Emotion Recognition using Speech Processing

Melissa Vega, Samara Marrow, Jon Lee

San Diego State University

10 May 2023

Abstract The identification of emotions from speech has become a significant area of research in recent times. In this work, we will review existing literature on emotion recognition using speech processing. The process of this paper is to recognize the different types of models used for emotion recognition through speech, the issues related to speech emotion recognition, and the different types of features used in recognizing emotions from speech. Different classification techniques are discussed in the literature reviewed.

Keywords Emotion recognition, Classification models, Natural speech, Feature extraction

1. Introduction

As humans, we are able to recognize, understand, and use emotions. Speech is one of the main ways humans express themselves. If you hear someone's voice, there is a great chance that you will recognize what they are feeling. Now, have you imagined a computer doing the same? Lately, this has emerged as a crucial area of research. Although it is not an easy task, machines are gaining ground in recognizing, understanding, and analyzing patterns in speech. Speech emotion recognition (SER) is concerned with recognizing human emotional states conveyed in speech, without taking into account the meaning of the words being spoken.

In this paper, we will discuss and review existing literature on emotional speech literature. The information presented in this work considers the necessary tools and features to implement an SER model, the

efficiency of emotion recognition using speech, the issues that might arise when building this type of model, and why SER is relevant to the modern world. This paper also discusses the different approaches used in emotion recognition along with their advantages and disadvantages.

The purpose of this paper is to analyze and understand the importance of emotion detection based on human speech. We will review existing works that present solutions to the challenging task of detecting emotions using speech. Due to the fact that emotions are subjective, it is difficult to achieve a model that resembles a human, but it is possible to achieve a model that is able to mimic the human understanding of emotions. This topic is important because, in today's world, we interact with machines every single day. Think about smart assistants, such as Alexa or Siri. If they are able to recognize how we feel when interacting with them, the communication between humans and machines will be much more effective. One of the limitations of this study is that we are not recollecting data but instead, we are only reviewing existing literature, therefore our data will be limited to the one presented in the works being analyzed. In addition, our only previous knowledge about our topic is the information given to us in class and the data found online, meaning we are not experts in the subject.

2. Review of Related Literature

We reviewed multiple studies and articles related to emotion recognition. Among them, we reviewed an article published by the International Journal of Soft Computing (IJSCE) published in 2012, "Speech Emotion Recognition." This paper reviews the different classifiers used for emotion recognition and how they are implemented. The classifiers discussed in this particular

paper are k-nearest neighbors, hidden Markov models, artificial neural networks, and support vector machines. This paper also discusses the required features extracted from emotional speech samples that, in this case, are necessary to create an emotion recognition model: mel frequency cepstrum coefficients (MFCC), energy, pitch, and linear prediction cepstrum coefficient (LPCC). Lastly, this paper reviews the possibility of combining different methods and extracting more effective features from speech signals to improve and develop more efficient models.

Another relevant study to our project was published by the Institute for Neural Computation at the University of California, San Diego, USA. This paper shares a few similarities with the study published by the IJSCE. For example, extracted features like pitch and mel frequency cepstrum coefficients, although this study also included other features, such as log energy and added velocity of the pitch. These features were analyzed by utilizing quadratic discriminant analysis and support vector machine. This paper concluded that the most determinant features in recognizing emotions were pitch and energy. It was also deduced that further study is needed to explore and implement new features and classifiers to create more sophisticated and accurate approaches.

Another important piece of literature reviewed for this study was “Speech Emotion Recognition Using Deep Learning Techniques: A Review” published by the Institute of Electrical and Electronics Engineers. This paper also discusses feature extraction and the classification phase. However, this paper focuses on deep learning techniques. Deep learning is a field of machine learning that has become very popular in recent years. Some of the

advantages of using deep learning methods over traditional ones include the capability to automatically learn useful features from data without the need for manual feature extraction and the ability to deal with both labeled and unlabeled data. Being able to train speech recognition models using unlabeled data is useful for learning representations of the underlying structure of the data, which can then be finetuned on smaller datasets to achieve a higher level of accuracy. Finally, the paper discusses how multimodal emotion recognition has recently become a research topic. Multimodal emotion recognition may use input data like speech and images at the same time to train a more complex and elegant model.

The last piece of literature we reviewed was "Emotion recognition from speech: a review" by the International Journal of Speech Technology. This paper discusses the issues related to emotional speech corpora and the different features and classifiers used for developing speech emotion recognition systems. It also discusses a few concerns when doing research about emotion speech recognition, such as the influence of language and culture, how emotion can be subjective, and how speech emotion recognition systems should be robust enough to identify emotions in real-life conversations.

