КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА
Факультет інформаційних технологій
Кафедра інтелектуальних технологій

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня «магістр»

НА ТЕМУ:

Інтелектуальний модуль розпізнавання емоцій за голосом

Галузь знань: 12 «Інформаційні технології»
Спеціальність: 122 «Комп'ютерні науки»
Освітньо-наукова програма «Технології штучного інтелекту»

Виконав:
студент 2 курсу магістратури, групи ТШІ-21

Астахов Антон Кирилович
(ПІБ)

Науковий керівник: Гайна Георгій Анатолійович
(ПІБ)
кандидат технічних наук,
професор кафедри інтелектуальних технологій
(науковий ступінь, вчене звання)

Засвідчую, що в цій кваліфікаційній роботі немає запозичень з праць інших авторів без відповідних посилань

Студент _____
підпис

Кваліфікаційна робота допущена до захисту
рішенням кафедри *інтелектуальних технологій*

Протокол № ____ від «____» травня 2021 р.

Зав. кафедри _____ доц. Іларіонов О.Є.
підпис

КИЇВ 2021

ЗМІСТ

ВСТУП

Для людей мова є основним засобом комунікації. Причому люди з мовлення можуть отримувати не тільки семантичну інформацію. Для людей процес сприйняття емоцій є інтуїтивним, однак можливість автоматизації цього процесу з використанням інформаційних технологій є актуальною темою досліджень.

Розпізнавання емоцій за голосом людини — це процес виділення емоційних аспектів мовлення без урахування семантичної інформації. Вирішення даної задачі є актуальним для таких галузей, як надання психологічної допомоги, розробка систем безпеки, виявлення брехні, аналіз зв'язків з клієнтами, розробка відеоігор.

Головна проблема, яка постає при вирішенні даної задачі, полягає у суб'єктивній природі емоцій. В психології існують багато моделей емоційних станів, і не існує єдиного консенсусу щодо того, як емоції вимірювати та класифікувати.

Розпізнавання емоцій за голосом складається з таких основних етапів, як вибір моделі емоційних станів, підготовка набору даних, передобробка даних, вилучення ознак, класифікація. Більшість робіт описують поєднання різними способами відомих методів на вищезгаданих етапах.

Есперимантальні дослідження в даній предметній області почали активно проводитися відносно нещодавно, більшість робіт датовані 21-м століттям. Основними методами вирішення задачі є приховані марковські моделі, метод опорних векторів, глибинні нейронні мережі. Середня точність класифікаторів в проаналізованих роботах становить *0,75*.

Провідними науковцями, які здійснили значний внесок в дослідження розпізнавання емоцій за голосом є Шулер, Тигер, Лоу, Екман.

Об'єкт дослідження: процес розпізнавання емоцій людини за голосом з використанням технологій штучного інтелекту.

Предмет дослідження: застосування методів глибинного навчання для розпізнавання емоцій людини за голосом.

Мета дослідження: розробити інтелектуальний модуль розпізнавання емоцій людини за голосом з використанням методів глибинного навчання.

Завдання дослідження:

1. провести аналіз методів та підходів до вирішення задачі розпізнавання емоцій за голосом людини;
2. створити модель інтелектуального модуля та згорткової нейронної мережі;
3. розробити інтелектуальний модуль розпізнавання емоцій за голосом;
4. оцінити ефективність отриманих результатів дослідження.

В якості методів дослідження використовувались порівняльний аналіз науково-практичної інформації для отримання теоретичних результатів та експерименти для отримання практичних результатів. Для досягнення практичних результатів методика дослідження складалася із застосування методів штучного інтелекту для розробки інтелектуального модуля розпізнавання емоцій за голосом та вимірювання точності і похибки для оцінки якості результатів дослідження.

Новизна даного дослідження виражається в модифікації існуючих методів вирішення науково-дослідницьких завдань. Було реалізовано інтелектуальний модуль розпізнавання емоцій за голосом на основі згорткових нейронних мереж та наборів даних RAVDESS і TESS. За результатами програмної реалізації модуля розпізнавання емоцій за голосом збільшено точність валідації до *0,8*.

Результати роботи мають практичне значення, оскільки можуть бути застосовані у сфері практичної діяльності. Зокрема, основними сценаріями використання розробленого інтелектуального модуля є оптимізація роботи кол центрів, виконання психологічної оцінки, оптимізація роботи інтелектуальних голосових помічників.

Для апробації результатів дослідження автор роботи приймав участь в VII Міжнародній науково-практичній конференції "Інформаційні технології та взаємодії", де виступив з доповіддю на тему "Analysis of speech emotion

recognition methods". Також автор подав для публікації до журналу статтю "Розробка інтелектуального модуля розпізнавання емоцій за голосом" .

Відповідно до мети та завдань дослідження робота складається зі вступу, 3 розділів, 18 підрозділів, висновків, списку використаних джерел із 54 найменувань та 2 додатків. Загальний обсяг роботи 80 сторінок.

РОЗДІЛ 1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Актуальність розпізнавання емоцій за голосом

Для людей мова є найбільш природним способом самовираження та передачі інформації. Найкраще ми усвідомлюємо її важливість, коли нам доводиться використовувати інші способи спілкування, такі як електронна пошта чи текстові повідомлення. У такому разі нам одразу стає важче зрозуміти співрозмовника, адже ми звикли отримувати не тільки семантичну інформацію зі слів, а й мати справу з емоціями людини, які допомагають нам краще зрозуміти ситуацію.

Тому логічним є поширення цього розуміння і на комп'ютери. Розпізнавання мови вже присутнє у нашому повсякденному житті завдяки розумним мобільним пристроям, які здатні приймати голосові команди та відповідати на них синтезованою мовою.

Задача розпізнавання емоцій за голосом існує вже більше двох десятиліть [1], є програми для взаємодії людина-комп'ютер [2], а також роботи [3], мобільні послуги [4], кол центри [5], комп'ютерні ігри [6] та психологічна оцінка [7, 8]. Дана задача є вкрай складною через те, що емоції суб'єктивні. Не існує єдиного консенсусу щодо того, як їх виміряти або класифікувати.

Система розпізнавання емоцій за голосом — це сукупність методологій, які обробляють та класифікують мовні сигнали для виявлення в них емоцій. Система має такі складові:

1. модель емоцій:
    1. дискретна;
    2. багатовимірна;
2. набір даних:
    1. датасети, що складаються з емоцій, які відіграні акторами;
    2. датасети з записами емоцій, що були штучно викликані;

3. датасети з записами природних емоцій;

3. процеси передобробки:
    1. фреймінг;
    2. віконне перетворення;
    3. нормалізація;
    4. зменшення шуму;
4. ознаки:
    1. просодичні:
        1. висота голосу;
        2. інтенсивність;
        3. тривалість;
    2. спектральні:
        1. мел-частотні кепстральні коефіцієнти (MFCC);
        2. коефіцієнти лінійного прогнозування;
        3. коефіцієнти частоти гаматонів;
        4. форманти;
    3. якості голосу:
        1. джитер (тремтіння);
        2. мерехтіння;
        3. гармоніка;
        4. якісні мовні ознаки;
    4. TEO-ознаки;
5. класифікатор:
    1. традиційні класифікатори:
        1. метод опорних векторів;
        2. прихована марковська модель;
        3. гаусівська змішана модель;
        4. штучна нейронна мережа;
        5. дерево рішень;
    2. класифікатори на основі глибинного навчання;

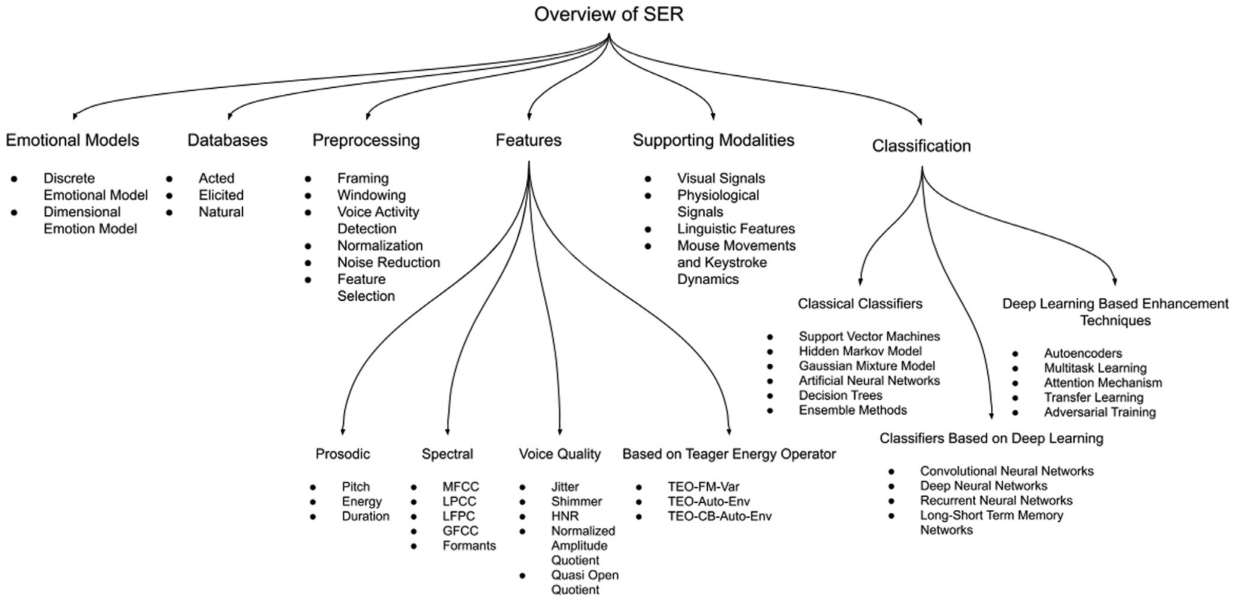


Рисунок 1 — Огляд систем розпізнавання емоцій за голосом [9]

1. згорткові нейронні мережі;
2. рекурентні нейронні мережі;
3. методи машинного навчання для покращення класифікації.

В роботі [9] наведено аналіз сучасного стану проблеми розпізнавання емоцій за голосом людини. Схема представлена на рисунку 1.

На рисунку 2 зображено структуру системи розпізнавання емоцій за голосом. На вході системи запис мовлення, далі відбувається обробка сигналу, накладання моделі емоцій, виділення ознак та розпізнавання емоцій. Всі складові будуть деталізовані та ретельно проаналізовані в наступних пунктах роботи.

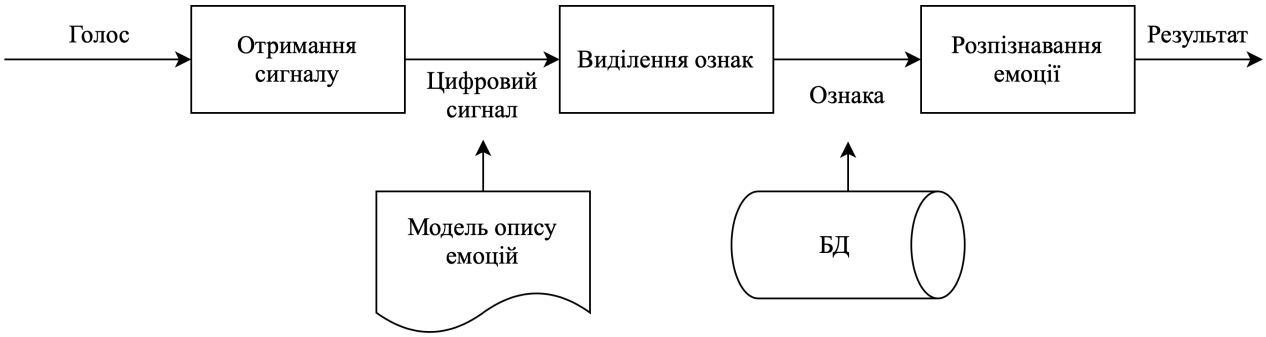


Рисунок 2 — Система розпізнавання емоцій за голосом

1.2 Огляд літератури

Зрозуміло, що вже є публікації, в яких описуються існуючі дослідження з розпізнавання емоцій за голосом. Були розглянуті в основному ті дослідження, що опубліковані порівняно недавно, адже попередні публікації не могли включати останні досягнення та тенденції, такі як глибокі нейронні мережі. Згідно з дослідженнями [9, 10], у Таблиці 1 підсумовано публікації, включаючи набори даних, ознаки, класифікатори та результати експериментів. Середня точність роботи класифікаторів у проаналізованих дослідженнях склала *0,759*.

Таблиця 1 — Перелік досліджень

Публікація	**Набір даних**	**Ознаки**	**Класифікатор**	**Точність (%)**
Albornoz et al. (2011)	Berlin Emo DB	MFCC, просодичні ознаки	Прихована марковська модель, гаусівська змішана модель, багатошаровий перцептрон	71.5%

Bitouk et al. (2010)	LDC, Berlin Emo DB	Спектральні	Метод опорних векторів	46.1%
Borchert and Dusterhoft (2005)	Berlin Emo DB	Форманти, спектральний розподіл енергії в різних частотних смугах, HNR, джитер, мерехтіння	Метод опорних векторів, дерева рішень	90%
Busso et al. (2009)	EPSAT, EMA, GES, SES, WSJ	Ознаки на основі базової частоти	Гаусівська змішана модель	77%
Deng et al. (2013)	AVIC, EMODB, eNTERFACE, SUSAS, VAM	LLD, інтенсивність, висота, HNR, MFCC	Авто-кодувальник	62.7%
Deng et al. (2014)	AIBO DB, ABC DB, SUSAS DB	Низькорівневі дескриптори	Авто-кодувальник, метод опорних векторів	64.18%
Han et al. (2014)	IEMOCAP DB	MFCC, висота голосу	Глибинні нейронні мережі, машина екстремального навчання	54.3%
Hu et al. (2007)	8 носіїв китайської мови	Спектральні ознаки	Гаусівська змішана модель, метод опорних векторів	82.5%
Kwon et al. (2003)	SUSAS DB, AIBO DB	Просодичні та спектральні ознаки	Метод опорних векторів, прихована марковська модель	92.2%
Lee et al. (2011)	AIBO DBUSC IEMOCAP DB.	Інтенсивність, висота, гармоніка, MFCC	Дерево рішень	58.46%

Luengo et al. (2005)	База даних емоційного мовлення країни Басків	Просодичні та спектральні ознаки	Метод опорних векторів, гаусівська змішана модель	98%
Mirsamadi et al. (2017)	Корпус IEMOCAP	Коефіцієнти лінійного прогнозування, інтенсивність, MFCC	Рекурентні нейронні мережі	63.5%
Mao et al. (2014)	SAVEE DB, Berlin EMO DB, DES DB, MES DB	Автоматичне навчання на основі згорткових нейронних мереж	Згорткові нейронні мережі	73.6%
Nakatsu et al. (1999)	100 висловлювань, 50 чоловіків, 50 жінок	Висота та інтенсивність голосу	Нейронні мережі	50%
Nogueiras et al. (2001)	Іспанський корпус INTERFACE, Emotional Speech Synthesis Database	Просодичні та спектральні ознаки	Прихована марковська модель	70%
Nwe et al. (2003b)	3 жінки та 3 чоловіки носії бірманської мови; 3 жінки та 3 чоловіки носії китайської	LFPC	Прихована марковська модель	78.5%
Rao et al. (2013)	Корпус Telugu	Просодичні ознаки	Метод опорних векторів	66%
Rong et al. (2009)	Корпус китайської мови	Висота, інтенсивність, спектральні ознаки	Метод *k*-найближчих сусідів	66.24%
Sato and Obuchi (2007)	Linguistic Data Consortium Database	MFCC	Прихована марковська модель	66.4%

Schuller et al. (2003)	5250 висловлювань англійською та німецькою, 5 спікерів	Висота та інтенсивність голосу	Прихована марковська модель	86.8%
Schuller et al. (2005)	База даних на основі 3947 фільмів, 35 спікерів	Висота, інтенсивність та довгота звуків	Метод опорних векторів, дерева рішень, метод *k*-найближчих сусідів	63.51%
Schuller et al. (2005)	Berlin EmoDb	Висота, форманти, спектральні та лінгвістичні ознаки	Метод опорних векторів, дерева рішень	76.23%
Schuller (2011)	VAM DB	Низькорівневі дескриптори, контур сигналу, форманти, висота голосу, MFCC, HNR, лінгвістичні ознаки	Опорно-векторна регресія	66.7%
Shen et al. (2011)	Berlin Emo DB	Інтенсивність, висота, MFCC, коефіцієнти лінійного прогнозування	Метод опорних векторів	66.02%
Ververidis and Kotropoulos (2005)	1300 висловлювань	Статистичні властивості формантів, висоти, інтенсивності	Гаусівська змішана модель	48.5%
Wang et al. (2015)	Berlin EMO DB, CASIA DB, (EESDB)	Параметри Фур'є, MFCC	Метод опорних векторів	88.88%
Wollmer et al. (2010)	Sensitive Artificial Listener (SAL) database	Висота голосу, MFCC, інтенсивність, HNR, лінгвістичні ознаки	LSTM, метод опорних векторів, рекурентні нейронні мережі	51.3%

Wu and Liang (2011)	Два корпуси, 2033 речень, 8 волонтерів	Висота, інтенсивність, форманти 1–4, джитер, мерехтіння, ознаки гармоніки, MFCC	Мета-дерева прийняття рішень, що складаються за допомогою методу опорних векторів, гаусівської змішаної моделі, багатошарового перцептрону	80%
Wu et al. (2011)	Berlin Emo DB, VAM DB	Просодичні ознаки, швидкість мовлення, TEO-ознаки	Метод опорних векторів	91.3%
Yang and Lugger (2010)	Berlin Emo DB	Просодичні, спектральні та ознаки якості голосу	Байєсів класифікатор	73.5%
Zhang et al. (2011)	ABC, AVIC, DES, eNTERFACE, SAL, VAM	Інтенсивність, висота, якість голосу, спектральні ознаки, MFCC	Навчання без учителя	66.8%
Середня точність				**0,759**

1.3 Аналіз моделей емоційних станів

Для успішної реалізації системи розпізнавання емоцій за голосом потрібно вірно визначати та моделювати емоції. Однак єдиного консенсусу щодо визначення емоції не існує, і це все ще залишається відкритою проблемою в психології. У XX столітті було запропоновано понад дев'яносто визначень емоцій [11]. Емоції є деякими психологічними станами, що складаються з декількох компонентів, таких як особистий досвід, фізіологічні, поведінкові та комунікативні реакції. На основі цих визначень у розпізнаванні емоцій стали поширеними дві моделі: дискретна модель емоцій та багатомірна модель емоцій.

Дискретна модель емоцій базується на шести категоріях основних емоцій: смуток, радість, страх, гнів, відраза та здивування [12, 13]. Ці вроджені та незалежні від культури емоції переживаються протягом короткого періоду часу [14]. Інші емоції отримуються поєднанням основних. Більшість існуючих систем розпізнавання емоцій зосереджені на цих основних емоційних категоріях. Однак, даними категоріями емоційних станів не можна вичерпно описати деякі складні емоційні стани.

Багатовимірна модель емоцій — це альтернативна модель, яка використовує невелику кількість вимірів, таких як привабливість, збудження, контроль, влада [15, 16]. Однією з найвживаніших моделей є двовимірна модель, яка використовує збудження чи активацію в якості одного виміру, а валентність, чи оцінку в якості іншого. Валентність визначає, чи є емоція позитивною або негативною, і змінюється від неприємної до приємної. Збудження визначає силу почуттів, і може коливатися від нудьги до шаленого хвилювання [17]. Тривимірна модель включає вимір домінування або влади, що характеризує сили людини, і знаходиться в діапазоні між слабкістю і силою. Наприклад, за даним виміром можна відрізняти гнів від страху [18].

Існує декілька недоліків багатовимірного зображення. По-перше, дана модель недостатньо інтуїтивна, і для позначення кожної емоції може знадобитися спеціальна підготовка [19]. Крім того, деякі емоції, наприклад, здивування, важно віднести до певної категорії, адже дані емоції можуть мати позитивну чи негативну валентність в залежності від контексту.

1.4 Аналіз методів створення наборів даних

Набори даних для розпізнавання емоцій за голосом можна розділити на три категорії:

1. датасети, що складаються з емоцій, які відіграні акторами;
2. датасети з записами емоцій, що були штучно викликані;
3. датасети з записами природних емоцій.

Перша категорія наборів даних складається з емоцій, що відтворені професійними акторами в звуконепроникних студіях. Створити такий набір даних порівняно простіше, однак такий спосіб може неадекватно передавати силу емоцій, часто призводячи до перебільшення. Це знижує точність розпізнавання у реальному житті.

Для створення другої категорії наборів даних дослідники відтворюють ситуації, що штучно провокують емоції. Такий спосіб більш точно відтворює емоційний стан людини.

Набори даних з записами природних емоцій здебільшого отримують із ток-шоу, записів кол центрів, радіопереговорів та подібних джерел. Отримати такі дані важче, в першу чергу через етичні та юридичні проблеми.

Після вибору методу створення набору даних розглядаються інші питання проектування, такі як вік та стать. Більшість баз даних містять дорослих ораторів, але також існують бази даних дітей та людей старшого віку.

Наприклад, загальновживаний Берлінський датасет містить сім емоцій, відіграних десятьма професійними акторами, половина з яких є чоловіками, половина жінками [20]. Кожне висловлювання повторюється різними акторами та з різними емоціями.

Перелік датасетів наведений в Таблиці 2.

