152118033 - PATTERN RECOGNITION (A)

Term Project

Project Title:

Speech Emotion Recognition

Project Members:

Mustafa Furkan BİLEN - 152120141013

Elif GENÇ - 152120151004

Sena KALAY – 152120151054

May 2020

Abstract- Automatic recognition of emotion is important for facilitating seamless interactivity between a human being and intelligent robot towards the full realization of a smart society. The methods of signal processing and machine learning are widely applied to recognize human emotions based on features extracted from facial images, video files or speech signals. [2] Recently, attention of the emotional speech signals research has been boosted in human machine interfaces due to availability of high computation capability. There are many systems proposed in the literature to identify the emotional state through speech. [3] Selection of suitable feature sets, design of a proper classifications methods and prepare an appropriate dataset are the main key issues of speech emotion recognition systems.[3] The purpose was to find an optimal combination of methods and group of features for stress detection in human speech. [4]

1. Introduction

Emotion plays an important role in the daily interpersonal interactions and is considered an essential skill for human communication [2]. Emotion recognition in spoken dialogues has been gaining increasing interest all through current years. Speech Emotion Recognition (SER) is a hot research topic in the field of Human Computer Interaction (HCI).[1] Huawei intelligent video surveillance systems, for instance, can support real-time tracking of a person in a distressed phase through emotion recognition. The capability to recognize human emotions is considered an essential future requirement of intelligent systems that are inherently supposed to interact with people to a certain degree of emotional intelligence [4]. The necessity to develop emotionally intelligent systems is exceptionally important for the modern society of the internet of things (IoT) because such systems have great impact on decision making, social communication and smart connectivity [2].

In general, the speech emotion recognition(SER) is a computational task consisting of two major parts: feature extraction and emotion machine classification. [1] The questions that arise here: What is the optimal feature set? What combination of acoustic features for a most robust automatic recognition of a speaker's emotion? Which method is most appropriate for classification? Thus came the idea to using a RNN method with the basic method MLP.[1]

2. Materials and Methods

Fig 1 is shown block diagram of speech emotion recognition system. This section is explaining method ve materials used.

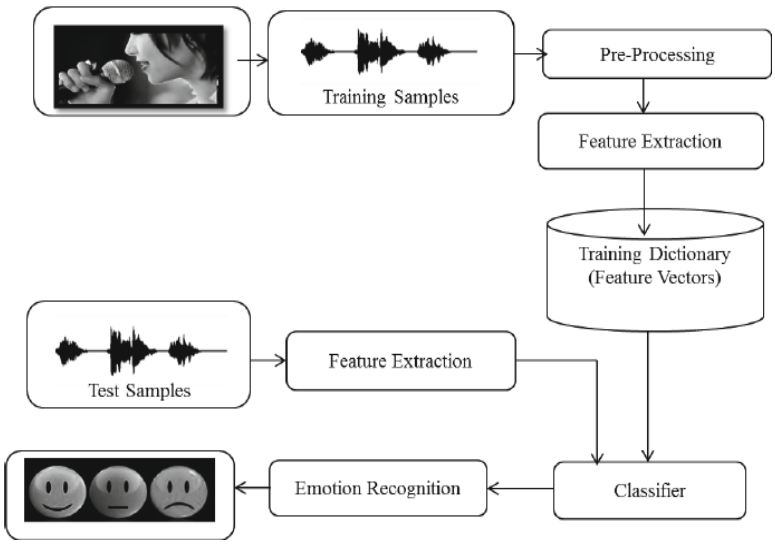


Figure 1. *Block diagram of speech emotion recognition system*

2.1. Database

The performance and robustness of the recognition systems will be easily affected if it is not well-trained with suitable database. Therefore, it is essential to have sufficient and suitable phrases in the database to train the emotion recognition system and subsequently evaluate its performance.[1] In this section, we detail the one emotional speech database used in our experiments: Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) Database[1].

It contains 1440 utterances spoken by 24 actors (12 female, 12 male) in 8 simulated emotions (neutral, calm, happy, sad, angry, fearful, disgust and suprised).

