Speech emotion recognition using suprasegment MFCC features

Daniil Krasnoproshin
Computer engineering department of
Belarussina State University
of Informatics and Radioelectronics
Minsk, Belarus
daniil.krasnoproshin@gmail.com

Maxim Vashkevich
Computer engineering department of
Belarussian State University
of Informatics and Radioelectronics
Minsk, Belarus
vashkevich@bsuir.by

***Abstract*—This study explores Speech Emotion Recognition (SER) using Mel-frequency cepstral coefficients (MFCCs) and Support Vector Machines (SVM) on the RAVDESS dataset. MFCCs are employed for acoustic feature extraction, capturing vital audio characteristics. SVM, a robust classifier, is used for recognizing emotions. The RAVDESS dataset offers diverse emotional states, making it ideal for training and evaluation. Our approach includes data preprocessing, MFCC feature extraction, and SVM-based classification.**

Experimental results demonstrate promising accuracy, showcasing the potential of this approach for applications like voice assistants, virtual agents, and mental health diagnostics. This study advances SER by emphasizing feature extraction and machine learning for improved human-machine interaction.

***Index Terms*—audio emotion recognition; human–computer interaction; computational paralinguistics; MFCCs; support vector machine; speech emotion recognition;**

I. Introduction

The field of computer speech emotions recognition began to develop rapidly in the last decade due to the growth in the performance of computational resources and the wide interest of researchers in the field of neurology [**], psychology [**], psychiatry and computer science. Emotions often influence decision-making processes, so emotion recognition may be of interest in order to build more effective communication, including dialogue systems (voice assistants, chat bots).

The problem of emotion recognition is currently a relevant and applied task of artificial intelligence. Its solution allows, for example, in the field of communication to build an effective relationship between a computer and a human, in the field of medicine(interfaces based on speech technologies for disabled, blind or visually impaired users), in decision-making tasks(recognition of negative emotions such as stress, anger, fatigue is an important aspect in terms of ensuring road safety with the use of intelligent vehicles, as it allows them to respond to the emotional state of the driver) etc.

There are various approaches to solving this problem. This largely depends on the information that captures the manifestation of human emotions. These can be facial images that capture facial expressions corresponding to different emotions or a speech signal that corresponds to the emotional state of the speaker. Therefore, the solution of the problem is reduced to the processing of the appropriate information.

Identify applicable funding agency here. If none, delete this.

In this paper, we consider an approach to solving the problem based on the processing of speech signals. At the same time, one of the main problems of this approach is related to the definition of a set of features that effectively describe this type of emotion. And thus, the construction of a feature space in which objects objects corresponding to different classes of emotions can be separated.

II. Extracting features from speech

To build a system for recognizing emotions from speech, it is necessary to preprocess the initial data[1]. The main task of preprocessing is to remove noise, increase the high frequencies of the signal and obtain a flat frequency spectrum of signals, as well as frequency characteristics [1] [2**][3].

Among the problems associated with speech processing, the selection and selection of features occupies a special place. Various audio features allow you to describe various aspects of the audio signal for solving various kinds of applied problems.

An analysis of the available approaches for feature categorization showed that the technique based on the calculation of Mel-frequency cepstral coefficients [2] is the most suitable for the purposes of the study [4**]. These indicators are widely used in the recognition of emotions in speech and are extremely effective tools for building various machine learning models[5**].

A. Cepstral representation of the voice in psychoacoustic scales

In this section, we consider the cepstral representation of a voice signal, obtained on the basis of a spectral analysis of the signal in a psychoacoustically motivated frequency scale. The mel-frequency cepstral representation, which is widely used to describe the voice signal, is analyzed [6].

The process of extracting Mel-frequency cepstral coefficients includes the following steps:

a) Short-time Fourier transform (STFT): : This is a special kind of Fourier transform that can be used to see how the frequencies in a signal change over time. It works by cutting your signal into many small segments and Fourier

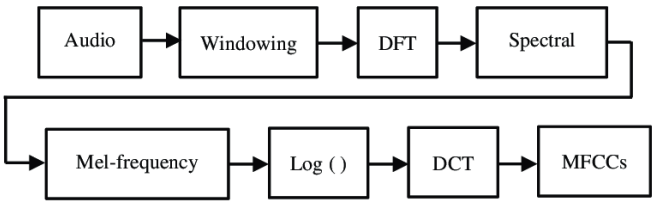

Fig. 1. Scheme for calculating mel-frequency cepstral coefficients (MFCCs)

transforming each one. The result is usually a waterfall plot showing frequency versus time.

