Formatting Instructions For NeurIPS 2023

Coauthor
Affiliation
Address
email

Coauthor
Affiliation
Address
email

Coauthor
Affiliation
Address
email

Coauthor
Affiliation
Address
email

Abstract

Our project is to build a deep learning model that can recognize emotions in human speech. By identifying emotions like happiness, sadness, anger, and surprise, this could potentially allow virtual assistants or music apps to adjust their responses based on how a person feels. To achieve this, we propose to use a type of deep learning model called a Convolutional Recurrent Neural Network. The idea is that our CRNN combines layers that pick up important sound features with layers that track changes over time, making it a good fit for picking up patterns in speech. Moreover, our training data will come from the Toronto Emotional Speech Set. This data set has over 2800 audio samples spoken by two voice actresses. Before feeding this data, we plan to apply data preprocessing techniques like removing background noise and converting the audio to visual spectrograms. Then we plan to implement a dynamic learning rate scheduler. This essentially allows us to adjust the learning rate of our CRNN during the training process, which optimizes our training speed. Lastly, some extra techniques we plan to use to help generalize speech patterns are pitch shifting and time stretching. Pitch shifting changes the tone of the sound, making it higher or lower. While time stretching changes how fast or slow the sound is, but keeps the tone untouched.

1 Introduction

Every day, millions of people interact with digital devices that can do amazing things - playing music, answering questions, and helping us stay organized. Although these technologies play a huge part in our lives, they are unable to understand how humans feel.

For this reason, we strive to develop a deep learning model that can classify emotions in human speech. This would allow us to have more enhanced human-machine interactions. For example, a person's favorite playlist could change their song choice by adapting to their emotions in real-time. Another example is that virtual assistants could respond to users based on their emotional state. Being able to offer mental health support to people at their lowest point can potentially save lives and prevent disasters such as school shootings or suicides.

While steering toward the future, we are inevitably becoming more reliant on digital interfaces in our daily lives. For this reason, the ability to detect emotions is essential if we want to create more enriched lifestyles for the majority of the populace.

To achieve this, our general approach is to use a CRNN to analyze audio data. This type of model allows us to automate the feature extraction process and also handle the complexities of audio signals. Moreover, many previous studies have shown that CRNNs perform well in similar tasks of audio recognition.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

2 Background and Related Works

Speech signals are the natural medium of human communication that convey information about the speaker's message, perspective and identity as well as their emotions. Speech Emotion Recognition (SER) is a classification problem that aims to infer the emotional state of a speaker from speech signals. Although recognizing emotions is easy for humans, it is a challenging task for machines. This is because speech itself is a complex signal that differs from person to person and can change depending on context and intonation [4] [5]. In recent years, there have been successful attempts at using Deep Neural Networks (DNN) and Convolutional Neural Networks (CNN) to recognize emotions from speech [7] [9]. These attempts have also stressed the importance of finding meaningful features through feature extraction in SER.

2.1 Feature Extraction for SER

Emotions in speech can be expressed through subtle variations in pitch, tone, energy, and timing. The goal of this step is to find features that can be extracted from input audio to differentiate between different emotions. Previous works have made use of a combination of acoustic features, which include prosodic and spectral features, for SER classification. Language information may be used as well but is often paired with acoustic features [7] [9].

Prosodic features are features that are related to pitch, loudness, rhythm, tempo, and intonation, which give clues about how something is said rather than what is said. These features provide a way to differentiate emotions of different arousal (intensity) levels, such as happy and sad. However, they can't differentiate between emotions of the same arousal level but different valence (positivity), such as happy and angry. Spectral features are features that capture frequency of speech signals. The most common of these features are Mel-Frequency Cepstral Coefficients (MFCCs) [4]. Although these handcrafted features seem meaningful, they might not be enough for SER [5]. In this case, deep learning can be trained on Mel-spectrograms.

