Urdu Speech Emotion Recognition

Anand Kumar ak05173, Ghani Haider gh05177, Salman Muhammad Younus sy04351
CS351 – Artificial Intelligence, Fall 2021
Habib University

Problem

Speech Emotion Recognition (SER) can be defined as the task of recognizing or extracting the emotional state of the speaker from the speech signal, irrespective of the semantic contents. It can be viewed as a classification problem in which you discriminate between different emotions. This is done by extracting different speech information which identifies specific types of tones and pitches in the voice of a speaker which is then mapped to a state of emotion respectively.
Building a model to detect emotions from speech is a challenging task due to different limitations in datasets, feature extraction, and classifiers. There have been various attempts at building SER models in the English language, but few attempts have been done in building an Urdu speech emotion recognition model [1] [2]. Therefore, the aim of our project is to build a speech emotion recognition model that can detect emotions from Urdu speech and audio files.

Related Works

1. Emotion recognition from speech: A Review
   In this paper, the recent literature on speech emotion recognition has been presented considering the issues related to emotional speech corpora, different types of speech features and models used for recognition of emotions from speech [3].
2. Cross-Lingual Speech Emotion Recognition: Urdu vs. Western Languages
   This paper investigates the problem of cross-lingual emotion recognition for Urdu language and contribute URDU—the first ever spontaneous Urdu-language speech emotion database [1].
3. SEMOUR: A Scripted Emotional Speech Repository for Urdu
   This paper presents SEMOUR, the first scripted database of emotion-tagged speech in the Urdu language, to design an Urdu Speech Recognition System [4].

Solution and Approach

In order to correctly classify the audio inputs to emotions, we will build our own feedforward Artificial Neural Network (Multilayer Perceptron) model. The input to the model will be the audio features extracted on which we will train the classifier.

Libraries

1. Sklearn: Using it for splitting our dataset into training and test, as well as for evaluating the predicted results.
2. Tensorflow: For building the MLP classifier.
3. SoundFile: Mainly using for I/O interaction with audio files and as a dependency for the Librosa library.
4. Librosa: Using for audio analysis. Allows to extract features and decompose spectrograms in order to give appropriate inputs, i.e real numbers, to the machine learning model.

Dataset

Following are the two dataset that we are utilizing to make our speech emotion detection model.

1. SEMOUR: The dataset contains 15,000 instances, around 7 hours of audio data spoken in Urdu language by eight actors containing eight emotions. https://doi.org/10.1145/3411764.3445171
2. Urdu Language Speech Dataset: Contains 400 Urdu speech audio samples classified into four different emotions; Angry, Happy, Neutral, and Emotion. https://www.kaggle.com/bitlord/urdu-language-speech-dataset

Methodology

In order to train the model and classify the correct emotion of any given speech, we would first need to extract useful features from the given audio. The audio features we will use for training the model are Mel-Frequency Cepstral Coefficients (MFCCS), Chromagram, and Mel-spectrogram. These features will then be fed into the Neural Network Classifier and trained to correctly map the audio to the eight emotions labelled as output.

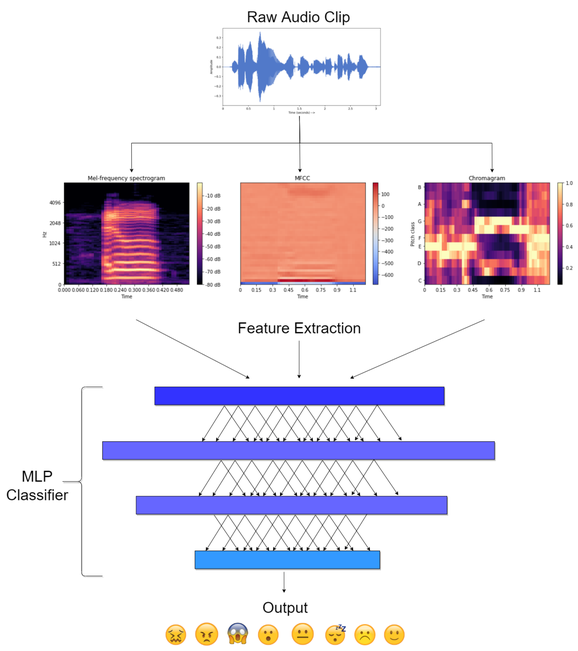

Figure 1. The complete procedure involved which include the acquisition of audio data, feature extraction and the classification to predict the complex emotion

Result

We constructed a five layer MLP classifier for training and classification. We trained our model for 50 epochs on a randomly selected 70% dataset (10, 500 instances). Once trained, we then test it on the remaining 4,500 instances and obtained an average accuracy of 91%. To validate the experiment, we also calculated precision and recall for each emotion as well as the accuracy for individual emotion class was also analyzed for the variance. We observed that our model performed exceptionally well on the Neutral and Boredom emotions with a f1-score of 95%, and 96%, respectively. The worse performing emotions were Fearful and Happiness which were identified with a score of 86%, and 85%, respectively. The results are visualized in figures below.

