Speech emotion recognition using SVM classifier with suprasegmental MFCC features

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

Maxim Vashkevich
Computer engineering department
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 (SVMs) classifier on the RAVDESS dataset. We proposed a model which uses 80-component suprasegmental MFCC feature vector as an input downstream by SVM classifier. To evaluate the quality of the model, unweighted average recall (UAR) was used. We evaluate different kernel functions for SVM (such as linear, polynomial and radial basis)and different frame size for MFCC extraction (from 20 to 170 ms). Experimental results demonstrate promising accuracy(UAR = 48%), showcasing the potential of this approach for applications like voice assistants, virtual agents, and mental health diagnostics.**

***Keywords*—emotion recognition; speech signal; MFCC; support vector machine; speech emotion recognition;**

I. Introduction

The field of computer speech emotions recognition (SER) began to develop rapidly in the last decades 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 [1], [2]. 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.

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 [1], [3]–[5]. And thus, the construction of a feature space in which objects corresponding to different classes of emotions can be separated.

In order to solve such a non-trivial problems two main techniques were: mel-frequency cepstral coefficients (MFCC) extraction as the basis for feature engineering pipeline and support vector machines (SVM) as a classification algorithm.

MFCC are a widely adopted and effective feature extractiontechnique for speech emotion recognition [1], [4]. MFCC replicate the human auditory system's response to sound,capturing relevant acoustic information [6]. By converting the audio signal into a frequency domain representation, MFCC highlight the essential characteristics of speech, such as spectral shape and pitch. This technique reduces the dimensionality of the data while retaining critical features, making it suitable for machine learning algorithms like SVM. Moreover, MFCCs are robust to noise and variations in speaking styles, ensuring that subtle emotional nuances in speech are preserved. As a result, MFCC serve as a valuable tool in speech emotion recognition, enabling models to discern emotional states accurately and reliably from audio signals.

At the same time, SVM offer a promising approach for speech emotion recognition, combining robust classification capabilities with adaptability to high-dimensional feature spaces. SVM are based on the principle of finding the optimal hyperplane that maximally separates different classes in feature space [7]. In the context of speech emotion recognition,this means SVM can effectively distinguish between various emotional states [4].Additionally, SVM can handle non-linear relationships through kernel functions, allowing them to capture intricate patterns in speech data. Their ability to generalize well and mitigate overfitting makes SVM suitable for the often noisy and nuanced nature of emotional speech.

II. Feature Extraction

The first stage of the SER system is the preprocessing of the input speech data [1], [4].

An analysis of the available approaches for feature categorization showed that the technique based on the calculation of MFCCs [6] is the most suitable for the purposes of the study. These indicators are widely used in the recognition of emotions in speech and are extremely effective tools for building various machine learning models [5], [8].

A. MFCC calculation

In this section, we consider the MFCC calculation. The steps of MFCC calculation is given in Fig. 1.

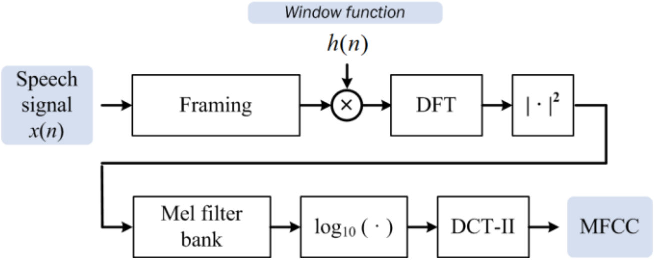


Fig. 1. Scheme for calculating mel-frequency cepstral coefficients

The process of MFCC extracting includes the following steps:

a) Short-time Fourier transform (STFT): This is a special kind of Fourier transform that is used to see how the amplitudes of the frequency components of a signal change over time. It works by splitting the signal into short-time segments and applying discrete Fourier transform (DFT) to each one. STFT is widely used for the analysis, modification and synthesis of audio signals [9]. The STFT can be viewed as a sliding window transform that has the form:

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

where *x(t)* is the input signal, *N* is a frame size, *h(n)* is the window function and $\omega_k = 2\pi k/M$, $k = 0, 1, \dots M - 1$ is the frequency index, *L* is the time step between adjacent frames(hop size), and *l* is the index of analysis frame. It is easy to see that (1) is the calculation of the DFT for the signal *h(n)x(n + lL)*. Thus, the representation resulting from the STFT is a sequence of time-localized spectra. Fig. 2 shows an example of a speech signal from the RAVDESS database and Fig. 3 shows the spectrogram (output of the STFT).