A Mel-spectrogram is a feature that provides a way to visually reprsent changes in loundless and frequency in a speech signal over time. The finer details of this feature aren't necessary to know but can provide useful context. The Mel part of the name comes from scaling frequency in a way that matches the human auditory perception. The spectrogram part means that different parts a speech signal input is mapped to from a time domain to a frequency domain using fast Fourier transform before being stacked together. The result of stacking is then a graph that maps time-frequency to loudness [8].

2.2 CRNN for SER

Most SER models use deep learning as they provide better results than traditional machine learning models and can learn features from complex raw data. These models often include Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) or a combination of both [4]. CNNs are effective at capturing important patterns that occur in Mel-spectrograms. Recurrent Neural Networks (RNN) , specifically Long Short-Term Memory (LSTM) networks are effective at handling sequential input data such as audio.

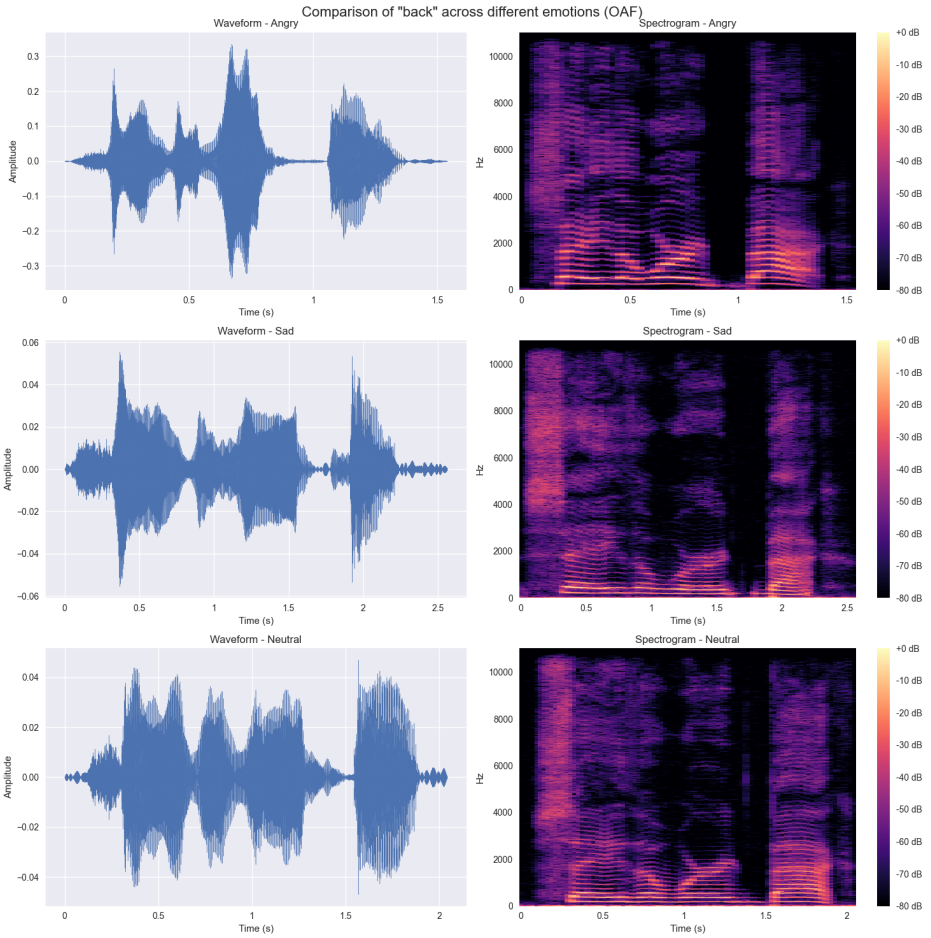
2.3 Attention Mechanism

Not all features of the speech signal might be relevant in recognizing emotions [4]. For example, a speaker might sound neutral in the beginning and ending of a sentence but sound angry in the middle. The classifier should focus on the features that occur in the middle of the sequence. This leads to using attention mechanisms to improve the performance of LSTM networks by putting more emphasis on certain parts of the input sequence.