Таблиця 2 — Перелік наборів даних

Датасет	**Емоції**	**Розмір**	**Мова**	**Доступність**
Surrey Audio-Visual Expressed Emotion (SAVEE)	Гнів, відраза, страх, радість, смуток, здивування, нейтральна	14 ораторів (чоловіки), 120 висловлювань	Англійська	Безкоштовно
Toronto Emotional SpeechDatabase (TESS)	Гнів, відраза, нейтральна, страх, радість, смуток, приємність, здивування	2 оратори (жінки), 2800 висловлювань	Англійська	Безкоштовно

eNTERFACE'05 Audio-Visual Emotion Database	Гнів, відраза, страх, радість, смуток, здивування	42 оратори (34 чоловіки, 8 жінок) 14 національностей 1116 відео	Англійська	Безкоштовно
SAMAINE Database	Валентність, сила, очікування, загальна емоційна напруженість	150 ораторів, 959 розмов	Англійська, грецька, іврит	Безкоштовно
TUM AVIC Database	5 рівнів інтересу; 5 нелінгвістичних вокалізацій (дихання, згода, шум, вагання, сміх)	21 оратор (11 чоловіків, 10 жінок), 3901 висловлювань	Англійська	Безкоштовно
AFEW Database	Гнів, відраза, здивування, страх, радість, нейтральна, смуток	330 ораторів, 1426 висловлювань з фільмів та теле-шоу	Англійська	Безкоштовно
Berlin Emotional Database (EmoDB)	Гнів, нудьга, відраза, страх, радість, смуток, нейтральна	7 емоцій х 10 ораторів (5 чоловіків, 5 жінок) х 10 висловлювань	Німецька	Вільний доступ
Chinese Emotional Speech Corpus (CASIA)	Здивування, радість, смуток, гнів, страх, нейтральна	6 емоцій х 4 оратори (2 чоловіки, 2 жінки) х 500 висловлювань	Китайська	Платний доступ
The Interactive EmotionalDyadic Motion CaptureDatabase (IEMOCAP)	Радість, гнів, смуток, розчарування, нейтральна	10 ораторів (5 чоловіків, 5 жінок)1150 висловлювань	Англійська	Ліцензований доступ
Chinese Annotated Spontaneous Speech corpus (CASS)	Гнів, страх, радість, смуток, здивування, нейтральна	7 ораторів (2 чоловіки, 5 жінок), 6 годин мовлення	Китайська	Платний доступ

Chinese Natural Emotional Audio–Visual Database(CHEAVD)	Гнів, тривога, відраза, радість, нейтральна, смуток, здивування, занепокоєння	238 ораторів (від дітей до літніх) емоційні моменти з фільмів та теле-шоу	Китайська	Безкоштовно для досліджень
Danish Emotional SpeechDatabase (DES)	Нейтральна, здивування, гнів, радість, смуток	4 оратори (2 чоловіки, 2 жінки)10 хв мовлення	Датська	Безкоштовно
Chinese Elderly Emotional Speech Database (EESDB)	Гнів, відраза, страх, радість, нейтральна, смуток, здивування	16 ораторів (8 чоловіків, 8 жінок), 400 висловлювань з телегри	Китайська	Безкоштовно для досліджень
Electromagnetic Articulography Database (EMA)	Гнів, радість, смуток, нейтральна	3 оратори (1 чоловік, 2 жінки)	Англійська	Безкоштовно
Italian Emotional Speech Database (EMOVO)	Відраза, радість, страх, гнів, здивування, смуток, нейтральна	6 ораторів (3 чоловіки, 3 жінки) х 14 речень х 7 емоцій — 588 висловлювань	Італійська	Безкоштовно
Keio University Japanese Emotional Speech Database (Keio-ESD)	Гнів, радість, відраза, веселість, стурбованість, ніжність, полегшення, обурення тощо (47 емоцій)	71 оратор (чоловіки) 940 висловлювань	Японська	Безкоштовно
LDC Emotional Speech Database	Гнів, відраза, страх, презирство, радість, смуток, нейтральна, паніка, гордість, відчай, піднесення, інтерес, сором, нудьга	7 ораторів (4 чоловіка, 3 жінки), 470 висловлювань	Англійська	Платний доступ

RECOLA Speech Database	5 соціальних форм поведінки, збудження і валентність	46 ораторів (19 чоловіків, 27 жінок) 7 год мовлення	Французька	Безкоштовно
Speech Under Simulated and Actual Stress Database (SUSAS)	Декілька станів під впливом стресу: нейтральний, злий, кричущий	32 оратори (19 чоловіків, 13 жінок), 16000 висловлювань, включаючи розмови пілотів гелікоптерів Апачі	Англійська	Платний доступ
Vera Am Mittag Database (VAM)	Валентність, активація та домінування	47 оратори з ток-шоу, 947 висловлювань	Німецька	Безкоштовно
FAU Aibo Emotion Corpus	Гнів, нудьга, рішучість, безпорадність, радість, моторошність, нейтральність, доганяючий, розслабленість, здивування, вразливість	51 дитина, 9 годин мовлення	Німецька	Платний доступ
RAVDESS	Гнів, відраза, нейтральна, страх, радість, смуток, спокій, здивування	24 актори (12 жінок, 12 чоловіків)	Англійська	Безкоштовно

1.5 Новизна дослідження

У попередніх підрозділах було проведено аналіз літератури в області систем розпізнавання емоцій людини за голосом. З розглянутих досліджень можна зробити висновки, що більшість існуючих методів вже були експериметнально досліджені, тому малоймовірним є використання якогось

нового методу в даній предметній області. Також вже були досліджені майже всі можливі поєднання ознак та класифікаторів.

Тому логічним рішенням є вибрати такі класифікатор та набір ознак, що продемонстрували найкращі результати в попередніх роботах, після чого знайти спосіб покращити процес класифікації.

В якості класифікатора одним з найкращих рішень можна назвати використання глибинних нейронних мереж. А в якості ознак були обрані мел-частотні кепстральні коефіцієнти.

Пропонується покращити процес класифікації за допомогою ускладнення архітектури нейронної мережі та використання декількох наборів даних. В якості класифікатора було обрано згорткову нейронну мережу. Даний вид глибинних нейромереж не так часто використовувався для вирішення задачі розпізнавання емоції за голосом, як інші методи. Крім того, згорткові нейронні мережі дозволяють виконувати гнучке та точне налаштування, що дає більші можливості для покращення результатів.

В якості наборів даних пропонується вибрати датасети RAVDESS та TESS. Вони безкоштовні для наукового використання, записані англійською мовою, мають розгорнуту документацію та варіативність у даних.

Використання двох датасетів має на меті покращити процес навчання нейронної мережі, уникаючи перенавчання. Два набори даних містять аудіозаписи з різними параметрами (наприклад, у датасеті RAVDESS вік акторів 21–33 роки, а в датасеті TESS — 26–64 роки). Параметр віку є вкрай важливим для систем розпізнавання емоцій, тому що вираження тієї чи іншої емоції під час мовлення спричинено фізичними особливостями будови голосового тракту людини. Через це з віком змінюється як сам голос, так і патерни, що дозволяють визначати емоції. Тому використання двох датасетів дозволить урізноманітнити навчальну вибірку та збільшити точність класифікації.

Новизною даної роботи є поєднання згорткових нейронних мереж, мел-частотних кепстральних коефіцієнтів та наборів даних RAVDESS і TESS, оскільки проведений аналіз доступних літературних джерел не виявив подібних

рішень, і дана розробка проводиться вперше. Окрім того, середня точність класичних класифікаторів становить не більше *0,75*, а запропонована модифікація збільшує точність.

РОЗДІЛ 2 ТЕХНОЛОГІЯ РОЗПІЗНАВАННЯ ЕМОЦІЙ ЗА ГОЛОСОМ

2.1 Етап передобробки

Передобробка — це перший крок після збору даних, який буде виконано для навчання класифікатора в системі розпізнавання емоцій за голосом. Процес передобробки складається з фреймінгу, віконного перетворення, нормалізації, зменшення шуму. Фреймінг, також відомий як сегментація мови, виконується для розділення неперервної мови на сегменти з фіксованою довжиною. Емоції можуть змінюватися в процесі мовлення, однак залишаються незмінними протягом короткого проміжку часу. Цим зумовлена довжина сегментів у 20–30 мс.

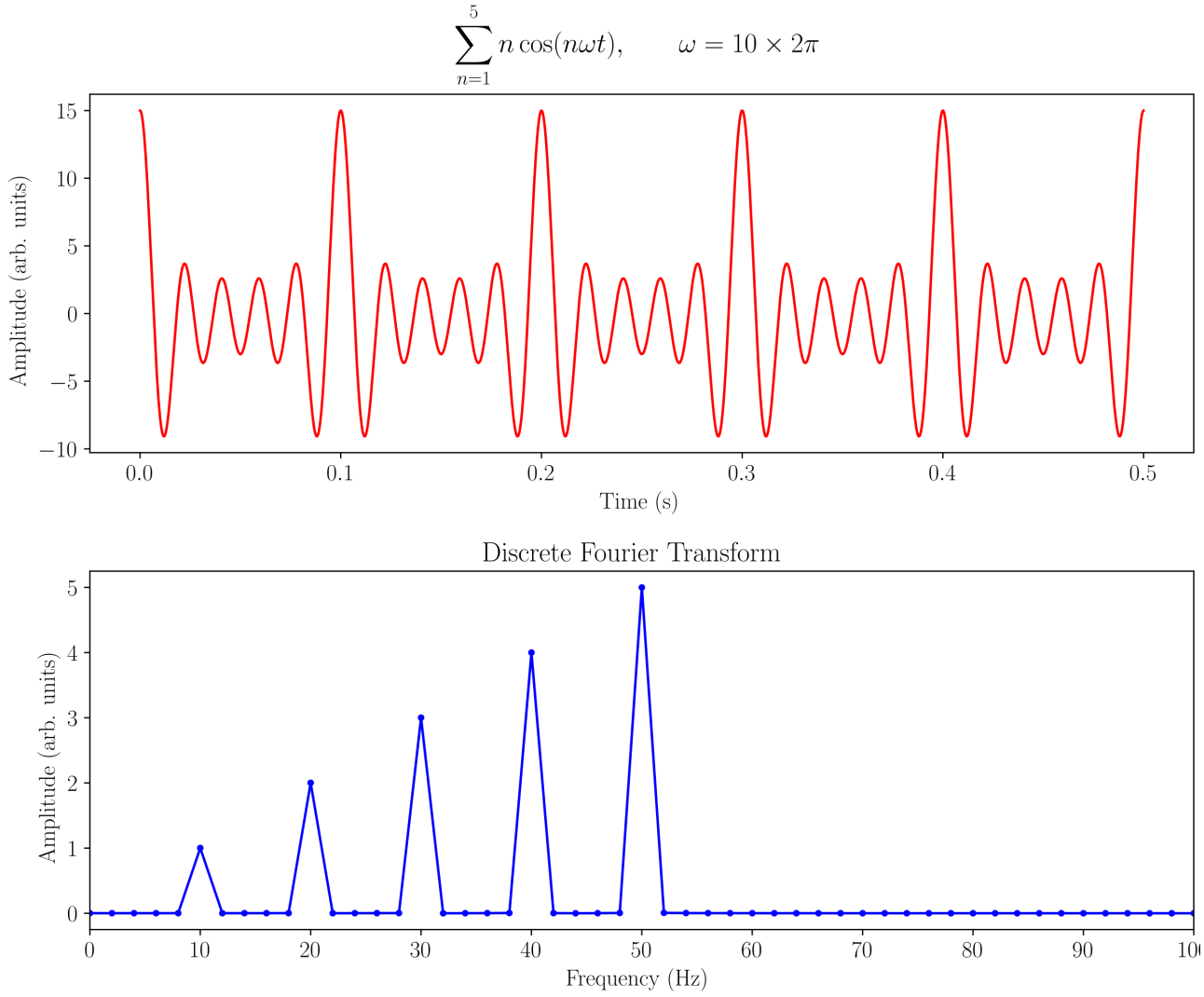


Рисунок 3 — Приклад перетворення Фур’є

Після цього виконується віконне перетворення Фур’є (рисунок 3).

Для цього використовується метод зважування вікном Хеммінга:

$$w(n) = 0.54 - 0.46cos(\frac{2\pi n}{M-1}),\ 0 \leq n \leq M-1,$$

де M — розмір вікна $w(n)$.

Будь-який аудіозапис мовлення складається з вокалізованого мовлення, невокалізованого мовлення та тиші. Вокалізована мова породжується вібрацією голосових зв'язок, що створює періодичне збудження голосових шляхів під час вимови фонем, які сприймаються нами як одиниці звуку, що відрізняють одне слово від іншого.

З іншого боку, невокалізоване мовлення є результатом проходження повітря через звуження в голосовому тракті. Причому вокалізоване мовлення є періодичним, на відміну від невокалізованого мовлення. Розпізнавання вокалізованого мовлення серед інших звуків та тиші називається виявленням кінцевих точок. Ефективність алгоритму виявлення кінцевих точок впливає на точність системи. Зазвичай використовується показник переходу через нуль (zero-crossing rate), який визначає, як швидко сигнал змінюється з позитивного на негативний.

Для етапу нормалізації використовується метод z-нормалізації. При середньому значенні μ та стандартному відхиленні σ, нормалізація розраховується за формулою:

$$z = \frac{x - \mu}{\sigma}$$

В реальному житті шум, присутній у навколишньому середовищі, фіксується разом із мовним сигналом. Це впливає на швидкість та точність розпізнавання, отже, для усунення або зменшення шуму необхідно використовувати спеціальні методи. Мінімальна середньоквадратична похибка

(MMSE) та логарифмічно-спектральні амплітудні MMSE (LogMMSE) — найбільш успішно застосовувані методи зменшення шуму [21].

Метод MMSE розраховує вихідний сигнал на основі даної функції зашумленого сигналу. Для цього потрібно знати апріорну інформацію про спектр мовлення та шуму. Метою методу є мінімізація показника спотворення між чистим та оціночним сигналом.

2.2 Вилучення ознак

Вилучення ознак є вкрай важливим процесом, адже таким чином вдається перетворити аудіофайли у формат, зрозумілий для моделей. Ретельно розроблений набір ознак, які успішно характеризують кожну емоцію, збільшує рівень розпізнавання. Не дивлячись на те, що для систем розпізнавання емоцій використовуються різні набори ознак, загальновизнаного консенсусу не існує. Всі існуючі зараз дослідження були виключно експериментальними.

Мовлення є неперервним сигналом різної довжини, який містить як інформацію, так і емоції. Тому глобальні чи локальні ознаки можна вилучати в залежності від підходу. Глобальні ознаки, які також називаються довгостроковими або надсегментними, по-суті є статистичними показниками, такими як середнє, мінімальне/максимальне значення та стандартне відхилення. Локальні ознаки, також відомі як короткострокові або сегментарні ознаки, мають за мету наближення до деякого стаціонарного стану. Ці стаціонарні стани є важливими, оскільки емоції не розподілені рівномірно по всьому мовленню [22]. Наприклад, такі емоції, як гнів, переважають на початку висловлювання, тоді як здивування переважно передається в кінці висловлювань.

Ці локальні та глобальні ознаки розділяються на чотири категорії:

1. просодичні ознаки;
2. спектральні ознаки;
3. ознаки якості голосу;
4. ознаки, що базуються на операторі енергійності Тигера (TEO).

Найчастіше в системах розпізнавання емоцій за голосом використовують просодичні та спектральні ознаки. Деякі ознаки різні дослідники відносять до різних категорій. Ознаки TEO спеціально розроблені для розпізнавання стресу та гніву.

2.2.1 Просодичні ознаки

Просодичні ознаки можуть бути сприйняті людиною. До них відносять інтонацію та ритм. Типовим прикладом для англійської мови є підвищення інтонації у висловлюванні, яке є питанням. Також ці ознаки відомі як паралінгвістичні, тому що вони є властивостям великих одиниць мови, таких як слова, словосполучення та речення. Через це просодичні ознаки є довгостроковими.

Найбільш широко використовувані просодичні ознаки засновані на фундаментальній частоті, інтенсивності та тривалості. Фундаментальна частота задається вібраціями в голосовому каналі. Вона задає ритмічні та тональні характеристики мови. Зміна фундаментальної частоти у висловлюванні задає її межі, що надалі може бути використано як ознака. Інтенсивність мовного сигналу, яку іноді називають гучністю або "енергією", вказує на коливання амплітуди мовних сигналів у часі. Дослідники припускають, що емоції сильного збудження, такі як гнів, радість чи здивування, призводять до збільшення інтенсивності, тоді як відраза та смуток призводять до її зменшення [23]. Тривалість — це кількість часу, яка потрібна на побудову складів, слів та інших конструкцій, присутніх у мовленні. Швидкість мовлення, тривалість мовчання, тривалість вокалізованих та невокалізованих звуків — одні з найбільш широко використовуваних характеристик, пов’язаних з тривалістю.

Існують кореляційні зв’язки між просодичними ознаками та емоційними станами. Просодичні ознаки вказують на зміни під час емоційного мовлення. Наприклад, під час демонстрацій таких емоцій, як гнів, страх та радість, фундаментальна частота та інтенсивність мають тенденцію до збільшення.

Емоція, для якої характерне низьке збудження, наприклад, смуток, дає нижчі значення фундаментальної частоти. А тривалість вираження гніву коротша, ніж тривалість вираження смутку [24, 25].

2.2.2 Спектральні ознаки

Коли людина вимовляє звуки, вони змінюються в залежності від форми голосового тракту. Справа в тому, що характеристики голосового тракту напряму впливають на частоту звуку [22]. Спектральні ознаки отримують за допомогою перетворення Фур'є. Вони вилучаються з мовних сегментів довжиною від 20 до 30 мілісекунд.

Ознака MFCC (мел-частотні кепстральні коефіцієнти від англ. "mel-frequency cepstral coefficients") — це короткострокова спектральна величина потужності мовного сигналу. Для отримання MFCC висловлювання діляться на сегменти, потім кожний сегмент перетворюється в частотну область за допомогою дискретного перетворення Фур'є. Далі розраховуються значення логарифма кожної області, після чого виконується дискретне косинусне перетворення. Коефіцієнтами MFC будуть амплітуди результуючого спектру. Блок схема отримання кепстральних коефіцієнтів із аудіосигналу зображена на рисунку 4 (наступна сторінка).

Мел-частотні кепстральні коефіцієнти часто використовують у різноманітних системах розпізнавання мовлення, наприклад, для систем розпізнавання цифр, що промовляються в телефон.

Згідно з аналізом [26], мел-частотні кепстральні коефіцієнти є найпоширенішою спектральною ознакою під час використання машинного навчання для роботи з аудіофайлами.

Ознака LPCC (Linear Prediction Cepstral Coefficients) також базується на характеристиках голосового тракту мовця. Ці характеристики вказують на різницю в емоційному стані. LPCC може бути безпосередньо отримана за допомогою рекурсивного методу лінійного коефіцієнта прогнозування (LPC).

Коефіцієнти частоти гаматонів (GFCC) також являються спектральною ознакою, яку отримують аналогічною методикою до вилучення MFCC. Однак для даної ознаки застосовують фільтр гаматонів.

Форманти — це частоти акустичного резонансу голосового тракту. Вони обчислюються як вершини амплітуд в частотному спектрі звуку. Вони визначають фонетичну якість голосних звуків, тому використовуються для розпізнавання голосних.

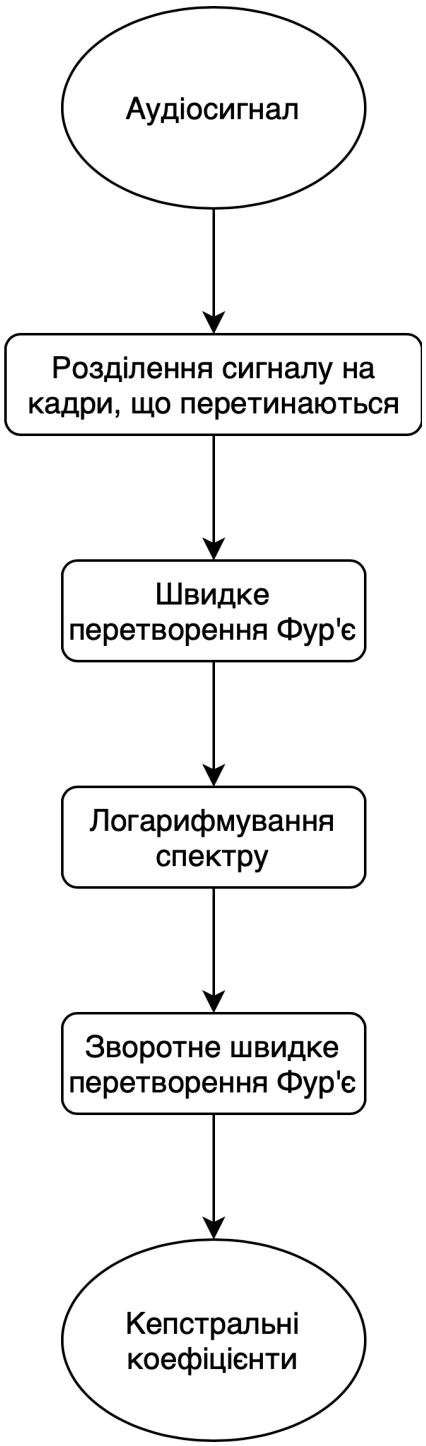


Рисунок 4 — Блок-схема отримання кепстральних коефіцієнтів

2.2.3 Ознаки якості голосу

Якість голосу визначається фізичними властивостями голосового тракту. Незначні зміни можуть стати причиною такого мовного сигналу, за яким можливо визначити емоцію, використовуючи такі властивості, як джитер (тріпотіння), мерехтіння та гармоніку відносно шуму. Існує сильна кореляція між якістю голосу та емоційним змістом мови [2].

Джитер — це спотворення фундаментальної частоти, тоді як мерехтіння є зміною амплітуди. Тому джитер — міра частотної нестабільності, а мерехтіння — нестабільність амплітуди. Відношення гармонік до шуму показує відносний рівень шуму в частотному спектрі голосних звуків.

Серед інших ознак якості голосу, що зустрічаються в літературі, варто виділити коефіцієнт нормалізованої амплітуди (NAQ), різницю в амплітуді

перших двох гармонік (H1H2), коефіцієнт максимальної дисперсії (MDQ), параболічний спектральний параметр (PSP) та параметр форми моделі Лілдженкрантса-Фанта (Liljencrants-Fant model) [27].

Існують також якісні мовні ознаки, такі як жорсткість, суворість, напруженість, важке дихання, шепіт, скрипучість голосу. Ці ознаки мають високу кореляцію зі сприйманими емоціями. Однак, ці ознаки дуже важко вилучати та інтерпретувати [28]. В роботі Лавера [29] демонструється, як взаємопов'язані важке дихання з близькістю, жорсткий голос з гнівом, шепіт з конфіденційною діяльністю, скрипучий голос з нудьгою. Напружений голос пов'язаний з гнівом, страхом та радістю, а слабкий голос зі смутком [30].

2.2.4 TEO-ознаки

Оператор енергійності Тигера (TEO) використовується для виявлення стресу мовця [31, 32]. Як вважає Тигер, мова створюється завдяки нелінійній взаємодії з вихровим потоком у гортанній системі людини. Стресова ситуація впливає на напругу м'язів мовця, що призводить до зміни потоку повітря під час формування звуку.

На основі теорії Тигера були запропоновані три нові ознаки, які все ще використовують ідею зміни потоку повітря в голосовому тракті [33]. Ці ознаки були порівняні з ознаками MFCC. Загалом, використання цих ознак показало гірші результати за MFCC та використання висоти (рівня) звуку в якості ознаки. Проте ознака TEO-CB-Auto-Env показала кращі результати за інші просодичні, спектральні ознаки, ознаки якості голосу, а також їх комбінації [8].

Отже, можна зробити висновок, що найуживанішими методами вилучення ознак є MFCC, LPCC, LFPC, Log-Mel спектрограми, адже вони дозволяють ефективно працювати саме з аудіосигналами. Однак, в більшості досліджень найкраще себе проявили саме MFCC ознаки. З ними зручно працювати, вони не потребують залучення професіоналів в галузі психології, а також їх підтримують сучасні бібліотеки, тому в роботі використовуються саме вони.

2.3 Аналіз класифікаторів

Системи розпізнавання емоцій за голосом класифікують основні емоції даного висловлювання. Для виконання цього завдання використовують традиційні класифікатори, алгоритми глибинного навчання, багато алгоритмів машинного навчання. Однак, як і з будь-якою складною проблемою, не існує загальновизнаного алгоритму машинного навчання, який можна було б використовувати. Поточні дослідження, як правило, є емпіричними. У таблиці 1 вже були наведені дослідження та використані в них бази даних, ознаки, класифікатори і результати експериментів. Далі наводиться розширений огляд методів класифікації.

