https://doi.org/10.31649/1997-9266-2026-186-3-52-60

УДК 004.9

М. А. Овчаренко[1]
В. Ю. Каштан[1]

НЕЙРОМЕРЕЖЕВИЙ МЕТОД РОЗПІЗНАВАННЯ ЕМОЦІЙНОГО СТАНУ МОВЛЕННЯ В СИСТЕМАХ КОНТАКТ-ЦЕНТРІВ НА ОСНОВІ АРХІТЕКТУРИ CNN-BiLSTM З МОДИФІКОВАНИМ МЕХАНІЗМОМ УВАГИ

[1]Національний технічний університет «Дніпровська політехніка»

Актуальність дослідження зумовлена необхідністю підвищення ефективності систем підтримки прийняття рішень у контакт-центрах для автоматизованого аналізу емоційного стану мовлення операторів і клієнтів. Визначення емоційної напруженості у голосі дозволяє своєчасно коригувати взаємодію, підвищувати якість обслуговування та ефективність роботи операторів. Традиційні методи обробки аудіосигналів, що базуються на мел-частотних кепстральних коефіцієнтах (MFCC), мають обмеження щодо збереження повної акустичної інформації, що знижує точність розпізнавання емоцій. У роботі запропоновано нейромережевий метод, який поєднує згорткові нейронні мережі (CNN) та двонаправлені мережі довгострокової короткочасної пам'яті (BiLSTM) з модифікованим механізмом уваги (Attention). Першим етапом здійснено завантаження українськомовного аудіонабору та попередню обробку даних, що включає нормалізацію амплітуди, фільтрацію шумів, сегментацію мовлення, перетворення у мел-спектрограму та вилучення низькорівневих дескрипторів (LLD), таких як енергія і частота основного тону. Вхідні дані формуються у тензори фіксованої розмірності для нейромережевого аналізу. На етапі вилучення ознак CNN автоматично визначає локальні спектральні характеристики сигналу, включно з інтенсивністю, частотними компонентами та інтонаційними сплесками. Кожен згортковий блок доповнено пакетною нормалізацією для стабільного навчання та пришвидшення збіжності. Для моделювання часової динаміки емоційного стану застосовано двонаправлені шари BiLSTM, що враховують контекст попередніх і наступних фрагментів сигналу. Механізм уваги формує контекстний вектор як зважену суму ознак, визначаючи відносну значущість окремих часових кадрів, який передається до повнозв'язного шару з функцією активації tanh. Модифікований механізм уваги у цій роботі означає інтеграцію метаданих аудіозапису (тривалість сигналу, бітрейт 640 кбіт/с, технічні ідентифікатори) у процес формування контекстного вектору через систему мультимодального зважування (MWS). Це дозволяє одночасно враховувати локальні спектральні ознаки, часову динаміку аудіосигналу та релевантність окремих фрагментів мовлення. Наукова новизна полягає у розробленні нейромережевого методу, що поєднує CNN–BiLSTM–Attention з механізмом мультимодального зважування, який інтегрує метадані аудіозапису у формування контекстного вектора. Така архітектура забезпечує підвищену точність розпізнавання емоційної напруженості. Проведено порівняльний аналіз традиційних MFCC та LLD, який показав перевагу LLD: базова точність CNN зросла з 82,42 % до 91,00 %, а інтеграція шару Attention додатково підвищила точність на 1,5...2 %. Найвищий результат 93,48 % досягнуто у разі поєднання CNN-BiLSTM-Attention з багатомодальною системою зважування та ознаками LLD.

Ключові слова: розпізнавання емоцій мовлення, згорткові нейронні мережі, двонаправлені мережі довгострокової короткочасної пам'яті, мультимодальне зважування, аудіоаналіз.

Вступ

Контакт-центри є складними організаційно-технічними системами, що забезпечують надання інтерактивних послуг через телефонні комунікації та цифрові канали взаємодії. Класичне визначення call center розглядає його як спеціалізовану інфраструктуру з багатьма робочими місцями операторів, які здійснюють прийом та обробку вхідних і вихідних дзвінків, координуючи комунікацію між компанією та клієнтами [1]. У реальних умовах функціонування такі системи характеризуються високим рівнем надходження викликів, що відбуваються випадковим чином, тривалість

розмов є змінною, частина клієнтів може перервати очікування через обмежений час терпіння, а кадрові ресурси підлягають варіативності через людський фактор. Управління контакт-центром передбачає ухвалення складних рішень від прогнозування обсягів звернень, формування змін та графіків операторів, визначення правил маршрутизації викликів, забезпечення необхідного рівня сервісу за мінімізації витрат. Показник рівня сервісу визначається як довгострокова частка дзвінків, що оброблені в межах заданого порогового часу очікування.