Each of the 1440 files has a unique filename. The filename consists of a 7-part numerical identifier (e.g., 03-01-06-01-02-01-12.wav). These identifiers define the stimulus characteristics:

- Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
- Vocal channel (01 = speech, 02 = song).
- Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
- Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion.
- Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- Repetition (01 = 1st repetition, 02 = 2nd repetition).
- Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

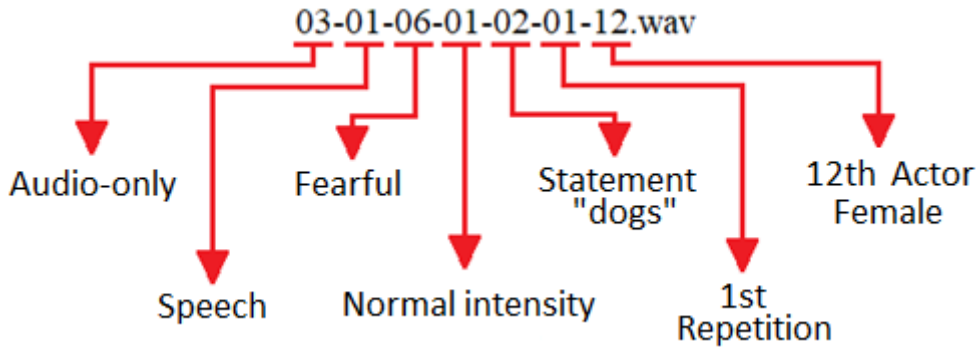


Figure 2. *Filename Example*

We used 70% of data for training, %10 of data for validation and 20% for testing.

2.2. Feature Extraction

The speech signal carries a large number of useful information that reflects emotion characteristics such as gender, age, stuttering and identity of the speaker.[2] The speech signal contains a large number of parameters hat reflect the emotional characteristics. One of the sticking points in emotion recognition is what features should be used. [1] Feature extraction is an important mechanism in audio processing to capture the information in a speech signal and most of the related studies have emphasized on the extraction of low-level acoustic features for speech recognition. [2] Fig 3. shows the features of the sound.

[1] https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio

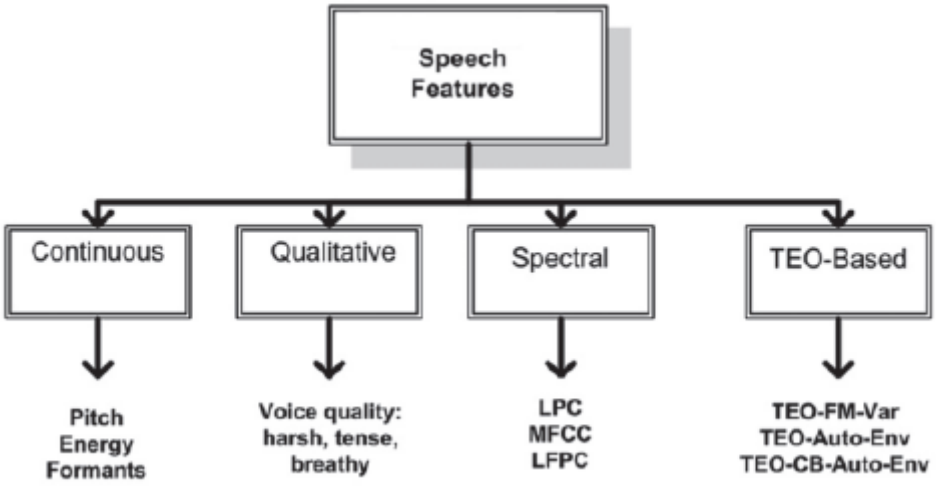


Figure 3. *Feature Categories[3]*

2.2.1. Spectral Feature

Spectral features of this study include timbral features that have been successful in music recognition. Timbral features define the quality of a sound and they are a complete opposite of most general features like pitch and intensity. It has been revealed that a strong relationship exists between voice quality and emotional content in a speech. Siedenburg et al. considered spectral features as significant in distinguishing between classes of speech and music. They claimed that the temporal evolution of the spectrum of audio signals mainly accounts for the timbral perception. Timbral features of a sound help in differentiating between sounds that have the same pitch and loudness and they assist in the classification of audio samples with similar timbral features into unique classes. The application of spectral timbral features like spectral centroid, spectral roll-of point, spectral flux, time domain zero crossings and root mean squared energy amongst others has been demonstrated for speech analysis. [2]