3 Data

The Toronto Emotional Speech Set (TESS) is a well-established dataset designed for emotion recognition in speech. The dataset consists of 2,800 audio samples of 200 target words spoken in seven distinct emotions (angry, disgust, fear, happy, pleasant, surprised, sad, and neutral) by two

actors: a younger adult (YAF) and an older adult (OAF). This creates a balanced dataset with 400 samples per emotion and 1,400 samples per speaker. This balance is crucial for training unbiased models, as it ensures equal representation of each emotional class.



Comparison of waveforms and spectrograms of the word “back” spoken with angry, sad, and neutral emotions by the older adult female speaker (OAF). Left column shows the time-domain waveforms, the right column shows the corresponding spectrograms with intensity in dB.

Figure 3 reveals several key patterns in how emotions affect speech characteristics. The waveforms show clear differences in amplitude, with angry speech having the highest amplitude (±0.3), while sad and neutral speech show much lower amplitudes (±0.06 and ±0.04 respectively). In the spectrograms, we can observe that angry speech displays more intense energy (shown by brighter colors) especially in the 0-5 kHz range, while sad and neutral expressions show more diffused energy patterns.

The temporal characteristics also vary notably across emotions. As seen in the time axes, the sad utterance is the longest (around 2.5 seconds), while the angry utterance is more compact (1.5 seconds), and the neutral utterance takes a moderate duration (2 seconds). Looking at the spectrograms’ frequency distribution, angry speech shows distinct, high-intensity patterns (visible as bright red-purple regions), whereas sad and neutral expressions share similar, more subtle spectral patterns. These acoustic similarities between certain emotion pairs, combined with varying temporal patterns, suggest our model will need to consider both timing and frequency characteristics to effectively classify emotions.

4 Ethical Considerations

In utilizing the Toronto Emotional Speech Set (TESS), we acknowledge several important ethical considerations regarding privacy, bias, and potential applications. While the dataset includes proper

consent from the two female actors and is publicly available, we recognize that voice data is inherently personal and requires careful handling. The dataset has notable representation limitations: it only includes female voices, is limited to two age groups, and contains only English recordings from Canadian speakers, potentially affecting our model's generalization capabilities. Additionally, since the emotions are performed rather than naturally occurring, this may limit the applicability of our results in the real world. We also acknowledge that emotion recognition technology could potentially be misused for unauthorized surveillance or biased decision-making.To address these concerns, we commit to the following:

1. Bias Mitigation: Recognizing the dataset's demographic limitations, we aim to use data augmentation techniques to simulate a broader range of voices. However, we will be transparent about the limitations that these techniques impose, as simulated diversity cannot truly replace real-world diversity in voice samples.
2. Fairness Evaluation: We will, to the best of our ability, assess the model's performance across various demographic and linguistic settings, using external datasets, if possible.
3. Usage and Transparency: We commit to using the data solely for emotion classification research, openly addressing the challenges of using acted emotion data and single-gender samples.

5 Work Division

Each team member meet every week Monday 9-10pm to discuss the progress and next steps of the final project. At each meeting, individual members are assigned a task that they are responsible for completing by the end of the week. Collaboration is done asynchronously through an online video and messaging platform. with writing done on a shared document.

5.1 Project Proposal

The project proposal responsibilities were equally divided as follows:

- Jai: Wrote the abstract and introduction
- David: Wrote the background and related work
- Fan: Wrote the data section and collaborated with Adam on the ethical considerations section
- Adam: Wrote the model architecture and figure and collaborated with Fan on the ethical considerations section

5.2 Final Report

Similarly, the tentative work division responsibilities for the final report are:

- Jai: Work on Results section
- David: Work on Discussion section
- Fan: Work on Limitations section
- Adam: Work on Conclusion section

6 Model Architecture

After a thorough investigation of the TESS Dataset, we propose a Convolutional Recurrent Neural Network (CRNN) architecture.