Column1	Sadness	Anger	Happy	Neutral	Surprise	Boredem	Fearful	Disgust
Sadness	526	2	11	0	0	4	24	3
Anger	4	505	14	0	14	0	6	4
Happy	10	9	515	2	9	2	31	3
Neutral	3	0	6	493	1	12	6	5
Surprise	9	13	33	0	466	5	25	10
Boredem	1	0	0	5	0	561	3	2
Fearful	17	6	11	0	7	5	517	15
Disgust	3	1	5	2	2	3	9	517

Figure 2. Confusion Matrix for the predicted dataset

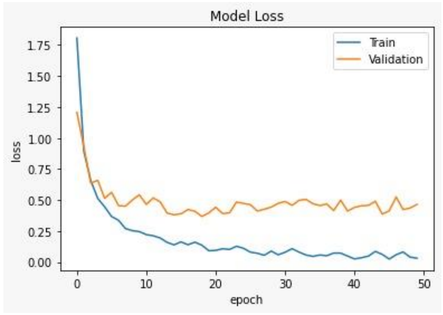

Figure 3. The overall loss of the training and validation dataset of the speech emotion detection model

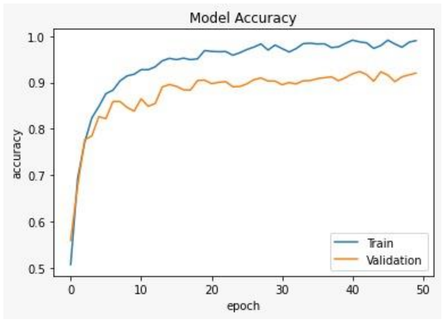

Figure 4. The overall accuracy of the training and validation dataset of the speech emotion detection model

	precision	recall	f1-score	support
Sadness	0.923	0.927	0.925	606
Anger	0.942	0.923	0.933	547
Happiness	0.866	0.886	0.876	581
Neutral	0.982	0.937	0.959	526
Surprise	0.934	0.831	0.879	561
Boredom	0.948	0.981	0.964	572
Fearful	0.833	0.894	0.862	578
Disgust	0.925	0.954	0.939	542
accuracy			0.916	4513
macro avg	0.919	0.917	0.917	4513
weighted avg	0.918	0.916	0.917	4513

Figure 5. Precision, recall and f1-score of all the eight emotions for the predicted dataset

Conclusion

The task of a speech emotion recognition system is to correctly identify the emotion of a conversation in a natural setting. However, SEMOUR dataset is an acted and a studio-recorded dataset which limits the capabilities of our trained model to correctly recognize the emotion in an actual conversation setting as compared to a natural dataset which can generalize the emotion in a better way [5].
It is also found that sometimes acted emotions tend to be more expressive than real ones which then affects the performance of the our emotion detection system [3]. Most of the datasets used for emotion recognition models rely on annotations of the audio clips by humans. This adds a subjective opinion of the person tagging the audio dataset [6] as well. Urdu has four dialects and various accents that vary depending on the speaker's demographics. However, SEMOUR dataset is limited to only one dialect that is spoken by the people of Lahore. A corpus consisting of more dialects and speaker accents can result in better emotion outcomes [4].

References

[1] Latif, Siddique, et al. "Cross-lingual speech emotion recognition: Urdu vs. western languages." 2018 International Conference on Frontiers of Information Technology (FIT). IEEE, 2018. https://arxiv.org/pdf/1812.10411.pdf
[2] Zehra, Wisha, et al. "Cross corpus multi-lingual speech emotion recognition using ensemble learning." Complex & Intelligent Systems (2021): 1-10. https://doi.org/10.1007/s40747-020-00250-4
[3] Koolagudi, Shashidhar G., and K. Sreenivasa Rao. "Emotion recognition from speech: a review." International journal of speech technology 15.2 (2012): 99-117. https://doi.org/10.1007/s10772-011-9125-1
[4] Zaheer, Nimra, et al. "SEMOUR: A Scripted Emotional Speech Repository for Urdu." Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. 2021. https://doi.org/10.1145/3411764.3445171
[5] Richard T Cauldwell. 2000. Where did the anger go? The role of context in interpreting emotion in speech. In ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion. International Speech Communication Association, Newcastle, Northern Ireland, UK, 5 pages.
[6] Ellen Douglas-Cowie, Nick Campbell, Roddy Cowie, and Peter Roach. 2003. Emotional speech: Towards a new generation of databases. Speech communication 40, 1-2 (2003), 33-60.