2.2.1.1. Mel-Frequency Cepstral Coefficients (MFCC)

Mel-Frequency Cepstrum coefficient is the most used representation of spectral property of voice signals. These are the best for speech recognition as it takes human perception sensitivity with respect to frequencies into consideration. For each frame, the Fourier transform and the energy spectrum were estimated and mapped into the mel-frequency scale.[1]

2.2.1.2. Linear Predictive Cepstral Coefficients (LPCC)

Linear prediction cepstral coefficients (LPCC) are cepstral coefficients derived from LPC calculated spectral envelope. LPCC are the coefficients of the Fourier transform illustration of the logarithmic magnitude spectrum of LPC. Cepstral analysis is commonly applied in the field of speech processing because of its ability to perfectly symbolize speech waveforms and characteristics with a limited size of features. It was observed by Rosenberg and Sambur that adjacent predictor coefficients are highly correlated and therefore, representations with less correlated features would be more efficient, LPCC is a typical example of such. The relationship between LPC and LPCC was originally derived by Atal in 1974. In theory, it is relatively easy to convert LPC to LPCC, in the case of minimum phase signals. [5]

2.3. Classification

Individual research shows that it cannot be said which classifier for emotion recognition is the best. Each classifier or their combination achieved some results accuracy, which depends on several factors.The success of classifier is directly dependent on the data. This is derived from the fact that the accuracy varies with the data character such as the uantity, density distribution of each class (emotions), and the language too. One classifier has different results with acted database,where the density of each emotion is equitabl and different with real data from call center where normal (calm) emotion state occupies 85 to 95 percent of all data. Appropriate choice of parameters has a considerable effect on the accuracy of these classifiers.The following subsections escribe the used classification methods.[4]

3.1. Multi-Layer Perceptron (MLP)

A multilayer perceptron is a neural network connecting multiple layers in a directed graph, which means that the signal path through the nodes only goes one way. Each node, apart from the input nodes, has a nonlinear activation function. An MLP uses backpropagation as a supervised learning technique. Since there are multiple layers of neurons, MLP is a deep learning technique. MLP is widely used for solving problems that require supervised learning as well as research into computational neuroscience and parallel distributed processing. Applications include speech recognition, image recognition and machine translation.[6]

3.2. Support Vector Machines (SVM)

Support Vector Machines (SVM) is an optimal margin classifiers in machine learning. It is also used extensively in many studies that related to audio emotion recognition which can be found in (A. et al., 2013), (Peipei et al., 2011) and (Pan et al., 2012). It can have a very good classification performance compared to other classifiers especially for limited training data (G.S. et al., 2016). SVM theoretical background can be found in (Gunn, 1998).

3. Experimental Results

The aim of the experiment was to clarify the significance of chosen groups of features, as well as classification ability of selected classification methods for speech emotion recognition system. Samples of examination were formed from recordings of human speech with various emotional characters. The following settings and features were used in the experiment:[4]

(i) input samples—RAVDESS database of emotional utterances:

(ii) feature extraction—computing of input vectors

(speech parameters):