Short-time Fourier transform (STFT) is widely used for the analysis, modification and synthesis of audio signals [4]. The STFT can be viewed as a sliding window transformation that has the form [5]:

$$X(k,l) = \sum_{n=0}^{N-1} h(n)x(n+lL)e^{-j\omega_k n} \quad (1)$$

where $X(t)$ is the input signal, $h(n)$ is the time-limited window function and $\omega_k = 2\pi k/M$, $k = 0, 1, ...M-1$ is the frequency index, L is the time step between adjacent frames, and l is the analysis frame number. It is easy to see that (1) is the calculation of the discrete Fourier transform (DFT) for the signal $h(n)x(n+lL)$. Thus, the representation resulting from the STFT is a sequence of time-localized spectra.

b) M-filter set calculation: used to model the properties of human hearing during the feature extraction phase, which can improve model performance. Therefore, we will use the Mel's scale to compare the actual frequency with the frequency that people perceive.

Define the mel filter bank: create a set of triangular filters spaced according to the mel-frequency scale. These filters are used to convert the power spectrum into the mel-frequency domain.

Apply the filters: Multiply each power spectrum frame with the mel filter bank to obtain mel-scale magnitudes.

Note that human hearing is less sensitive to changes in the energy of an audio signal at higher energy than at lower energy. The logarithmic function also has a similar property, with a low value of the input x, the gradient of the logarithmic function will be higher, but with a high value of the input gradient, the value will be smaller. So we apply log to the Mel filter output to simulate human hearing.

c) Discrete Cosine Transform (DCT): The problem with the resulting melspectrogramm coefficients are highly correlated. DCT is used to decorrelate these coefficients.

As a result, we get a set of numbers that are Mel-frequency cepstral coefficients (MFCCs).

III. Audio Dataset

During the study, the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [7] was used as the initial data set. RAVDESS contains 7356 entries from 24 actors (12 males, 12 females). All actors produced 104 different vocalizations, consisting of 60 spoken utterances and 44 song utterances. Each of the 104 vocalizations was exported to create three separate modal audio conditions: audio-video (face and voice), video-only (face but no voice), and audio-only (voice but no face). There were 312 files per actor (104×3). Recordings of one participant were lost for technical reasons (132 files). So 24 × 312–132 = 7356 files. This set consists of 4320 speech recordings and 3036 songs. The actors voiced two different phrases (in speech and song). Two phrases were pronounced with eight emotional colors (neutrality, calmness, happiness, sadness, anger, fear, surprise and disgust). In the case of the songs, six emotional colors were used (neutrality, calmness, happiness, sadness, anger and fear). All emotional states, except for the neutral one, were voiced at two levels of emotional loudness (normal and increased). The actors repeated each vocalization twice.

As part of this work, only a part of the RAVDESS dataset will be used, namely RAVDESS Emotional speech audio. This part of RAVDESS contains 1440 wav files (16bit, 48kHz): 60 entries for each of 24 professional actors (12 males, 12 females). Phrases with a neutral North American accent. Speech emotions include expressions of neutrality, calmness, happiness, sadness, anger, fear, surprise, and disgust. All emotional states, except for the neutral one, were voiced at two levels of emotional loudness (normal and increased). The actors repeated each vocalization twice.

Fig. 2 shows an example of a speech signal from the RAVDESS database:

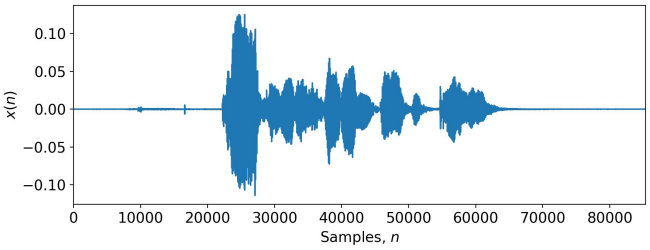

Fig. 2. Representation of the speech signal expressing anger

Fig. 3 shows the spectrogram of the speech signal:

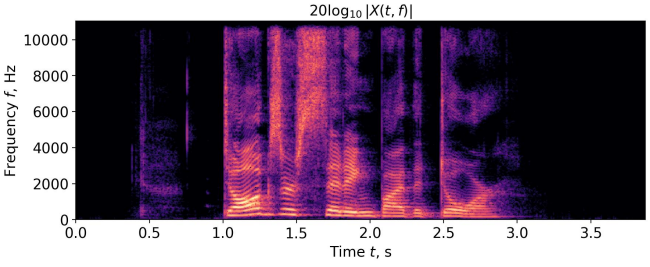

Fig. 3. Spectrogram of a speech signal expressing anger

Frame-calculated mel-frequency cepstral coefficients (MFCCs):

