УДК 004.93
DOI https://doi.org/10.32782/tnv-tech.2024.4.13

РОЗПІЗНАВАННЯ ЕМОЦІЙ УЧАСНИКІВ ВІДЕОКОНФЕРЕНЦІЙ З ВИКОРИСТАННЯМ МУЛЬТИМОДАЛЬНОГО АНАЛІЗУ

Савчук Т. О. *– кандидат технічних наук, професор кафедри комп'ютерних наук Вінницького національного технічного університету*
ORCID ID: 0000-0002-0061-6206

Пастух І. П. *– аспірант кафедри комп'ютерних наук Вінницького національного технічного університету*
ORCID ID: 0000-0003-1080-1736

У роботі описано актуальність задачі аналізу емоцій у відеокомунікаціях, підкреслено важливість розуміння настроїв учасників для поліпшення якості взаємодії у сучасному світі, де відеозустрічі стають нормою у бізнесі, освіті та особистих контактах. Ефективне розуміння емоцій сприяє адаптації комунікації, вирішенню конфліктів на ранніх стадіях та покращенню загального сприйняття взаємодії. Незважаючи на наявність потужних інструментів для розпізнавання емоцій, як FaceReader та Microsoft Oxford Project, їхня ефективність обмежена через фокусування виключно на аналізі виразів облич. Точність таких систем часто поступається через недоліки в розпізнаванні емоцій, що вимагає удосконалення методів аналізу. Робота пропонує новітній підхід до розпізнавання емоцій учасників відеоконференцій через мультимодальний аналіз, що поєднує обробку фізичних характеристик голосу та виразів облич. Використання згорткових нейронних мереж дозволяє з високою точністю ідентифікувати емоційні стани, враховуючи різноманітні спотворення вхідних даних. Методика передбачає аналіз голосових даних, їх нормалізацію та перетворення у спектрограми для подальшої обробки нейронною мережею. Особлива увага приділяється процесу навчання мережі, що базується на методі градієнтного спуску, для підвищення точності розпізнавання емоцій. Результати експериментів демонструють перевагу запропонованого методу над існуючими програмними засобами, з підвищенням точності розпізнавання емоцій до 79%, що є значним поліпшення. Запропонований мультимодальний аналіз, що включає в себе комплексний підхід до аналізу звукових та візуальних характеристик, відкриває нові можливості для розвитку інструментів відеокомунікації та покращення міжособистісного спілкування. Висновки роботи підкреслюють значення інтегрованого підходу до розпізнавання емоцій, наголошуючи на потенціалі застосування згорткових нейронних мереж для ефективної обробки емоційних станів у реальному часі, що є ключовим для розширення можливостей відеокомунікацій.

Ключові слова: *розпізнавання емоцій, згорткові нейронні мережі, математична модель, відеоконференції, спектрограми, аналіз голосу.*

Savchuk T. O., Pastukh I. P. Participants emotion recognition in video conferences using multimodal analysis

The paper describes the relevance of the task of analyzing emotions in video communications, emphasizing the importance of understanding the moods of participants to improve the quality of interaction in the modern world, where video meetings are becoming the norm in business, education, and personal contacts. Effective understanding of emotions helps to adapt communication, resolve conflicts at early stages, and improve the overall perception of interaction. Despite the availability of powerful emotion recognition software such as FaceReader and Microsoft Oxford Project, their effectiveness is limited due to their focus on analyzing facial expressions alone. The accuracy of such systems is often inferior due to shortcomings in emotion recognition, which requires improved analysis methods. This paper proposes a novel approach to recognizing emotions of video conference participants through multimodal analysis that combines the processing of physical voice characteristics and facial expressions. The use of convolutional neural networks allows for high accuracy in identifying emotional states, taking into account various distortions of the input data. The technique involves analyzing voice data, normalizing it, and converting it into spectrograms for further processing by a neural network. Special attention is paid to the process of training the network based on the gradient descent

method to improve the accuracy of emotion recognition. Experimental results demonstrate the advantage of the proposed method over existing software tools, with an increase in emotion recognition accuracy of up to 79%, which is a significant improvement.