(a) MFCC coefficients

(b) LPC coefficients,

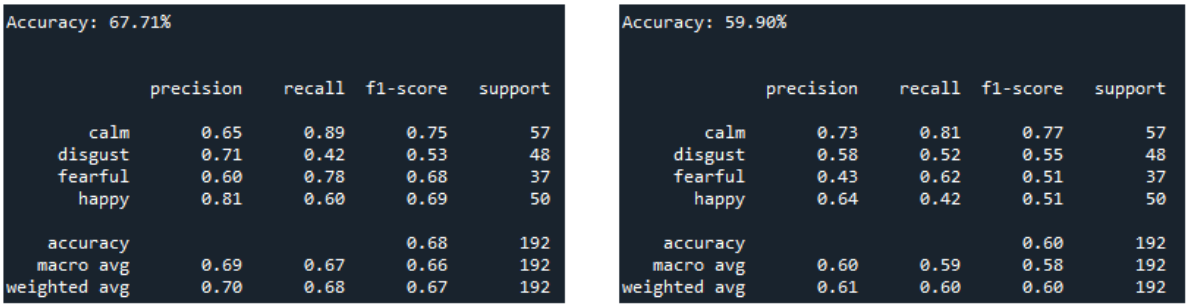
(iii) emotion classification:

(a) Multi-layer Perceptron

(b) Support Vector Machines

In this section, we describe the experiment environment and report the recognition accuracy of using MLR and SVM classifiers on RAVDESS database. Fig x is shown this. In conclusion accuracy depend

on feature, chosen number of features, classification algorithm, database and hyperparameters of classification models.



Accuracy: 67.71%

	precision	recall	f1-score	support
calm	0.65	0.89	0.75	57
disgust	0.71	0.42	0.53	48
fearful	0.60	0.78	0.68	37
happy	0.81	0.60	0.69	50
accuracy			0.68	192
macro avg	0.69	0.67	0.66	192
weighted avg	0.70	0.68	0.67	192

Accuracy: 59.90%

	precision	recall	f1-score	support
calm	0.73	0.81	0.77	57
disgust	0.58	0.52	0.55	48
fearful	0.43	0.62	0.51	37
happy	0.64	0.42	0.51	50
accuracy			0.60	192
macro avg	0.60	0.59	0.58	192
weighted avg	0.61	0.60	0.60	192

Figure 4. *Percentage average precision, recall, F1-score and accuracy with confidence intervals of learning algorithms trained with MFCC and LPCC*

4. Conclusion

A lot of uncertainties are still present for the best algorithm to classify emotions. Different combinations of emotional features give different emotion detection rate. The researchers are still debating for what features influence the recognition of emotion in speech.[1] The primary contribution of this study is the construction and validation of a set of hybrid acoustic features based on prosodic and spectral features for improving speech emotion recognition. In conclusion accuracy depend on feature, chosen number of features, classification algorithm, database and hyperparameters of classification models.

KAYNAKÇA

[1] Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, and et al., “Speech Emotion Recognition: Methods and Cases Study”, In Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018) - Volume 2, pages 175-182

[2] Kudakwashe Zvarevashe and Oludayo Olugbara, “Ensemble Learning of Hybrid Acoustic Features for Speech Emotion Recognition”, Algorithms 2020, 13, 70.

[3] Babak Basharirad and Mohammadreza Moradhaseli, “Speech Emotion Recognition Methods: A Literature Review”, AIP Conference Proceedings, Volume 1891, Issue 1, ID: 020105

[4] Pavol Partila, Miroslav Voznak, and Jaromir Tovarek, “Pattern Recognition Methods and Features Selection for Speech Emotion Recognition System”,Scientific World Journal Volume 2015, Article ID: 573068, pages 7

[5] Sabur Ajibola Alim and Nahrul Khair Alang Rashid,”Some Commonly Used Speech Feature Extraction Algorithms”, From Natural to Artificial Intelligence - Algorithms and Applications, Ricardo Lopez-Ruiz, IntechOpen, DOI: 10.5772/intechopen.80419.

[6] https://www.techopedia.com/definition/20879/multilayer-perceptron-mlp#:~:text=A%20multilayer%20perceptron%20(MLP)%20is,the%20input%20and%20output%20layers.&text=MLP%20is%20a%20deep%20learning%20method.
(Access Date: 29.05.2020)