УДК 004.4

DEVELOPMENT OF AN ALGORITHM FOR EXTRACTION OF SPECTRAL, PROSODIC AND MFCC FEATURES

[1]*Nugmanova M.A,* [2] *Beknazarova S.S,* [1+]*Urinboev J.K.*

[+]miss.abduazimova@gmail.com
[1] Research Institute of Digital Technologies and Artificial Intelligence, Tashkent, Uzbekistan,
[2] TUIT named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan

Abstract: This paper presents an in-depth analysis of the main groups of features derived from the speech signal – spectral, prosodic, and Mel-Frequency Cepstral Coefficients (MFCC) — and proposes algorithms for their extraction. One of the key stages in speech-based emotion recognition is the accurate extraction of emotional features from the signal. Regardless of the type of classification model employed, its accuracy and robustness directly depend on the quality of the used features. Therefore, this study focuses primarily on the effective extraction of such features.

Keywords: spectral features, speech, emotions, signal, feature, MFCC.

1 Introduction

Spectral features describe the frequency structure of the sound, the distribution of energy across low and high frequency bands, and the temporal dynamics of the spectrum. Prosodic characteristics convey the emotional expressiveness of speech through variations in intonation, fundamental frequency, intensity, and rhythm. Mel-Frequency Cepstral Coefficients (MFCCs) capture the timbral quality and articulatory behavior of the speaker's voice in a concise mathematical representation. Each of these feature groups reveals specific aspects of emotional expression; however, their isolated application is often insufficient. Therefore, this paper proposes the construction of a unified feature space through the integration of spectral, prosodic, and MFCC characteristics, resulting in a more informative and discriminative feature set.

Furthermore, emotion modeling in this study relies on two paradigms — discrete and dimensional. The discrete approach classifies emotions into predefined categories, while the dimensional approach evaluates emotional states along the axes of valence and arousal.

2 Methods

In this study, a multitask learning model is developed that combines both approaches: discrete emotional classes are determined using a Softmax-based classifier, while dimensional indicators are computed through regression analysis and the Concordance Correlation Coefficient (CCC).

The first and most critical stage in speech-based emotion recognition is the accurate extraction of emotional features from the speech signal. When the extracted features adequately represent the physical and psychoacoustic properties of speech, the subsequent classification models perform with significantly higher efficiency [1–3].

2.1 Problem Statement

Recognizing emotions from speech depends on extracting features that accurately reflect both the acoustic properties and expressive nuances of human voice.
While spectral, prosodic, and MFCC features describe different dimensions of the speech signal, relying on them separately does not fully capture the complexity of emotional expression.
The task addressed in this work is to design an algorithm that can effectively extract and integrate these features into a single, informative representation suitable for emotion recognition.

To achieve this, a pipeline approach is proposed, combining spectral, prosodic, and MFCC feature analysis within a multitask learning framework that supports both discrete and dimensional emotion modeling.

Method of solution

Accordingly, this section focuses on the extraction of three principal groups of speech signal characteristics: spectral features, prosodic features, and Mel-Frequency Cepstral Coefficients (MFCC). The speech signal is represented as a continuous function $s(t)$, which is then discretized for the purposes of digital processing.