В умовах зростання конкуренції та підвищення вимог до якості клієнтського досвіду традиційні підходи до управління контакт-центрами, що базуються переважно на кількісних показниках (тривалість обробки виклику, середній час очікування, коефіцієнт завантаження операторів) є недостатніми для повноцінної оцінки ефективності взаємодії [2]. Важливого значення набувають якісні характеристики комунікації, зокрема емоційний стан клієнта під час розмови. Негативні емоції (роздратування, тривога, агресія) можуть бути індикаторами ризику відтоку клієнтів, ескалації конфліктів або зниження рівня задоволеності послугами. У цьому контексті автоматизоване розпізнавання емоційного стану мовлення на основі методів глибокого навчання розглядається як інструмент підвищення інтелектуалізації систем підтримки прийняття рішень контакт-центрів. Актуальність дослідження зумовлена необхідністю підвищення якості обслуговування клієнтів в контакт-центрах на основі автоматизованого розпізнавання емоційного стану мовлення з подальшим використанням отриманої інформації в системах підтримки прийняття рішень. Додатковим чинником актуальності є обмежена представленість україномовних наборів даних емоційного мовлення та відсутність адаптованих нейромережевих моделей, навчання яких здійснюється з урахуванням фонетичних, просодичних та інтонаційних особливостей української мови. Використання моделей, навчених переважно на англомовних наборах даних, знижує точність класифікації в умовах україномовного середовища контакт-центрів, що зумовлює необхідність розроблення спеціалізованих підходів.

Метою роботи є підвищення точності автоматизованої ідентифікації емоційного стану мовлення українською шляхом розроблення нейромережевого методу на основі поєднання згорткових і двонаправлених рекурентних шарів з модифікованим механізмом уваги, що реалізує мультимодальне зважування для інтеграції акустичних та контекстних ознак у системах підтримки прийняття рішень контакт-центрів.

Аналіз останніх публікацій

Аналіз наявних досліджень свідчить, що задача автоматизованого розпізнавання емоцій мовлення набула значного розвитку завдяки впровадженню методів глибокого навчання [3]. Згорткові нейронні мережі [4] продемонстрували високу ефективність у вилученні локальних спектрально-часових ознак зі спектрограм мовлення, що відображають інтонаційні, енергетичні та частотні характеристики емоцій. Їх перевагою є здатність формувати дискримінаційні просторові представлення ознак без необхідності ручного вилучення ознак експертом. Подальшим розвитком стали повністю згорткові мережі, які дозволяють працювати з вхідними даними змінної довжини та досягати високих результатів у задачах аналізу часових рядів. Проте такі моделі обмежені у відображенні довгострокових часових залежностей, що є важливими для аналізу динаміки емоцій у процесі розмови. Для моделювання часової структури мовлення широко застосовуються рекурентні нейронні мережі та їх модифікації LSTM [5] і BiLSTM [6], які здатні враховувати контекстні залежності між послідовними фрагментами сигналу. Проте використання лише рекурентних архітектур не завжди забезпечує достатньо ефективне первинне вилучення спектральних ознак і може супроводжуватися значними обчислювальними витратами та проблемами затухання градієнта під час роботи з довгими послідовностями. Окреме застосування згорткових або рекурентних моделей не дозволяє повною мірою врахувати як просторово-частотні, так і часові характеристики мовлення, що зумовило розвиток гібридних архітектур типу CNN-LSTM [7] та CNN-BiLSTM [8].

Вищеописані підходи мають низку обмежень. Більшість наявних нейромережевих моделей не враховують специфіку телефонних каналів зв’язку, наявність фонового шуму, перекриття мовлення та багатомовність, характерні для реальних умов функціонування контакт-центрів. До того ж значна частина моделей є обчислювально складною та потребує значних апаратних ресурсів, що ускладнює їх інтеграцію в системи реального часу. Окремою проблемою є мовна адаптація моделей, оскільки переважна більшість досліджень використовує англомовні дані, тоді як питання розпізнавання емоцій українського мовлення залишається недостатньо дослідженим. Відсутність репрезентативних україномовних датасетів та адаптованих моделей знижує точність класифікації

в національному сегменті контакт-центрів. Також недостатньо опрацьованим є питання інтеграції результатів розпізнавання емоцій безпосередньо в контури систем підтримки прийняття управлінських рішень.

Нейромережевий метод ідентифікації емоційного стану мовлення

У цьому дослідженні запропоновано нейромережевий метод розпізнавання емоційного стану мовлення в системах підтримки прийняття рішень контакт-центрів на основі архітектури CNN-BiLSTM з модифікованим механізмом уваги, структурну схему якого показано на рис. 1.