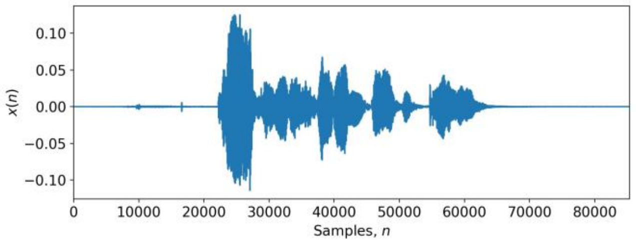


Fig. 2. Representation of the speech signal expressing anger

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

Mel filter bank is a set of triangular filters that have uniform spaced in the mel-frequency scale. These filters are used to convert the power spectrum into the mel-frequency domain.

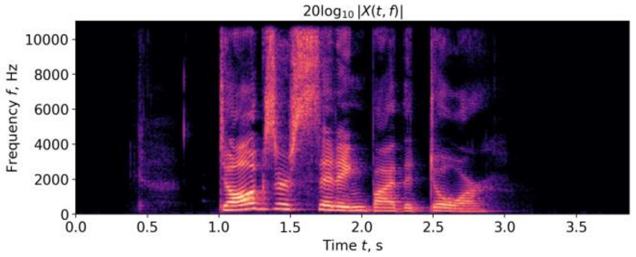


Fig. 3. Spectrogram of a speech signal expressing anger

Mel filter bank is applied to STFT output $X(k,l)^2 \vee$ to obtain mel-scale spectrogram.

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, 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 bank 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. Fig. 4 shows time-sequence of MFCC calculated for signal given in Fig. 2.

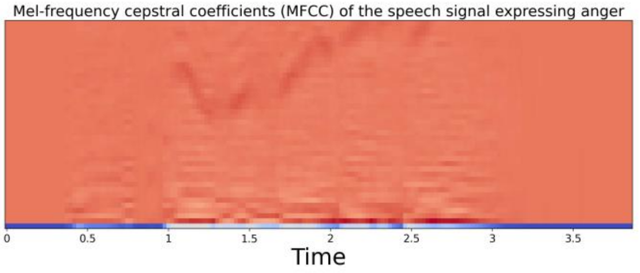


Fig. 4. Time-sequence of MFCC

In this work the speech signals with 48 kHz sampling rateare used. STFT is calculated using the following set of framesizes *N = {1024, 2048, 4096, 8192}*. The hop size *L* is set to *N/2*. From each *N* -sample frame we extract 40 MFCCs usingthe Librosa library in Python. After processing of one audio file we get MFCCs matrix M of size 40 × N_{frames} , where N_{frames} is a number of time frames. To get uniform feature vector for each audio file we calculate mean and std values for MFCCs in matrix M along time axis, thus for each audio file we obtain 80-component vector of suprasegmental MFCC features.

III. Audio Dataset

In this study the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [10] dataset was used. We used only a part of the RAVDESS dataset, 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.

IV. SVM Classifier

The SVM was used to solve the problem of recognizing emotions in speech. Classification using SVM is achieved by constructing a linear (or non-linear) separating surface in the feature space [7]. The idea of this approach is to transform (using the kernel function) the original features into a higher dimensional space. And already in the new transformed feature space to achieve an optimal classification in a certain sense.

Any symmetric, positive (semi-)definite function *K* can be considered as a kernel. This function computes ”scalar product” of the feature vectors x_i and x_j transformed to the higher dimensional space using function ϕ:

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$

K(x_i*,* x_j*)* characterizes the measure of similarity between x_i and x_j . In our research we used the following kernel functions:

- linear kernel:

$$K(x_i, x_j) = \langle x_i^T x_j \rangle$$

which corresponds to the classifier on the support vectors in the original space.

- Polynomial kernel with degree *p*:

$$K(x_i, x_j) = \left(1 + \gamma x_i^T x_j\right)^p$$

- Gaussian kernel with radial basis function (RBF):

$$K(x_i, x_j) = exp\left(\gamma \|x_i - x_j\|^2\right)$$

The parameter γ is hyperparameter that is chosen using grid search procedure. The SVM also has hyperparameter C that controls the the ”budget” of violating the margin boundary. Hyperparameter C also selected using greed search procedure.