Key words: *emotion recognition, convolutional neural networks, mathematical model, video conferencing, spectrograms, voice analysis.*

Вступ. Задача аналізу емоцій учасників є актуальною у сучасному світі, де відеокомунікація стала стандартом для бізнесу, освіти та особистих взаємодій. Розуміння емоцій учасників допомагає покращити якість відеокомунікації [1]. Вивчення настроїв взаємодії людини з людиною може допомогти машинам ідентифікувати та реагувати на невербальне спілкування людини. Аналіз емоцій учасників відеоконференцій дає можливість ідентифікувати настрої людей, адаптуватись під них, та впливати, що у свою чергу позитивно відображається на загальній картині сприйняття відеоконференції за рахунок вирішення конфліктів та інших негативних ситуацій на початкових етапах їх утворення [2].

На даний момент існує велика кількість інструментів для розпізнавання емоцій, найпопулярнішими з яких є FaceReader та Microsoft Oxford Project Emotion Recognition. Зазначені інструменти мають потужний функціонал і високу точність розпізнавання у багатьох випадках [3, 4]. Проте, існуючі додатки мають не високу точність розпізнавання через те, що проводять розпізнавання лише по виразу обличчя. Отож, доцільним є підвищення точності при розпізнаванні емоцій учасників відеоконференцій, що дозволить отримувати кращі результати роботи.

Математична модель та удосконалений метод розпізнавання емоцій учасників відеоконференцій з використанням мультимодального аналізу. Математичну модель процесу аналізу емоцій учасників відеоконференцій можна подати у такому вигляді.

$$f(b) = (p_1 + q_1, p_2 + q_2, \ldots, p_n + q_n)), \quad (1)$$

де $f(b)$ – найбільш ймовірна емоція на аналізуємому часовому інтервалі; b – блок звукового сигналу учасника відеоконференції довжиною в 1 секунду; $p_1, p_2, \ldots, p_n$ – ймовірність належності звукового сигналу до i-го класу емоції; $q_1, q_2, \ldots, q_n$ – ймовірність належності зображення обличчя до i-го класу емоції; e – функція, що повертає i-у емоцію яка має найбільшу ймовірність; n – кількість класів емоцій.

Для вирішення задачі розпізнавання емоцій за фізичними характеристиками голосу введемо блочний аналіз, що передбачає розбиття звукового сигналу спікера на блоки визначеної довжини, наприклад, 1 секунда, що дасть можливість аналізу коротких відрізків голосу учасників відеоконференції. Також проведемо нормалізацію значень амплітуди звуку в діапазоні від -1 до 1 для підвищення точності обробки сигналу. Загальний приклад розбиття голосового сигналу на блоки зображено на рисунку 1.

Для кожного блоку застосуємо алгоритм дискретного перетворення Фур'є, яке дає можливість визначити частотну складову дискретного сигналу [5].

Дискретне перетворення Фур'є голосового відбитку можна подати у матричному вигляді, яке у свою чергу може бути використане для утворення спектрограм вхідних голосових сигналів учасників відеоконференції короткої довжини для подальшого аналізу їх на емоційну забарвленість [5]. Отже, процес перетворення звукового сигналу учасників відеоконференції у спектрограму включає кроки, що представлені на рисунку 2.