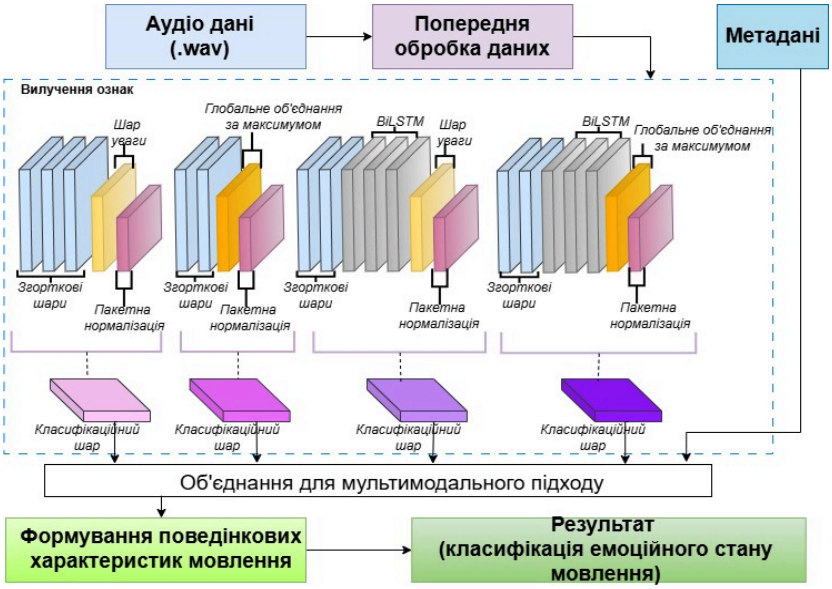

Рис. 1. Схема нейромережевого методу розпізнавання емоційного стану мовлення на основі архітектури CNN-BiLSTM

Першим етапом є завантаження аудіоданих у форматі .wav. Використано реальні аудіозаписи українськомовного набору даних DRSSU (Dataset of Real and Synthesized Speech in Ukrainian).

На другому етапі попередньої обробки [9] виконано нормалізацію амплітуди сигналу, фільтрацію шумів, сегментацію мовлення, перетворення сигналу у мел-спектрограму, вилучення низькорівневих дескрипторів (LLD), енергія (RMS), частота основного тону (F_0) та формування вхідних тензорів фіксованої розмірності. Попередня обробка спрямована на підвищення інформативності даних та їх адаптацію до подальшого нейромережевого аналізу.

Після попередньої обробки аудіодані подаються до блоку вилучення ознак, реалізованого на основі згорткових нейронних мереж. Згорткові шари (ЗШ) використовуються для автоматизованого виявлення локальних спектральних та характеристик мовлення таких як інтенсивність, частотні компоненти, інтонаційні сплески. У моделі застосовано конфігурацію ЗШ = 256/64 фільтрів для аудіосигналу. Як функцію активації використано tanh. Кожний згортковий блок доповнюється шаром пакетної нормалізації для стабільного навчання нейронної мережі та пришвидшує збіжність моделі. Після згорткових шарів застосовується або шар глобального максимального об’єднання або механізм уваги.

Для визначення часової динаміки емоційного стану використано двонапрямлені шари довгострокової короткочасної пам’яті (BiLSTM = 128). Вони дозволяють враховувати контекст як із попередніх, так і з наступних часових фрагментів сигналу, що є критично важливим для аналізу зміни інтонації та емоційної експресії протягом мовленнєвого сегмента. Механізм уваги (Attention = 500/128) використовується для визначення відносної значущості окремих часових кадрів мовлення. Послідовність векторів ознак, сформованих ЗНМ (або ЗНМ та BiLSTM), передається до шару уваги, де для кожного кадру обчислюється ваговий коефіцієнт з використанням функції Softmax [10]

$$\alpha_t = \frac{\exp\left(f\left(\mathbf{x_t}\right)\right)}{\sum_{j=1}^{T} \exp\left(f\left(\mathbf{x}_j\right)\right)}, \tag{1}$$

де α_t — коефіцієнт (вага) уваги для часового кроку t; $\mathbf{x}_t$ — вхідний вектор ознак на часовому кроці

t, що надходить від попередніх шарів (від згорткових блоків або BiLSTM); x_j — функція оцінки (score function).

Контекстний вектор формується як зважена сума ознак [10]

$$C = \sum_{t=1}^{T} a_t x_t. \tag{2}$$

У цій роботі модифікований механізм уваги означає інтеграцію метаданих аудіозапису (тривалість, бітрейт, технічні ідентифікатори) у формування контекстного вектора через систему мультимодального зважування (MWS).

На відміну від стандартних моделей розпізнавання емоцій у мовленні, у запропонованому методі реалізовано мультимодальне зважування, що дозволяє інтегрувати метадані (тривалість запису, бітрейт 640 кбіт/с, технічні ідентифікатори) у процес формування остаточного представлення. Це здійснюється шляхом модифікації контекстного вектора

$$\tilde{C} = C \otimes \sigma\left(W_m m + b_m\right), \tag{3}$$

де σ — сигмоїдна функція активації; $\otimes$ — поелементне множення; $\tilde{C}$ — мультимодально зважений контекстний вектор; W_m, b_m — параметри повнозв'язного шару об'єднання ознак, що визначаються під час навчання моделі.

Після агрегування ознак формується інтегрований вектор представлення, який подається до повнозв'язного шару (Dense = 564) з функцією активації tanh.