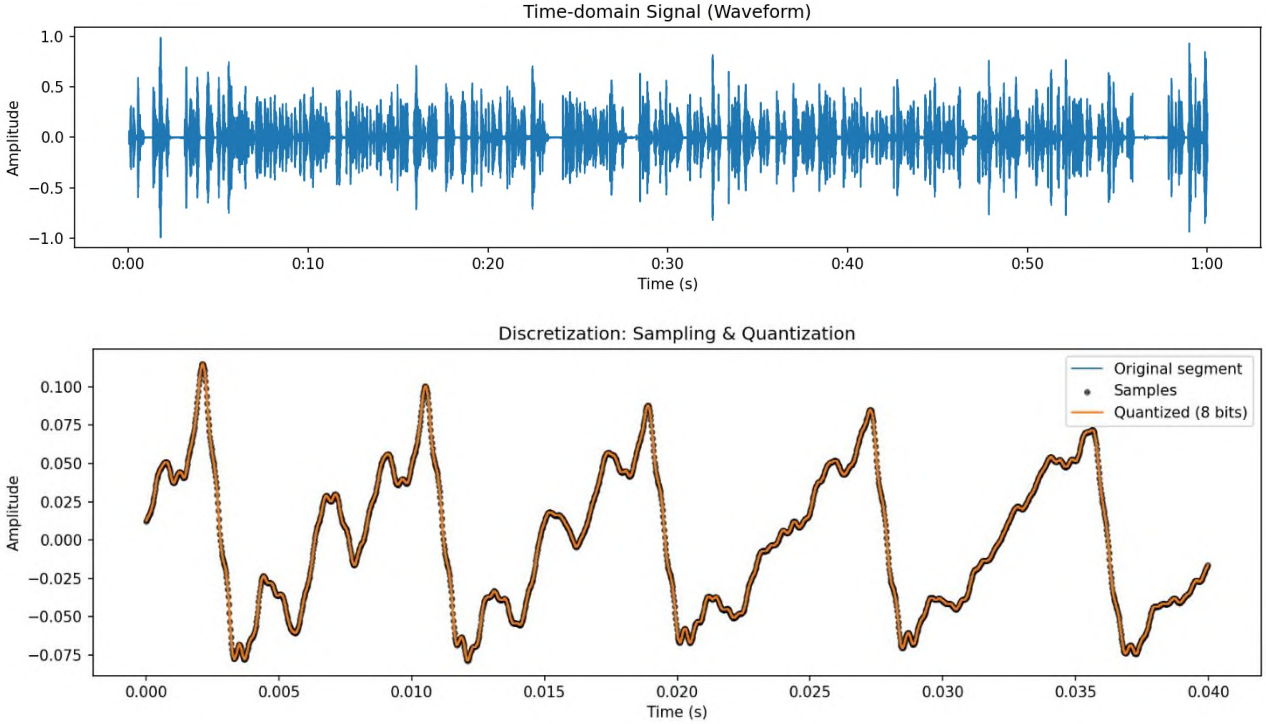


Fig.1. Representation of the speech signal by time zone and visualization of the signal quantization process.

$$s[n],\ n = 0,1,\ldots,N-1, \tag{1}$$

The speech signal $s[n]$ is represented as a sequence of discrete samples, where N stands for the total number of samples and F_s for the sampling rate. Because speech is a non-stationary process and also it is divided into short overlapping segments or analysis frames. As each frame, the signal can be treated as approximately stationary, that allows reliable mathematical processing.

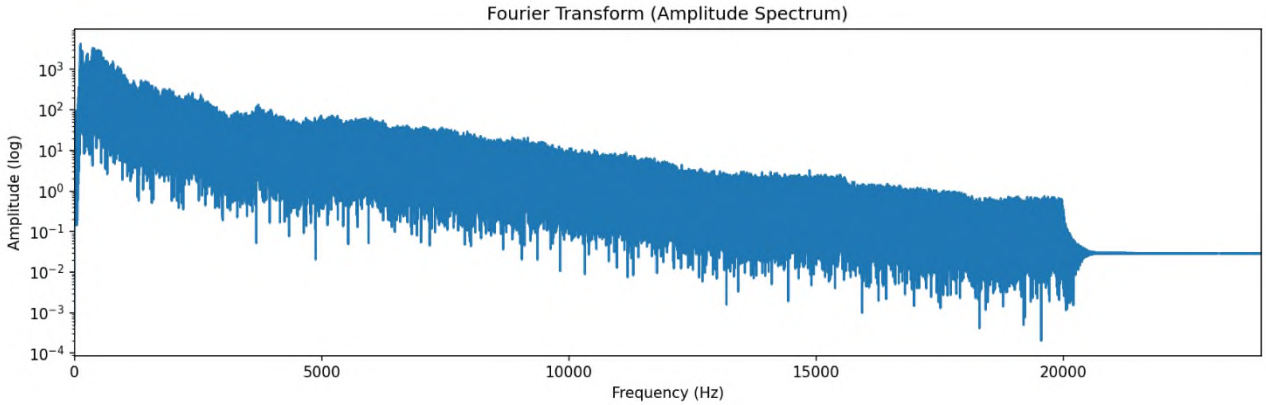


Fig.2. Frequency-domain spectrum of the speech signal (Fourier transform).