V. Evaluation Design

For testing the model performance, the k-fold cross-validation (CV) method was used [7]. The k-fold CV includes the following steps:

1) shuffle the dataset in a random way;

2) divide the dataset into *k* groups;

3) for each unique group do the steps:

a) select a group as test set;

b) take the remaining groups as training set;

c) train the model on training set and evaluate its performance on test set;

d) save score value and reset model to initial state for next iteration;

4) calculate the average score.

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

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

This splitting pattern proposed and explained in [2].

To evaluate the quality of the model, unweighted average recall (UAR) was calculated. UAR is a metric used to measure the overall performance of a multi-class classification model. It calculates the average recall across all classes, giving equal importance to each class without considering the class imbalance. The formula for Unweighted Average Recall (UAR) is given by:

$$UAR = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{A_{ii}}{\sum_{j=1}^{N_c} A_{ij}}$$

where *A* – confusion matrix, N_c – number of classes. The UAR value is in the range from 0 to 1.

The experiment was carried out in three stages:

1) training sample preparation;

2) training and testing of the classifier using a different kernel function and different speech analysis parameters;

3) model evaluation using UAR metric.

VI. Results and discussion

The experiments conducted on the RAVDESS dataset using SVM classifiers with various kernels and hyperparameters, including RBF, linear, and polynomial kernels, along with different frame lengths for MFCC extraction, yielded valu- able insights into emotion recognition. We used grid search technique in order to tune the and find best hyperparameters for a given kernel.

The table I gives a short summary of all the conducted experiments.

TABLE I. The Resulting Uar For SVM Classifier With Different Kernels

Frame size	Linear Kernel	Polynomial Kernel	RBF Kernel
1024	0.435 (C =1.25)	0.434 (C = 0.01, γ= 10, deg= 1)	0.462 (C = 4.33, γ= 7e-3)
2048	0.443 (C =1.05)	**0.442** (C = 0.01, γ= 10, deg= 1)	0.464 (C = 8.22, γ= 15e-4)
4096	0.445 (C =1.05)	0.437 (C = 0.01, γ= 10, deg= 1)	**0.480** (C = 15.2, γ= 4e-3
8192	**0.447** (C =1.05)	0.439 (C = 0.01, γ= 10, deg= 1)	0.469 (C = 15.2, γ= 4e-3)

The best UAR value *48%* is reached using SVM with RBF kernel and suprasegmental MFCC features calculated based on frames with size 4096 samples. UAR surface calculated during the grid search for this model is given in Fig. 5. It can be seen that higher value of *C* parameters results in more flexible classifier with higher performance.

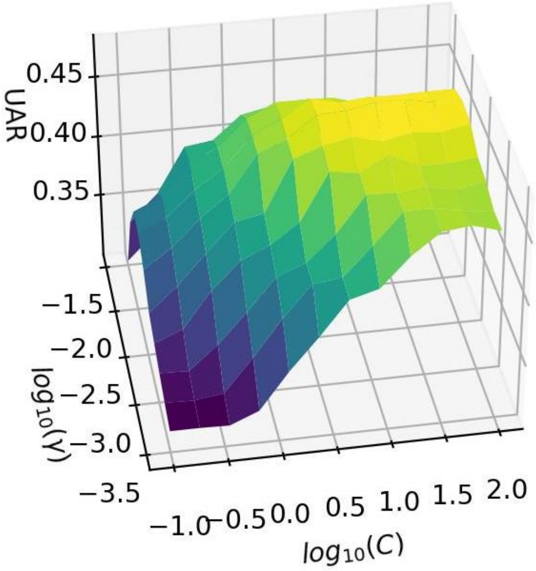
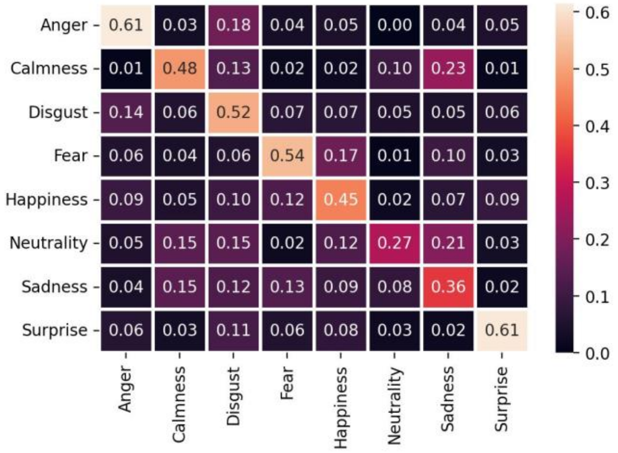