На завершальному етапі архітектури результативний шар виконує фінальну обробку отриманих мультимодальних ознак. Вихідний шар із сигмоїдною активацією забезпечує ідентифікацію наявності емоційної напруженості в мовленні [11]. Такий підхід дозволяє трансформувати складні акустичні вектори у зрозумілий для менеджера контакт-центру індикатор: наявність або відсутність конфліктної ситуації/стресу у діалозі. Математично робота механізму уваги реалізує багатомодальну систему зважування, описується формулою (4), яка враховує не лише акустичний контекст, а й вагові коефіцієнти метаданих [10]

$$\alpha_t = \frac{\exp\left(f\left(\mathbf{x}_t\right)\right)\cdot\beta}{\left\|\exp\left(f\left(\mathbf{x}_t\right)\right)\right\|\cdot\left\|\beta\right\|}, \tag{4}$$

де β — вектор параметрів мультимодального зважування, що відповідає за інтеграцію метаданих (тривалість, бітрейт, ідентифікатори) у процес прийняття рішення.

Результати дослідження

Першим етапом проведення експериментальних досліджень стало формування та підготовка набору даних для навчання нейронної мережі. У роботі використано реальні аудіозаписи з україномовного набору даних DRSSU. Цей набір даних розроблено з метою покращення алгоритмів штучного інтелекту, орієнтованих на аналіз емоцій та верифікацію аудіоконтенту в українському лінгвістичному контексті. Для проведення експериментального дослідження з загального набору даних відібрано категорію реальних записів мовлення (Real audio recordings), що налічує понад 30000 зразків. Ці дані включають записи новинних блоків, інтерв'ю та публічних виступів, що забезпечує високу варіативність голосів, тембрів, інтонацій та діалектів.

Технічні специфікації використаного набору даних збережені у форматі wav без втрати даних із частотою дискретизації, що забезпечує бітрейт 640 кбіт/с та тривалістю записів від 2 до 30 секунд. Розподіл тривалості (рис. 2) демонструє значну концентрацію записів у діапазоні до 5 секунд, проте реальне мовлення у вибірці характеризується високою варіативністю параметрів порівняно з синтезованим. Кожний файл супроводжується метаданими, що містять унікальний ідентифікатор та класифікаційні теги, які інтегруються в модель через блок об'єднання для мультимодального підходу. Критерій — енергія (RMS) та якість обслуговування, виявлено сильний негативний зв'язок (–0,74), що доводить гіпотезу про вплив гучності та напруженості голосу на зниження ефективності комунікації. Висока позитивна кореляція (тривалість та паузи = 0,73) підтверджує, що тривалі діалоги супроводжуються збільшенням кількості пауз, що враховується шарами BiLSTM для розуміння темпоритму розмови.

Для проведення мультимодального аналізу виконано попередню обробку даних за допомогою

розробленого програмного інструментарію на мові Python. Результати екстракції параметрів лише для репрезентативної вибірки (фрагмента) подано у табл. 1.

