

Toward end-to-end interpretable convolutional neural networks for waveform signals

Linh Vu

Monash University

linh.vu@monash.edu

Thu Tran

Singapore Management University

ndttran.2019@phdcs.smu.edu.sg

Lim Wern Han

Monash University

lim.wern.han@monash.edu

Raphaël C.-W. Phan

Monash University

raphael.phan@monash.edu

Abstract—This paper introduces a novel convolutional neural networks (CNN) framework tailored for end-to-end audio deep learning models, presenting advancements in efficiency and explainability. By benchmarking experiments on three standard speech emotion recognition datasets with five-fold cross-validation, our framework outperforms Mel spectrogram features by up to seven percent. It can potentially replace the Mel-Frequency Cepstral Coefficients (MFCC) while remaining lightweight. Furthermore, we demonstrate the efficiency and interpretability of the front-end layer using the PhysioNet Heart Sound Database, illustrating its ability to handle and capture intricate long waveform patterns. Our contributions offer a portable solution for building efficient and interpretable models for raw waveform data.

Index Terms—speech recognition, audio classification, signal processing, interpretable neural networks, convolutional neural networks

I. INTRODUCTION

The use of deep learning and representation learning has significantly impacted the signal processing field, similar to the success observed in the fields of Natural Language Processing (NLP) and Computer Vision (CV). Deep neural networks can automatically learn hierarchical representations from raw data. However, simply using deep learning as a one-tool-fit-all solution often leads to poor interpretability, i.e., monitoring how the model works is difficult.

Recent developments have led to the creation of efficient and easy-to-understand models for processing raw waveform signals. Cheuk et al. [1] have introduced nnAudio, a framework that uses 1D convolutional neural networks to extract spectrograms on-the-fly. Leiber et al. [2] have introduced a differentiable modification of the Short-Time Fourier Transform (STFT) that makes it possible to optimize the window length parameter through gradient descent. These works primarily revolve around rethinking the fusion of spectrograms and neural networks, placing emphasis on employing an adapted spectrogram as the initial layer for input signals. Seki et al. [3] showed that Gaussian filters with three trainable parameters of gain, bandwidth, and centre frequency can outperform Mel filterbank in both clean and noise-corrupted environments. Inspired by standard filtering in digital signal processing, Ravanelli and Bengio [4] proposed a novel CNN called SincNet that uses rectangular band-pass filters as kernels for the first convolution layer with two learnable parameters: high and low cut-off frequencies. Right from the first layer, SincNet

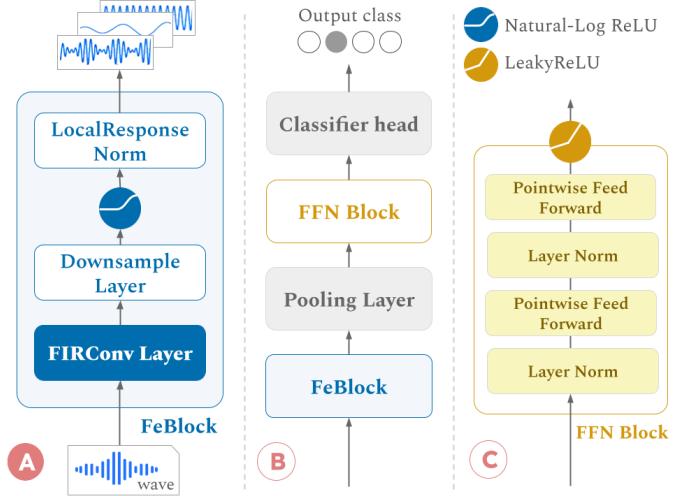


Fig. 1: The proposed IConNet architecture for end-to-end audio classification: A- the front-end block containing the FIRconv layer; B- the proposed general architecture for end-to-end audio classification; C- the classifier used in the experiments.

can already learn meaningful filters and thus converge faster than the conventional CNN. In a comparison between different parametric modulated kernel-based filters including SincNet (rectangular filters), Sinc2Net (triangular filters), GammaNet (gammatone filters) and GaussNet (Gaussian filters); it is found that SincNet is the best for the speech recognition task [5].