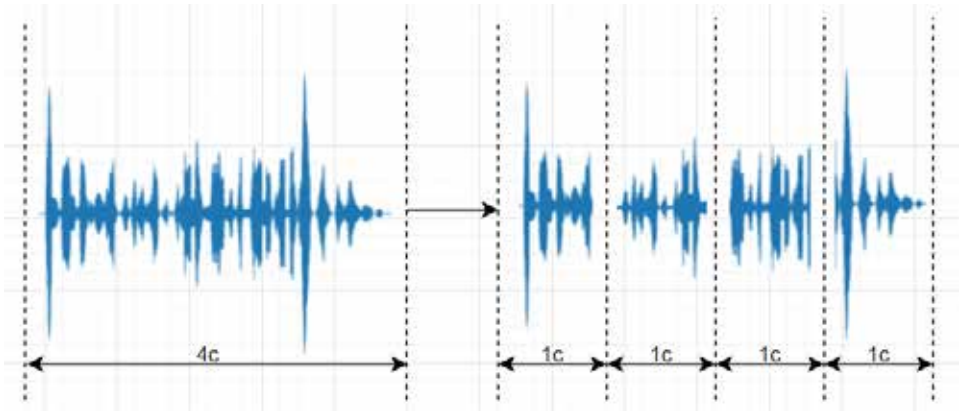

Рис. 1. Розбиття голосового сигналу на рівні блоки

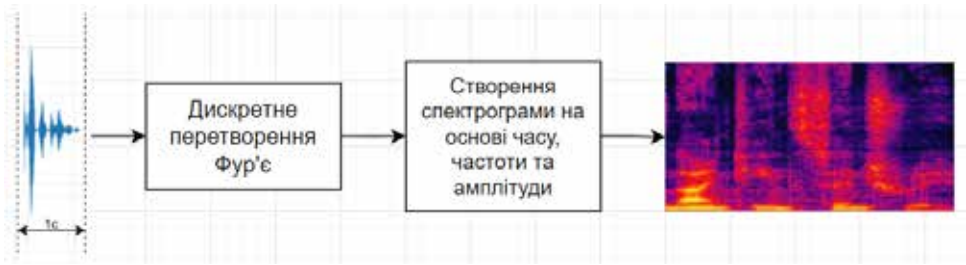

Рис. 2. Процес перетворення звукового сигналу учасників відеоконференції у спектрограму

Для вирішення задачі розпізнавання емоцій учасників відеоконференції на основі спектрограм їх голосу та зображення обличчя застосуємо метод згорткових нейронних мереж, який являє собою особливий клас багатошарового персептрона з двовимірною структурою і добре підходить для обробки зображень з високим ступенем інваріантності до зміщення, поворотів, масштабування та інших спотворень вхідних даних.

Структура згорткової нейронної мережі для задачі розпізнавання емоції людини за фізичними характеристиками голосу та виразом обличчя являє собою послідовність із двох типів шарів: згорткові, які виконують операцію згортки, та агрегувальні, які зменшують розмірність карт ознак обличчя та звукових характеристик за рахунок їх агрегування. Кожен шар складається з набору площин, які у свою чергу складаються з нелінійних нейронів [6].

Функціонування нелінійної моделі нейрону l можна описати такими рівняннями [4]:

$$u_l = \sum_{j=0}^{d} w_{lj} s_j, \tag{2}$$

де $s_1, s_2, \ldots, s_d$ – вхідні сигнали, ознаки або характеристики зображення спектрограми та обличчя; $w_1, w_2, \ldots, w_d$ – синаптичні ваги нелінійного нейрону l; u_l – лінійна комбінація ознак спектрограми та обличчя; d – кількість нейронів.

Кожен нейрон згорткового шару має зв'язок із невеликою групою нейронів попереднього шару (локальне рецептивне поле). Основною задачею згорткового шару є виявлення і витягнення важливих ознак з зображень спектрограм та облич, що сприяє покращенню точності класифікації емоцій. Фільтр згорткового шару рухається по вхідному зображенню і виконує поелементне множення з послідуючим сумуванням результатів, що визначається формулою [4]:

$$G[M,N]=\sum_{w}\sum_{z}h[w,z]g[M-w][N-z], \tag{3}$$

де $G[m,n]$ – вихідна матриця ознак спектрограм та облич; g – вхідне зображення спектрограми та обличчя; h – фільтр; M,N – розмір вхідного зображення; w,z – розмір фільтра.

Після згорткового шару розташовується агрегувальний шар, який забезпечує часткову інваріантність нейронної мережі до зміни масштабу вхідного зображення, що покращує точність розпізнавання емоцій. Основна мета агрегувального шару полягає в зменшенні розміру виходу попереднього згорткового шару та витягненні найважливіших ознак спектрограм та облич, зберігаючи при цьому інформацію про певні шаблони та характеристики зображення. Функціонування агрегувального шару визначається такою формулою:

$$a(v)_{i,j,o}=\left(v_{is_x+q,js_y+r,o}\right), \tag{4}$$

де v – вхідне зображення спектрограми чи обличчя після операції згортки; i,j – індекси виходу; o – індекс каналу; q,r – розмір фільтру ознак спектрограми та обличчя; s_x,s_y – значення кроку в горизонтальному та вертикальному напрямках.