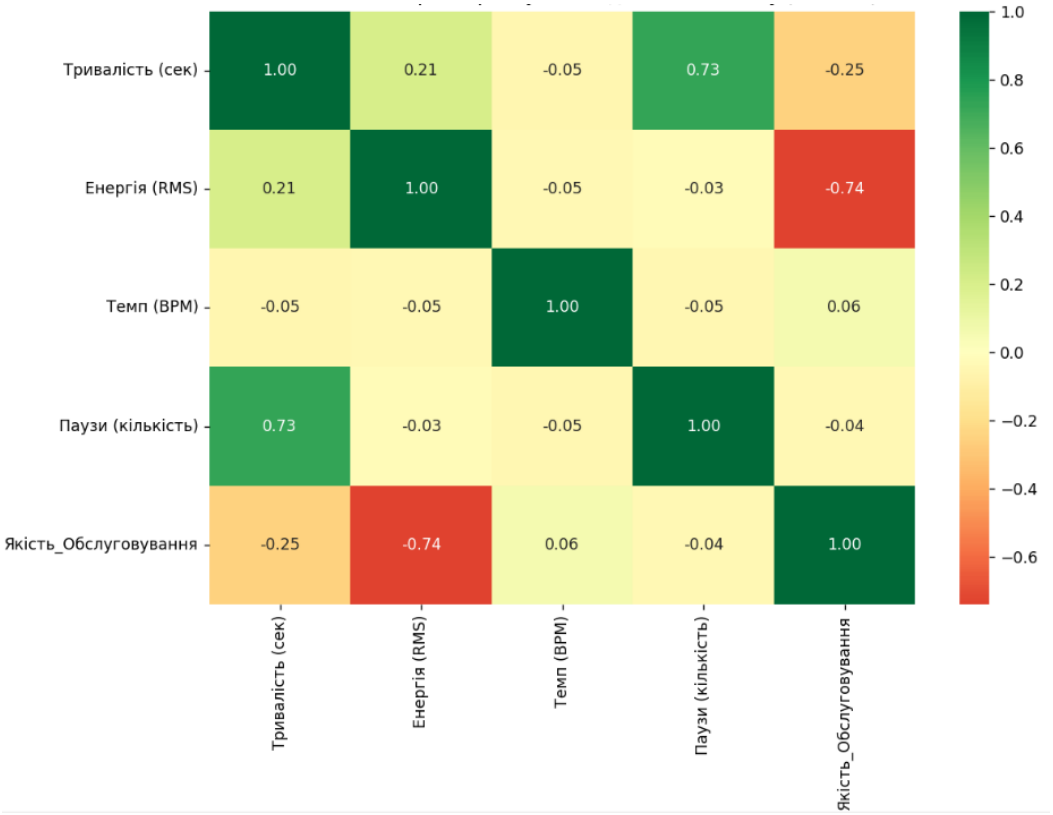

Рис. 2. Матриця кореляційного зв'язку між параметрами мультимодального аналізу емоційного стану мовлення (для 1000 записів)

Таблиця 1

Фрагмент результатів попередньої обробки та екстракції ознак з набору даних DRSSU

Назва файлу	Тривалість (сек)	Енергія (RMS)	Темп (BPM)	Кількість пауз
1.wav	2,89	0,1356	123,05	1
10.wav	4,11	0,0544	95,70	2
100.wav	04,04	0,0805	80,75	2
1000.wav	04,02	0,0649	89,10	2
10000.wav	6,59	0,0102	129,31	4

Візуальний аналіз часових та частотних характеристик типового сигналу (рис. 3) дозволив ідентифікувати ключові емоційні маркери. Осцилограма сигналу демонструє амплітудні коливання, що відповідають активним фазам мовлення, тоді як мейл-спектрограма відображає інтенсивний розподіл енергії в діапазоні до 8192 Гц, що є критичним входом для згорткових шарів CNN. Особливу увагу приділено динаміці енергії мовлення (RMS), де зафіксовані пікові сплески понад 0,20 безпосередньо вказують на зони емоційної напруженості.

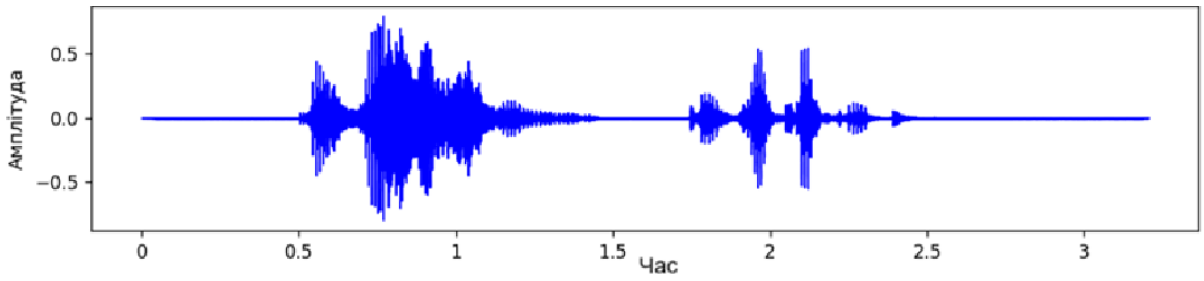

Рис. 3. Осцилограма аудіодзвінка

Мейл-спектрограма показана на рис. 4 є двовимірним представленням аудіосигналу, яке трансформує часовий ряд у візуальний образ, придатний для обробки згортковими шарами.

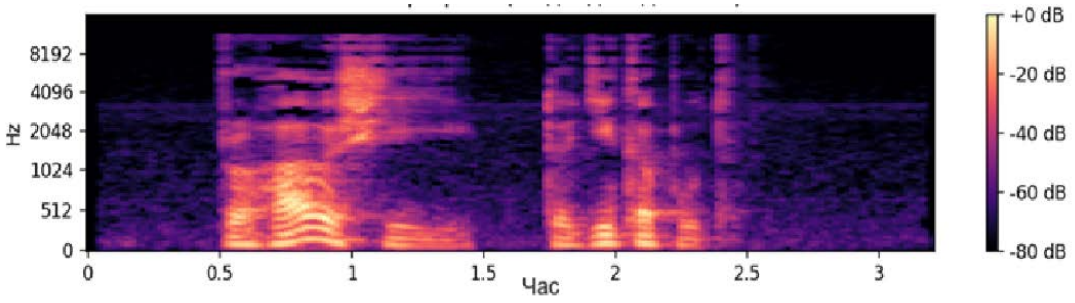

Рис. 4. Мейл-спектограма (вхідні дані для ЗНМ)

Використання широкого спектра до 8192 Гц дозволяє охопити не лише основний тон голосу, а й високочастотні гармоніки та шумові компоненти, які часто супроводжують емоційні сплески, такі як різкі крики або напружений шепіт. Візуалізація амплітуди в логарифмічній шкалі децибел (від –80 до 0 dB) забезпечує високий контраст між нейтральним фоном та активними мовленнєвими ділянками. Це дозволяє першим шарам ЗНМ виділяти локальні ознаки. Яскраві горизонтальні смуги на спектрограмі відповідають формантам резонансним частотам мовленнєвого тракту. Зміна відстані між ними та їх інтенсивності є прямим індикатором зміни емоційного стану (тембрального голосу), що фіксується нейромережею як специфічний набір ознак. Так як спектрограма зберігає як часову (вісь X), так і частотну (вісь Y) інформацію, блоки ЗШ = 256 та ЗШ = 64 у запропонованій в роботі архітектури здатні виявляти складні кореляції між зміною висоти голосу та тривалістю окремих звуків.