2.3.1 Традиційні класифікатори

Системи розпізнавання емоцій за голосом зазвичай використовують класифікаційні алгоритми. Алгоритм класифікації вимагає введення *X*, виходу *Y* та функції, яка відображає *X* у *Y*, тобто □ *(□)* = □. Алгоритм навчання наближає функцію відображення, яка допомагає передбачити клас нового введення. Алгоритму навчання потрібні марковані дані, що ідентифікують зразки та їх класи. Після закінчення навчання дані, які не використовувались під час навчання, використовуються для перевірки ефективності класифікатора.

Найбільш часто використовуваними алгоритмами є приховані марковські моделі, змішані гаусівські моделі (Gaussian mixture model), метод опорних векторів та штучні нейронні мережі. Існують також методи класифікації, засновані на деревах прийняття рішень, метод *k*-найближчих сусідів, метод *k*–середніх та наївний байєсівський класифікатор. Інколи також використовуються поєднання декількох класифікаторів для досягнення кращих результатів.

Прихована марковська модель — це загальновживаний метод розпізнавання мови, який також був успішно використаний для розпізнавання емоцій. Як слідує з назви, модель розглядається як марковський процес із неспостережуваними (прихованими) станами. Марковський процес — це

випадковий процес, конкретні значення якого для будь-якого заданого часового параметру $t + 1$ залежать від значення у момент часу t, але не залежать від його значень у момент часу $t - 1$. Тобто «майбутнє» процесу залежить лише від «поточного» стану, але не залежить від «минулого». [34]

Ногейрас [35] використовував низькорівневі ознаки висоти, інтенсивності та їх контури з прихованими напівнеперервними марковськими моделями. Було отримано точність розпізнавання понад 0,7 для 6 класів емоцій: радість, гнів, задоволення, страх, відраза, смуток.

Шулер [36] порівняв два методи. У першому методі висловлювання класифікуються за допомогою гаусівської змішаної моделі, використовуючи глобальні статистики ознак, отриманих із вихідного тону та контуру інтенсивності мовного сигналу. У другому методі прихована марковська модель застосовується з використанням миттєвих ознак низького рівня, а не глобальних статистик. Середня точність розпізнавання семи дискретних класів емоцій перевищувала 0,86, використовуючи глобальну статистику, тоді як рівень розпізнавання того ж самого корпусу випадковими людьми (волонтерами), становив 0,79.

Було показано [37], що ознаки LFPC в прихованій марковській моделі дають кращі показники, ніж MFCC чи LPCC. Було досягнуто рівня розпізнавання 0,77 та 0,89 для середнього та найкращого рівня розпізнавання відповідно, тоді як точність розпізнавання тих ще аудіофайлів людьми становила 0,65.

В роботі [38] використовувались приховані марковські моделі та метод опорних векторів для класифікації п’яти емоцій: гнів, радість, смуток, здивування та нейтральний стан. Точність марковських моделей виявилася вищою за метод опорних векторів.

Гаусівська змішана модель — це ймовірнісний метод, який можна розглядати як окремий випадок неперервної марковської моделі, що містить лише один стан. Ідея даної моделі полягає у моделюванні даних з точки зору змішування декількох компонентів, де кожен компонент має просту параметричну форму, таку як функція Гауса. Передбачається, що кожна точка

даних належить одному з компонентів, і робиться припущення про розподіл для кожного компонента окремо. При порівнянні ознак MFCC та висоти голосу на гаусівській змішаній моделі, було створено окрему модель для кожної з ознак (MFCC, низькорівневі MFCC, ознаки висоти голосу). Було виявлено, що найкращі результати дає комбінація класифікаторів з використанням акустичних ознак [39].

Штучні нейронні мережі є загальновживаним методом для кількох видів класифікаційних задач. Базова архітектура складається з вхідного шару, одного чи декількох прихованих шарів та вихідного шару. Шари складаються з нейронів, причому кількість нейронів вхідного та вихідного шарів залежить від даних та позначених класів, а приховані шари можуть мати значно більшу (за потреби), кількість нейронів. Кожен шар поєдано з наступним за допомогою ваг, які спочатку були вибрані випадковим чином. Коли обрано навчальні дані, їх значення завантажуються на вхідний рівень, а потім пересилаються на наступний рівень. На вихідному рівні ваги оновлюються за допомогою алгоритму зворотного розповсюдження помилки. Після закінчення навчання очікується, що ваги зможуть класифікувати нові дані.

Метод опорних векторів — це класифікатор для навчання з вчителем, який знаходить оптимальну гіперплощину для лінійного відокремлення класів. Основна ідея методу — переведення вихідних векторів в простір більш високої розмірності і пошук розділяючої гіперплощини з найбільшим проміжком в цьому просторі. Дві паралельні гіперплощини будуються по обидва боки гіперплощини, що розділяє класи. Розділяючою гіперплощиною в такому разі буде гіперплощина, що створює найбільшу відстань до двох паралельних гіперплощин. Алгоритм заснований на припущенні, що чим більша різниця або відстань між цими паралельними гіперплощинами, тим меншою буде середня похибка класифікатора.

Ансамбль методів — це поєднання алгоритмів для підвищення ефективності прогнозування. Як правило, алгоритми поєднуються за допомогою процедури голосування. Ефективність ансамблю зазвичай вища за ефективність

окремих класифікаторів. Існують різні види архітектур ансамблевих методів, наприклад, подача на вхід одних і тих самих даних кожному окремому класифікатору, після чого результати порівнюються і вибирається кращий класифікатор. Іншим підходом є використання ієрархічного класифікатора, коли вхідні дані надходять до одного алгоритму, потім результат першого надходить на вхід до іншого типу класифікатора і т. д.

2.3.2 Класифікатори на основі глибинного навчання

Більшість алгоритмів глибинного навчання базуються на штучних нейронних мережах, тому їх зазвичай називають глибинними нейронними мережами. Така назва походить від наявності прихованих шарів, так як їх кількість може сягати сотень, в той час як традиційна нейронна мережа містить зазвичай не більше пари прихованих шарів. В останні роки ефективність алгоритмів глибинного навчання перевершує традиційні алгоритми машинного навчання, тому дослідники почали приділяти більшу увагу саме їм, та ця тенденція в дослідженнях розпізнавання емоцій за голосом зберігається і нині. Перевага деяких з цих алгоритмів полягає у відсутності необхідності в етапах вилучення ознак. Ознаки вибираються частково автоматично за допомогою алгоритмів глибинного навчання. Найбільш широко використовуваними алгоритмами глибинного навчання є згорткові нейронні мережі та рекурентні нейронні мережі.

Рекурентні нейронні мережі — це клас штучних нейронних мереж, в яких зв’язки між елементами утворюють направлену послідовність, тому дані мережі спеціалізуються на обробці послідовних даних. Використовуючи внутрішню пам’ять, вони можуть запам’ятовувати отримані вхідні дані та робити точні припущення. Рекурентні мережі успішно використовуються для таких послідовних даних, як мова, текст, відео. Останнім часом набули широкого поширення мережі з довгою короткочасною пам’яттю та мережі на основі вентильних рекурентних вузлів.

Коли нейрон рекурентної мережі видає вихід, дані пересилаються до наступного нейрону, а також циклічно повертають вихід назад. Як результат, така мережа має два типи вхідних даних: поточний вхід і вхід із недавнього минулого. Вхідні дані недавнього минулого важливі, оскільки послідовність даних містить важливу інформацію про те, що буде в майбутньому.

Рекурентні мережі мають короткочасну пам’ять, однак, використовуючи архітектуру з довгою короткочасною пам’яттю, мережа може отримати доступ і до довгострокової пам’яті. Такі мережі мають спеціальні "комірки", які мають внутрішню рекурентність, окрім зовнішньої рекурентності всієї мережі. До того ж, крім стандартних вхідних і вихідних даних, мережа має більше параметрів і блоків керування з сигмоїдною функцією, що контролює потік інформації. Мережі з довгою короткочасною пам’яттю мають три типи вентилів: вхідні, вентилі "забування" та вентилі "запам’ятовування". Відкриваючи та закриваючи ці вентилі, комірка приймає рішення про те, що зберігати, чи коли дозволяти входи, виходи та видалення.

Згорткові нейронні мережі — це особливі типи нейронних мереж, призначені для обробки даних, що мають сітчасту топологію, таких як зображення. Застосовуючи кілька відповідних фільтрів, така мережа може успішно фіксувати часові та просторові залежності. Вхідні дані зменшуються у розмірі без втрати своїх ознак, таким чином обчислювальна складність зменшується, а рівень точність алгоритму збільшується. Згорткова мережа зазвичай складається з декількох шарів: шару згортки, шару агрегування та повнозв’язного шару.

Шар згортки використовується для вилучення високорівневих ознак із вхідних даних. З математичної точки зору, згортка означає поєднання двох функцій для отримання третьої. На виході шару згортки отримується карта ознак.

Шар агрегування (pooling) використовується для зменшення розміру згорткових ознак заради зменшення обчислювальної складності за рахунок

зменшення розмірності. Це корисно для вилучення головних ознак із вхідних даних.

В роботі [54] наводиться схема (рисунок 5) архітектури згорткової нейронної мережі для роботи з аудіоданими. В якості ознак мережа на вхід приймає значення кепстральних коефіцієнтів.

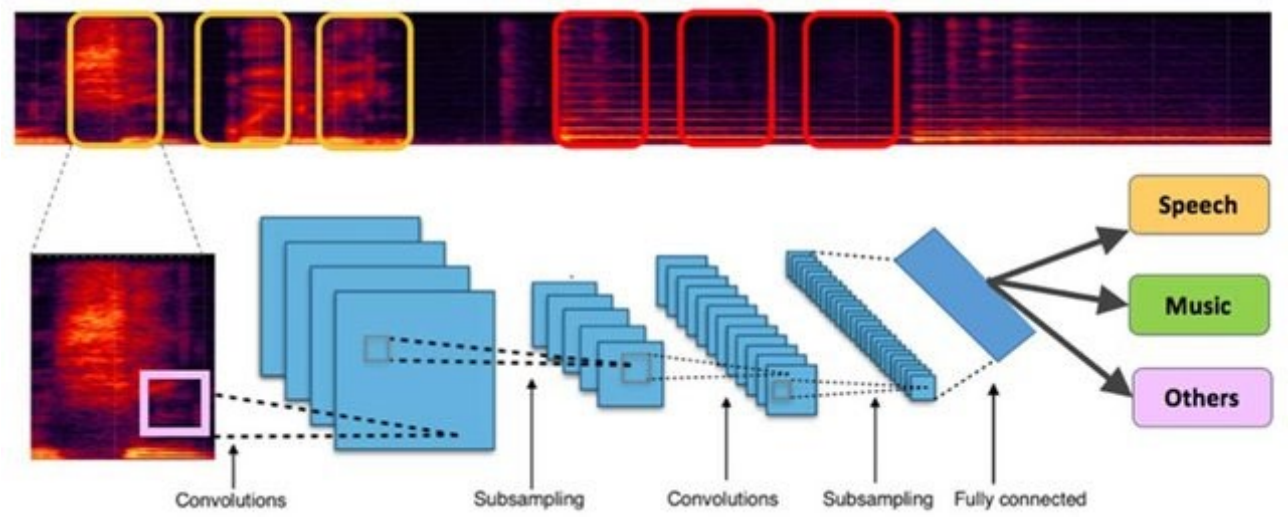


Рисунок 5 — Архітектура згорткової мережі для роботи з аудіоданими [54]

Після передачі вхідних даних через декілька шарів згортки та агрегування і вилучення високорівневих ознак, отримані ознаки використовуються як вхідні дані для повнозв’язного шару шляхом зглажування двовимірних даних до одновимірного масиву та надсилання його до мережі прямої передачі похибки, яка працює як звичайна нейронна мережа.

В роботі [40] запропоновано систему, яка поєднує згорткову нейронну мережу з мережею з довгою короткочасною пам’яттю, де згорткова мережа використовується для автоматичного вивчення найкращих описових характеристик мовного сигналу безпосередньо з "сирих" даних. Спочатку голосові сигнали сегментуються, потім виконується зменшення шуму в якості передобробки. Після цього за допомогою згорткових мереж вилучаються акустичні ознаки, і, нарешті, отримані ознаки передаються в мережу з довгою короткочасною пам’яттю.

Подібний підхід було застосовано Лім [41]. Були порівняні згорткові нейронні мережі з рекурентними. Поєднанням двох методів вдалося отримати точність 0,88. В роботі [42] вдалося покращити базову мережу використанням тривимірних згорткових нейромереж.

2.3.3 Методи машинного навчання для покращення класифікації

Пошук розмічених даних є складним завданням для дослідження та застосування систем розпізнавання емоцій за голосом. Крім того, навіть якщо отримані марковані дані розмічені, немає жодних гарантій щодо коректності емоційних міток, оскільки не існує єдиної стандартизації. В останні роки автокодувальники привернули увагу дослідників завдяки можливості навчання без учителя.

Автокодувальник — це штучна нейронна мережа, що дозволяє виконувати навчання без учителя в поєднанні з методом зворотного розповсюдження похибки. Автокодувальники складаються з трьох шарів, як і інші нейронні мережі: вхідного й вихідного шарів однакового розміру та прихованих шарів, які містять менше нейронів, ніж вхідні та вихідні шари. Основний принцип роботи і навчання мережі автокодувальника — отримати на вихідному шарі результат, найбільш близький до вхідного. Щоб рішення не виявилося тривіальним, на проміжний шар автокодувальника накладають обмеження: проміжний шар повинен бути або меншої розмірності, ніж вхідний і вихідний шари, або штучно обмежується кількість одночасно активних нейронів проміжного шару — розріджена активація. Ці обмеження змушують нейромережу шукати узагальнення і кореляцію у вхідних даних, виконувати їх стиснення. Таким чином, мережа автоматично навчається виділяти з вхідних даних загальні ознаки, які кодуються в значеннях ваг штучної нейронної мережі. Так, при навчанні мережі на наборі різних вхідних зображень, мережа може самостійно навчитися розпізнавати лінії і смуги під різними кутами. Загальна схема (архітектура) роботи базового автокодувальника зображена на рисунку 6. Ліва

частина (вхідний шар разом з прихованими шарами) являє собою кодувальник, а права (приховані шари з вихідним шаром) є декодувальником.

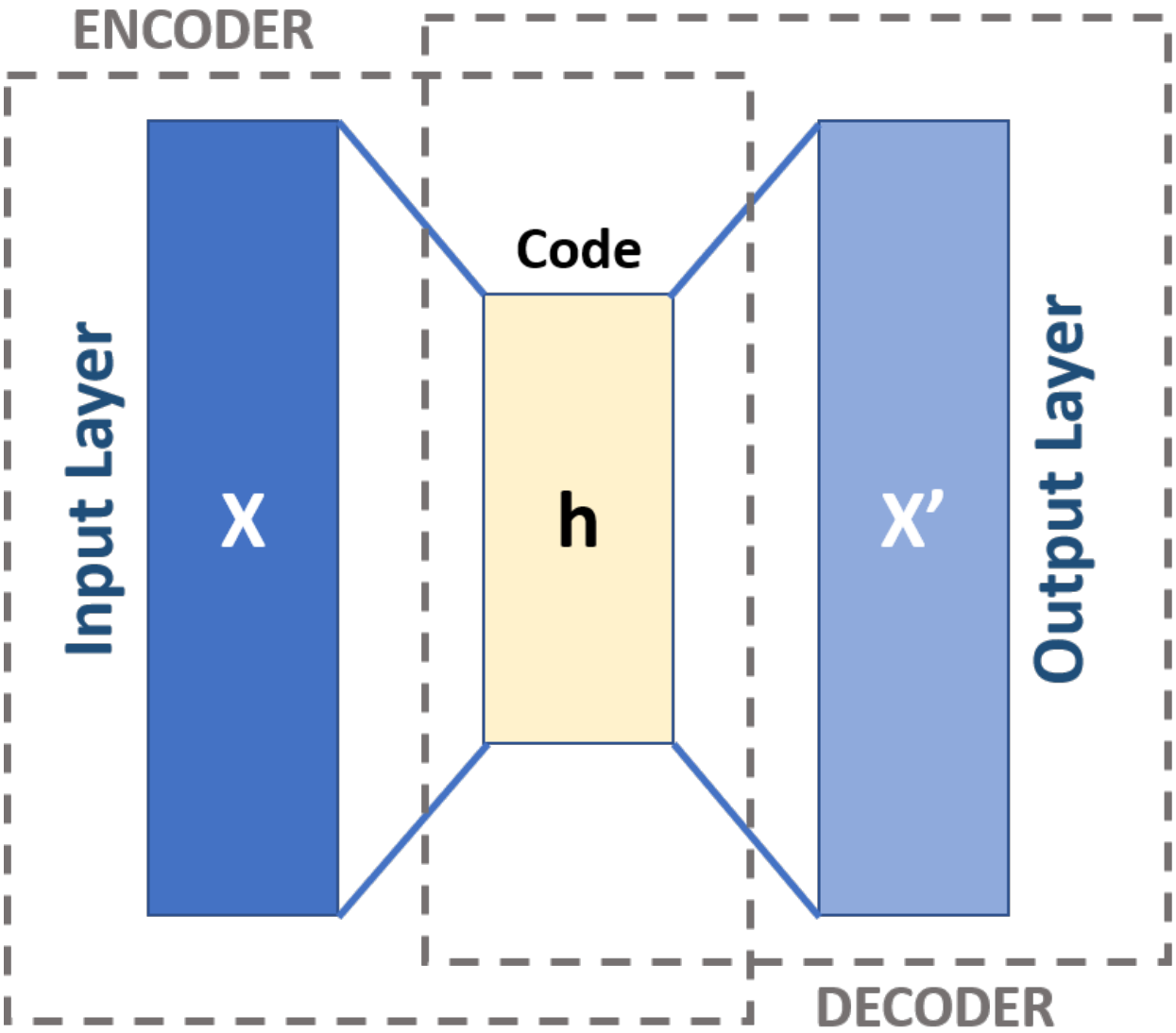


Рисунок 6 — Схема базового автокодувальника

Таким чином, автокодувальник працює в якості мережі, що вилучає ознаки, а не в якості класифікатора, тому лише після навчання автокодувальник підключається до класифікатора. Існує кілька типів автокодувальників: знешумлювальний автокодувальник, розріджений автокодувальник, варіаційний автокодувальник.

Знешумлювальний автокодувальник намагається досягти "гарного" представлення — такого, що буде отримано надійно навіть із зашумленого входу, і може стати корисним для відновлення первинного входу.

Розріджений автокодувальник може включати більше прихованих нейронів, аніж вхідних, проте лише невеликій частині нейронів дозволяється бути активними одночасно. Таке обмеження змушує модель реагувати на статистичні особливості навчальних даних.

Більшість систем розпізнавання емоцій за голосом зосереджені на навчанні з одним завданням, яке спрямоване на вивчення та прогнозування емоцій за висловлюванням. Однак деякі дослідження [43, 44, 45] показують, що багатозадачне навчання може покращити рівень розпізнавання. Багатозадачне навчання — це розділ машинного навчання, де кілька завдань розв’язуються одночасно. Використовується подібність у завданнях для покращення узагальнень. Це називається індуктивним перенесенням. Як правило, розпізнавання емоцій визначається як основне завдання, а кілька інших завдань, таких як стать чи спонтанність, вибираються в якості допоміжних. Результативність багатозадачного навчання напряму залежить від вибору підзадач.

В останні роки механізм уваги почав активно застосовуватися для глибинного навчання при вирішенні задачі розпізнавання емоцій людини за голосом [46, 47, 48]. Цей метод гарантує, що класифікатор "звертає увагу" на конкретне розташування даних зразків на основі ваги уваги для кожного набору вхідних даних. Як вже згадувалось, емоції під час мовлення розподіляються нерівномірно, а зосереджуються в окремих висловлюваннях. При розпізнаванні емоцій за голосом механізм уваги використовується для фокусування на емоційно значущій частині висловлювання.

Механізм уваги в машинному навчанні бере за основу когнітивну увагу. Ідея полягає у тому, щоб направити обчислювальні можливості комп’ютера на ті дані, які варти уваги. Які дані є більш важливими залежить від контексту, а тренування відбувається методом градієнтного спуску.

На рисунку 7 наведено діаграму, на якій зображено приклад автокодувальника з механізмом уваги. Це приклад моделі, яка займається перекладом з англійської мови на французьку.

Механізм уваги є повнозв'язною мережею. Ліва частина діаграми (чорним кольором) — це кодер-декодер, середня (помаранчевим кольором) — механізм уваги, права (сірим та іншими кольорами) — обчислювальні дані.

