. Sato and Y. Obuchi, "Emotion recognition using mel-frequency cepstral coefficients," *Information and Media Technologies*, vol. 2, no. 3, pp. 835–848, 2007.

[7] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents speech during family interactions," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 574–586, 2010.

[8] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech communication*, vol. 52, no. 7-8, pp. 613–625, 2010.

[9] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.

[10] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, "Towards speech emotion recognition" in the wild" using aggregated corpora and deep multi-task learning," *arXiv preprint arXiv:1708.03920*, 2017.

[11] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[12] 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.

[13] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2021.

[14] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[15] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram & phoneme embedding." in *Interspeech*, 2018, pp. 3688–3692.

[16] A. Nediyanchath, P. Paramasivam, and P. Yenigalla, "Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7179–7183.

[17] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Sixteenth annual conference of the international speech communication association*, 2015.

[18] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms." in *Interspeech*, 2017, pp. 1089–1093.

[19] G. Ramet, P. N. Garner, M. Baeriswyl, and A. Lazaridis, "Context-aware attention mechanism for speech emotion recognition," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 126–131.

[20] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, "Attention based fully convolutional network for speech emotion recognition," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 1771–1775.