Fig. 5. UAR surface

In Fig. 6 a multiclass confusion matrix is presented for the best SVM-RBF model. The confusion matrix analysis of the RAVDESS dataset using an SVM classifier reveals insightful patterns in emotion recognition. Among the emotions, it was observed that the most frequently misclassified emotion was Neutrality (*27%*). Interestingly, this emotion appeared to be frequently confused with Sadness, suggesting some similarities in their acoustic characteristics. Conversely, Surprise demonstrated a high recognition accuracy (*61%*) and was seldom misclassified as another emotion, indicating distinctive features in its acoustic profile. These findings shed light on the challenges faced by the classifier in distinguishing subtle emotional nuances and underscore the importance of feature engineering and model refinement in improving emotion recognition performance.



	Anger	Calmness	Disgust	Fear	Happiness	Neutrality	Sadness	Surprise
Anger	0.61	0.03	0.18	0.04	0.05	0.00	0.04	0.05
Calmness	0.01	0.48	0.13	0.02	0.02	0.10	0.23	0.01
Disgust	0.14	0.06	0.52	0.07	0.07	0.05	0.05	0.06
Fear	0.06	0.04	0.06	0.54	0.17	0.01	0.10	0.03
Happiness	0.09	0.05	0.10	0.12	0.45	0.02	0.07	0.09
Neutrality	0.05	0.15	0.15	0.02	0.12	0.27	0.21	0.03
Sadness	0.04	0.15	0.12	0.13	0.09	0.08	0.36	0.02
Surprise	0.06	0.03	0.11	0.06	0.08	0.03	0.02	0.61

Fig. 6. Multiclass confusion matrix

Our findings demonstrate that the choice of kernel has asignificant impact on classification accuracy. The RBF kernel exhibited robust performance across multiple emotions, while the linear kernel excelled in distinguishing certain emotionalstates. Notably, the frame size used for MFCCs extractionplayed a significant role in the overall accuracy of the system,with shorter frames providing finer temporal details and longerframes capturing broader contextual information. These results emphasize the importance of fine-tuning the SVM classifier's kernel and considering the trade-offs associated with frame size when designing emotion recognition systems.

VII. Conclusion

In the realm of human-computer interaction, the accurate recognition of emotions from speech is a pivotal factor. This work presented an approach to speech emotion recognition problem based on SVM classifier and MFCC surpasegmental features. The best results (UAR = *48%*) is obtained using SVM-RBF with MFCC features calculated based on 85 ms frames. Comparing to the other works [2]–[4] there is a room for improvement.

References

[1] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," Biomedical Signal Processing and Control, vol. 59, 2020.

[2] C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero, and F. Fernández-Martínez, "Multimodal emotion recognition on RAVDESS dataset using transfer learning," Sensors, vol. 21, no. 22, pp. 1–29, 2021.

[3] S. Sadok, S. Leglaive, and R. Séguier, "A vector quantized masked autoencoder for speech emotion recognition," arXiv preprint arXiv:2304.11117, 2023.

[4] A. Bhavan, P. Chauhan, R. R. Shah et al., "Bagged support vector machines for emotion recognition from speech," Knowledge-Based Systems, vol. 184, pp. 1–7, 2019.

[5] M. Baruah and B. Banerjee, "Speech emotion recognition via generation using an attention-based variational recurrent neural network," Proc. Interspeech 2022, pp. 4710–4714, 2022.

[6] X. Huang, A. Acero, H.-W. Hon, and R. Foreword By-Reddy, Spoken language processing: A guide to theory, algorithm, and system development. Prentice hall PTR, 2001.

[7] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction. Springer, 2009.

[8] 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.

[9] M. M. Goodwin, "The STFT, sinusoidal models, and speech modification," Springer Handbook of Speech Processing, pp. 229–258, 2008.

[10] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," PloS one, vol. 13, no. 5, p. e0196391, 2018.