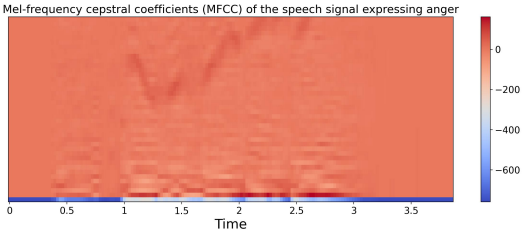

Fig. 4. Frame-based mel-frequency cepstral coefficients (MFCCs)

IV. Evaluation Design

When conducting experiments and testing the effectiveness of the MFCCs, the support vector machine (SVM) was used to solve the problem of recognizing emotions in speech. The SVM performs the classification by constructing N-dimensional hyperplanes that optimally separate the data into distinct categories.

Classification using SVM is achieved by constructing a linear (or non-linear) separating surface in the input data space. The idea of this approach is to transform (using the kernel function) the original dataset into a higher dimensional feature space. And already in the new feature space to achieve an optimal classification in a certain sense.

Any symmetric, positive semi-defining matrix K is used as a kernel, which consists of x_i and x_j pairs of vectors from scalar power elements, where $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ characterizes the measure of their adjustment. And ϕ is the result of the magnification transformation that forms the kernel. In particular, examples of such functions are:

- linear kernel:
$$K(x_i, x_j) = x_i^T x_j, \tag{2}$$
which corresponds to the classifier on the support vectors in the original space
- polynomial kernel with degree p:
$$K(x_i, x_j) = (1 + x_i^T x_j)^p, \tag{3}$$
- Gaussian kernel with radial base function (RBF):
$$K(x_i, x_j) = \exp(\gamma \|x_i - x_j\|^2). \tag{4}$$

RBF function was chosen as the kernel function of the model based on the SVM. The value of the parameter C (cost) (permissible penalty for violating the gap boundary) was equal to ... γ parameter was equal to... The construction of a classifier on support vectors using the kernels listed above can, in particular, be carried out using the sklearn library written in Python.

For training, testing and validation of the model, the k-fold cross-validation method was used [8]. The k-block cross-validation method includes the following steps [9]:

1) Shuffle the data set in a random (pseudo-random) way;
2) Divide the set into k groups;
3) For each unique group:
   a) select a group of records as test data (test data)
   b) take the remaining groups as training data (train data)
   c) train the model on training and evaluate its performance on test data
   d) save score value and reset model to initial state for next iteration
   e) set the average skill level of the model.

In this paper, the data was split into blocks as follows (in parentheses are the indices of the actors):

- block 0: (2, 5, 14, 15, 16)
- block 1: (3, 6, 7, 13, 18)
- block 2: (10, 11, 12, 19, 20)
- block 3: (8, 17, 21, 23, 24)
- block 4: (1, 4, 9, 22)

To evaluate the quality of the model, unweighted average recall was calculated. Unweighted average recall is a metric used to measure the overall performance of a classification model in a multi-class setting. It calculates the average recall across all classes, giving equal importance to each class without considering the class imbalance. To calculate the unweighted average recall, the following steps should be accomplished:

1) Calculate the recall for each individual class by dividing the number of true positives for that class by the sum of true positives and false negatives.
$$Recall = \frac{TP}{TP + FN}, \tag{5}$$
where TP – true positive, FN – false negative.
2) Add up the recall values for all classes.
3) Divide the sum by the total number of classes to get the unweighted average recall.

The UAR value is in the range from 0 to 1.

Characteristics of the machine on which the experiment was carried out:

1. Processor AMD Ryzen 7 5700U with Radeon Graphics; 2.Video card AMD Radeon RX Vega 8 (Ryzen 4000/5000) (- 1900 MHz); 3. RAM 16GB DDR4-2400; 4. OS Ubuntu 20.04.5 LTS;

The experiment was carried out in three stages:

1) training sample preparation;
2) training and testing of the classifier using a different number of features.
3) model evaluation using UAR metric

The formula for Unweighted Average Recall (UAR) is given by:

$$\text{UAR} = \frac{1}{k} \sum_{i=1}^{k} \frac{A_{ii}}{\sum_{j=1}^{k} A_{ij}} \tag{6}$$

where A – confusion matrix, k – number of classes.