Таким чином, агрегувальний шар зменшує розмірність площин попереднього шару вдвічі, що зменшує обчислювальну складність виявлення ознак обличчя та характеристик звуку. Послідовно чергуючись один за одним, розміри площин зменшуються, але їхня кількість збільшується. Чергування шарів дозволяє формувати різні карти характеристик емоцій, що наділяє згорткові нейронні мережі здатністю до ідентифікації складніших ієрархічних ознак спектрограм та облич. Поступово під час проходження кількох шарів карта ознак емоцій вироджується у вектор. Таким чином, площини згорткових нейроних мереж є фільтрами, кожен з яких здійснює пошук індивідуальних характерних ознак вхідного зображення, які збільшують точність класифікації емоції. Це дозволяє згорткової нейронної мережі запам'ятовувати взаємозв'язок просторово-залежних областей зображення спектрограм та облич. Характерні ознаки емоцій, які отримують тій чи іншій площиною, визначаються у процесі навчання. Процес навчання згорткової нейронної мережі відбувається за допомогою методу градієнтного спуску [6].

Задачею навчання згорткової нейронної мережі є мінімізація значення середньоквадратичної похибки, що дозволяє моделі розпізнавати емоції на зображені з високою точністю.

Перевагами застосування згорткових нейронних мереж для вирішення задачі розпізнавання емоцій за фізичними характеристиками голосу та виразом облич є:

1. Здатність до роботи зі зображеннями: згорткові нейронні мережі оптимально підходять для обробки зображень завдяки своїм згортковим та пулінговим шарам. Вони можуть виявити шаблони, текстури та емоції на зображеннях.

2. Здатність до виділення важливих ознак: моделі вчаться виділяти важливі ознаки на зображеннях, такі як контури, кутові точки, текстури, форми тощо, що корисно для виділення емоцій.

3. Висока точність: згорткові нейронні мережі показують високу точність у завданні розпізнавання емоцій на зображеннях спектрограм та облич, за рахунок використання глибокого навчання та згорткових шарів.

4. Інваріантність до зсувів та масштабування: згорткові шари виявляють шаблони емоцій незалежно від їх положення на зображенні, що робить модель інваріантною до невеликих зсувів та змін масштабу.

Наведений математичний підхід може бути реалізовний удосконаленим методом розпізнавання емоцій учасників відеоконференцій, що включатиме об'єднання етапів розпізнавання емоцій учасників за фізичними характеристиками голосу та розпізнавання емоцій учасників за виразом обличчя.

Загальний алгоритм розпізнавання емоцій учасників відеоконференцій за фізичними характеристиками голосу зображений на рисунку 3, буде містити такі кроки:

1. Отримання медіа даних голосу учасника відеоконференції.
2. Розбиття звукового сигналу на блоки рівної довжини.
3. Нормалізація значень амплітуди звуку в діапазоні від -1 до 1.
4. Застосування алгоритму дискретного перетворення Фур'є до звукових блоків.
5. Створення спектрограм звукових блоків.
6. Обробка спектрограм згортковою нейронною мережею.
7. Вибір найбільш ймовірної емоції.