Класифікація емоцій за допомогою матриці помилок (рис. 5) підтверджує високу селективну здатність моделі: нейтральний стан розпізнається з найвищою точністю (66 випадків), активні емоції (злість/радість) та стани збудження чи втоми демонструють стабільно високі показники (по 47 та по 42 правильних розпізнавань відповідно). Мінімальна похибка між полярними емоційними станами свідчить про надійність комбінованої архітектури та доцільність використання механізму уваги (формула (4) для виділення найінформативніших акустичних фрагментів.

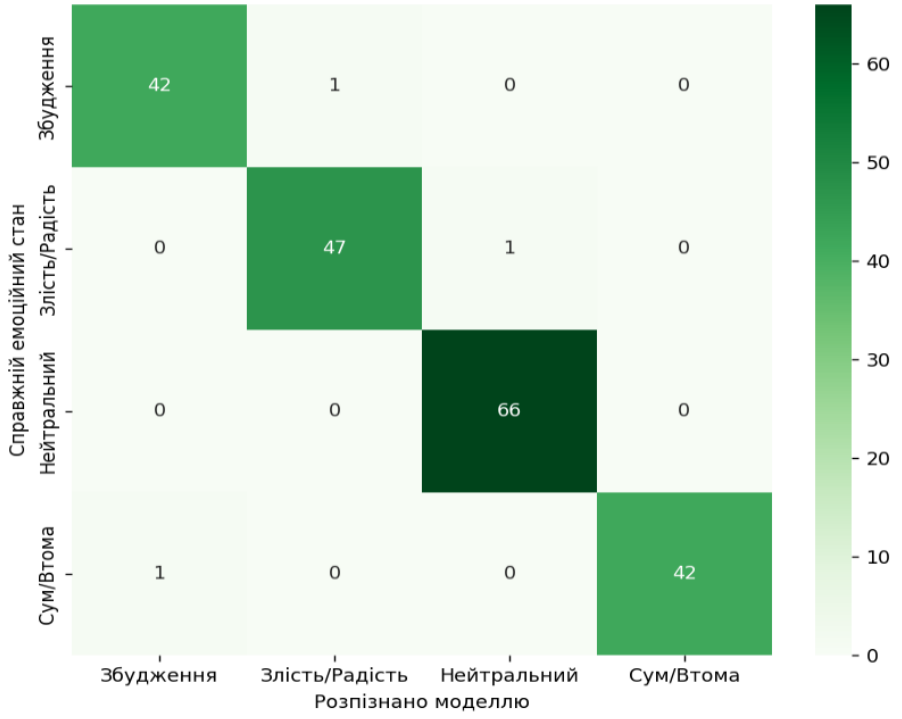

Рис. 5. Матриця помилок для розпізнавання емоцій

Для оцінки якості запропонованого нейромережевого підходу розпізнавання емоційного стану мовлення проведено серію порівняльних експериментів з використанням різних архітектур ней-

ронних мереж та типів вхідних ознак. Результати порівняння за метрикою точність (Accuracy) [12] подано в табл. 2.

Таблиця 2

Результати метрики точність (Accuracy) методів класифікації мовлення

Тип архітектури	Тип ознак	Точність (Accuracy) , %
CNNs	MFCC	82,42
CNNs-Attention	MFCC	83,92
CNNs-BiLSTMs-Attention	MFCC	83,90
CNNs	LLD	91,00
CNNs-Attention	LLD	92,99
CNN-BiLSTMs-Attention	LLD	92,98
Запропонована (CNN-BiLSTM-Attention + MWS)	**LLD**	**93,48**

Аналіз даних свідчить про суттєву перевагу використання низькорівневих дескрипторів (LLD) порівняно з традиційними мейл-кепстральними коефіцієнтами (MFCC). Перехід до LLD дозволив підвищити базову точність моделі CNN з 82,42 % до 91,00 %, що підтверджує важливість збереження первинної акустичної інформації (енергії, висоти тону) для ідентифікації емоційної напруженості. Інтеграція шару уваги до архітектури CNN забезпечує приріст точності приблизно на 1,5...2 % для обох типів ознак. Це математично підтверджує ефективність використання механізму уваги для зважування найінформативніших фрагментів аудіосигналу. Найвищий показник точності (93,48 %) продемонструвала запропонована архітектура CNNs-BiLSTMs-Attention у поєднанні з багатомодальною системою зважування (MWS) та ознаками LLD. Це доводить, що спільне використання згорткових шарів для вилучення спектральних ознак та шарів BiLSTM для аналізу часових залежностей (пауз, темпу) є оптимальним для задач розпізнавання емоцій в україномовному сегменті.