According to Ravanelli and Bengio [4], the frequency gain of the SincNet filters is only defined in subsequent layers of the neural network, which are conventional CNN layers. To further contribute to this research direction, we propose a new interpretable CNN architecture called IConNet that utilizes a finite impulse response (FIR)-based kernel with learnable window functions. FIR filters play a crucial role in signal processing as they enable the extraction of critical information; and different filter designs can enhance or mitigate unwanted effects. With adaptive window functions, the filters can dynamically adjust to varying signal profiles depending on the data and the problem. The primary benefit of this approach is the transparency in the way the model learns – which frequency bands it focuses on and which will be cut off.

We illustrate the effectiveness of this approach in two health-related problems: speech emotion recognition and abnormal heart sound detection.

In the next section II, we will describe the proposed method. Section III is about the Speech Emotion Recognition (SER) experiment, followed by section IV about abnormal heart sound detection.

II. THE ICONNET ARCHITECTURE

The proposed Interpretable Convolutional Neural Network (ICONNet) architecture in this research is designed to leverage insights from audio signal processing; aiming to improve end-to-end deep neural networks' ability to extract features and patterns from raw waveform signals. The foundation of our method draws inspiration from the standard signal processing process, in which the input audio signal undergoes a windowing operation to segment it into smaller frames and simultaneously mitigate spectral leakage to improve frequency resolution. A key novelty lies in using the Generalized Cosine Window function as parametrization for the convolution kernels to enable the neural networks to choose the most suitable shape for each frequency band.

The convolution layer of the proposed model has restricted-shaped kernels which is a band-pass filter defined by non-learnable low cut-off frequency f_0 and frequency bandwidth f_δ . The filter shape and frequency gain are determined by the window function W with p learnable parameters ϕ_p . Let $H = \{h_k : k = 1, \dots, K\}$ denotes the kernel (filter) of width K in the time domain. V_n denotes the output of the convolution layer for each n input time-domain value. $H(k, f_0, f_\delta, \phi)$ is parametrically modulated by *sinc* as a non-learnable function and W with learnable parameters.

$$V_n = X \cdot H = \sum_{k=1}^K X_k \cdot h_{n-k} \quad (1)$$

$$H = T * W = T(k, f_0, f_\delta) * W(k, \phi), \quad 0 \leq k < K \quad (2)$$

$$\begin{aligned} T(k, f_0, f_\delta) &= 2(f_0 + f_\delta) \text{sinc}(2\pi k(f_0 + f_\delta)) \\ &- 2f_0 \text{sinc}(2\pi kf_0), \quad \text{sinc}(a) = \sin(a)/a \end{aligned} \quad (3)$$

$$W(k, \phi) = \left\{ w_k \mid w_k = \sum_{i=0}^p (-1)^i \phi_i \cos \frac{2\pi i k}{K-1} \right\} \quad (4)$$

Thus, for each kernel, there are only p parameters ϕ_p and two band parameters f_0 and f_δ to train via gradient descent. f_0 and f_δ can also be non-learnable, which is how we can incorporate prior domain knowledge into the deep neural network design to solve specific problems and control what information is fed into the model to prevent unwanted effects from the first layers.

After separating signals into different channels, we use a downsampling layer to reduce the data in the time domain to a

lower sample rate. This compression step decreases the data's dimensions, making it easier to process in subsequent steps while retaining the essential features needed for precise analysis and classification. It's worth noting that the convolution filters from the previous step also serve as anti-aliasing filters like traditional signal downsampling. In the second-to-last step, we use an NLRelu activation function recommended in [6] to mimic the amplitude-to-decibel conversion since human hearing functions on a logarithmic scale. Finally, we apply the Local Respond Normalization function [7] to the output before forwarding it to the next layer. This entire process can be repeated by stacking these blocks on top of each other to achieve further pattern decorrelation from previous decomposed channels while maintaining a more compressed representation. The outputs of the front-end blocks can be combined together as long as they are resampled to the same sampling rate to preserve the temporal characteristic. Depending on the task, these front-end blocks can be incorporated into any deep neural network architecture.