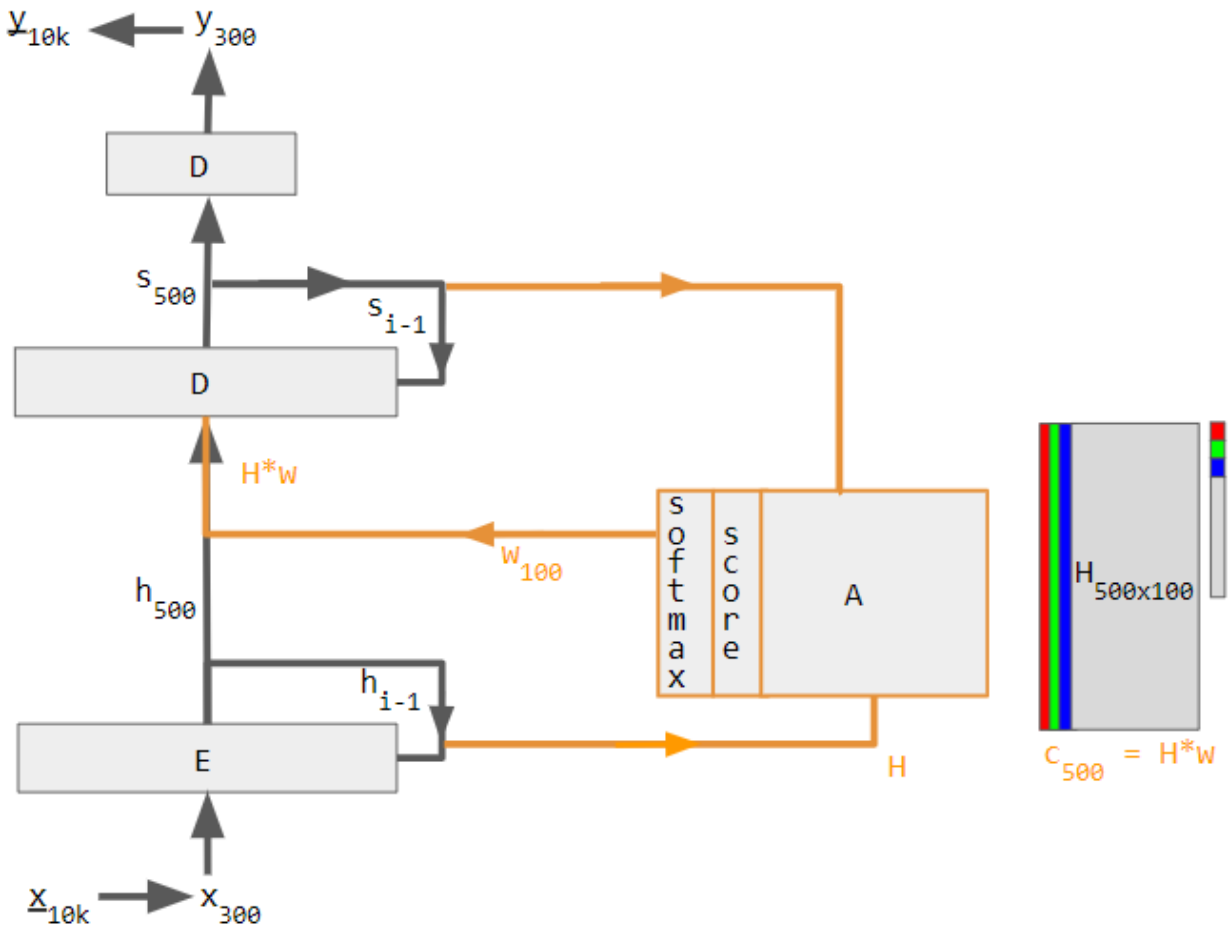


Рисунок 7 — Приклад кодувальника з механізмом уваги

Пошук розмічених даних, які будуть використані в навчанні для розпізнавання емоцій за голосом, є доволі складним процесом порівняно з таким завданням, як автоматичне розпізнавання мови. Низька кількість даних негативно впливає на швидкість розпізнавання через велику дисперсію. Одним

із методів, призначених для вирішення цієї проблеми, є передавальне навчання. Це техніка машинного навчання, де знання, отримані при вирішенні одного завдання, передаються для використання як відправна точка цільової моделі для іншого, але пов’язаного із попереднім завдання. Однак, щоб використовувати передавальне навчання, вихідна модель повинна бути досить загальною. Найпоширенішим підходом до навчання за допомогою трансферу є підготовка вихідної моделі з набором вихідних даних або використання попередньо навченої моделі, а потім використання вивчених знань як відправної точки для відповідного завдання. Проте така модель може потребувати додаткового опрацювання.

Техніка передавального навчання застосовується в таких алгоритмах, як приховані марковські моделі та байєсовські мережі. Галузі, в яких даний підхід вже було випробувано на практиці, включають розробку інтелектуальних помічників, класифікацію текстів, фільтрацію спам-повідомлень, розпізнавання медичних знімків.

В останні роки шкідливе (adversarial) навчання привернуло велику увагу дослідників машинного навчання [49, 50]. Шкідливе машинне навчання — це технологія у машинному навчання, яка намагається "обдурити" модель, подаючи на вхід оманливі дані. Таке навчання використовується для підвищення точності розпізнавання емоцій за голосом шляхом тренування моделей як на "гарних", так і на шкідливих даних. Значні зміни на виході моделі "караються" алгоритмом, а незначні стають частиною готового рішення.

Прикладами практичних галузей, в яких такий підхід знайшов своє застосування, є протидія спам-розсилкам (спам-повідомлення спеціально подають на вхід нейромережі в якості шкідливих даних), забезпечення безпеки комп’ютерних мереж (для навчання використовуються підроблені сертифікати безпеки), захист під час біометричної ідентифікації (спроба зловмисника скопіювати деякі ознаки та видати себе за іншу людину).

Такий підхід до машинного навчання набув популярності після декількох інцидентів, коли, наприклад, глибинні мережі виявились вразливими до зміни

одного пікселя вхідних даних, а автопілот Tesla помилився з вибором швидкості на 80 км/год, коли на знак обмеження швидкості наклеїли 5 см чорної плівки.

2.4 Аналіз можливостей бібліотек TensorFlow та Keras

Для розробки штучної нейронної мережі доцільно скористатися вже існуючим програмним пакетом, який дозволить спростити на пришвидшити розробку. Одним з найпопулярніших пакетів для машинного навчання є TensorFlow.

TensorFlow — це відкрита (open source) бібліотека для машинного навчання. Основне призначення — навчання глибинних нейронних мереж. Створена в компанії Google командою Google Brain та випущена для публічного доступу у 2015 році.

Схема екомистеми TensorFlow зображена на рисунку 8. Бібліотека Tensorflow доступна в чотирьох варіантах:

1. розробка на мові Python;
2. розробка на мові JavaScript;
3. розробка для мобільних платформ (iOS та Android) та вбудованих пристроїв (для IoT — Інтернету речей);
4. розширена версія TensorFlow для великих підприємств.

Крім того, TensorFlow можна використовувати за допомогою C API з багатьма мовами програмування: C++, Go, Java, Swift. Враховуючи простоту та велику кількість бібліотек, в якості мови програмування обрана мова Python та відповідна їй версія TensorFlow.

Для початку роботи з TensorFlow потрібно встановити пакет за допомогою команди:

pip install tensorflow

Далі у проекті на мові Python достатньо здійснити імпорт бібліотеки TensorFlow:

import tensorflow as tf

Після цього бібліотека TensorFlow повністю встановлена та готова до роботи.

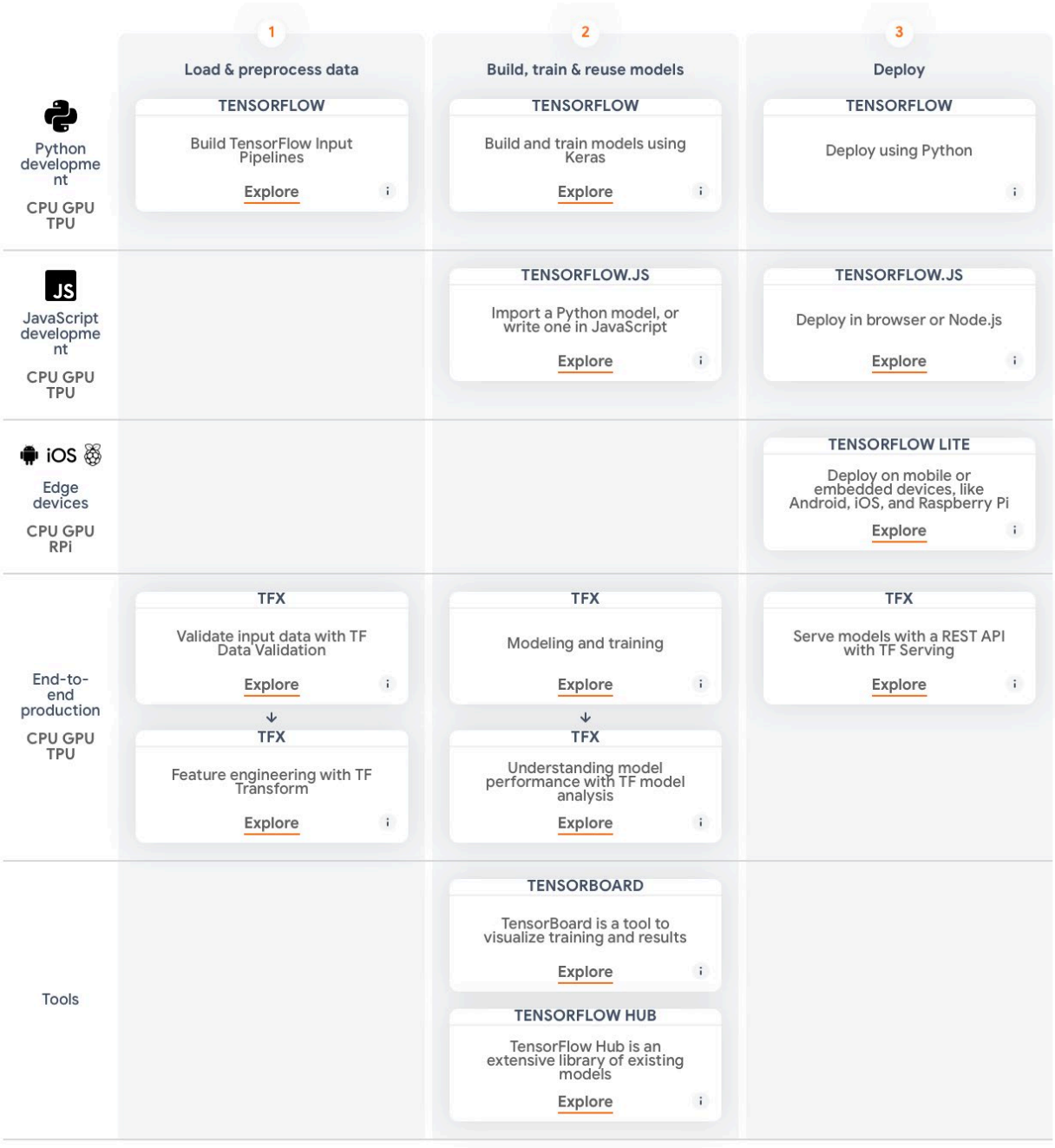


Рисунок 8 — Екосистема TensorFlow

Для роботи зі штучними нейронними мережами на мові Python найкращим рішенням буде використання поверх TensorFlow спеціальної бібліотеки Keras. Keras — відкрита бібліотека, що надає інтерфейс мовою Python для роботи зі

штучними нейронними мережами. Дана бібліотека служить інтерфейсом між TensorFlow та Python.

Перевагами бібліотеки Keras є:

1. вузька спеціалізація саме на глибинних нейронних мережах;
2. тісна інтеграція з TensorFlow;
3. модульність;
4. швидкість налаштування та простота роботи;
5. можливість швидко перевіряти ідеї та використовувати короткі ітерації в процесі розробки;
6. підтримка згорткових та рекурентних нейронних мереж.

Для роботи з Keras достатньо в існуючому проекті на мові Python та при завантаженому TensorFlow додати наступний код:

```
from tensorflow import keras
from tensorflow.keras import layers
```

Для створення згорткової нейронної мережі за допомогою бібліотеки Keras найкраще підходить модель Sequential (послідовна). Як зрозуміло з назви, вона дозволяє створювати послідовний набір шарів. Схематично, модель має наступний вигляд:

```
model = keras.Sequential(
  [
    layers.Dense(2, activation="relu"),
    layers.Dense(3, activation="relu"),
    layers.Dense(4),
  ]
)
```

В даному прикладі до моделі додано 3 повнозв’язні шари з двома, трьома та чотирма нейронами відповідно В якості функції активації виступає ReLU.

Також додавати шари до даної моделі можна інкрементально:

```
model = keras.Sequential()
model.add(layers.Dense(2, activation="relu"))
```

В цьому прикладі спочатку була створена порожня модель, а потім до неї був доданий повнозв’язний шар (в якості функції активації також використана ReLU).

2.5 Аналіз наборів даних RAVDESS та TESS

Для навчання нейронної мережі було вирішено використовувати набори даних RAVDESS [51] та TESS [52]. Розглянемо детальніше їх особливості.

Набір даних (або датасет) RAVDESS (The Ryerson Audio-Visual Database of Emotional Speech and Song) складається з аудіо- та відеозаписів, виконаних професійними акторами, що працюють в Торонто, Канаді. Записи зроблені 24-ма акторами. Половина акторів — чоловіки, половина — жінки. Середній вік акторів — 26 років (від 21 року до 33 років). 20 акторів ідентифікують себе як представників європеоїдної раси, двоє — як східно-азійців, двоє — змішаний тип.

Створення набору даних відбувалось в декілька етапів:

1. Прослуховування акторів (було прослухано 58 акторів).
2. Найм 24 акторів із 58 (оцінка проводилась трьома фахівцями).
3. Студійний запис:
    1. запис діалогів;
    2. запис емоційних висловлювань;
    3. повторний запис.
4. Вибір найкращих записів трьома експертами.
5. Пост-продакшн:
    1. вивантаження файлів в Adobe Premiere;
    2. нормалізація аудіодоріжок;
    3. автокорекція тону для записів співу (не більше 5%);
    4. експорт в аудіо- та відео-формати.
6. Валідація з допомогою 247-ми волонтерів.

Всього датасет містить 7356 файлів (обсягом 24.8 Гб). Діаграма структури набору даних представлена на рисунку 9.

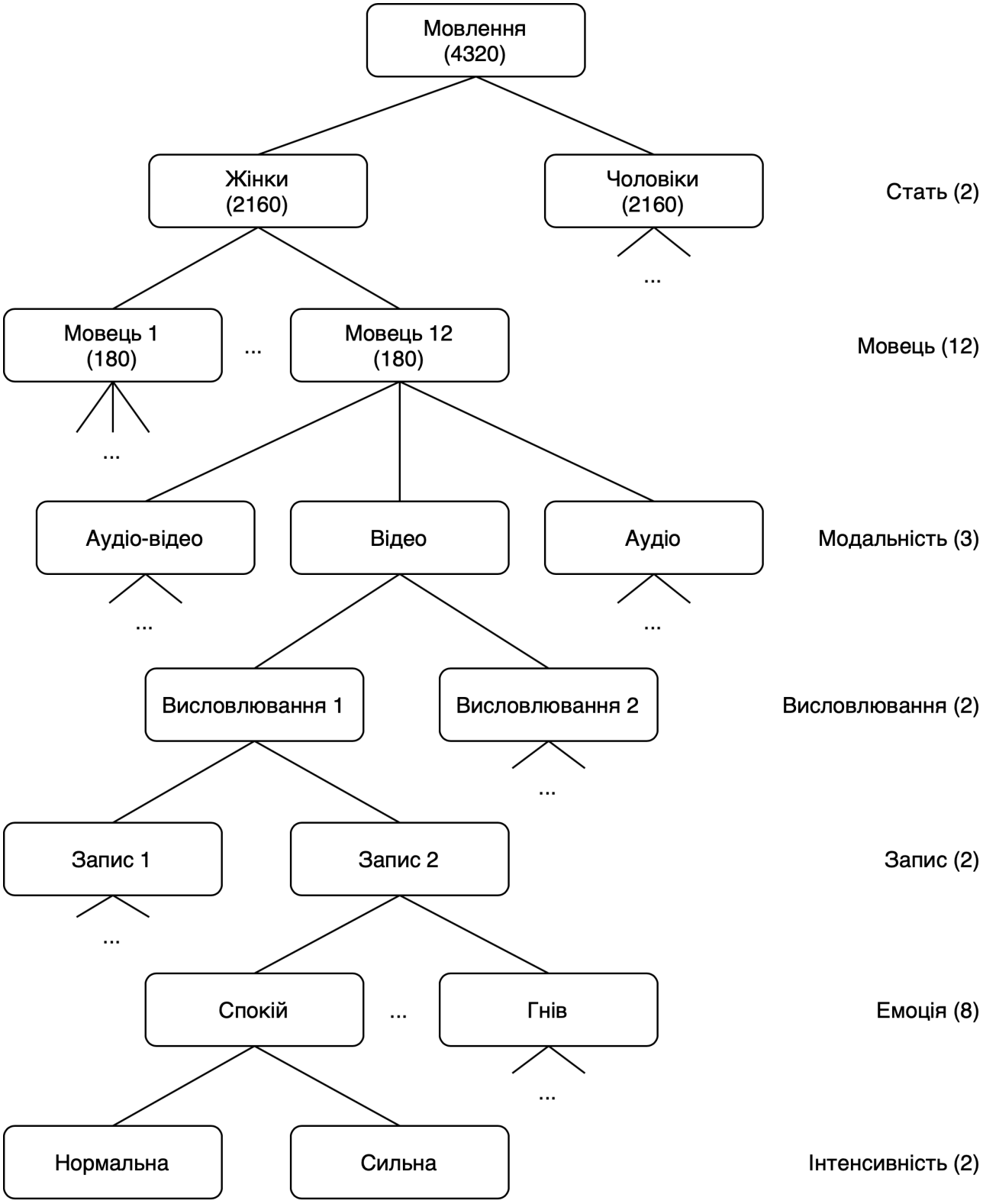


Рисунок 9 — Деревоподібна діаграма набору даних

Кожен файл датасету кодифікується за наступними параметрами:

- Модальність (01 = аудіо-відео, 02 = відео, 03 = аудіо).
- Голосовий канал (01 = мовлення, 02 = спів).
- Емоція (01 = нейтральна, 02 = спокій, 03 = радість, 04 = смуток, 05 = гнів, 06 = страх, 07 = відраза, 08 = здивування).
- Емоційна інтенсивність (01 = нормальна, 02 = висока), причому для нейтральної емоції відсутня висока інтенсивність.
- Висловлювання (01 = "Kids are talking by the door" ("Діти розмовляють біля дверей"), 02 = "Dogs are sitting by the door" ("Собаки сидять біля дверей")).
- Повторення (01 = перший повтор, 02 = другий повтор).
- Актор/акторка (від 01 до 24; непарні числа позначають чоловічі голоси, а парні — жіночі).

Всі записи зроблені на професійній студії звукозапису, схему якої представлено на рисунку 10. Відеозаписи зроблені в якості 1080i, роздільна здатність дорівнює 1920x1080 пікселів, частота кадрів 30 fps (від англ. "frames per second" — кадри за секунду). Файли записувалися у форматі AVCHD, пізніше аудіофайли будуть експортовані у формат WAV. Мікрофон було розташовано на відстані 20 см від актора. Аудіозапис проводився за допомогою інструмента Pro Tools 8 при рівні дискретизації 48 кГц, 16 біт.

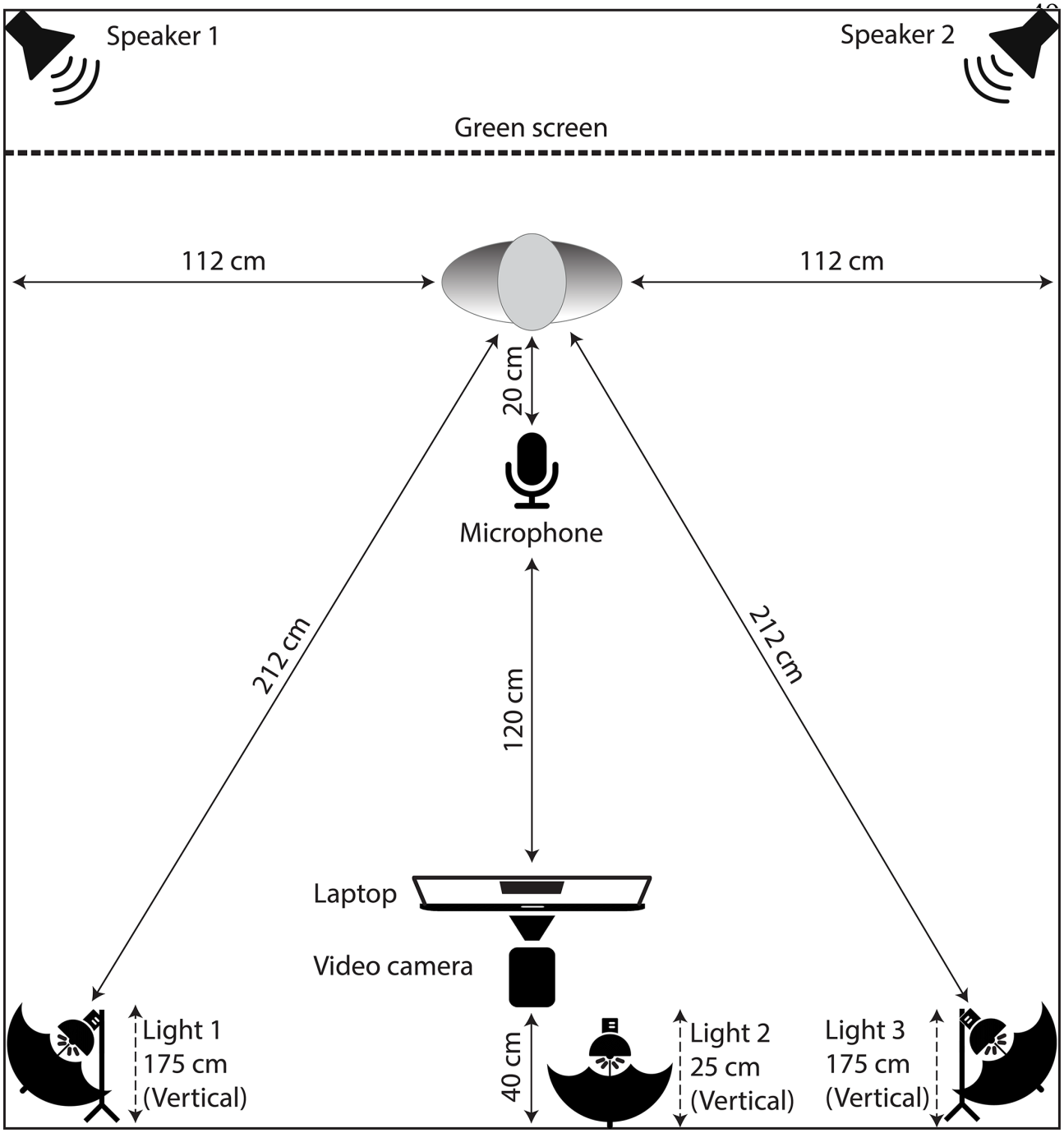


Рисунок 10 — Схема студії звукозапису для створення датасету RAVDESS

Для валідації та вимірювання точності були створені матриці невідповідностей, що показують середнє відношення цільових та нецільових міток до кожного емоційного стану. Візуалізація цих даних подана на рисунку 11, де представлено два канали: (a) — мовлення (n = 43200 оцінок), а також (b) — спів (n = 30360 оцінок).