Висновки

У роботі розроблено нейромережевий метод розпізнавання емоційного стану мовлення для застосування в системах контакт-центрів на основі поєднання згорткових та двонапрямлених рекурентних шарів з модифікованим механізмом уваги, який реалізує мультимодальне зважування для інтеграції акустичних та контекстних ознак аудіозапису. Проведена попередня обробка аудіоданих включала нормалізацію амплітуди, фільтрацію шумів, сегментацію мовлення, перетворення сигналу у мел-спектрограму та вилучення низькорівневих дескрипторів (LLD), що дозволило сформувати вхідні тензори фіксованої розмірності для нейромережевого аналізу та підвищити інформативність даних. Аналіз мел-спектрограми показав, що запропонована архітектура здатна виявляти локальні спектральні та тембральні ознаки сигналу, зокрема форманти, високо-частотні гармоніки та шумові компоненти, які корелюють із зміною емоційного стану. Використання блоків CNN із 256 та 64 фільтрами дозволило моделі ефективно виділяти складні взаємозв’язки між висотою голосу, тривалістю звуків та інтенсивністю мовлення. Використання механізму уваги та системи мультимодального зважування забезпечило виділення найрелевантніших часових фрагментів аудіосигналу, що дозволяє враховувати одночасно локальні спектральні характеристики, часову динаміку та контекстні ознаки. Порівняльний аналіз традиційних MFCC та низькорівневих дескрипторів показав перевагу LLD: базова точність CNN зросла з 82,42 % до 91,00 %, а інтеграція шару Attention забезпечила додаткове підвищення точності на 1,5...2 %. Найвищий показник точності 93,48 % досягнуто у разі поєднання CNN-BiLSTM-Attention з багатомодальною системою зважування та ознаками LLD. Аналіз матриці помилок підтвердив високу точність розпізнавання моделі — нейтральний стан розпізнається з найвищою точністю, активні емоції (злість/радість) та стани збудження чи втоми демонструють стабільно високі показники розпізнавання. Мінімальна похибка між полярними емоційними станами свідчить про надійність комбінованої архітектури та доцільність використання модифікованого механізму уваги. Проведені експериментальні результати підтверджують ефективність методу для автоматизованого розпізнавання емоційного стану

мовлення та його практичну цінність у системах підтримки рішень контакт-центрів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] N. Dhanpat, F. D. Modau, P. Lugisani, R. Mabojane, and M. Phiri, “Exploring employee retention and intention to leave within a call center,” *SA J. Hum. Resour. Manag.*, vol. 16, pp. 1-13, 2018. [Online]. Available: https://www.researchgate.net/publication/323869688_Exploring_employee_retention_and_intention_to_leave_within_a_call_centre .

[2] B. Karakus, and G. Aydin, “Call center performance evaluation using big data analytics,” in *Proc. 2016 International Symposium on Networks, Computers and Communications (ISNCC)*, Hammamet, Tunisia, 2016, pp. 1-6. https://doi.org/10.1109/ISNCC.2016.7746116 .

[3] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, “Deep neural networks for emotion recognition combining audio and transcripts,” *arXiv preprint arXiv:1911.00432*, 2019. https://doi.org/10.21437/Interspeech.2018-2466 .

[4] X. Liu, J. van de Weijer, and A. D. Bagdanov, “Exploiting unlabeled data in cnns by self-supervised learning to rank,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, pp. 1862-1878, 2019, https://doi.org/10.48550/arXiv.1902.06285 .

[5] F. Karim, S. Majumdar, and H. Darabi, “Insights into LSTM fully convolutional networks for time series classification,” *IEEE Access*, vol. 7, pp. 67718-67725, 2019, https://doi.org/10.1109/ACCESS.2019.2916828 .

[6] N. Shabrina, F. Kasyidi, and R. Ilyas, “A BiLSTM-Based Approach for Speech Emotion Recognition in Conversational Indonesian Audio using SMOTE,” *Jurnal Teknik Informatika (Jutif)*, vol. 6, pp. 3173-3187, 2025, https://doi.org/10.52436/1.jutif.2025.6.5.5183 .

[7] S. Ayadi, and Z. Lachiri, “A combined CNN-LSTM Network for Audio Emotion Recognition using Speech and Song attributs,” in *Proc. 2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, Sfax, Tunisia, 2022, pp. 1-6, https://doi.org/10.1109/ATSIP55956.2022.9805924 .