Figure 1 illustrates the proposed architecture, with the part **A** on the left describing the front-end block. The middle part **B** of the figure 1 is the high-level deep neural network architecture that incorporates the front-end block to process raw waveform signal data. Part **C** depicts a simple classifier block consisting of a pooling layer, a two-layer feed-forward neural network with layer norm and a LeakyRelu activation function. The following experiments are designed to demonstrate the feature extraction ability of the front-end blocks and evaluate the effectiveness of this new architecture.

We have selected two classification problems and designed a classifier consisting of a pooling layer, a two-layer feed-forward neural network with layer norm and a LeakyRelu activation function. In both experiments, we use k-fold cross-validation settings and report three different metrics: unweighted accuracy (UA), unweighted F1 (UF1) and weighted-F1 (F1). For training the neural networks, we employed RAdam optimizer [8] with OneCycleLR learning rate scheduler [9], Cross Entropy loss and trained each model to up to 60 epochs on each fold.

III. SPEECH EMOTION RECOGNITION

A. Background

Emotions play a pivotal role in revealing essential cues about a speaker's intentions, attitudes, and mental state, particularly in the context of AI health. This significance is underscored by a recent study from Gheorghe et al. [10] where a system leveraging deep neural networks successfully identifies depression from speech samples with an unweighted accuracy of 91.25% using Mel-frequency cepstral coefficients (MFCC). MFCC is the most popular feature based on the short-time Fourier Transform spectrogram and the Mel filterbank, which was designed based on human perception of our hearing system. As humans perceive sound on a logarithmic scale, Mel filterbank uses narrower bandwidths at the lower frequencies to capture more information. Its filters have a triangular shape and are used in many tasks. However, according to [3], [11], in

the data-driven filterbanks, each filterbank's centre frequencies and shapes are different depending on the tasks and the presence of background noise.

Based on these insights, we've developed three IConNet variants detailed in Table I. These variants assess the influence of adjustable bands versus adaptive windows on CNN model classification outcomes. The front-end kernels in all IConNet models are initialized with the Mel-filterbank, with additional filters allocated to the lower frequency range, prioritizing crucial speech signals.

B. Experiment setup

1) *Datasets*: Our first experiment incorporates the proposed method, evaluating its performance on three standard Speech Emotion Recognition (SER) datasets with similar settings described in [12]¹. Firstly, the **RAVDESS** dataset [13] features high-quality audio speech with eight emotional expressions from 24 North-American actors. Secondly, the **CREMA-D** dataset [14] comprises recordings from 91 speakers with diverse races and ethnicities, representing real-world audio recordings where the recordings are often noise-corrupted. Lastly, **IEMOCAP** [15] is a popular SER dataset with 12-hour conversational speech audio from 10 speakers. We use four classes of data, namely *happy, sad, angry, neutral*, which are shared between the aforementioned three datasets for a fair comparison.

2) *Classifiers and evaluation*: For the classifier, we use a dense neural network consisting of two layers with 512 nodes in each layer followed by layer norm, as described in the diagram 1. To evaluate the performance of the proposed models, we adopted 5-fold cross-validation with stratified train-test splitting to ensure the proportion of each class in both sets. We trained each model for up to 60 epochs with early stopping, then evaluated the model on the test set and reported the average metrics across folds. We benchmarked IConNet with different settings against Mel and MFCC features extracted using the *TorchAudio* library [16] with different resolutions as described in the table I.