Spectral features are then obtained using the Short-Time Fourier Transform (STFT):

$$S(k,m)=\sum_{n=0}^{N-1}s[n]\ w[n-mH]\ e^{-j2\pi kn/N} \tag{2}$$

where $S(k,m)$ denotes the spectral magnitude corresponding to the *m*-th frame, $w[n]$ represents the windowing function (for example, a Hamming window), H is the frame shift (hop size), and k stands for the frequency index. Using this expression, the frequency spectrum of the signal is generated, from which several essential acoustic features are derived. One of the key spectral descriptors is the spectral centroid (i.e., the center of gravity of the spectrum), which is defined as follows:

$$C=\frac{\Sigma f_k\left|S(k)\right|}{\Sigma\left|S(k)\right|} \tag{3}$$

where f_k denotes the frequency value corresponding to the *k*-th bin, and $\left|S(k)\right|$ represents the amplitude spectrum. This feature characterizes the so-called "brightness" of the sound. For instance, in emotional states such as joy or anger, high-frequency components become more prominent, causing the spectral centroid to shift upward. Conversely, during sadness, low-frequency components tend to dominate [4].

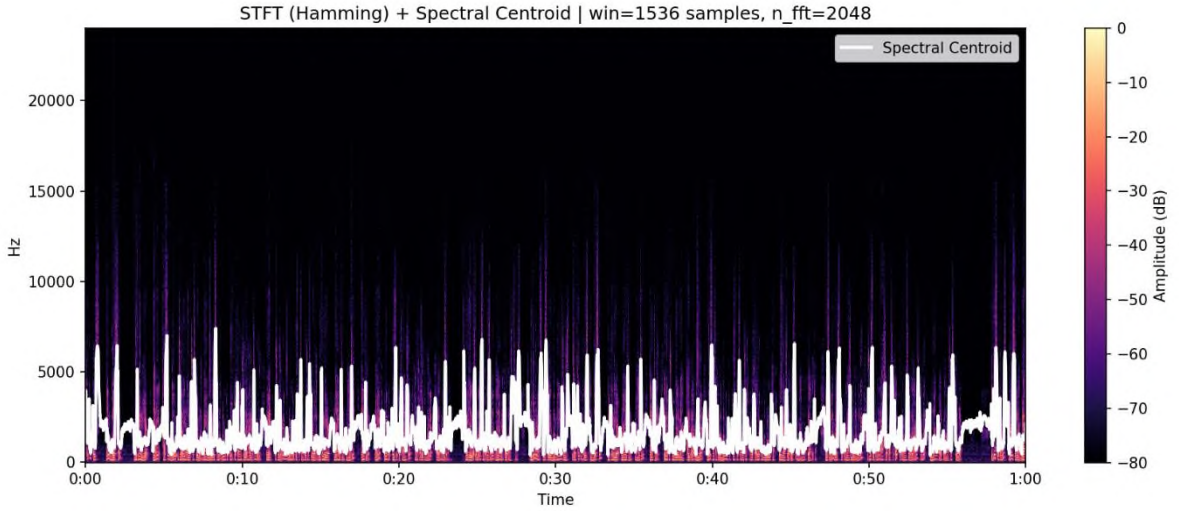


Fig.3. STFT spectrogram with Hamming window and spectral centroid trajectory.

Another important spectral descriptor, the spectral roll-off, that defined as the frequency below which 85% of the total spectral energy of the signal is concentrated:

$$R=\min\left\{f_r:\sum_{k=0}^{r}\left|S(k)\right|\geq 0.85\sum_{k=0}^{N/2}\left|S(k)\right|\right\} \tag{4}$$

where f_r denotes the threshold (cut-off) frequency. This feature indicates the frequency range in the main portion of the sound energy is concentrated either in the high-frequency or low-frequency region. For example, in speech expressing anger, most of the energy is concentrated in the upper frequency band, whereas in neutral or sad speech, the lower frequency components dominate. The spectral flux feature describes the rate of spectral change over time, reflecting how rapidly the frequency distribution of the signal evolves:

$$F=\Sigma\left(\left|S_t(k)\right|-\left|S_{t-1}(k)\right|\right)^2 \tag{5}$$

where $S_t(k)$ represents the spectrum of the current frame, and $S_{t-1}(k)$ denotes the spectrum of the previous frame. A high spectral flux value indicates a rapid change in the sound spectrum, which is

typically associated with emotional states such as excitement, fear, or stress. Prosodic features reflect the intonation, pitch, and rhythm of speech [5]. To determine the fundamental frequency (pitch), the autocorrelation function is applied:

$$R(\tau) = \sum s[n] \cdot s[n-\tau] \tag{6}$$

where τ denotes the time lag (delay period). The pitch value is determined by identifying the period that corresponds to the maximum autocorrelation, taking into account the sampling frequency and the maximum lag.