6.1 Input Processing

In order to ensure uniformity across the TESS dataset, we apply a standardized preprocessing pipeline to the raw audio signals. First, each sound bite is resampled to a fixed rate and then normalized to

a consistent amplitude. Second, silence segments are removed to minimize irrelevant input (noise) [1], and segmented to a fixed length; shorter samples are padded while longer samples are trimmed. Finally, each processed sample is converted into a Mel-spectogram. In effect, this transformation reduces the dimensionality of the audio data while preserving essential information about the temporal and spectral characteristics of the speech [1].

6.2 Convolutional Layers

The CRNN architecture is initiated by a series of three-dimensional convolutional layers designed to extract progressively abstract features from the spectogram input. The first block employs local feature detection, using small kernels to capture micro-level emotional indicators and basic acoustic patterns. As the layers deepen, the receptive fields gets larger to capture broader, higher-level abstractions, such as intonation patterns or general timbral characteristics of certain emotions. Each convolution is followed by a max-pooling layer which reduces the spatial dimensions of the feature maps. Moreover, batch normalization and dropout is applied after each convolutional layer to stabilize learning and prevent overfitting. These convolutional layers help the model "learn affective salient features for SER and show excellent performances on several benchmark datasets" [3].

6.3 Recurrent Layers

Following the convolutional layers, the model incorporates two bidirectional Long Short-Term Memory (BiLSTM) layers. Given that the audio data in TESS is relatively short and sequential, BiLSTMs are a practical choice as they capture dependencies across time steps, and allow for a more nuanced understanding of emotional expression. Additionally, the bidirectional design of the model helps capture the full emotional arc of an utterance by analyzing past and future time frames [1]. Thus, the RNN module serves as a temporal feature extractor [3], supporting the objective of emotion identification over the course of the audio sample.

6.4 Attention Mechanism

To help the model focus on the most salient temporal features for emotion recognition, an attention mechanism is incorporated after the BiLSTM layers [6], [3]. More precisely, the attention module calculates a weighted sum of the hidden states from the BiLSTM, where the weights are determined by both current hidden state and the overall sequence of hidden states [3]. This can be exceptionally useful since it allows the model to dynamically weight the importance of different time steps, emphasizing parts of the audio where emotional cues are the strongest, thus possibly improving model performance.