Table I: Models used in the SER experiment

Model	Description
Mel- K	Mel spectrogram with K Mel frequency bins
MFCC- K	MFCC from Mel- K
IConNet-B- K	IConNet with K learnable-band filters
IConNet-W- K	IConNet with K learnable-window filters
IConNet-BW- K	IConNet with K learnable-band-window filters

C. Experiment results

Table II provides detailed experiment result on the three datasets mentioned above when using early stopping. On the RAVDESS dataset, the IConNet model with 456 adaptive-window FIR kernels achieved an unweighted accuracy of

¹It is worth mentioning that in [12], Vu et al. benchmarked the performance of handcrafted features after 500 epochs on the 20% test set after 10-fold cross validation on 80% of the train set.

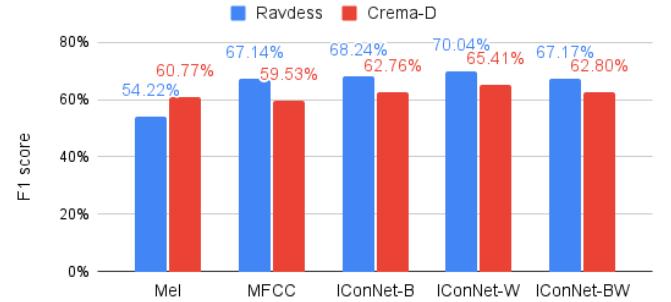


Fig. 2: Result on RAVDESS and CREMA-D datasets after 60 epochs

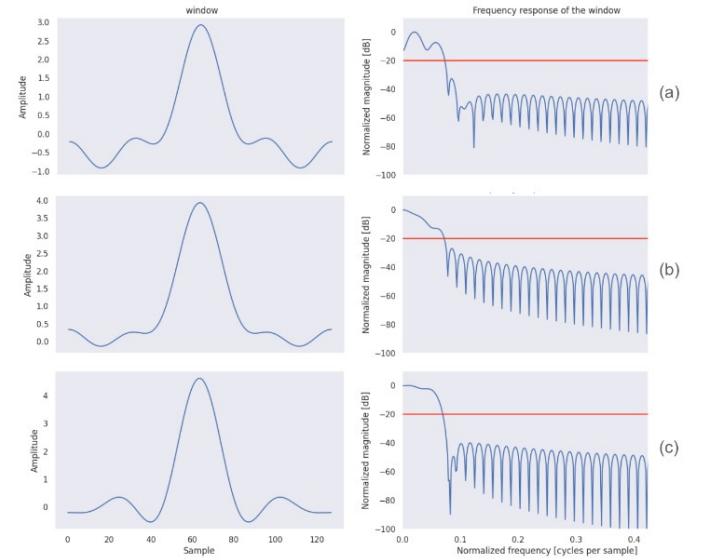


Fig. 3: Comparison of Window Shape and Frequency Response of Filters from Different Bands. The chart displays the frequency response of low-range (a), mid-range (b), and high-range (c) frequency bands. The red line at -20dB represents the threshold at which noise is perceived as not noticeable.

66.83%, which is 4.83% higher than the adjustable-band-FIR model with the same number of kernels. When the number of kernels was reduced to 256, the UA of the IConNet-W dropped 1%, while the IConNet-B increased 1.73%. For the IConNet-W models and Mel-spectrogram models, higher resolution helped the models make better predictions. The size of Mel-256/IConNet-256 and Mel-456/IConNet-456 models is 1.6 MB and 2 MB, respectively. Despite having the same size, Mel models performed poorly compared to all IConNet models, even after being trained for more epochs.

The bar chart 2 compares the F1 scores of Mel-456, MFCC-456, and IConNet-456 models on the RAVDESS and CREMA-D datasets. After 60 epochs, the performance of MFCC on the RAVDESS dataset increased from 46.13% to 67.14%, which is a more than 21% improvement. However, the gain for MFCC-456 model on CREMA-D is less than 1%. The IConNet-W model still gave the highest F1 score on both the RAVDESS

Table II: Results of hand-crafted features and IConNet on three datasets in percentage with early stopping (%)