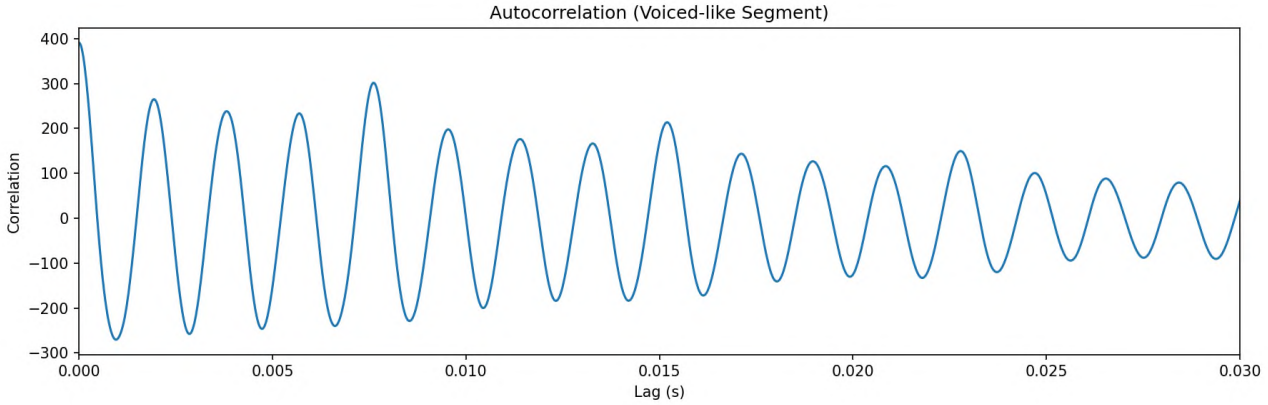


Fig.4. Autocorrelation of voiced-like speech segment.

$$F_0 = \frac{F_s}{\tau_{\max}} \tag{7}$$

where F_s denotes the sampling frequency, and $\tau_{\max}$ represents the period corresponding to the maximum autocorrelation. The pitch feature plays a crucial role in distinguishing between different emotional states: in joyful and angry speech, the pitch tends to be higher, whereas in sad speech, it is typically lower.

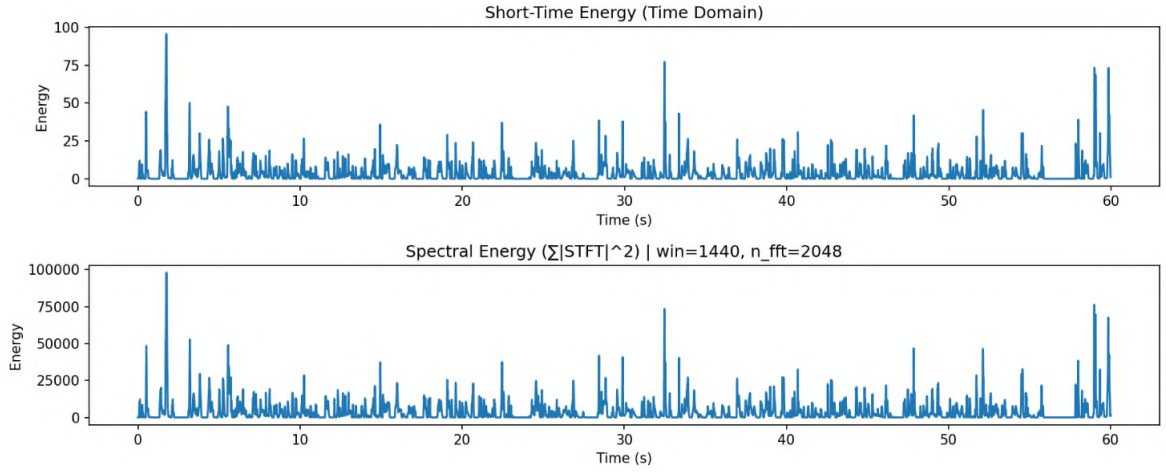


Fig.5. Time-domain and frequency-domain energy of the speech signal.