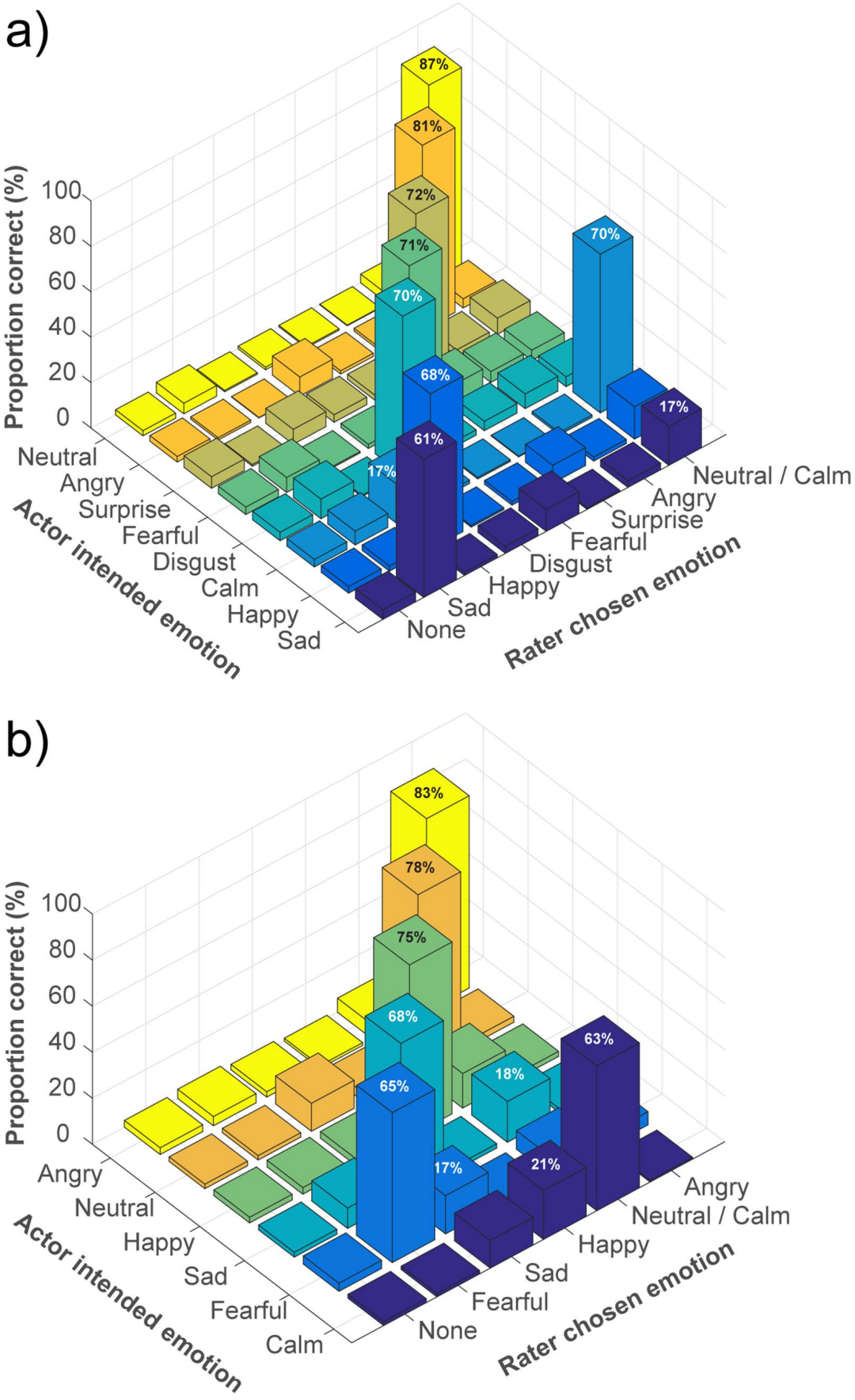


Рисунок 11 — Матриці невідповідностей датасету RAVDESS

Набір даних TESS (Toronto emotional speech set) складається з набору з 200 слів, що були вимовлені у складі цільового виразу "Say the word _". Датасет записано двома акторками, віком 26 та 64 роки. Набір даних містить 8 емоцій: гнів, радість, страх, відраза, здивування, задоволення, смуток та нейтральна емоція. Всього в датасеті 2800 аудіофайлів, що подані у форматі WAV. Схема створення датасету подібна по RAVDESS, тому немає потреби повторно її описувати.

Варто відмітити, що в наборі даних присутні записи, що зроблені не тільки молодими людьми, як в датасеті RAVDESS (від 21 до 33 років), а й акторкою більш старшого віку (64 роки). Справа в тому, що людський голос з часом змінюється, тому при розпізнаванні емоцій важлива варіативність не тільки за статтю, а й за віком. Це дозволить урізноманітнити аудіозаписи, що має призвести до покращення результатів розпізнавання порівняно з випадком, коли використовується лише один набір даних.

2.6 Теоретичний опис згорткових нейронних мереж

Згорткова нейронна мережа — це клас глибинних штучних нейронних мереж прямого поширення. Ідея таких мереж полягає у чергуванні згорткових шарів та шарів субдискретизації. Такі мережі у 1989 році вперше запропонував Ян Лекун [53]. Типова архітектура мережі зображена на рисунку 12.

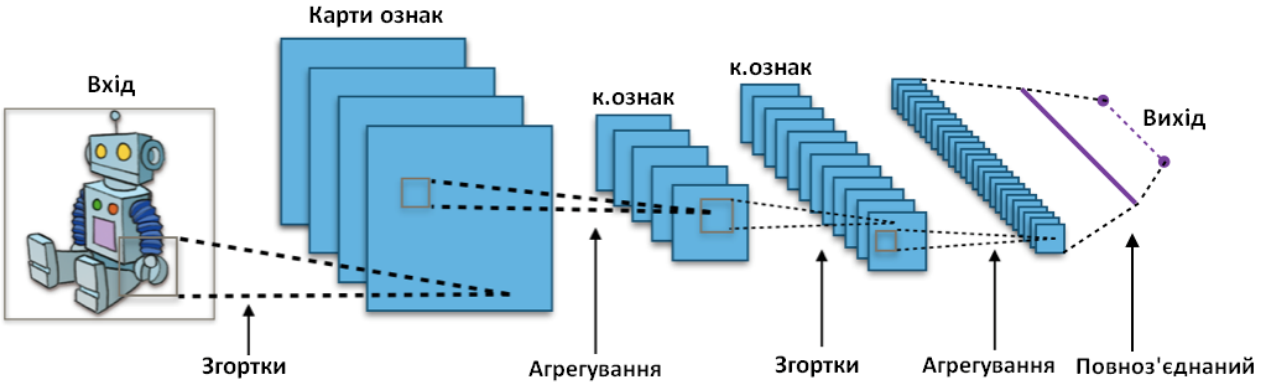


Рисунок 12 — Типова архітектура згорткових нейронних мереж

Основною операцію в згорткових нейронних мережах є операція згортки. Згортка — це операція над парою матриць A (розміру $n_x \times n_y$) та B (розміру $m_x \times m_y$), результатом якої є матриця $C=A \times B$, розмір якої становить $(n_x - m_x + 1) \times (n_y - m_y + 1)$. Кожний елемент результату обраховується як скалярний добуток матриці B та деякої підмножини матриці A такого ж розміру. Тобто, використовується формула:

$$C_{i,j} = \sum_{u=0}^{m_x-1} \sum_{v=0}^{m_y-1} A_{i+u,j+v} B_{u,v}$$

Зміст такої згортки полягає у тому, що чим більша величина елемента згортки, тим більше ця частина матриці A була схожа на матрицю B. Тому матрицю A називають зображенням, а матрицю B — фільтром.

В структурі мережі основними типами шарів є згортковий, агрегувальний та повнозв'язний.

Згортковий шар нейронної мережі представляє собою застосування операції згортки до виходів з попереднього шару, де ваги ядра згортки є тренувальними параметрами. Ще одна тренована вага використовується в якості константного зсуву (bias). Причому в одному згортковому шарі може бути декілька згорток. В такому випадку для кожної згортки на виході буде своє зображення. Якщо розмірність входу складала $w \times h$, а шар містив n згорток з ядром розмірності $k_x \times k_y$, тоді на виході розмірність становитиме $n \times (w - k_x + 1) \times (h - k_y + 1)$.

Ядра згортки можуть бути тривимірними. Згортка тривимірного входу з тривимірним ядром відбувається аналогічно, просто скалярний добуток рахується ще й по всіх шарах зображення.

Шар агрегування потрібен для зменшення розмірності. Початкове зображення ділиться на блоки розміром $w \times h$ і для кожного блоку обчислюється деяка функція. Найчастіше використовується функція максимуму, середнього чи зваженого середнього. Основними задачами шару агрегування є зменшення зображення та пришвидшення розрахунків.

2.7 Проектування архітектури за допомогою UML-діаграм

В першу чергу створюється діаграма, яка має показати загальну архітектуру. Такою UML-діаграмою є діаграма компонентів, що зображена на рисунку 13. Вона включає в себе два компоненти наборів даних (RAVDESS та TESS), два компоненти сторонніх бібліотек (TensorFlow та Keras), основний класифікатор та інтерфейс вводу/виводу для роботи з класифікатором.

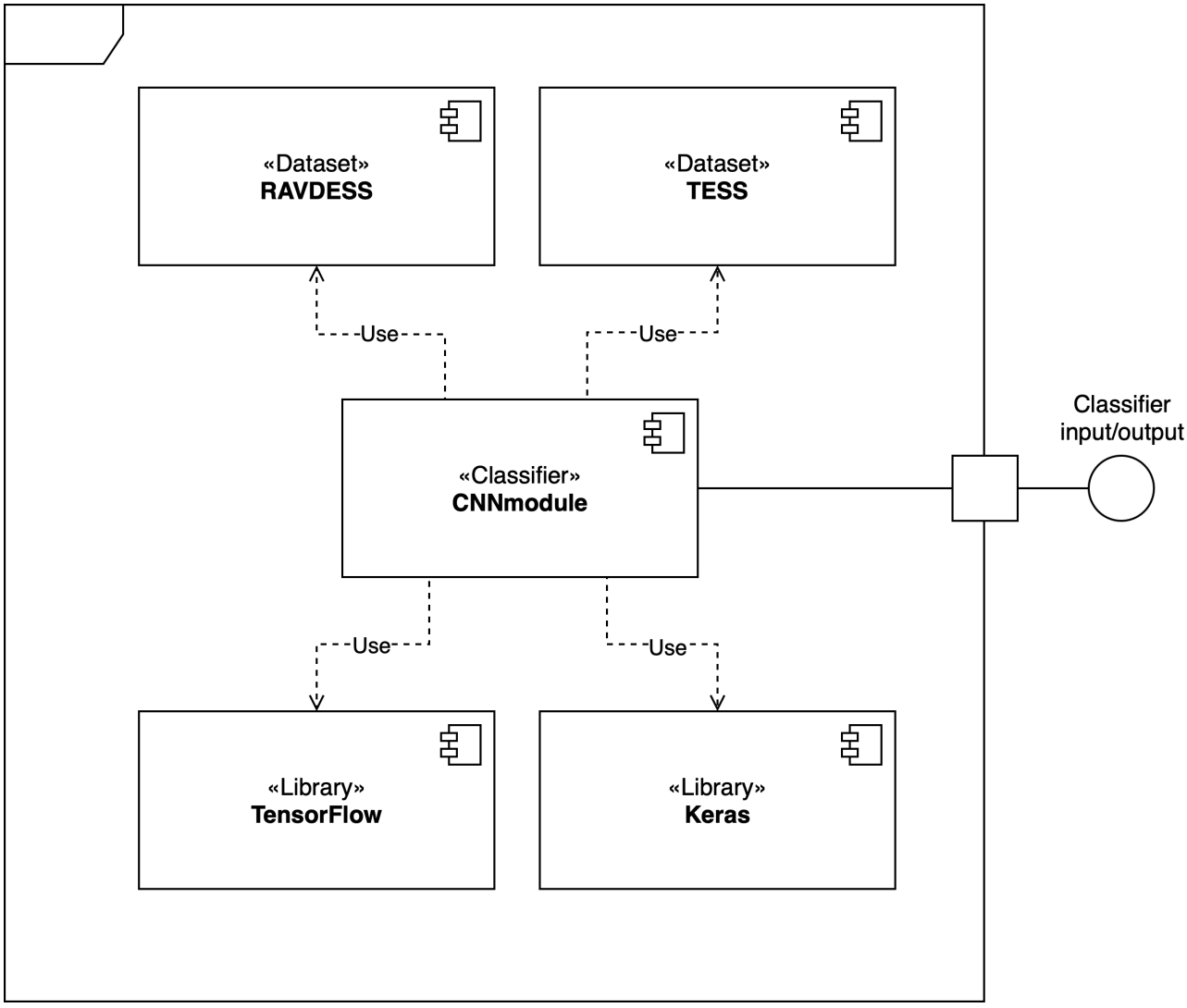


Рисунок 13 — Діаграма компонентів

Після цього була розроблена діаграма класів. Вона деталізує центральний компонент попередньої діаграми, а саме модуль розпізнавання емоцій за голосом. Діаграма класів зображена на рисунку 14.

Діаграма включає такі класи:

1. Datasets — виконує обробку датасетів, створює структуру даних, що "зрозуміла" для модуля;
2. Features — виконує вилучення ознак; на вхід приймає файли з набору даних, на виході створює масиви ознак;
3. CNN — будує згорткову нейронну мережу; на вхід приймає ознаки, використовує пакети TensorFlow і Keras, на виході — готова модель;
4. Testing — виконує тестування мережі; на вхід подаються аудіофайли, а на виході отримується один з емоційних станів;
5. Plotting — будує графіки, потрібні для візуалізації.

Крім того, на діаграмі присутні два пакети (TensorFlow та Keras), що допомагають у побудові нейронної мережі, та два файли (RAVDESS і TESS), що представляють набори даних.

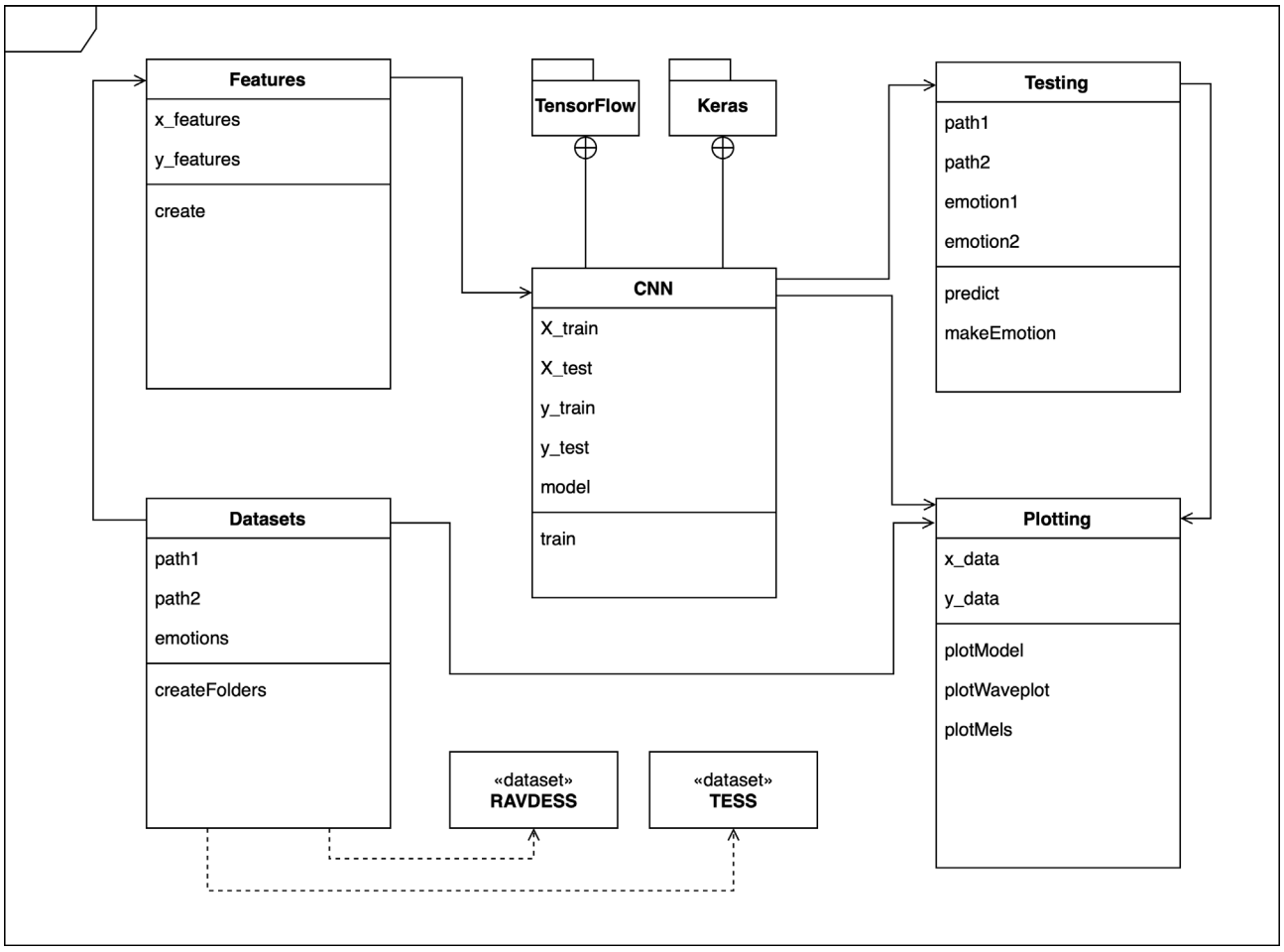


Рисунок 14 — Діаграма класів

2.8 Моделювання згорткової нейронної мережі

Оскільки було вирішено розробляти інтелектуальний модуль з використанням згорткових нейронних мереж на базі бібліотек TensorFlow та Keras, то й моделювання потрібно виконувати виходячи з обмежень та можливостей пакету TensorFlow. В якості архітектури для згорткової нейронної мережі якнайкраще підійде послідовна модель (The Sequential model). Головне обмеження даної моделі у тому, що вона може мати лише одне вхідне значення. Але це не буде проблемою, адже інтелектуальний модуль на вхід за раз отримуватиме лише один аудіофайл.

Отже, на вхід моделі буде подаватися аудіофайл. В першу чергу створюється основний та обов'язковий шар згортки. В якості функції активації зазвичай виступає ReLU.

Оскільки повнозв'язні шари використовують більшість параметрів, вони мають схильність до перенавчання (overfitting). Одним з методів, який дозволяє попереджати перенавчання, є використання так званого дропауту (від англ. "dropout"). На кожному етапі навчання окремі нейрони вилучаються з мережі з деякою ймовірністю $1 - p$. Зв'язки, що вели до нейрона та з нейрону також вилучаються. Нейрони, які залишились в мережі, приймають участь у подальшому навчанні. Потім вилучені нейрони повертають до мережі з початковими вагами.

Далі додається шар Flatten, метою якого є приведення даних до одномірного вектора для того, щоб їх можна було передати на наступний шар.

Наступним шаром є повнозв'язний шар, як у звичайного перцептрона. Це стандартний останній шар, який на виході має кількість нейронів, рівну кількості класів. В якості функції активації часто використовують Softmax.

На рисунку 15 зображена блок-схема нейронної мережі.

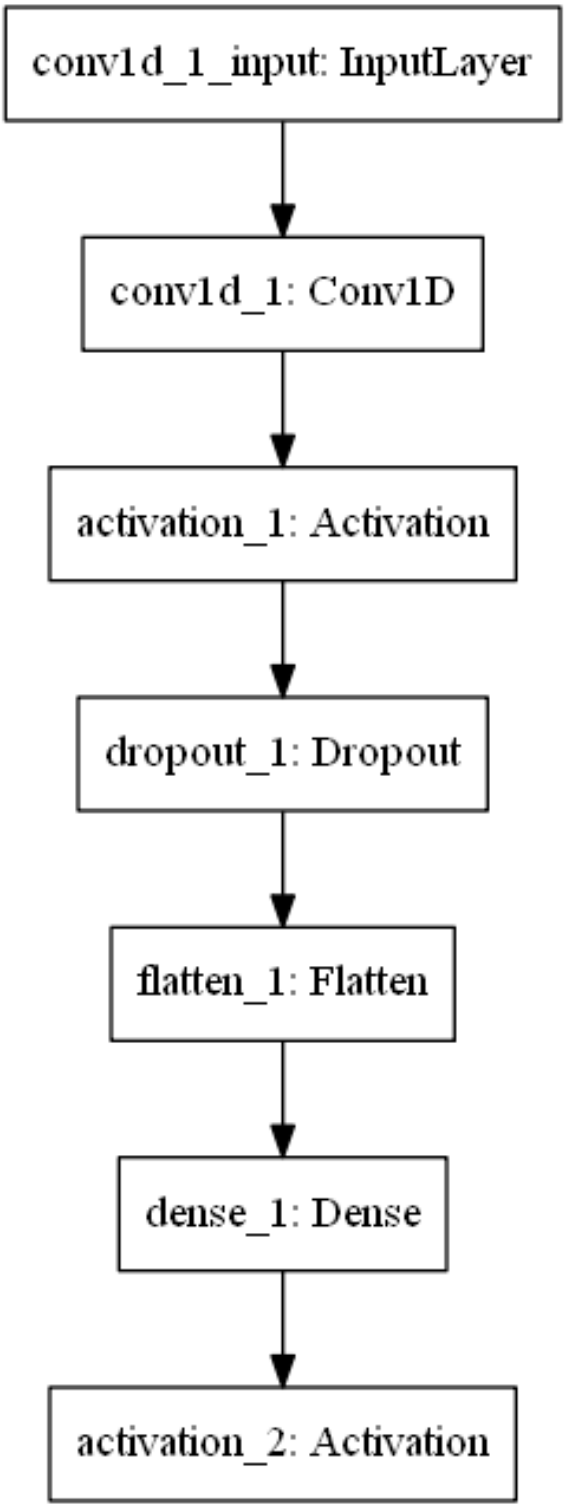


Рисунок 15 — Блок-схема згорткової нейронної мережі

РОЗДІЛ 3 РОЗРОБКА ІНТЕЛЕКТУАЛЬНОГО МОДУЛЯ

3.1 Реалізація інтелектуального модуля

Перед безпосередньою реалізацією моделі, виконуються останні кроки підготовки даних. Набір даних розділяється на навчальну та тестову вибірки у відношенні 2:1 за допомогою стандартної функції *train_test_split()* бібліотеки

scikit-learn (безкоштовна бібліотека для машинного навчання) та стандартних параметрів:

X_train, X_test, y_train, y_test = train_test_split (X, y, test_size=0.33)

Далі програмується сама модель. Оскільки в якості моделі TensorFlow використовується послідовна мережа (Sequential model), для її створення в першу чергу додається базовий шар згортки (Convolutional layer 1). Цей шар створює ядро згортки. Розмір шару 40x1, тому що аудіофайл, тривалістю 2 секунди, розбивається на 40 інтервалів.

```
tf.keras.layers.Conv1D(
    filters,
    kernel_size,
    strides=1,
    padding="same",
    data_format="channels_last",
    dilation_rate=1,
    groups=1,
    activation=None,
    use_bias=True,
    kernel_initializer="glorot_uniform",
    bias_initializer="zeros",
    kernel_regularizer=None,
    bias_regularizer=None,
    activity_regularizer=None,
    kernel_constraint=None,
    bias_constraint=None,
    input_shape=(40,1)
)
```

Наступним додається шар активації (Activation layer 1) з функцією активації ReLU, яка визначається як:

$$g(z) = max\{0, z\},$$

де *z* — вхідне значення нейрона.