Dataset	RAVDESS			CREMA-D			IEMOCAP			Average		
Model	UA	UF1	F1									
Mel-256	51.59	51.73	53.22	59.07	59.07	58.83	50.85	50.85	51.13	54 ±5	54 ±5	54 ±4
Mel-456	52.00	49.09	51.67	59.98	59.98	59.60	52.85	52.85	53.03	55 ±4	54 ±6	55 ±4
MFCC-256	49.29	46.01	47.58	56.47	56.27	56.70	56.68	56.41	56.25	54 ±4	53 ±6	54 ±5
MFCC-456	45.21	42.27	46.13	56.70	56.57	56.86	56.34	56.07	56.14	53 ±7	52 ±8	53 ±6
IConNet-B-256	63.73	63.40	65.01	61.67	61.31	61.53	53.72	53.37	53.37	60 ±5	59 ±5	60 ±6
IConNet-B-456	62.00	62.46	64.65	61.40	61.67	61.94	53.17	53.12	53.51	59 ±5	59 ±5	60 ±6
IConNet-W-256	65.83	65.02	66.65	62.30	62.30	62.06	56.44	55.94	56.17	62 ±5	61 ±5	62 ±5
IConNet-W-456	66.83	66.15	67.37	62.34	62.34	62.56	56.67	56.32	56.56	62 ±5	62 ±5	62 ±5
IConNet-BW-256	64.84	64.83	66.35	61.08	61.08	60.75	52.78	52.77	53.37	60 ±6	60 ±6	60 ±7
IConNet-BW-456	62.36	62.07	63.73	61.28	61.28	61.00	53.26	53.76	54.21	59 ±5	59 ±5	60 ±5

Notation: **bold-underline**: best results, **bold**: second-best results

and CREMA-D datasets at 70.04% and 65.41%, respectively.

The table in II clearly indicates that IConNet-W-456 and MFCC-256 attained the highest unweighted accuracy on IEMOCAP with only a 0.01% difference between them. On average, the IConNet-W models outperformed other IConNet variants by roughly 2%. Overall, IConNet models gave better results than Mel-spectrogram and MFCC models, especially on the CREMA-D dataset.

On the interpretable ability of the IConNet, Figure 3 demonstrates the alterations in the shape of windows that are used to extract important information from different frequency ranges. The window (*a*) is learned during the model training process for the low-frequency range and exhibits a complex shape, determining the frequency gain at each point. This complex shape enhances the robustness of the model because the low-frequency range contains crucial speech signals that are susceptible to noise. On the other hand, the window (*c*) is tailored for the higher frequency range, serving as a narrow-band filter that effectively mitigates spectral leakage with its ideal sidelobes.

In summary, the experiment results have proven the appropriate use of the proposed IConNet, especially the IConNet with adaptive-window FIR kernel for SER. Moreover, the IConNet end-to-end models are portable, with the model size being 30% smaller than the hand-crafted feature set models proposed by Vu et al. in [12].

IV. ABNORMAL HEART SOUND DETECTION

A. Background

Cardiovascular diseases are a leading global health concern, demanding improved diagnostic precision. A *heart murmur* is an unusual sound heard during the heartbeat cycle, indicating underlying cardiac conditions. Various ML approaches have been developed to identify heart murmurs from stethoscope recordings. As the recording duration is quite long (30 seconds on average), there are often many preprocessing steps, such as heartbeat segmentation to trim the waveform to a smaller size (3 to 5 seconds) and bandpass filtering as the most critical signals lay in the range from 10 Hz to 400 Hz. After that,

MFCC and its derivatives are extracted to apply CNN or LSTM models or a combination of both [17], [18].

B. Proposed model

This experiment mainly examines the potential role of the IConNet in refining abnormal heart sound detection. Our proposed model is an end-to-end lightweight neural network architecture that eliminates the need for heart sound segmentation, low-pass filtering and MFCC-based feature extraction. We designed an IConNet-W model with two front-end blocks with 128 and 32 kernels respectively, a max-pooling layer and a 2-layer FFN classifier with 256 nodes on each layer. The total number of parameters is below 200K.