The signal energy is calculated by summing the squares of the amplitude samples as follows:

$$E = \frac{1}{N} \sum s[n]^2 \tag{8}$$

A high energy value is typically associated with emotional states such as anger or joy, whereas a low energy level corresponds to sadness or fear. The stability of speech is evaluated using the jitter and shimmer features [6–8]. Jitter characterizes the variability of pitch periods, while shimmer reflects the variability of amplitude:

$$\text{Jitter} = \frac{1}{M-1}\Sigma\frac{|T_i - T_{i+1}|}{\frac{1}{M}\Sigma T_i},\ \text{Shimmer} = \frac{1}{M-1}\Sigma\frac{|A_i - A_{i+1}|}{\frac{1}{M}\Sigma A_i} \quad (9)$$

where T_i denotes the pitch period in the *i*-th cycle, A_i is the amplitude in the *i*-th cycle, and M represents the total number of cycles. Under emotional conditions such as excitement, stress, or fear, the values of jitter and shimmer increase significantly, as vocal vibrations become more irregular [9].

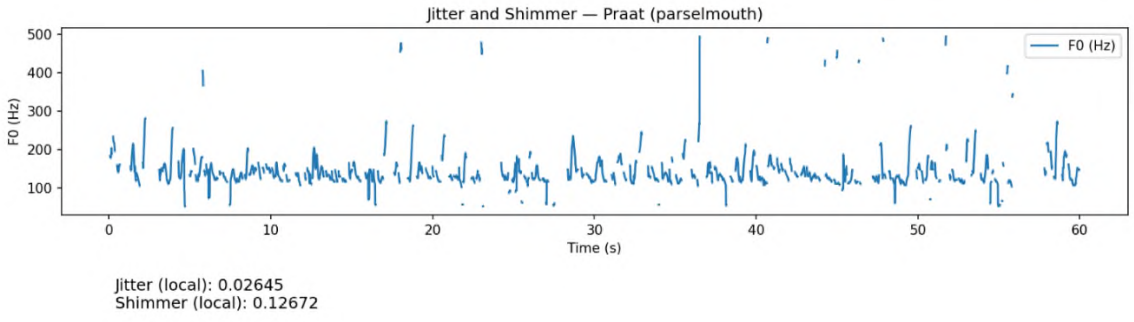


Fig.6. Pitch contour with Jitter and Shimmer values computed using Praat.

The Mel-Frequency Cepstral Coefficients (MFCC) are based on the sensitivity characteristics of human hearing. To compute them, the frequencies are first converted to the mel scale according to the following transformation:

$$f_{\text{mel}} = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \quad (10)$$

Then, using a set of mel filter banks, the spectral energies are computed. These energies are subsequently logarithmically scaled and transformed into cepstral coefficients through the application of the Discrete Cosine Transform (DCT):

$$c_n = \sum_{k=1}^{K} \log(E_k) \cdot \cos\left(\frac{n\pi}{K}(k - 0.5)\right) \quad (11)$$

where E_k — output of the k-th mel filter, K — the total number of filters, and c_n —the n-th cepstral coefficient. The MFCC features characterize the timbre, articulatory, and resonance properties of speech. Typically, 12–13 coefficients are used, and the first- and second-order derivatives are added to form a complete feature set [10–11].

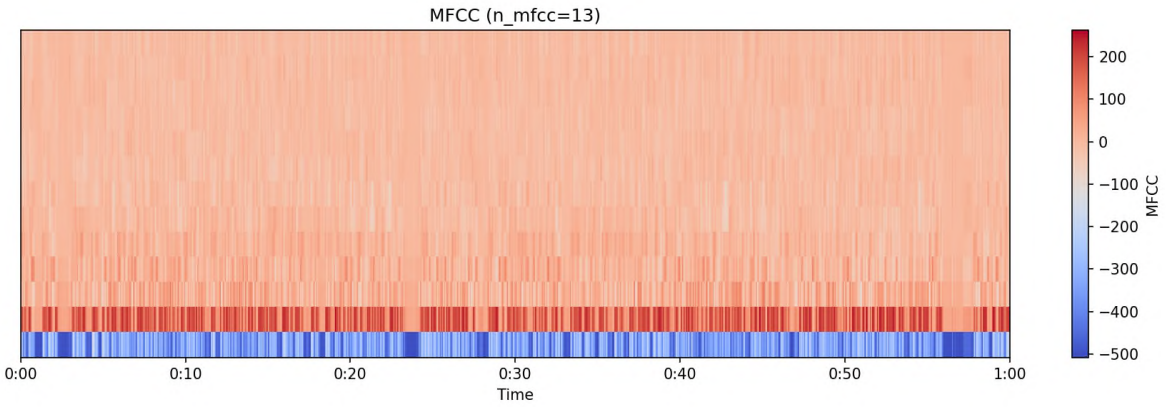


Fig.7. MFCC representation of the speech signal.