Дана функція є стандартною функцією активації для бібліотеки TensorFlow/Keras. Перевагою функції ReLU є те, що на великих наборах даних навчання глибинних нейронних мереж відбувається швидше, порівняно з сигмоїдом. Додається до моделі активаційний шар за допомогою команди:

model.add(Activation('relu'))

Далі додається шар дропауту (від англ. "dropout") або виключення. Він використовується для запобігання перенавчання нейронної мережі та сприяє збільшенню швидкості навчання. В якості величини дропауту зазвичай береться деяке евристичне значення, найчастіше близько 0,2–0,3. Варто зазначити, що шар дропауту застосовується тільки під час навчання, і не застосовується під час тестування. Шар було додано зі значенням 20% за допомогою наступної команди:

model.add(Dropout(0.2))

Після цього додається повнозв’язний шар (Dense layer). Для цього спочатку виконується операція:

model.add(Flatten())

Далі додається сам повнозв’язний шар за допомогою команди:

model.add(Dense(8))

Число 8 передається в якості параметра саме тому, що в моделі використовується саме 8 класів емоцій. Загалом, повнозв’язний шар має наступний вигляд:

```
tf.keras.layers.Dense (
    8,
    activation=None,
    use_bias=True,
    kernel_initializer="glorot_uniform",
    bias_initializer="zeros",
    kernel_regularizer=None,
    bias_regularizer=None,
```

activity_regularizer=None,

kernel_constraint=None,

bias_constraint=None,

)

На цьому етапі в якості функції активації виступає функція Softmax:

model.add(Activation('softmax'))

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_1 (Conv1D)            (None, 40, 64)            384
_________________________________________________________________
activation_1 (Activation)    (None, 40, 64)            0
_________________________________________________________________
dropout_1 (Dropout)          (None, 40, 64)            0
_________________________________________________________________
flatten_1 (Flatten)          (None, 2560)              0
_________________________________________________________________
dense_1 (Dense)              (None, 8)                 20488
_________________________________________________________________
activation_2 (Activation)    (None, 8)                 0
=================================================================
Total params: 20,872
Trainable params: 20,872
Non-trainable params: 0
```

Рисунок 16 — Архітектура мережі

Це стандартний вибір функції активації останнього шару мережі. Дана функція конвертує вектор дійсних чисел у вектор ймовірностей, тобто елементи вихідного вектора знаходяться в діапазоні (0, 1), а їх сума рівна 1. Значення softmax кожного вектора *x* обчислюється як:

exp(x) / tf.reduce_sum(exp(x))

Отримана архітектура мережі представлена на рисунку 16.

3.2 Опис результатів на основі контрольних прикладів

Розроблений інтелектуальний модуль було протестовано на декількох тестових зразках, що не були частиною початкового набору даних. Зокрема, було обрано зразок з промови активістки Ґрети Тунберг, яка відома своїми емоційними виступами. Також було взято зразок чоловічого голосу, який зображав відразу. Результати тестування інтелектуального модуля на цих зразках подані на рисунку 17. Як видно з рисунка, програма коректно розпізнала обидві емоції.

```
Емоція: гнів
Емоція: відраза
```

Рисунок 17 — Результати роботи модуля на тестових зразках

Далі пропонується переглянути візуальну різницю в емоційності мовлення. Нижче подано спектрограми, де чоловічим голосом зображені три емоційні стани: смуток (рисунок 18), гнів (рисунок 19) та нейтральна емоція (рисунок 20) в якості базового стану. Можна побачити, що на спектрограмі з нейтральним станом, яскраво вираженими є нижні частоти, а також відсутні значні коливання. На спектрограмі емоції смутку видно понижену та низхідну частоту.

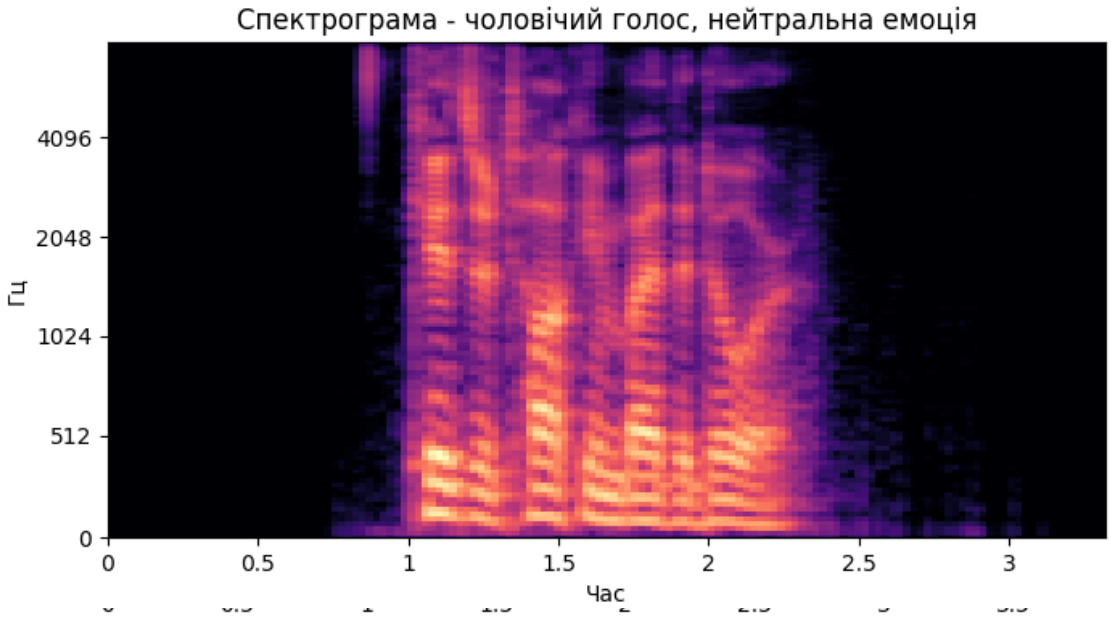


Рисунок 20 — Спектрограма нейтральної емоції

Рисунок 18 — Спектрограма емоції смутку

Як видно на рисунку 19, на спектрограмі емоції гніву наявні суттєві коливання голосу з яскраво вираженими піками.

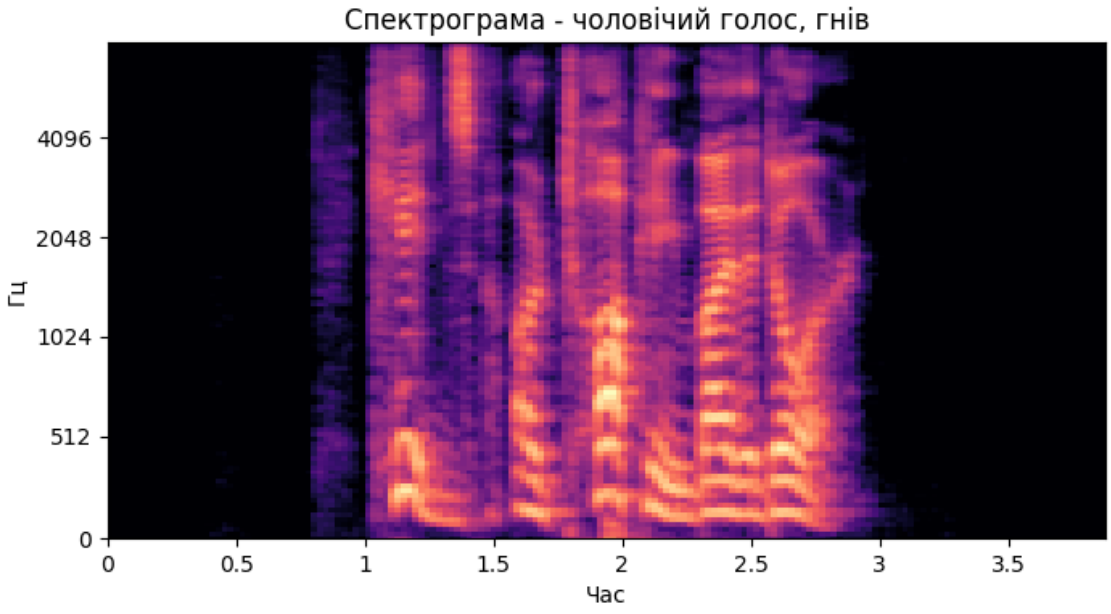


Рисунок 19 — Спектрограма емоції гніву

Ще більш яскраво ця різниця помітна на графіках звукової хвилі (waveplots). Для емоції гніву (рисунок 22) характерні часті гострі піки. У той час як для смутку (рисунок 21) є характерним низхідний графік. Нейтральний стан подано на рисунку 23 для порівняння.

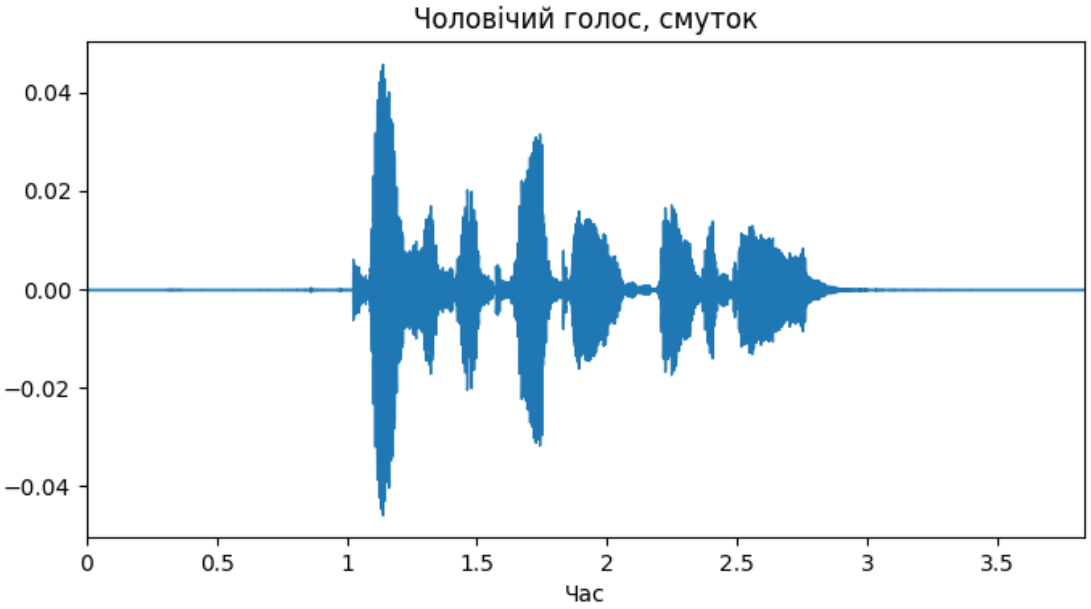


Рисунок 21 — Графік емоції смутку

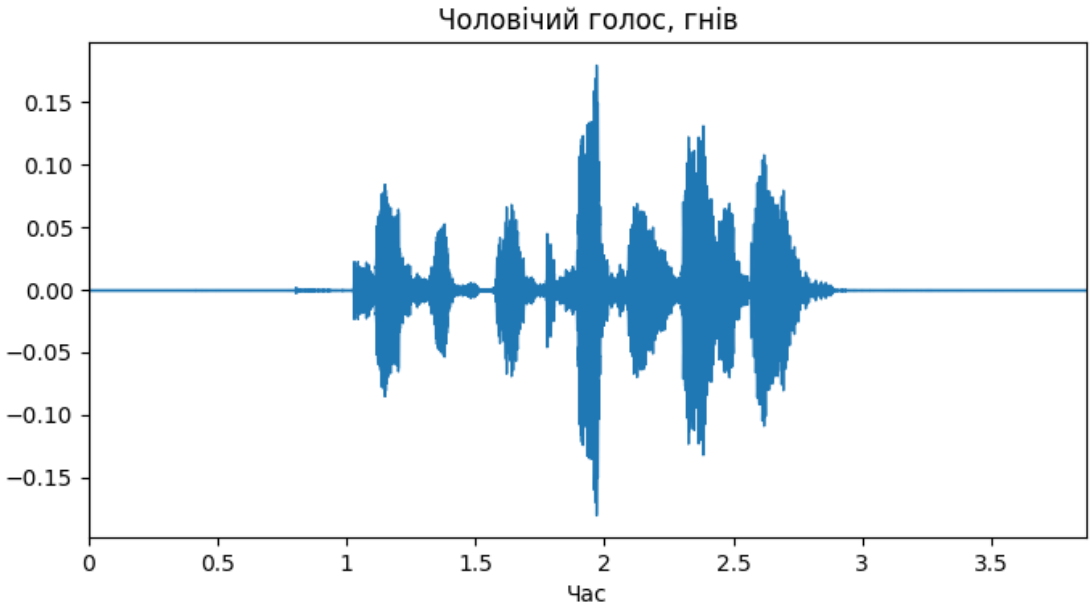


Рисунок 22 — Графік емоції гніву

Також вартим уваги моментом є різниця у вираженні емоцій між чоловіками та жінками. Насправді, емоційний паттерн жіночого і чоловічого голосів суттєво не відрізняється. Найбільш характерним є те, що жіночий голос вище чоловічого: це помітно по вищій максимальній частоті, а також по малій частці низьких частот.

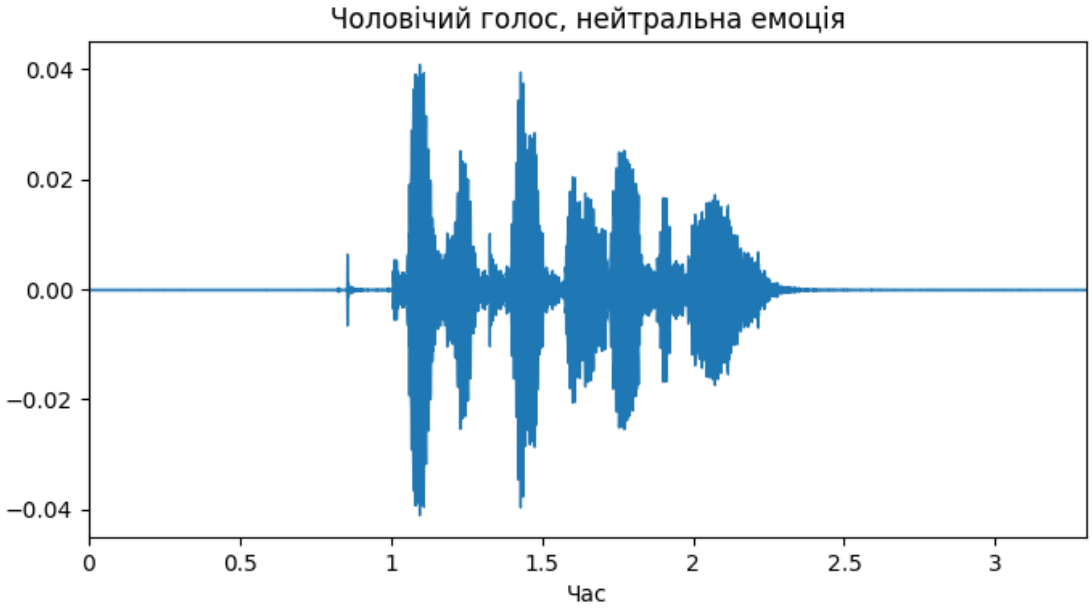


Рисунок 23 — Графік нейтральної емоції

На рисунках 24 та 25 показані спектрограми емоції гніву та нейтральної емоції відповідно, записані жіночими голосами. Добре видно, що низькі частоти менше виражені, у порівнянні з чоловічим голосом. Однак загальний патерн той самий: для емоції гніву характерним є нестабільний голос з вираженими

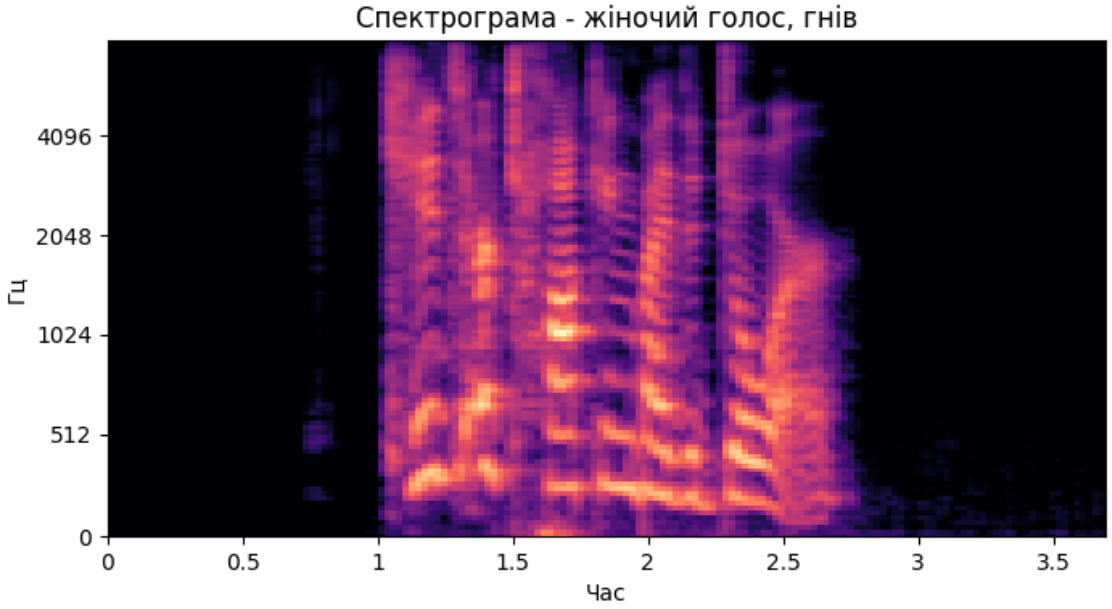


Рисунок 24 — Спектрограма емоції гніву, жіночий голос

імпульсивними піками.

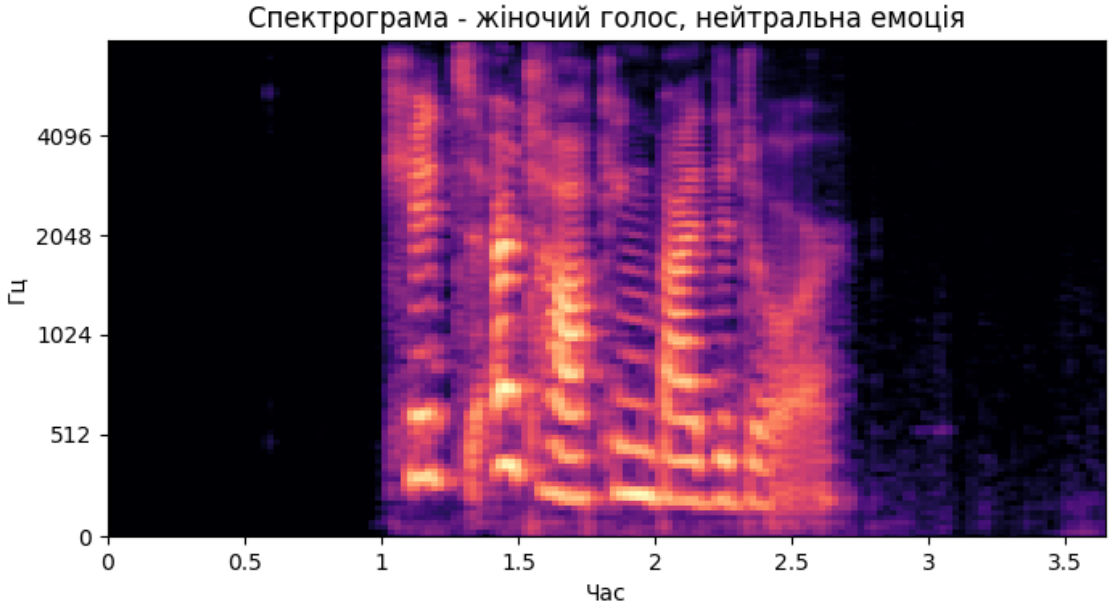


Рисунок 25 — Спектрограма нейтральної емоції, жіночий голос

На рисунках 26 та 27 зображені графіки звукових хвиль (waveplots) для записів, зроблених жіночим голосом. Як видно, закономірності ідентичні до чоловічого голосу: для емоції гніву характерна імпульсивність, тому на графіку бачимо значно вищі піки. Крім того, на графіку емоції гніву видно нетипове для спокійного стану (принаймні, для англомовного спікера) закінчення

висловлювання: замість слабо вокалізованого закінчення речення бачимо явне

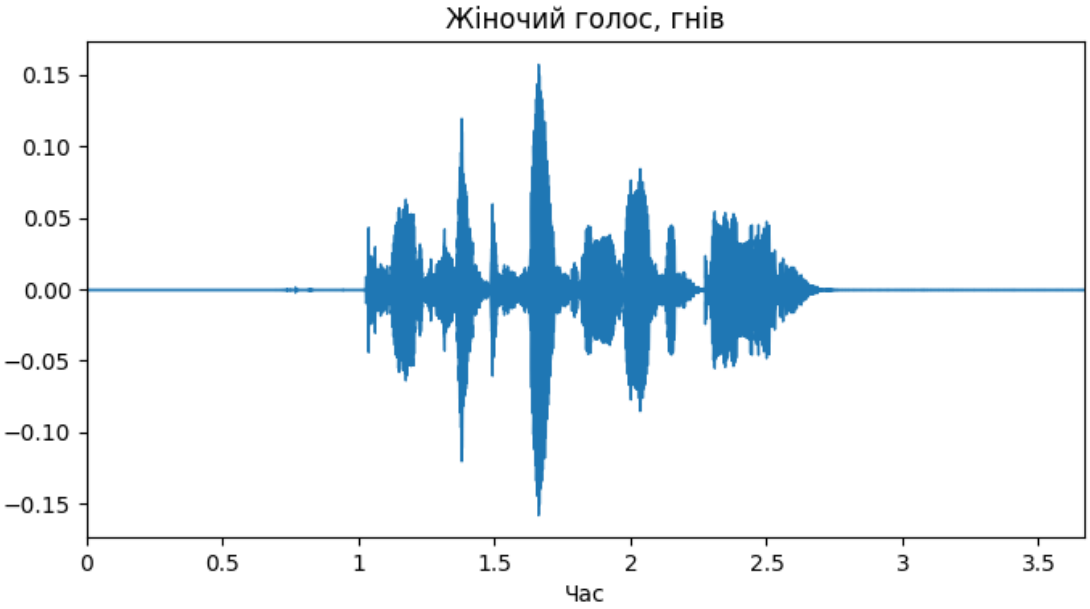


Рисунок 26 — Графік емоції гніву, жіночий голос

напруження у голосі.

Різницю в голосі в залежності від віку можна побачити на прикладі записів із датасету TESS. На рисунку 28 зображена спектрограма голосу акторки, якій виповнилося 64 роки. На рисунку 29 зображено графік звукових хвиль запису,

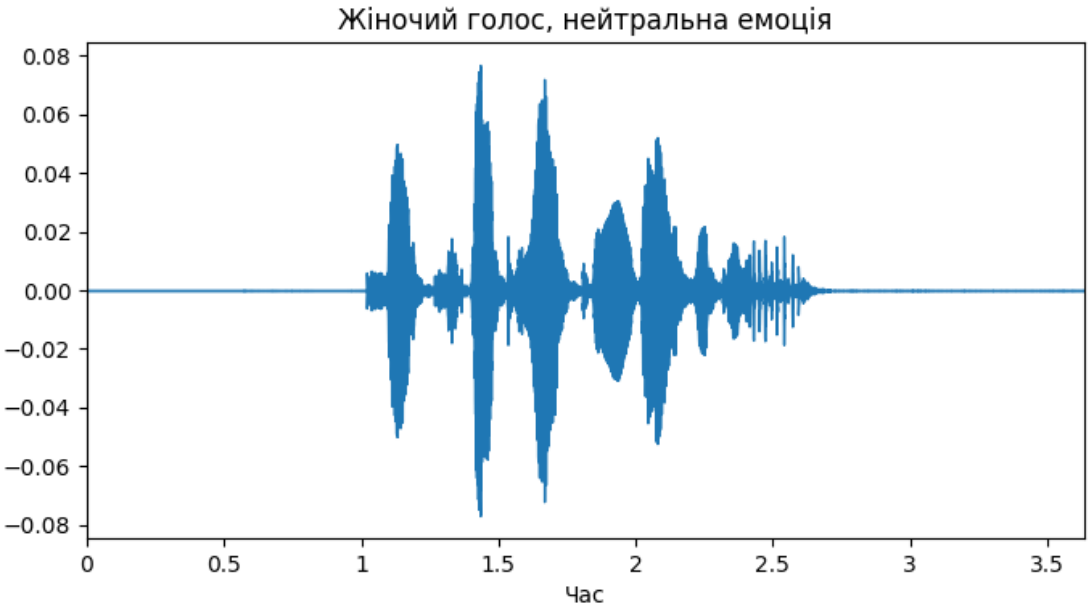


Рисунок 27 — Графік нейтральної емоції, жіночий голос

зробленого тією ж акторкою. Як видно, нижні частоти більш яскраво виражені,

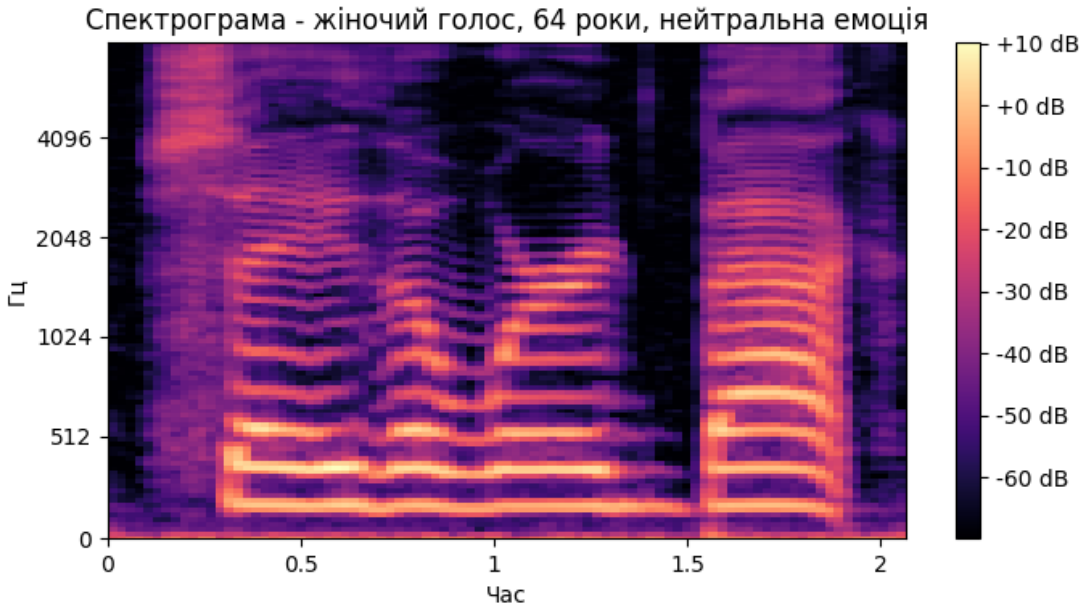


Рисунок 28 — Графік нейтральної емоції, жіночий голос, 64 роки

голос ближчий до чоловічого.

3.3 Аналіз ефективності результатів

Спочатку була перевірена точність розпізнавання лише однієї емоції. Всі вісім емоційних станів були закодовані наступним чином:

labels = { '0': 'нейтральна',

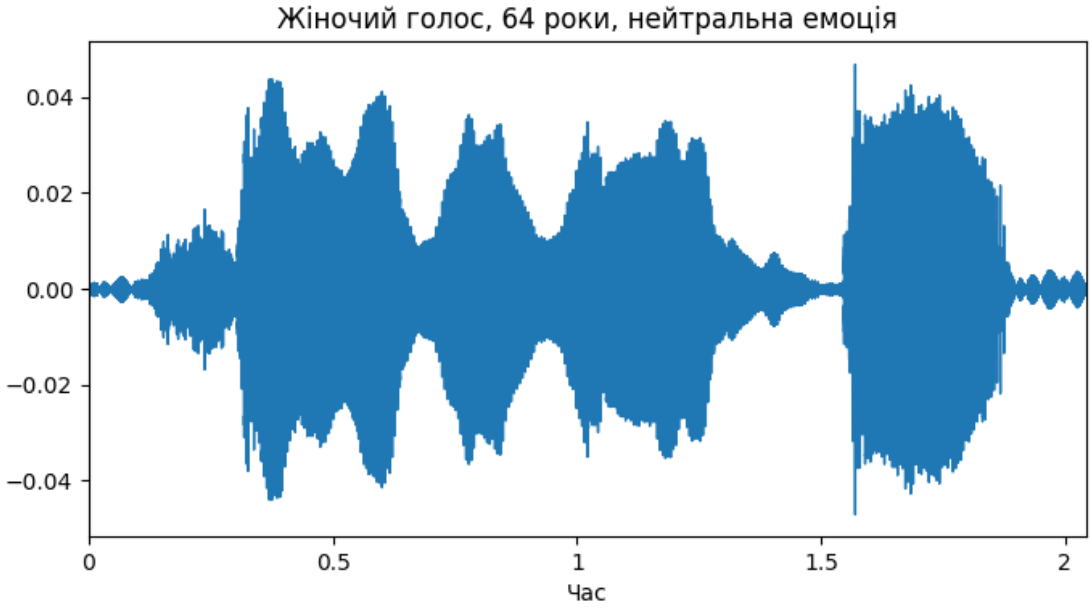


Рисунок 29 — Графік нейтральної емоції, жіночий голос, 64 роки

'1': 'спокій',

'2': 'радість',

'3': 'смуток',

'4': 'гнів',

'5': 'страх',

'6': 'відраза',

'7': 'здивування' }

Точність розпізнавання цих емоцій наведена в таблиці 3:

Таблиця 3 —Точність розпізнавання окремих емоцій

Емоція	**Точність**
Гнів	0.93
Радість	0.92
Нейтральна	0.91
Смуток	0.84
Спокій	0.96
Страх	0.92
Відраза	0.95
Здивування	0.90

Розроблений інтелектуальний модуль було протестовано на вибірці, яка склала 33% від загального набору даних. Навчання відбувалось протягом 50 епох, оскільки при збільшенні кількості епох не спостерігалося покращення точності. Це може означати, що при більшій кількості епох відбувалося перенавчання мережі (так званий "overfitting").

Графік точності наведено на рисунку 30. Точність валідації склала ≈ 0,8, що є кращим показником за середній показник точності в проаналізованих публікаціях (0,75)

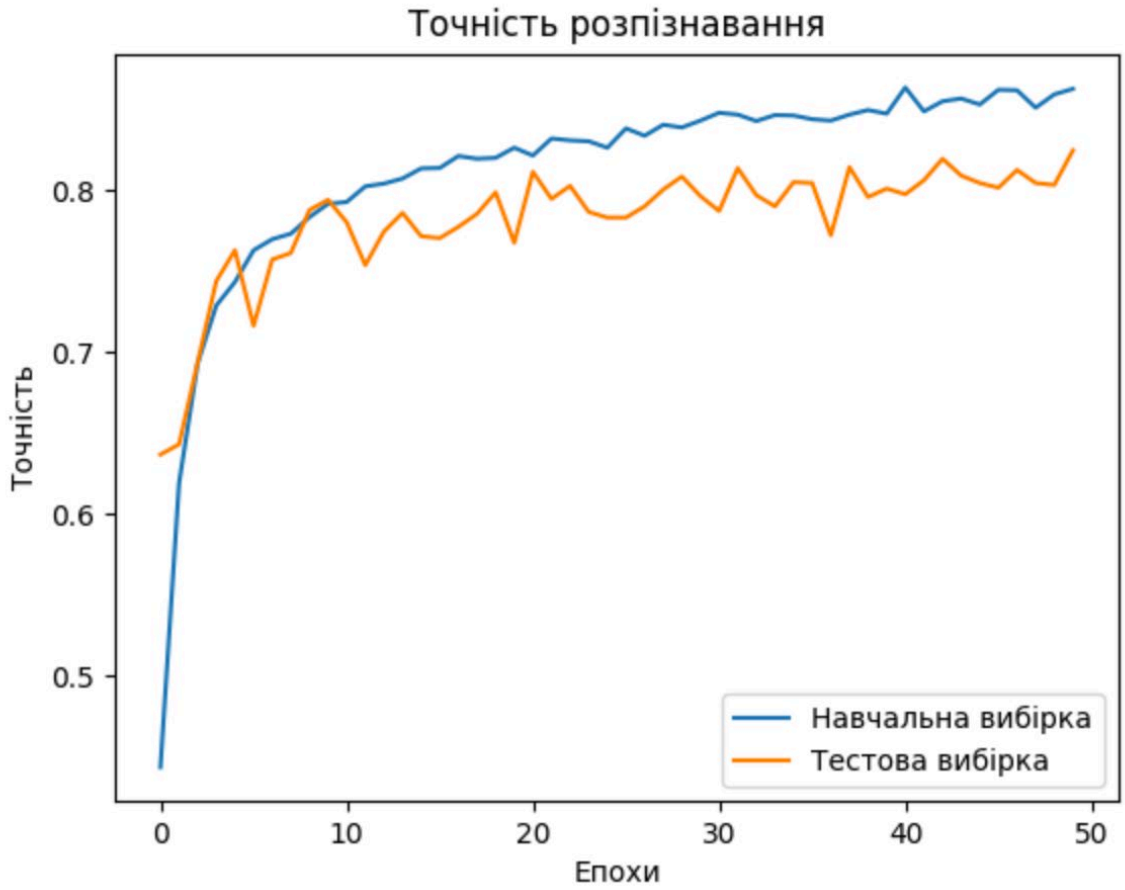


Рисунок 30 — Графік точності розпізнавання

На рисунку 31 наведено знімок екрану зі значенням точності (accuracy)