C. Experiment setup

We employ the widely-used **PhysioNet/CinC Challenge** dataset [19] for heart sound classification evaluation. This dataset comprises 2575 *normal* and 665 *abnormal* samples. To validate the effectiveness of the IConNet in identifying relevant features, we resample the waveform from 2000 Hz to 16000 Hz and conduct 4-fold cross-validation, reporting UA, UF1, and F1 metrics. Deng et al. [20] serve as our baseline, utilizing a preprocessing pipeline with a bandpass filter, MFCC features, and a CRNN model consisting of three 2D CNN layers and two LSTM layers. We also include the MFCC performance for comparison. Preprocessing for other models involves waveform trimming and downsampling, excluding the IConNet model.

D. Experiment results

Based on the results presented in Table III, it is clear that the baseline model [20] performed better than the MFCC + FFN model, thanks to its preprocessing steps that included bandpass filtering and the use of MFCC deltas. The baseline model achieved 90.06% F1 score. However, our proposed architecture surpassed both models with an F1 score of 92.05%, which is 2% higher than the baseline model. While this result still not yet outperforms the state-of-the-art Resnet result reported by Li et al. in [18], it has successfully demonstrated the effectiveness of our proposed method in classifying heart

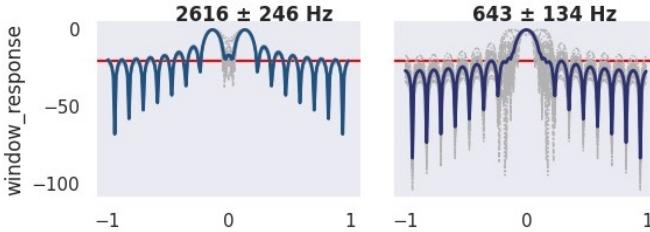


Fig. 4: Frequency response of filters from different bands

sound data. Furthermore, the visualization of the front-end filters confirms that it allocates band-pass filters that actively change the window shapes to extract essential information in the range of 643 ± 134 Hz. The windows have learned to transform into band-stop filter shapes for the high-frequency range above 2000 Hz, which only contain meaningless artifacts from the resampling step. Understanding the features utilized by the backbone neural network model, which influences its decisions, is vital for ensuring reliable outcomes, particularly in health applications.

Table III: Abnormal heart sound detection result on the **PhysioNet** dataset

Model	UA	UF1	F1
MFCC + FFN	82.98	83.97	88.68
MFCC deltas + CRNN [20]	85.67	80.21	90.60
IConNet	87.48	81.12	92.05

V. CONCLUSIONS

Our proposed framework introduces a novel method for constructing end-to-end audio deep-learning models, showcasing its efficacy in healthcare applications such as speech emotion recognition and abnormal heart sound detection. Our findings reveal that the proposed CNN framework surpasses traditional methods utilizing the Mel spectrogram, and potentially MFCC (further experimentation required for confirmation), for both tasks. Moreover, visualization of CNN kernels underscores the value of transparent CNN architectures in healthcare settings, shedding light on the features extracted from input signals in mission-critical tasks.

Insightful observations arise regarding the front-end layer within our proposed framework. Our results indicate that front-end layers featuring learnable windows demonstrate superior performance compared to those employing learnable bands, diverging from the predominant focus in existing literature. We suggest exploring methods to incorporate prior knowledge into the front-end layer to improve performance, especially with the learnable window model. While the proposed framework offers promising results, it requires slightly more parameters and, thus, more training resources.

In conclusion, this study emphasizes the advantages of employing interpretable CNN for analyzing raw waveform data, showcasing their potential benefits in healthcare settings to foster transparency and trust in medical AI applications.