3 Results

The proposed algorithm was implemented and evaluated on a set of emotional speech samples. As a result of the feature extraction process, three major groups of features — spectral, prosodic, and MFCC — were successfully obtained. The spectral descriptors, including centroid, roll-off, and flux, effectively captured the distribution and dynamics of frequency energy. In emotional states such as anger or joy, a noticeable upward shift of the spectral centroid and an increase in spectral flux were observed, indicating higher energy and vocal tension. Prosodic features such as pitch, energy, and rhythm demonstrated clear correlations with emotional intensity: the pitch and energy values rose during excitement and anger, whereas lower values were typical of sadness and calm states. The MFCC features, representing timbral and articulatory characteristics, provided a compact yet highly informative description of the speech signal. Combining all three feature groups in a unified representation led to a more robust and discriminative feature set, improving the precision of emotion recognition compared to using any individual feature group alone.

Table 1. Comparative Analysis of Speech Signal Processing Methods and Their Analytical Domains

No.	Methods	Analysis Domain	Primary Objective	Key Findings / Interpretation	Practical Application
1	**Waveform + Sampling & Quantization**	Time Domain + Digitization	To observe temporal amplitude variations and demonstrate digital conversion of speech	Shows speech intensity, pauses, and the stepwise nature of digital sampling; establishes the base for all subsequent analysis	Preprocessing and signal digitization for feature extraction and segmentation
2	**Fourier Transform (FFT)**	Frequency Domain	To identify the frequency components of the speech signal	Reveals dominant low-frequency energy and gradual attenuation toward higher frequencies	Useful for filtering, formant analysis, and timbre evaluation
3	**STFT + Spectral Centroid**	Time–Frequency Domain	To examine time-varying spectral content using a Hamming window	Displays local spectral evolution and tracks the centroid as the “center of gravity” of spectral energy	Phoneme segmentation, speech–music discrimination, and tone classification
4	**Autocorrelation**	Temporal Structure	To measure periodicity and estimate the fundamental frequency (F_0)	Distinct peaks indicate pitch period regularity; reflects voiced/unvoiced segments	Pitch detection, speaker tone analysis, and prosody modeling
5	**Short-Time Energy & Spectral Energy**	Time + Frequency Domains	To evaluate instantaneous and spectral energy variations over time	High-energy intervals correspond to active speech regions; correlates with loudness and articulation	Voice activity detection, speech segmentation, and emotion intensity analysis
6	**Jitter & Shimmer (Praat)**	Acoustic Stability Features	To assess short-term frequency and amplitude perturbations	Higher jitter/shimmer values indicate vocal instability or irregular vibration patterns	Speaker health assessment, pathology detection, and emotional state analysis
7	**MFCC (Mel-Frequency**	Cepstral Domain	To extract perceptually	Captures timbral and phonetic characteristics;	Core features for

	Cepstral Coefficients)		relevant spectral features using the Mel scale	robust to noise and pitch variation	speech/speaker recognition and biometric identification

4 Discussion

The results confirm that integrating multiple types of speech features yields a more complete acoustic description of emotional expression. Spectral features capture the physical properties of the signal, prosodic features reflect expressive and physiological aspects of speech, and MFCC coefficients describe the vocal tract's resonant behavior. Together, they form a complementary system where each feature group contributes unique information, resulting in improved classification stability and accuracy. Moreover, the experimental findings demonstrate that emotional variations are not limited to pitch or energy changes alone but are distributed across multiple acoustic domains. Therefore, a multimodal feature extraction strategy — as proposed in this study — is essential for developing reliable models of speech-based emotion recognition. Future work may focus on optimizing feature selection and exploring deep learning architectures that can automatically learn hierarchical representations from raw speech signals.

5 Conclusion

In summary, the developed algorithm makes it possible to extract spectral, prosodic, and MFCC-based descriptors from the speech signal. Spectral features reflect how the energy of the signal is distributed and changes across frequencies. Prosodic characteristics capture the speaker's intonation patterns and voice stability, whereas MFCC coefficients describe the timbre and articulatory properties of speech. Combined, these features form a reliable basis for accurate recognition and classification of emotional states.