```
Точність моделі: 0.8091118931770325
```

Рисунок 31 — Графік точності розпізнавання

отриманої моделі в середовищі PyCharm:

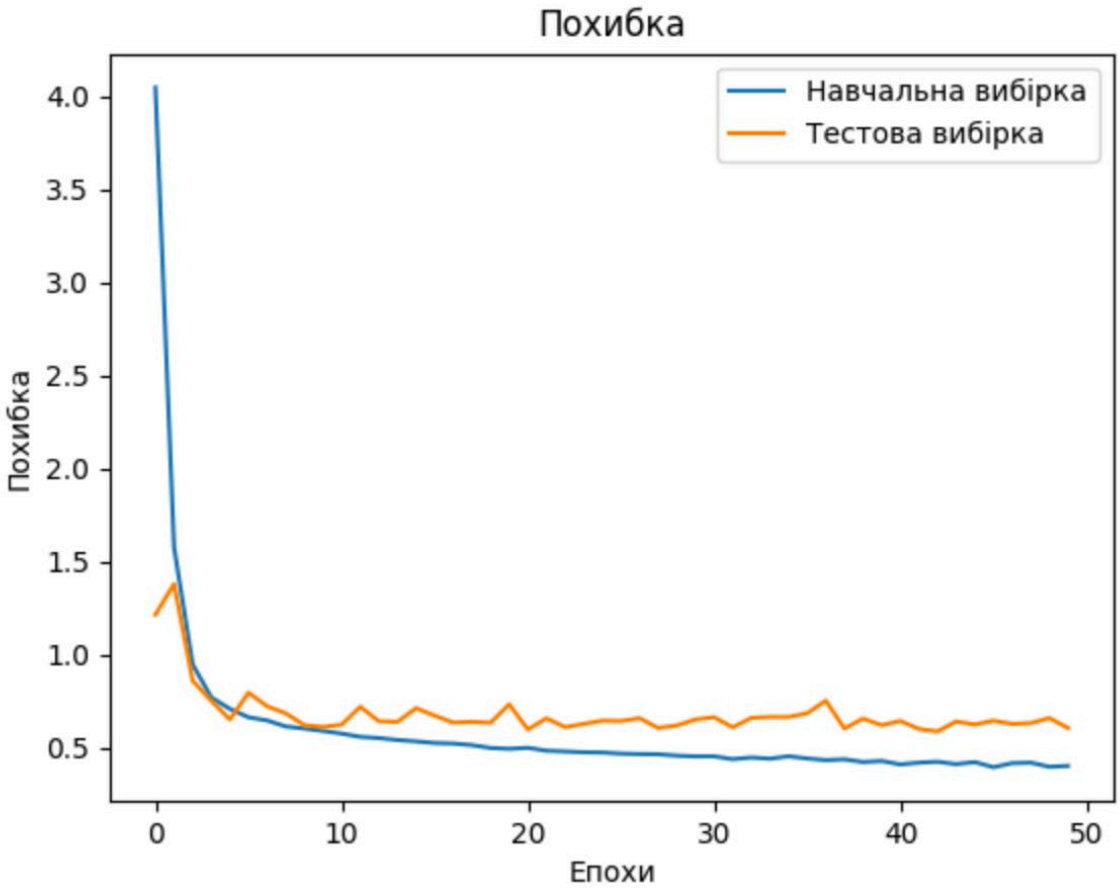


Рисунок 32 — Графік похибки мережі

На рисунку 32 наведено графік похибки мережі:

3.4 Сценарії використання інтелектуального модуля

Основними сценаріями використання розробленого інтелектуального модуля є:

1. оптимізація роботи кол центрів, сервіс-десків;
2. оптимізація мап в комп'ютерних іграх;
3. виконання психологічної оцінки;
4. оптимізація роботи інтелектуальних голосових помічників.

Зупинимося детальніше на кожному зі сценаріїв.

Робота кол центрів, як і будь-який бізнес-процес, потребує оцінки для розуміння ефективності. Однак, оцінювати кожне звернення та кожну

інтеракцію з клієнтом за допомогою людських ресурсів дуже дорого та вкрай неефективно. В цьому випадку на декілька чоловік команди підтримки потрібен один менеджер, який прослуховує розмови та оцінює роботу персоналу. Тому потрібен інструмент, який міг би автоматично виконувати таку оцінку. Оскільки критерієм якості роботи персоналу кол центру є задоволеність клієнта, а клієнти часто пропускають процедуру зворотного зв'язку (коли після розмови зі службою підтримки клієнта просять залишити відгук, натиснувши відповідну кнопку, більшість клієнтів просто кладе слухавку), було б слушно використовувати деяку інтелектуальну систему, яка могла б автоматично визначати, наскільки клієнт залишився задоволеним. Причому немає необхідності розробляти систему, що працюватиме в режимі реального часу. Достатньо мати програмний модуль, в який завантажуються записи розмов персоналу з клієнтами, а програмний модуль підкаже, грунтуючись на емоції клієнта, чи була його проблема вирішена. Інтелектуальний модуль, розроблений в даній роботі, ідеально підходить для використання саме за цим сценарієм.

Інший цікавий варіант використання модуля розпізнавання емоцій за голосом — моделювання мап в комп'ютерних онлайн-іграх. Ігровим дизайнерам потрібно проектувати такі мапи, що максимально вволікають гравця у процес гри, викликаючи в нього специфічні емоції. Якщо ігровий рівень буде запростим — користувачу швидко набридне; якщо заскладним — користувач засмутиться і не буде продовжувати гру. Знаючи емоції гравця у кожний момент гри, можна точніше моделювати складність рівнів та ігрових мап. В сучасних онлайн-іграх часто використовуються мікрофони для спілкування гравців між собою, а записи зберігаються на серверах ігрових компаній. Ці записи, якщо їх попередньо розмітити та поєднати з певним моментом гри, можуть бути вкрай корисними. Вже існують дослідження в області психології, де показано, яка емоційна поведінка сприяє найбільшому зануренню в процес. Використовуючи ці психологічні моделі, можна оптимізувати ігровий процес таким чином, щоб гравець отримував якумога більше задоволення від гри.

Набагато більш серйозним сценарієм використання модуля розпізнавання емоцій є психологічна оцінка. Дана область застосування потребує великої точності класифікатора, однак в перспективі дозволить значно вплинути на процес діагностики психологічних порушень. Можуть існувати багато сценаріїв для проведення такого оцінювання: від первинного анкетування людей, що звертаються за психологічною допомогою, до превентивного виявлення людей з ментальними розладами. Наприклад, людина телефонує, щоб записатись на прийом до психолога, і вже під час запису система може проаналізувати стан людини і запропонувати тип її розладу.

Ще одним сценарієм використання розробленого модуля є його інтеграція з голосовими інтелектуальними помічниками, які наразі набувають все більшої популярності. Домашні колонки з голосовим управлінням (наприклад, Amazon Alexa, Apple Home, Google Home) вже стали загально використовуваними в США і з'являються на вітчизняному ринку. Крім того, в сучасних смартфонах також є інтелектуальні голосові асистенти (Apple Siri, Google Assistant). Всі ці програми можуть використовувати модуль розпізнавання емоцій як для більш релевантних відповідей, так і для інтелектуальних рекомендацій. Крім того, модуль можна використовувати в маркетингових цілях, пропонуючи клієнтам певний товар в залежності від їх поточного стану.

3.5 Подальший розвиток інтелектуального модуля

Розроблений інтелектуальний модуль має широкі можливості щодо його покращення. Основними напрямками подальших робіт можуть бути:

1. використання різноманітних наборів даних для покращення процесу навчання мережі;
2. ускладнення самої структури згорткової нейронної мережі, додавання більшої кількості шарів;
3. перехід від дискретної до багатовимірної моделі емоційних станів для більш детального аналізу емоційного стану людини;

4. впровадження інтелектуального модуля в певну існуючу систему голосового спілкування чи систему аналізу мовлення.

Використання різних датасетів ймовірно покращить точність роботи мережі, якщо підібрані датасети будуть мати різні параметри (наприклад, стать, вік, національність, мова спікера). Також покращить точність розпізнавання емоцій в реальних умовах використання датасетів з емоціями, що не відіграні акторами, а записані наживо під час мовлення.

Ускладнення архітектури згорткової нейронної мережі, в першу чергу додавання більшої кількості шарів згортки та агрегування також може покращити результати класифікації. Крім того, варіювання параметра дропауту допоможе уникати перенавчання.

Перехід до багатовимірної моделі емоцій — значний крок у розвитку розпізнавання емоцій за голосом. Він потребує експертних знань в області психології, однак величезною перевагою такої моделі є можливість значно повніше та точніше ідентифікувати емоційний стан людини.

Логічним кроком є імплементація розробки в існуючу систему. Є два основні варіанти такого впровадження. Перший — використання розробленого модуля в системі оцінки якості роботи кол-центрів. Другий — інтеграція з інтелектуальними голосовими помічниками для оптимізації їх роботи за рахунок розпізнавання не тільки вербальної, а й емоційної інформації.

ВИСНОВКИ

В даній роботі було проведено аналіз методів та підходів до вирішення задачі розпізнавання емоцій людини за голосом, виконано проектування згорткової нейронної мережі, розроблено інтелектуальний модуль розпізнавання емоцій за голосом та виконана оцінка ефективності отриманих результатів дослідження.

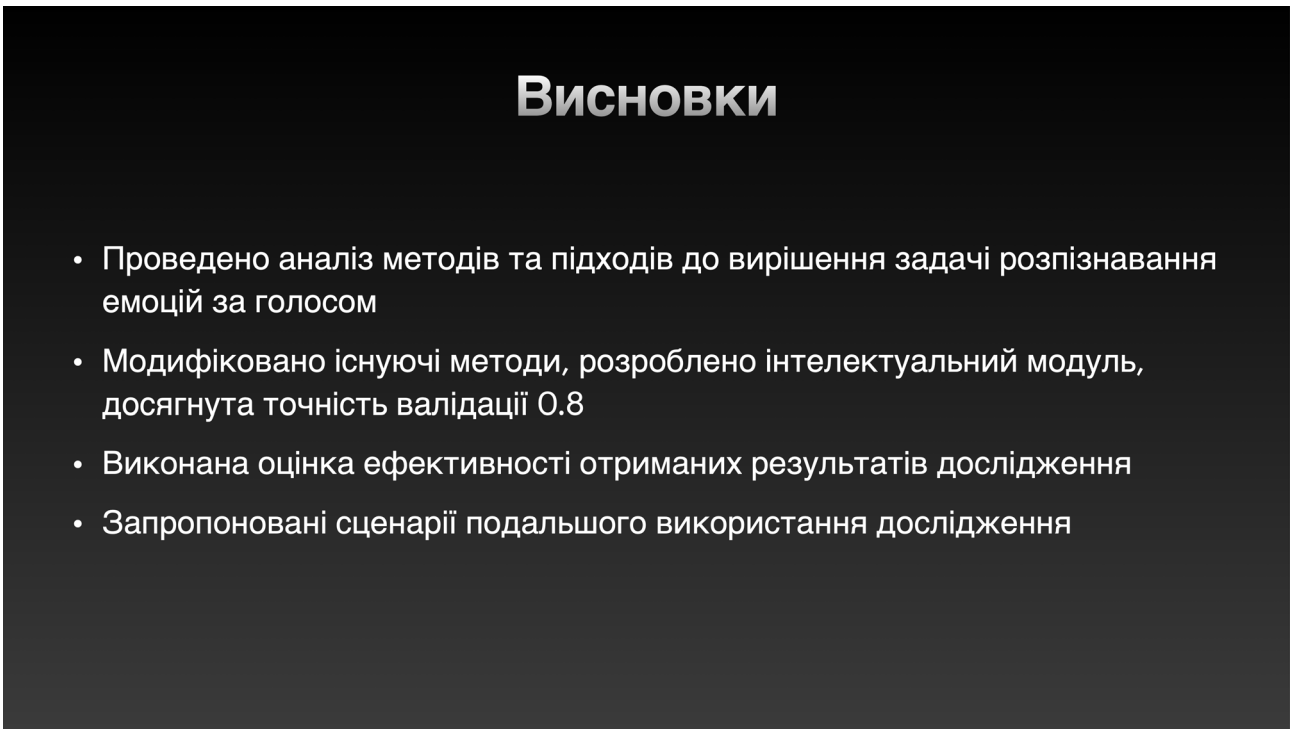
В першому розділі була розкрита актуальність та новизна роботи, проведено огляд літератури, проаналізовані моделі емоційних станів та методи створення наборів даних. В якості моделі емоційних станів було вирішено використовувати дискретну модель, адже з нею зручно працювати і вона не потребує експертних знань в галузі психології. В якості наборів даних було вирішено використовувати датасети RAVDESS та TESS. Поєднання двох наборів даних дозволяє урізноманітнити дані та покращити ефективність результатів.

В другому розділі роботи було детально розглянуто процес розпізнавання емоцій за голосом. Був розглянутий етап передобробки, виділення ознак та класифікації. Обрано метод мел-частотних кепстральних коефіцієнтів для вилучення ознак та розроблена модель вилучення ознак. Проведено підготовку та описано структуру набору даних. Були порівняні декілька варіантів класифікаторів, детально розглянуто їх принципи роботи. В якості класифікатора для інтелектуального модуля розпізнавання емоцій за голосом обрано згорткову нейронну мережу, адже аналіз літератури показав високу точність глибинних нейронних мереж, а їх структура є гнучкою та надає можливості до модифікації. Після цього було зроблено теоретичний опис роботи згорткових нейронних мереж та розроблено модель згорткової нейронної

мережі. Проаналізовані можливості бібліотек для роботи з глибинними нейронними мережами.

В третьому розділі було описано процес розробки інтелектуального модуля розпізнавання емоцій за голосом, деталізовано архітектуру згорткової нейронної мережі. Для розробки мережі використано бібліотеки TensorFlow та Keras, в якості базової моделі обрано послідовну нейромережну модель. Навчання проходило протягом 50 епох на 67% набору даних, відповідно тестування проводилось на 33% даних. Зроблено опис результатів на основі контрольних прикладів, причому в якості зразків були обрані аудіофайли з живою мовою, а не файли з набору даних. Не дивлячись на це, емоції мовлення були розпізнані точно. Проведено аналіз ефективності отриманих результатів дослідження. Отримана точність валідації *0,8*, що перевищує середню точність класифікаторів у проаналізованих дослідженнях (*0,75*). Також були розглянуті особливості вираження емоцій в залежності від віку та статі. Було відмічено зміни у вираженні емоцій голосом з віком. При порівнянні емоцій чоловіків і жінок суттєвої різниці виявлено не було, крім того, що частота жіночого голосу природно вища.

В кінці роботи були запропоновані сценарії використання розробленого інтелектуального модуля та ідеї щодо його подальшого розвитку. В основному розроблений модуль може бути застосований в практичній діяльності. Наприклад, для оптимізації роботи кол центрів, проведення психологічної оцінки та оптимізації роботи інтелектуальних голосових систем. Найкраще розроблений модуль підходить саме для використання у кол центрах, адже для психологічної оцінки бажано збільшити точність розпізнавання. Основними напрямками в подальшому розвитку інтелектуального модуля можуть бути використання різних додаткових наборів даних для оптимізації навчання, модифікація архітектури згорткової нейронної мережі, перехід до багатовимірної моделі емоційних станів. Перші два напрями ймовірно зможуть незначним чином збільшити точність розпізнавання. Перехід же від дискретної до багатовимірної моделі ймовірно зможе якісно покращити процес



розпізнавання емоцій завдяки можливості значно повніше та точніше ідентифікувати емоційний стан людини. Використання такої моделі дало б можливість виконувати розпізнавання емоцій за голосом не тільки в практичній, а й в науковій діяльності.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Schuller, B.W., 2018. Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. Commun. ACM 61 (5), 90–99. doi:10.1145/3129340.
2. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G., 2001. Emotion recognition in human-computer interaction. IEEE Signal Pro- cess. Mag. 18 (1), 32–80. doi:10.1109/79.911197.
3. Huahu, X., Jue, G., Jian, Y., 2010. Application of speech emotion recognition in intelligent household robot. In: 2010 International Conference on Artificial Intelligence and Computational Intelligence, 1, pp. 537–541. doi:10.1109/AICI.2010.118.
4. Yoon, W.-J., Cho, Y.-H., Park, K.-S., 2007. A study of speech emotion recognition and its application to mobile services. In: Indulska, J., Ma, J., Yang, L.T., Ungerer, T., Cao, J. (Eds.), Ubiquitous Intelligence and Computing. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 758–766.
5. Gupta, P., Rajput, N., 2007. Two-stream emotion recognition for call center monitoring. Proc. Interspeech 2007, 2241–2244.
6. Szwoch, M., Szwoch, W., 2015. Emotion recognition for affect aware video games. In: Choraś, R.S. (Ed.), Image Processing & Communications Challenges 6. Springer International Publishing, Cham, pp. 227–236.
7. Lancker, D.V., Cornelius, C., Kreiman, J., 1989. Recognition of emotional prosodic meanings in speech by autistic, schizophrenic, and normal children. Develop. Neuropsy- chol. 5 (2–3), 207–226.
8. Low, L.A., Maddage, N.C., Lech, M., Sheeber, L.B., Allen, N.B., 2011. Detection of clinical depression in adolescents 'speech during family interactions. IEEE Trans. Biomed. Eng. 58 (3), 574–586. doi:10.1109/TBME.2010.2091640.
9. Mehmet Berkehan Akçay, Kaya Oğuz, Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, Speech Communication, Volume 116, 2020.
10. AIP Conference Proceedings 1891, 020105 (2017)

11. Plutchik, R., 2001. The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Am. Sci. 89 (4), 344–350.
12. Ekman, P., Oster, H., 1979. Facial expressions of emotion. Ann. Rev. Psychol. 30 (1), 527–554.
13. Ekman, P., Friesen, W.V., Ellsworth, P., 2013. Emotion in the Human Face: Guidelines for Research and an Integration of Findings. Elsevier.
14. Ekman, P., 1971. Universals and cultural differences in facial expressions of emotion.. Nebraska symposium on motivation. University of Nebraska Press.
15. Watson, D., Clark, L.A., Tellegen, A., 1988. Development and validation of brief measures of positive and negative affect: the panas scales.. J. Personal. Soc. Psychol. 54 (6), 1063.
16. Russell, J.A., Mehrabian, A., 1977. Evidence for a three-factor theory of emotions. J. Res. Personal. 11 (3), 273–294.
17. Nicolaou, M.A., Gunes, H., Pantic, M., 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. IEEE Trans. Affect. Comput. 2 (2), 92–105.
18. Grimm, M., Kroschel, K., Narayanan, S., 2008. The vera am mittag german audio-visual emotional speech database. In: 2008 IEEE international conference on multimedia and expo. IEEE, pp. 865–868.
19. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S., 2009. A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Patt. Analy. Mach. Intell. 31 (1), 39–58.
20. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B., 2005. A database of German emotional speech.. In: Interspeech. ISCA, pp. 1517–1520.
21. Pohjalainen, J., Fabien Ringeval, F., Zhang, Z., Schuller, B., 2016. Spectral and cepstral audio noise reduction techniques in speech emotion recognition. In: Proceedings of the 24th ACM international conference on Multimedia. ACM, pp. 670–674.

22. Rao, K.S., Koolagudi, S.G., Vempada, R.R., 2013. Emotion recognition from speech using global and local prosodic features. Int. J. Speech Technol. 16 (2), 143–160.
23. Lin, J.-C., Wu, C.-H., Wei, W.-L., 2012. Error weighted semi-coupled hidden markov model for audio-visual emotion recognition. IEEE Trans. Multimed. 14 (1), 142–156.
24. Frick, R.W., 1985. Communicating emotion: the role of prosodic features. Psychol. Bull. 97 (3), 412.
25. Bachorowski, J.-A., 1999. Vocal expression and perception of emotion. Curr. Dir. Psychol. Sci. 8 (2), 53–57.
26. Kuchibhotla, S., Vankayalapati, H., Vaddi, R., Anne, K.R., 2014. A comparative analysis of classifiers in emotion recognition through acoustic features. Int. J. Speech Technol. 17 (4), 401–408.
27. Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., Pantic, M., 2016. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In: Proceedings of the 6th International Work- shop on Audio/Visual Emotion Challenge. ACM.
28. Gobl, C., Chasaide, A.N., 2003. The role of voice quality in communicating emotion, mood and attitude. Speech Commun. 40 (1–2), 189–212.
29. Laver, J., 1980. The Phonetic Description of Voice Quality / John Laver. Cambridge Uni- versity Press Cambridge [Eng.], New York.
30. Scherer, K.R., 1986. Vocal affect expression: a review and a model for future research. Psychol. Bull. 99 (2), 143.
31. Teager, H., Teager, S., 1990. Evidence for nonlinear sound production mechanisms in the vocal tract. In: Speech production and speech modelling. Springer, pp. 241–261.
32. Kaiser, J.F., 1990. On a simple algorithm to calculate the 'energy' of a signal. In: Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on. IEEE, pp. 381–384.

33. Zhou, G., Hansen, J.H., Kaiser, J.F., 2001. Nonlinear feature based classification of speech under stress. IEEE Trans. Speech Audio Process. 9 (3), 201–216.
34. Baum, L. E.; Petrie T. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. The Annals of Mathematical Statistics 37 (6): 1554–1563. doi:10.1214/aoms/1177699147
35. Nogueiras, A., Moreno, A., Bonafonte, A., Mariño, J.B., 2001. Speech emotion recognition using hidden markov models. In: Seventh European Conference on Speech Communication and Technology.
36. Schuller, B., Rigoll, G., Lang, M., 2003. Hidden markov model-based speech emotion recognition. In: Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on, 2. IEEE, pp. II–1.
37. Nwe, T.L., Foo, S.W., De Silva, L.C., 2003. Speech emotion recognition using hidden markov models. Speech Commun. 41 (4), 603–623.
38. Lin, Y.-L., Wei, G., 2005. Speech emotion recognition based on hmm and svm. In: 2005 international conference on machine learning and cybernetics, 8. IEEE, pp. 4898–4901.
39. Neiberg, D., Elenius, K., Laskowski, K., 2006. Emotion recognition in spontaneous speech using gmms. In: Ninth International Conference on Spoken Language Processing.
40. Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M.A., Schuller, B., Zafeiriou, S., 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In: Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, pp. 5200–5204.
41. Lim, W., Jang, D., Lee, T., 2016. Speech emotion recognition using convolutional and recurrent neural networks. In: 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). IEEE pp. 1–4.

42. Zhao, J., Mao, X., Chen, L., 2019. Speech emotion recognition using deep 1d & 2d cnn lstm networks. Biomed. Signal Process. Control 47, 312–323.
43. Caruana, R., 1997. Multitask learning. Mach. Learn. 28 (1), 41–75.
44. Kim, J., Truong, K.P., Englebienne, G., Evers, V., 2017. Learning spectro-temporal features with 3d cnns for speech emotion recognition. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). IEEE, pp. 383–388.
45. Mangalam, K., Guha, T., 2018. Learning spontaneity to improve emotion recognition in speech. Proc. Interspeech 2018, 946–950. doi:10.21437/Interspeech.2018-1872.
46. Mirsamadi, S., Barsoum, E., Zhang, C., 2017. Automatic speech emotion recognition using recurrent neural networks with local attention. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 2227–2231.
47. Chen, M., He, X., Yang, J., Zhang, H., 2018. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Process. Lett. 25 (10), 1440–1444.
48. Neumann, M., et al., 2018. Cross-lingual and multilingual speech emotion recognition on english and french. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5769–5773.
49. Goodfellow, I. J., Shlens, J., Szegedy, C., 2014. Explaining and harnessing adversarial examples. arXiv:1412.6572.
50. Sahu, S., Gupta, R., Sivaraman, G., Espy-Wilson, C., 2018. Smoothing model predictions using adversarial training procedures for speech based emotion recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4934–4938.
51. Livingstone SR, Russo FA (2018) 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 13(5): e0196391.

52. Pichora-Fuller, M. Kathleen; Dupuis, Kate, 2020, "Toronto emotional speech set (TESS)", https://doi.org/10.5683/SP2/E8H2MF, Scholars Portal Dataverse, V1
53. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel «Backpropagation Applied to Handwritten Zip Code Recognition» – 1989
54. Yung-Hsiang Shawn Chang, 2017, "Development of a Large-Scale Mandarin Radio Speech Corpus", National Taipei University of Technology, Taiwan.

ДОДАТКИ

Додаток А. Лістинг коду інтелектуального модуля мовою Python