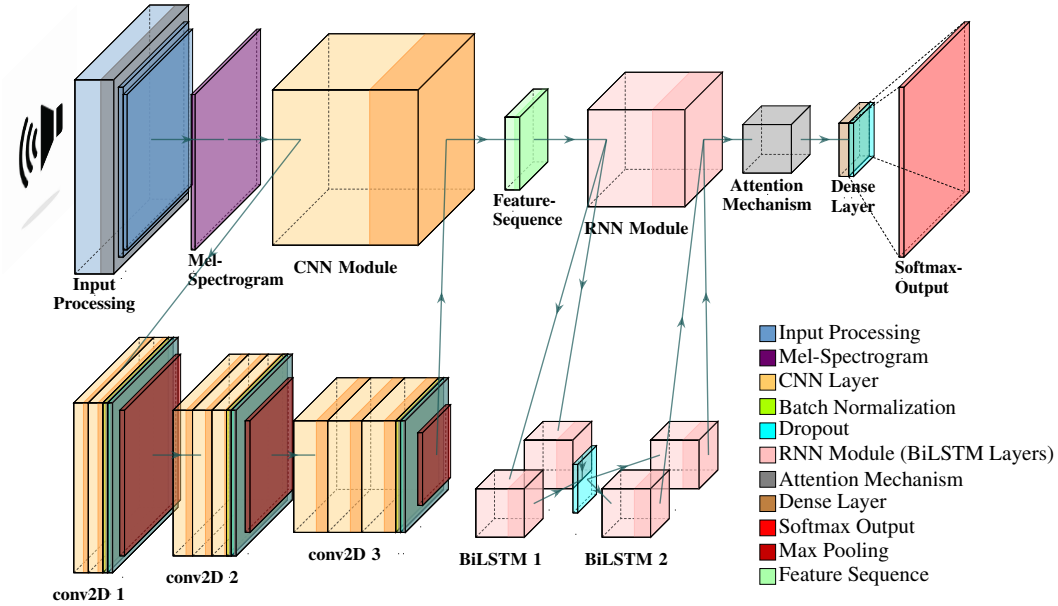
6.5 Dense Layers and Output Layer

Finally, after both the CNN and RNN modules, the resulting features are passed through a fully-connected dense layer. This integration layer is crucial since it combines our high-level temporal and spatial features into a single, unified representation [3]. To enhance generalization, dropout is applied within the layer. In turn, this refined representation helps transition our data into the final softmax output layer, producing a probability distribution across the seven emotion classes.

6.6 Training Protocol

The training strategy is specifically designed around the nuanced nature of emotional speech patterns. The model employs categorical cross-entropy loss (standard for multi-class classification), with the Adam optimizer for its ability to handle the inherent noise and variability in emotional expression [2]. Additionally, to enhance the robustness of the model, we apply audio-specific data augmentation. Augmentations such as random pitch shifting and time-stretching effectively expand our dataset by introducing slight variability in existing data, in turn making our model less prone to overfitting [2]. These protocols ensure model robustness, aiding us in our goal of building a truly useful and reliable voice-based emotion classifier.

CRNN Model Architecture



References

[1] O. Atila and A. Şengür. Attention guided 3d cnn-lstm model for accurate speech based emotion recognition. *Applied Acoustics*, 182:108260, 2021. ISSN 0003-682X. doi: https://doi.org/10.1016/j.apacoust.2021.108260. URL https://www.sciencedirect.com/science/article/pii/S0003682X21003546.

[2] S. Bhatlawande, V. Telgote, S. Agrawal, and S. Shilaskar. Decoding emotions from sound: A comprehensive approach to audio emotion recognition. In *2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS)*, pages 513–518, 2024. doi: 10.1109/ICC-ROBINS60238.2024.10533914.

[3] M. Chen, X. He, J. Yang, and H. Zhang. 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, 25(10):1440–1444, 2018. doi: 10.1109/LSP.2018.2860246.

[4] A. Hashem, M. Arif, and M. Alghamdi. Speech emotion recognition approaches: A systematic review. *Speech Communication*, 154:102974, Oct. 2023. ISSN 0167-6393. doi: 10.1016/j.specom.2023.102974. URL http://dx.doi.org/10.1016/j.specom.2023.102974.

[5] A. Koduru, H. B. Valiveti, and A. K. Budati. Feature extraction algorithms to improve the speech emotion recognition rate. *International Journal of Speech Technology*, 23(1):45–55, Jan. 2020. ISSN 1572-8110. doi: 10.1007/s10772-020-09672-4. URL http://dx.doi.org/10.1007/s10772-020-09672-4.

[6] Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang, and M. Akagi. Speech emotion recognition using 3d convolutions and attention-based sliding recurrent networks with auditory front-ends. *IEEE Access*, 8:16560–16572, 2020. doi: 10.1109/ACCESS.2020.2967791.

[7] N. T. Pham, D. N. M. Dang, N. D. Nguyen, T. T. Nguyen, H. Nguyen, B. Manavalan, C. P. Lim, and S. D. Nguyen. Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition. *Expert Systems with Applications*, 230:120608, Nov. 2023. ISSN 0957-4174. doi: 10.1016/j.eswa.2023.120608. URL http://dx.doi.org/10.1016/j.eswa.2023.120608.

[8] L. Roberts. Understanding the mel spectrogram, Mar 2020. URL https://medium.com/analytics-vidhya/understanding-the-mel-spectrogram-fca2afa2ce53.

[9] S. Zhang, S. Zhang, T. Huang, and W. Gao. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia*, 20(6):1576–1590, June 2018. ISSN 1941-0077. doi: 10.1109/tmm.2017.2766843. URL http://dx.doi.org/10.1109/TMM.2017.2766843.