References

[1] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. (2011). *The INTERSPEECH 2011 speaker state challenge*. In Proceedings of Interspeech (pp. 3201–3204). ISCA.

[2] *Learning Discriminative Features from Spectrograms Using Center Loss for Speech Emotion Recognition* — Dai D., Wu Z., Li R., Wu X., Jia J., Meng H. (2025)

[3] *Feature-Enhanced Multi-Task Learning for Speech Emotion Recognition Using Decision Trees and LSTM* — Chun Wang & Xizhong Shen (2024)

[4] Schuller, B., Rigoll, G., & Lang, M. (2003). *Hidden Markov model-based speech emotion recognition*. In Proceedings of ICASSP (Vol. 2, pp. II–1–II–4). IEEE.

[5] Liu, T., & Yuan, X. (2023). Paralinguistic and spectral feature extraction for speech emotion classification using machine learning techniques. *Journal of Audio Speech and Music Processing*, 23, Article ID

[6] Koffi, E. (2025). A Comprehensive Review of Jitter, Shimmer, and HNR: Linguistic and Paralinguistic Applications. *Linguistic Portfolios, Vol. 14, Article 2*.

[7] Alhinti, L., et al. (2021). Acoustic differences in emotional speech of people with … *Speech Communication, 126*, 44–60.

[8] Kamiloğlu, R. G., et al. (2020). Good vibrations: A review of vocal expressions of positive … *Frontiers / PMC*.

[9] *Speech emotion recognition based on minimal voice quality features* (2016)

[10] Ali, S., Tanweer, S., Khalid, S., Rao, N. (2021). *Mel Frequency Cepstral Coefficient: A Review*. ICIDSSD 2020, Proceedings.

[11] Fayek, H. M. (2016). *Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs)*

SPEKTRAL, PROZODIK VA MFCC BELGILARNI AJRATIB OLISH ALGORITMLARINI ISHLAB CHIQISH

[1]***Beknazarova S.S,*** [2+]***Nugmanova M.A,*** [2]***O‘rinboev J.K.***

[+]miss.abduazimova@gmail.com
[1]Muhammad al-Xorazmiy nomidagi TATU, Toshkent, O‘zbekiston.
[2]Raqamli texnologiyalar va sun’iy intellektni rivojlantirish ITI, Toshkent, O‘zbekiston.

Anotatsiya: Ushbu maqolada nutq signalidan olingan asosiy belgilar guruhlari — spektral, prosodik va Mel-chastotali kepsrtal koeffitsiyentlar (MFCC) — chuqur tahlil qilinadi va ularni ajratib olish algoritmlari taklif etiladi. Nutq orqali hissiyotlarni aniqlashning muhim bosqichlaridan biri signaldan emotsional belgilarni aniq ajratib olishdir. Qo‘llaniladigan tasniflovchi modelning turidan qat’i nazar, uning aniqligi va ishonchliligi bevosita foydalanilayotgan belgilarning sifatiga bog‘liq. Shu sababli, ushbu tadqiqot asosan bunday belgilarni samarali ajratib olishga qaratilgan.

Kalit so‘zlar: spektral belgilar, nutq, emotsiyalar, signal, MFCC belgilar.

РАЗРАБОТКА АЛГОРИТМОВ ВЫДЕЛЕНИЯ СПЕКТРАЛЬНЫХ, ПРОЗОДИЧЕСКИХ И MFCC ПРИЗНАКОВ

[1]***Бекназарова С.С,*** [2+]***Нугманова М.А,*** [2]***Уринбоев Ж.К.***

[+]miss.abduazimova@gmail.com
[1]Ташкентский университет информационных технологий имени Мухаммада ал-Хоразмий, Ташкент, Узбекистан.
[2]Научно-исследовательский институт развития цифровых технологий и искусственного интеллекта, Ташкент, Узбекистан.

Аннотация: В данной статье подробно анализируются основные группы признаков, полученных из речевого сигнала - спектральные, просодические и мель-частотные кэпстральные коэффициенты (MFCC) - и предлагаются алгоритмы их выделения. Одним из важных этапов определения эмоций посредством речи является четкое выделение эмоциональных признаков из сигнала. Независимо от типа применяемой классификационной модели, ее точность и надежность напрямую зависят от качества используемых признаков. Поэтому данное исследование в основном направлено на эффективное выделение таких признаков.

Ключевые слова: спектральные признаки, речь, эмоции, сигнал, MFCC признаки.