REFERENCES

- [1] K. W. Cheuk, H. Anderson, K. Agres, and D. Herremans, “nnaudio: An on-the-fly gpu audio to spectrogram conversion toolbox using 1d convolutional neural networks,” *IEEE Access*, vol. 8, pp. 161981–162003, 2020.
- [2] M. Leiber, A. Barrau, Y. Marnissi, and D. Abboud, “A differentiable short-time fourier transform with respect to the window length,” in *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 1392–1396.
- [3] H. Seki, K. Yamamoto, and S. Nakagawa, “A deep neural network integrated with filterbank learning for speech recognition,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5480–5484.
- [4] M. Ravanelli and Y. Bengio, “Interpretable convolutional filters with sincnet,” *Proc. of IRASL@NIPS*, 2019.
- [5] E. Loweimi, P. Bell, and S. Renals, “On Learning Interpretable CNNs with Parametric Modulated Kernel-Based Filters,” in *Proc. Interspeech 2019*, 2019, pp. 3480–3484.
- [6] Y. Liu, J. Zhang, C. Gao, J. Qu, and L. Ji, “Natural-logarithm-rectified activation function in convolutional neural networks,” in *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*. IEEE, 2019, pp. 2000–2008.
- [7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, 2012.
- [8] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, “On the variance of the adaptive learning rate and beyond,” *Proc. International Conference on Learning Representations 2020*, 2020.
- [9] L. N. Smith and N. Topin, “Super-convergence: Very fast training of neural networks using large learning rates,” in *Artificial intelligence and machine learning for multi-domain operations applications*, vol. 11006. SPIE, 2019, pp. 369–386.
- [10] M. Gheorghe, S. Mihalache, and D. Burileanu, “Using deep neural networks for detecting depression from speech,” in *2023 31st European Signal Processing Conference (EUSIPCO)*. IEEE, 2023, pp. 411–415.
- [11] S. Sarangi, M. Sahidullah, and G. Saha, “Optimization of data-driven filterbank for automatic speaker verification,” *Digital Signal Processing*, vol. 104, p. 102795, 2020.
- [12] L. Vu, R. C.-W. Phan, L. W. Han, and D. Phung, “Improved speech emotion recognition based on music-related audio features,” in *2022 30th European Signal Processing Conference (EUSIPCO)*. IEEE, 2022, pp. 120–124.
- [13] S. R. Livingstone and F. A. Russo, “The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english,” *PloS one*, vol. 13, no. 5, p. e0196391, 2018.
- [14] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, “Crema-d: Crowd-sourced emotional multimodal actors dataset,” *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.
- [15] 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, p. 335, 2008.
- [16] Y.-Y. Yang, M. Hira, Z. Ni, A. Chourdia, A. Astafurov, C. Chen, C.-F. Yeh, C. Puhrsch, D. Pollack, D. Genzel, D. Greenberg, E. Z. Yang, J. Lian, J. Mahadeokar, J. Hwang, J. Chen, P. Goldsborough, P. Roy, S. Narendhiran, S. Watanabe, S. Chintala, V. Quenneville-Bélair, and Y. Shi, “Torchaudio: Building blocks for audio and speech processing,” *arXiv preprint arXiv:2110.15018*, 2021.
- [17] T. Li, Y. Yin, K. Ma, S. Zhang, and M. Liu, “Lightweight end-to-end neural network model for automatic heart sound classification,” *Information*, vol. 12, no. 2, p. 54, 2021.
- [18] F. Li, Z. Zhang, L. Wang, and W. Liu, “Heart sound classification based on improved mel-frequency spectral coefficients and deep residual learning,” *Frontiers in Physiology*, vol. 13, p. 1084420, 2022.
- [19] C. Liu, D. Springer, Q. Li, B. Moody, R. A. Juan, F. J. Chorro, F. Castells, J. M. Roig, I. Silva, A. E. Johnson *et al.*, “An open access database for the evaluation of heart sound algorithms,” *Physiological measurement*, vol. 37, no. 12, p. 2181, 2016.
- [20] M. Deng, T. Meng, J. Cao, S. Wang, J. Zhang, and H. Fan, “Heart sound classification based on improved mfcc features and convolutional recurrent neural networks,” *Neural Networks*, vol. 130, pp. 22–32, 2020.