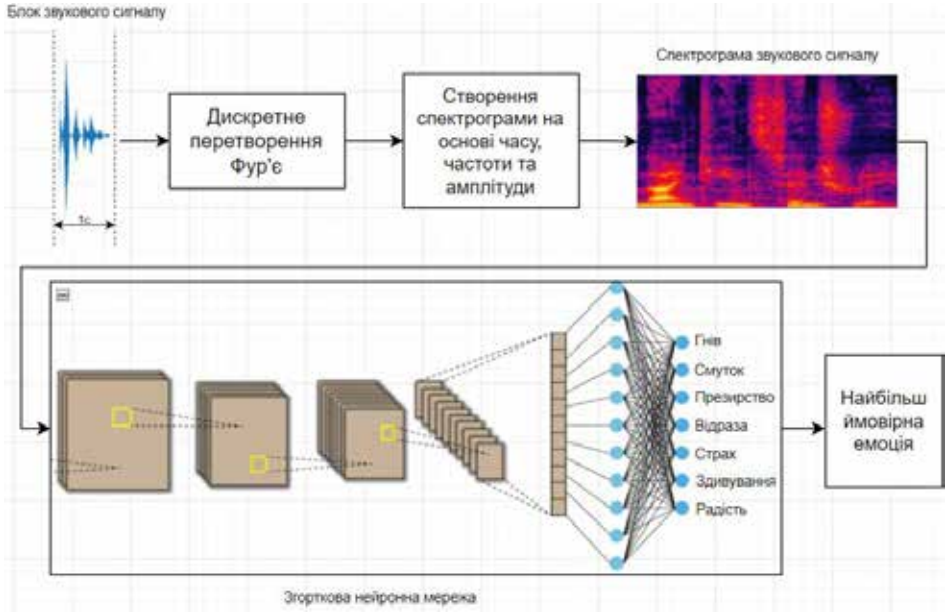

Рис. 3. Основні кроки алгоритму розпізнавання емоцій учасників відеоконференції за фізичними характеристиками голосу

Загальний алгоритм розпізнавання емоцій учасників відеоконференції за виразом обличчя зображений на рисунку 4, буде містити такі кроки:

1. Отримання зображення учасника відеоконференції.
2. Нормалізації зображення.
3. Розпізнавання обличчя
4. Виділення обличчя з зображення.
5. Обробка зображеня обличчя згортковою нейронною мережею.
6. Вибір найбільш ймовірної емоції.

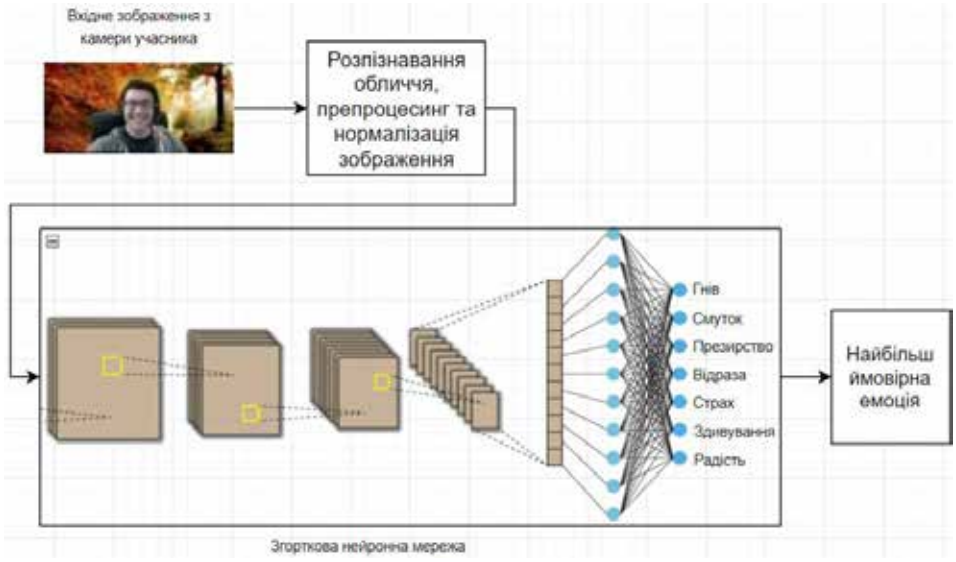

Рис. 4. Основні кроки алгоритму розпізнавання емоцій учасників відеоконференції за виразом обличчя

З метою підвищення точності розпізнавання емоцій було об'єднано алгоритм розпізнавання емоцій за фізичними характеристиками голосу та алгоритм розпізнавання емоцій на основі виразу обличчя в удосконалений метод розпізнавання емоцій учасників конференції. Такий мультимодальний аналіз з урахуванням описаних алгоритмів представлено на рисунку 5.