3. Presentation

Emotion recognition models consist of three fundamental components: signal preprocessing, feature extraction, and classification. Signal preprocessing identifies meaningful units in a signal. Feature extraction is then applied to determine the important features present in the signal. Some of the features employed to identify emotions are intensity, energy, pitch, rate of spoken words and variance, etc. The final step involves using classifiers

to map the extracted features to relevant emotions. The justification for choosing a certain classifier is not discussed very often in the literature. Typically, classifiers are selected using empirical evaluation. It is deduced that deep learning techniques like deep Boltzmann machines, recurrent neural networks, and deep convolutional neural networks perform better in emotion recognition than traditional techniques. For example, in a comparative analysis of different classifiers, the linear discriminant analysis showed 55% average accuracy for happiness, while the deep convolutional neural network showed 99% average accuracy.

Quality of speech emotional databases also plays a major role in the accuracy of speech emotion recognition models. There are multiple speech emotion databases available for free, like RAVDES, Emo-DB, and IEMOCAP. These databases are categorized into three types: Simulated, induced, and natural databases. The most common databases used in speech emotion recognition are simulated databases, in which the data is recorded by experienced performers. It is the simplest way to obtain a speech-based dataset of multiple emotions. Some of the issues with the speech emotional databases available at the moment are that they are not robust enough. Most of them are too small in size, containing very few emotions. In addition, there is a disparity among the available databases in terms of language. Very few databases are collected in languages like Swedish, Spanish, or Russian. The English language is the most predominant one among databases, which makes us wonder if language might be a factor when recognizing emotions accurately.

4. Conclusion

This paper reviews recent literature on speech emotion recognition and illustrates the different features and classifiers utilized to develop speech emotion recognition models. Both pitch and energy were shown to be the features that had the biggest role in recognizing emotions, while deep learning techniques were shown to be effective and easy to train. Limitations of speech emotion recognition include the predominant use of simulated databases to develop SER systems, which may not produce efficient results when recognizing speech from general public conversations. Another limitation is that expressing emotions is a multi-modal activity. Modalities like facial expressions, age, and gender may be used in addition to speech to develop superior speech emotion recognition systems. Finally, further study is needed to create models that resemble the human ability to identify emotions.

5. Appendix

This code is an implementation of an emotion recognition system based on audio data. The system uses a support vector machine (SVM) classifier to predict the emotion from the input audio file. The SVM classifier is trained on a dataset of audio files labeled with eight emotions: neutral, calm, happy, sad, angry, fearful, disgust, and surprised. This code uses the Librosa library to extract features from the audio files, particularly the Mel-frequency cepstral coefficients (MFCCs) which are commonly used in speech processing. The extracted features are then scaled using StandardScaler from the Scikit-learn library.

The project begins by defining two functions: load_data() and extract_features(). The load_data() function takes the path of the directory containing the audio files, iterates over the directories and audio files to load the audio files and their corresponding labels. The extract_features()

function extracts the MFCC features from the audio file using Librosa.

After loading and extracting features from the audio files, the data is split into training and testing sets using the train_test_split() function from Scikit-learn. The StandardScaler() is used to scale the training and testing data, which is then used to train an SVM model using the SVC() function. The trained model is then saved as a file in the current directory using the pickle.dump() function from the Pickle library.

When the script is run again, it checks if the saved files exist, and loads them if they do. If they do not exist, the script trains and saves new ones. The loaded scaler and model are used to scale the features extracted from an audio file using the transform() method, and predict the emotion using the predict_proba() method from the SVM model. The predicted emotion is then displayed on the console using the print() function.

Overall, this code demonstrates the process of loading, processing, training, and testing audio data to recognize emotions from audio files using machine learning algorithms, particularly SVM. The Pickle library is used to save and load trained models and scalers, allowing the script to reuse them later without retraining the model from scratch.

6. References

[1] Kwon, Oh-Wook, et al. "Emotion recognition by speech signals." *Eighth European conference on speech communication and technology*. 2003.

[2] Ingale, Ashish B., and D. S. Chaudhari. "Speech emotion recognition." *International Journal of Soft Computing and Engineering (IJSCE)* 2.1 (2012): 235-238.

[3] Koolagudi, S.G., Rao, K.S. Emotion recognition from speech: a review. *Int J Speech Technol* 15, 99–117 (2012). https://doi.org/10.1007/s10772-011-9125-1

[4] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124.