[8] A. Ahmed, S. Toral, K. Shaalan, and Y. Hifny, “Agent Productivity Modeling in a Call Center Domain Using Attentive Convolutional Neural Networks,” *Sensors*, vol. 20, p. 5489, 2020, https://doi.org/10.3390/s20195489 .

[9] В. Гнатушенко, В. Каштан, А. Іванько, і М. Овчаренко, «Аналіз неструктурованих даних контакт-центру для підтримки прийняття рішень,» *Електротехнічні та інформаційні системи*, № 106, с. 80-86, 2024, https://doi.org/10.32782/EIS/2024-106-14 .

[10] A. Norouzian, B. Mazoure, D. Connolly, and D. Willett, “Exploring attention mechanism for acoustic-based classification of speech utterances into system-directed and non-system-directed,” in *Proc. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 7310-7314, https://doi.org/10.48550/arXiv.1902.00570 .

[11] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, “A review of emotion recognition using physiological signals,” *Sensors*, vol. 18, p. 2074, 2018, https://doi.org/10.3390/s18072074 .

[12] T. Anvarjon, Mustaqeem, and S. Kwon, “Deep-Net: A Lightweight CNN-Based Speech Emotion Recognition System Using Deep Frequency Features,” *Sensors*, vol. 20, p. 5212, 2020, https://doi.org/10.3390/s20185212 .

Рекомендована кафедрою автоматизації та інтелектуальних інформаційних технологій ВНТУ

Дата надходження 27.02.2026
Дата прийняття до друку після рецензування 2.03.2026
Дата публікації 7.07.2026

Овчаренко Максим Анатолійович — аспірант кафедри інформаційних технологій та комп’ютерної інженерії, e-mail: ovcharenko.m.a@nmu.one . https://orcid.org/0009-0006-0730-0913 ;

Каштан Віта Юріївна — канд. техн. наук, доцент, доцент кафедри інформаційних технологій та комп’ютерної інженерії, e-mail: kashtan.v.yu@nmu.one . https://orcid.org/0000-0002-0395-5895 .

Національний технічний університет «Дніпровська політехніка», Дніпро

M. A. Ovcharenko[1]
V. Yu. Kashtan[1]

Neural Network Method for Emotion Speech State Recognition in Contact Centre Systems Based on CNN-BiLSTM Architecture with a Modified Attention Mechanism

[1]Dnipro University of Technology

The relevance of this research lies in the need to improve the efficiency of decision-support systems in contact centers for automated analysis of the emotional states of operators and clients. Detecting emotional tension in the voice enables

timely adjustments to interactions, enhancing service quality and operator performance. Traditional audio signal processing methods based on Mel-Frequency Cepstral Coefficients (MFCCs) have limitations in preserving complete acoustic information, thereby reducing the accuracy of emotion recognition. This work proposes a neural network method that combines Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) networks with a modified Attention mechanism. The first stage involves loading a Ukrainian-language audio dataset and performing preliminary data processing, including amplitude normalization, noise filtering, speech segmentation, conversion to Mel-spectrograms, and extraction of low-level descriptors (LLD), such as energy and fundamental frequency (F0). The input data are formed into fixed-dimension tensors for neural network analysis. During the feature extraction stage, the CNN automatically identifies local spectral characteristics of the signal, including intensity, frequency components, and intonational peaks. Each convolutional block is complemented with batch normalization to stabilize training and accelerate convergence. To model the temporal dynamics of the emotional state, bidirectional BiLSTM layers are applied, taking into account the context of preceding and subsequent signal segments. The Attention mechanism forms a context vector as a weighted sum of features, determining the relative importance of individual time frames, and then passes it to a fully connected layer with a tanh activation function. In this work, the modified Attention mechanism refers to the integration of audio recording metadata (signal duration, 640 kbps bitrate, technical identifiers) into the formation of the context vector via a Multi-Weighting System (MWS). It allows simultaneous consideration of local spectral features, temporal dynamics of the audio signal, and the relevance of individual speech segments. The scientific novelty lies in the development of a neural network method that combines CNN–BiLSTM–Attention with a multimodal weighting mechanism, integrating audio metadata into the formation of the context vector. This architecture provides increased accuracy in recognizing emotional tension. A comparative analysis of traditional MFCC and LLD was conducted, demonstrating the advantage of LLD: the baseline CNN accuracy increased from 82,42 % to 9,.00 %, and integrating the Attention layer further improved accuracy by 1.5...2%. The highest achieved accuracy was 93,48 %.

Keywords: speech emotion recognition, convolutional neural networks, bidirectional long-term memory networks, multimodal weighting, audio analysis.

Ovcharenko Maksym A. — Post-Graduate Student of the Chair of Information Technology and Computer Engineering, e-mail: ovcharenko.m.a@nmu.one . https://orcid.org/0009-0006-0730-0913 ;

Kashtan Vita Yu. — Cand. Sc. (Eng.), Associate Professor, Associate Professor of the Chair of Information Technology and Computer Engineering, e-mail: kashtan.v.yu@nmu.one . https://orcid.org/0000-0002-0395-5895