Результати досліджень. Було проведено аналіз результатів роботи розпізнавання емоцій з використанням сучасних програмних засобів, що базуються на відповідних методах, та з використанням мультимодального аналізу. Для доведення факту досягнення поставленої у роботи мети, а саме – підвищення точності розпізнавання – було проведено 100 експериментів, кожен з яких передбачав різну кількість учасників відеоконференції, різні комбінації ввімкнення та вимкнення камер, різні голосові та візуальні емоції. Середні значення отриманих результатів наведено у таблиці 1.

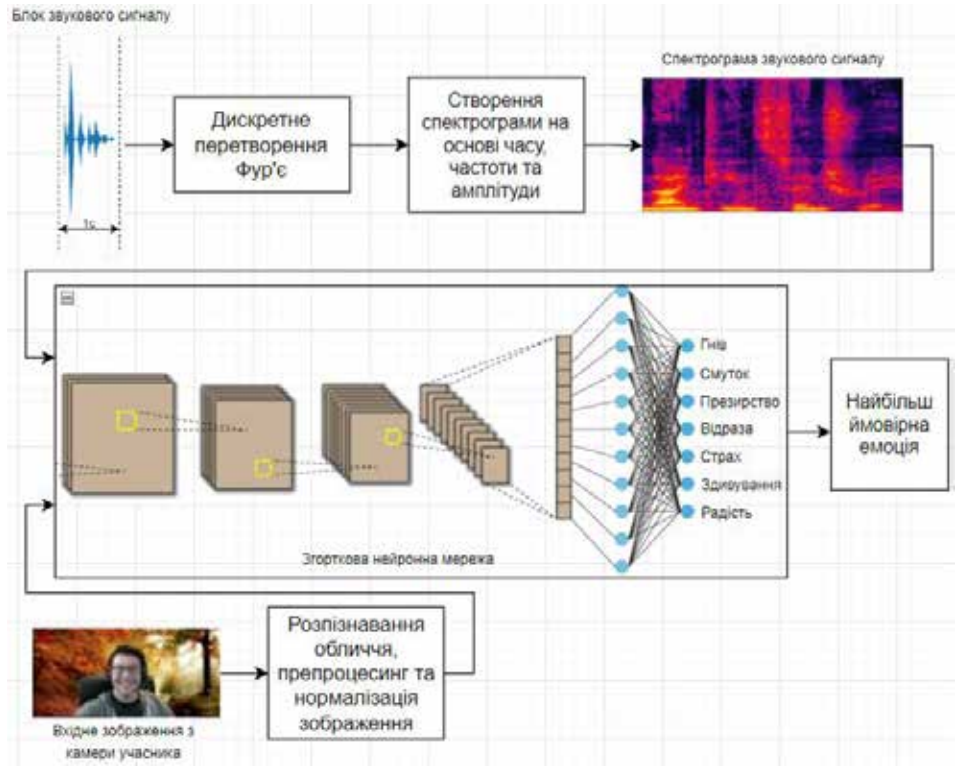

Рис. 5. Удосконалений метод розпізнавання емоцій учасників відеоконференції з використанням мультимодального аналізу

Таблиця 1

Результати аналізу розпізнавання емоцій з використанням сучасних програмних засобів та з використанням мультимодального аналізу

Програмний засіб	Відсоток розпізнаних емоцій	Точність розпізнавання емоцій
FaceReader	62%	74%
MOPER	62%	71%
Розроблена технологія	100%	79%

Із таблиці 1 видно, що відсоток розпізнаних емоцій у програм-аналогів є нижчою на 38% порівняно із відсотком розпізнаних емоцій з використанням запропонованої технології, що базується на мультимодальному аналізі фізичних характеристик голосу учасника та виразу обличчя. При цьому, точність розпізнавання емоцій складає 79%, що на 5-8% вище за результати роботи програм-аналогів. Це відкриває нові можливості для розвитку інструментів відеокомунікації та покращення міжособистісного спілкування.