As a result of building and training the model, a classifier was obtained, the prediction accuracy of which, when using the test data set and the above quality metric, reached 46.1%. Next, a multiclass confusion matrix will be presented, which is a table or diagram showing the accuracy of classifier prediction

in relation to two or more classes. The cells of the table are filled with the number of classifier predictions. Correct predictions go along the main diagonal from the top left corner to the bottom right.

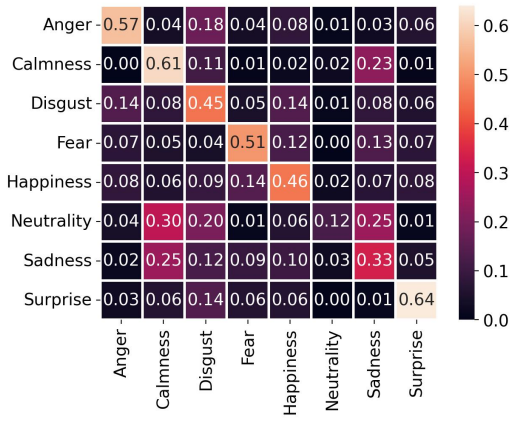

Fig. 5. Multiclass confusion matrix

V. Conclusion

In the realm of human-computer interaction, the accurate recognition of emotions from speech is a pivotal factor. This study has delved into the intricate world of Speech Emotion Recognition (SER), unveiling an approach that shows promise, though with an UAR of 46.1% ($C = 10$, $\gamma = 0.001$), there is room for improvement.

By harnessing Mel-frequency cepstral coefficients (MFCCs) and leveraging the power of Support Vector Machines (SVM), our research took a significant step toward recognizing emotions from audio data. The RAVDESS dataset, with its diverse emotional range, provided an excellent backdrop for training and testing our model.

Through meticulous data preprocessing, MFCC feature extraction, and SVM-driven classification, we demonstrated the efficacy of this approach. While our results hint at practical applications, including emotionally responsive voice assistants, empathetic virtual agents, and enhanced mental health diagnostics, they also underscore the need for further research and refinement.

In conclusion, this study marks a stride forward in Speech Emotion Recognition, highlighting the potential of robust feature extraction and machine learning techniques in amplifying human-machine interactions. However, it also humbly acknowledges that the journey to harness the emotional dimension of human communication requires continued exploration and enhancement.

References

[1] L. Xie, Z.-H. Fu, W. Feng, and Y. Luo, "Pitch-density-based features and an SVM binary tree approach for multi-class audio classification in broadcast news," *Multimedia systems*, vol. 17, pp. 101–112, 2011.

[2] C. K. On, P. M. Pandiyan, S. Yaacob, and A. Saudi, "Mel-frequency cepstral coefficient analysis in speech recognition," in *2006 International Conference on Computing & Informatics*, 2006, pp. 1–5.

References

[1] Pandiyan. Mel-frequency cepstral coefficient analysis in speech recognition. Computing and Informatics 2006, ICOCI'06, no. 2, pp. 2–6.

[2] L. Xie, Z.-H. Fu, W. Feng, and Y. Luo, "Pitch-density-based features and an SVM binary tree approach for multi-class audio classification in broadcast news," Multimedia Systems, vol. 17, pp. 101–112, 2011.

[3] Suliman, A., Omarov, B., Dosbayev, Zh. Detection of impulsive sounds in stream of audio signals. 2020 8th International Conference on Information Technology and Multimedia, ICIMU 2020, 2020, pp. 283–287.

[4] Rajkumar Palaniappan, K. Sundaraj. Respiratory Sound Classification using Cepstral Features and Support Vector Machine. 2013 IEEE Recent Advances in Intelligent Computational Systems (RAICS). 978-1-4799-2178-2/13.

[5] Nazarov MV, Prokhorov Yu. N. Methods of digital processing and transmission of speech signals. Moscow: Radio and communication, 1985. 176 p.

[6] Sorokin V.N. The Structure of the Problem of Automatic Speech Recognition // Information Technologies and Computing Systems, 2004, No. 2, pp. 25–40.

[7] Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Access url: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio?datasetId=107620 (Access Date: 21.05.2023).

[8] Issa, D.; Fatih Demirci, M.; Yazici, A. Speech emotion recognition with deep convolutional neural networks. Biomed. Signal Process. Control 2020, 59, 101894.

[9] Luna-Jiménez, Cristina, et al. "Multimodal emotion recognition on ravdess dataset using transfer learning." Sensors 21.22 (2021): 7665.