Висновки. Отже, у роботі була запропонована математична модель та удосконалений метод розпізнавання емоцій учасників відеоконференцій, що базується

на мультимодальному аналізі, який включає в себе розпізнавання за допомогою звукових характеристик голосу та виразу обличчя. Запропонована технологія передбачає використання згорткових нейронних мереж, що дає можливість розпізнавати емоції з високою точністю. Було розглянуто математичну модель розпізнавання емоцій за фізичними характеристиками голосу та за виразом обличчя. Розглянуто процес розбиття звукового сигналу на блоки, перетворення звукових блоків у спектрограми, функціонування згорткових нейронних мереж, згорткові та агрегувальні шари, процес навчання мережі для збільшення точності виявлення емоцій на зображеннях. Наведено математичну модель роботи нейронних мереж, шару згортки та агрегувального шару, що дозволить виявляти емоції на зображеннях спектрограм з високою точністю за рахунок використання глибокого навчання та згорткових шарів, які автоматично виявляють і витягують важливі ознаки з зображень спектрограм та агрегувальних шарів, які зменшують розмірність зображення при збереженні найважливіших ознак та шаблонів обличчя. Мету дослідження було досягнуто за рахунок використання мультимодального аналізу, який включає в себе розпізнавання за допомогою звукових характеристик голосу та виразу обличчя, і використання згорткових нейронних мереж. Удосконалений метод розпізнавання емоцій учасників конференції показав 79% правильно розпізнаних емоцій, що на 5-8% вище за результати роботи програм-аналогів. Наведений метод дозволяє точніше ідентифікувати настрої людей, що дає можливість адаптуватись під них та впливати, що у свою чергу позитивно відображається на загальній картині сприйняття відеоконференції за рахунок вирішення конфліктів та інших негативних ситуацій на початкових етапах їх утворення.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ:

1. Ekman, P. Basic emotions. Handbook of cognition and emotion, 1999. 45-60.
2. Fredrickson, B. L. The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. American psychologist, 2001. 56(3), 218-226.
3. Facereader. URL: https://www.noldus.com/facereader
4. Happy? Sad? Angry? This Microsoft tool recognizes emotions in pictures. URL: https://blogs.microsoft.com/ai/happy-sad-angry-this-microsoft-tool-recognizes-emotions-in-pictures/
5. Understanding Audio data, Fourier Transform, FFT and Spectrogram features for a Speech Recognition System. URL: https://towardsdatascience.com/understanding-audio-data-fourier-transform-fft-spectrogram-and-speech-recognition-a4072d228520
6. Савчук Т. О., Пастух І. П. Розпізнавання емоцій учасників відеоконференцій в Microsoft Teams. Таврійський науковий вісник. Серія: Технічні науки. Херсон: Видавничий дім «Гельветика», 2023. Вип. 6. С. 18-24. https://doi.org/10.32851/tnv-tech.2022.6.3

REFERENCES:

1. Ekman, P. (1999). Basic emotions. Handbook of cognition and emotion, 45-60.
2. Fredrickson, B. L. (2001). The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. American psychologist, 56(3), 218-226.
3. Noldus (electronic journal) Facereader. Retrieved from: https://www.noldus.com/facereader (accessed 29 August 2022).
4. Microsoft (electronic journal) Happy? Sad? Angry? This Microsoft tool recognizes emotions in pictures. Retrieved from: https://blogs.microsoft.com/ai/happy-sad-angry-this-microsoft-tool-recognizes-emotions-in-pictures/ (accessed 30 August 2022).

5. Towards Data Science (electronic journal) Understanding Audio data, Fourier Transform, FFT and Spectrogram features for a Speech Recognition System [Electronic journal] Retrieved from: https://towardsdatascience.com/understanding-audio-data-fourier-transform-fft-spectrogram-and-speech-recognition-a4072d228520

6. Savchuk T. O., Pastukh I. P. (2023). RECOGNIZING THE EMOTIONS OF PARTICIPANTS IN VIDEO CONFERENCES IN MICROSOFT TEAMS. Taurida Scientific Herald. Series: Technical Sciences, (6), 18-24. https://doi.org/10.32851/tnv-tech.2022.6.3