```
import matplotlib.pyplot as plt
import librosa.display
import librosa

z, t = librosa.load('angry.wav')
plt.figure(figsize=(8, 4))
librosa.display.waveplot(z, sr=t)
plt.title('Чоловічий голос, гнів')
plt.xlabel('Час')
plt.savefig('wave-angry.png')
melsSpecs = librosa.feature.melspectrogram(y=z, sr=t, n_mels=128, fmax=8000)
melsSpecs = librosa.power_to_db(melsSpecs)
librosa.display.specshow(melsSpecs, y_axis='mel', fmax=8000, x_axis='time')
plt.title('Спектрограма - чоловічий голос, гнів')
plt.xlabel('Час')
plt.ylabel('Гц')
plt.savefig('MelsSpecs-angry.png')

z, t = librosa.load('sad.wav')
plt.figure(figsize=(8, 4))
librosa.display.waveplot(z, sr=t)
plt.title('Чоловічий голос, смуток')
plt.xlabel('Час')
plt.savefig('wave-sad.png')
melsSpecs = librosa.feature.melspectrogram(y=z, sr=t, n_mels=128, fmax=8000)
melsSpecs = librosa.power_to_db(melsSpecs)
librosa.display.specshow(melsSpecs, y_axis='mel', fmax=8000, x_axis='time')
plt.title('Спектрограма - чоловічий голос, смуток')
plt.xlabel('Час')
plt.ylabel('Гц')
plt.savefig('MelsSpecs-sad.png')
```

```
z, t = librosa.load('neutral.wav')
plt.figure(figsize=(8, 4))
librosa.display.waveplot(z, sr=t)
plt.title('Чоловічий голос, нейтральна емоція')
plt.xlabel('Час')
plt.savefig('wave-neutral.png')
melsSpecs = librosa.feature.melspectrogram(y=z, sr=t, n_mels=128, fmax=8000)
melsSpecs = librosa.power_to_db(melsSpecs)
librosa.display.specshow(melsSpecs, y_axis='mel', fmax=8000, x_axis='time')
plt.title('Спектрограма - чоловічий голос, нейтральна емоція')
plt.xlabel('Час')
plt.ylabel('Гц')
plt.savefig('MelsSpecs-neutral.png')

z, t = librosa.load('woman_neutral.wav')
plt.figure(figsize=(8, 4))
librosa.display.waveplot(z, sr=t)
plt.title('Жіночий голос, нейтральна емоція')
plt.xlabel('Час')
plt.savefig('woman_neutral.png')
melsSpecs = librosa.feature.melspectrogram(y=z, sr=t, n_mels=128, fmax=8000)
melsSpecs = librosa.power_to_db(melsSpecs)
librosa.display.specshow(melsSpecs, y_axis='mel', fmax=8000, x_axis='time')
plt.title('Спектрограма - жіночий голос, нейтральна емоція')
plt.xlabel('Час')
plt.ylabel('Гц')
plt.savefig('MelsSpecs-woman_neutral.png')

z, t = librosa.load('woman_anger.wav')
plt.figure(figsize=(8, 4))
librosa.display.waveplot(z, sr=t)
plt.title('Жіночий голос, гнів')
plt.xlabel('Час')
plt.savefig('woman_anger.png')
melsSpecs = librosa.feature.melspectrogram(y=z, sr=t, n_mels=128, fmax=8000)
melsSpecs = librosa.power_to_db(melsSpecs)
librosa.display.specshow(melsSpecs, y_axis='mel', fmax=8000, x_axis='time')
plt.title('Спектрограма - жіночий голос, гнів')
plt.xlabel('Час')
plt.ylabel('Гц')
plt.savefig('MelsSpecs-woman_anger.png')
```

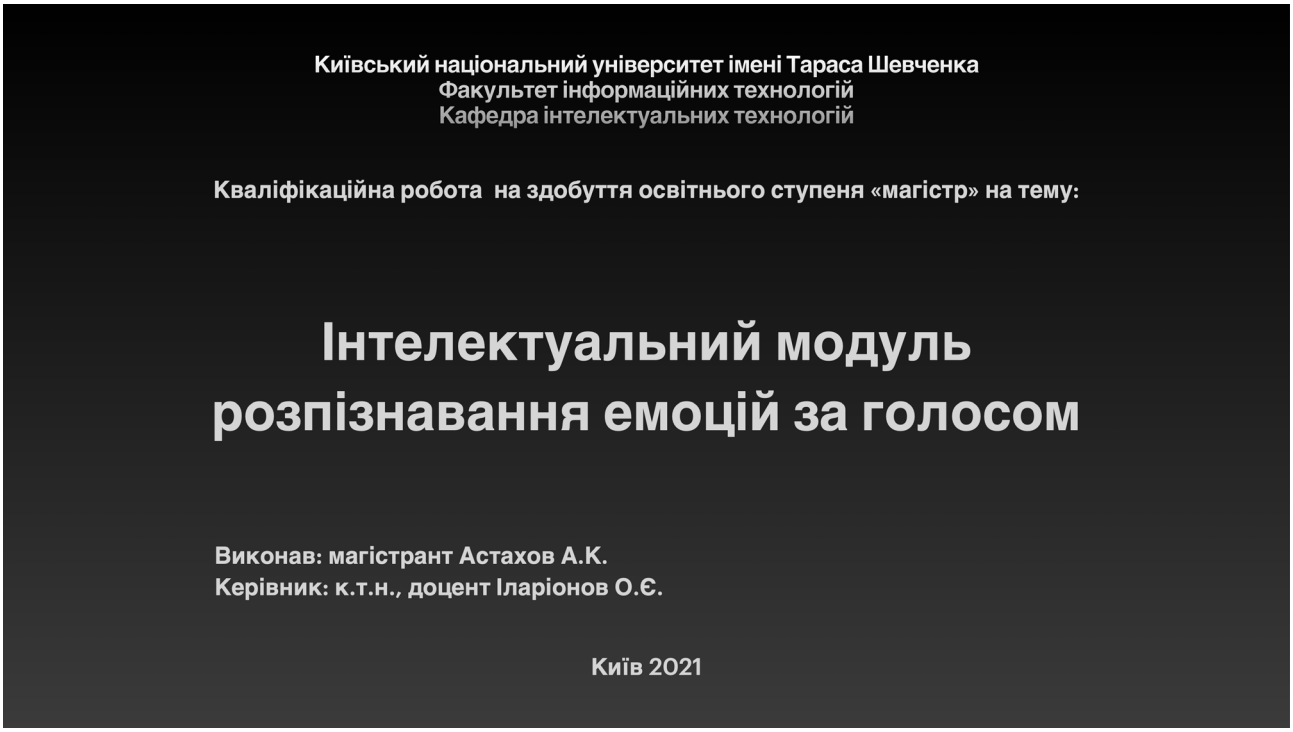
```
z, t = librosa.load('OAF_back_happy.wav')
plt.figure(figsize=(8, 4))
librosa.display.waveplot(z, sr=t)
plt.title('Жіночий голос, 64 роки, радість')
plt.xlabel('Час')
plt.savefig('OAF_back_happy.png')
melsSpecs = librosa.feature.melspectrogram(y=z, sr=t, n_mels=128, fmax=8000)
melsSpecs = librosa.power_to_db(melsSpecs)
librosa.display.specshow(melsSpecs, y_axis='mel', fmax=8000, x_axis='time')
plt.title('Спектрограма - жіночий голос, 64 роки, радість')
plt.xlabel('Час')
plt.ylabel('Гц')
plt.savefig('MelsSpecs-OAF_back_happy.png')

z, t = librosa.load('YAF_back_happy.wav')
plt.figure(figsize=(8, 4))
librosa.display.waveplot(z, sr=t)
plt.title('Жіночий голос, 26 років, радість')
plt.xlabel('Час')
plt.savefig('YAF_back_happy.png')
melsSpecs = librosa.feature.melspectrogram(y=z, sr=t, n_mels=128, fmax=8000)
melsSpecs = librosa.power_to_db(melsSpecs)
librosa.display.specshow(melsSpecs, y_axis='mel', fmax=8000, x_axis='time')
plt.title('Спектрограма - жіночий голос, 26 років, радість')
plt.xlabel('Час')
plt.ylabel('Гц')
plt.savefig('MelsSpecs-YAF_back_happy.png')

z, t = librosa.load('OAF_back_neutral.wav')
plt.figure(figsize=(8, 4))
librosa.display.waveplot(z, sr=t)
plt.title('Жіночий голос, 64 роки, нейтральна емоція')
plt.xlabel('Час')
plt.savefig('OAF_back_neutral.png')
melsSpecs = librosa.feature.melspectrogram(y=z, sr=t, n_mels=128, fmax=8000)
melsSpecs = librosa.power_to_db(melsSpecs)
librosa.display.specshow(melsSpecs, y_axis='mel', fmax=8000, x_axis='time')
plt.title('Спектрограма - жіночий голос, 64 роки, нейтральна емоція')
plt.xlabel('Час')
plt.ylabel('Гц')
plt.savefig('MelsSpecs-OAF_back_neutral.png')
```

```
z, t = librosa.load('YAF_back_neutral.wav')
plt.figure(figsize=(8, 4))
librosa.display.waveplot(z, sr=t)
plt.title('Жіночий голос, 26 років, нейтральна емоція')
plt.xlabel('Час')
plt.savefig('YAF_back_neutral.png')
melsSpecs = librosa.feature.melspectrogram(y=z, sr=t, n_mels=128, fmax=8000)
melsSpecs = librosa.power_to_db(melsSpecs)
librosa.display.specshow(melsSpecs, y_axis='mel', fmax=8000, x_axis='time')
plt.title('Спектрограма - жіночий голос, 26 років, нейтральна емоція')
plt.xlabel('Час')
plt.ylabel('Гц')
plt.savefig('MelsSpecs-YAF_back_neutral.png')


z, t = librosa.load('OAF_back_angry.wav')
plt.figure(figsize=(8, 4))
librosa.display.waveplot(z, sr=t)
plt.title('Жіночий голос, 64 роки, гнів')
plt.xlabel('Час')
plt.savefig('OAF_back_angry.png')
melsSpecs = librosa.feature.melspectrogram(y=z, sr=t, n_mels=128, fmax=8000)
melsSpecs = librosa.power_to_db(melsSpecs)
librosa.display.specshow(melsSpecs, y_axis='mel', fmax=8000, x_axis='time')
plt.title('Спектрограма - жіночий голос, 64 роки, гнів')
plt.xlabel('Час')
plt.ylabel('Гц')
plt.savefig('MelsSpecs-OAF_back_angry.png')


from keras.models import load_model
import joblib
from sklearn.model_selection import train_test_split
import numpy as np

myModel = load_model('speech_emotion_recognition_v2_1.h5')

X = joblib.load('X.joblib')
y = joblib.load('y.joblib')
X_training, X_testing, y_training, y_testing = train_test_split(X, y, test_size=0.33)

x_test = np.expand_dims(X_testing, axis=2)
```

```
print('\nТочність моделі:', myModel.evaluate(x_test, y_testing)[1])


import keras
import librosa
import numpy as np


class MyPrediction:

    def __init__(self, file):
        self.file = file
        self.path = 'speech_emotion_recognition_v2_1.h5'
        self.loaded_model = keras.models.load_model(self.path)

    def predict_now(self):
        data, sampling_rate = librosa.load(self.file)
        mfccs = np.mean(librosa.feature.mfcc(y=data, sr=sampling_rate, n_mfcc=40).T, axis=0)
        x = np.expand_dims(mfccs, axis=2)
        x = np.expand_dims(x, axis=0)
        predictions = self.loaded_model.predict_classes(x)
        print("Емоція:", self.number_emotion(predictions))

    def number_emotion(my_prediction):
        emotions = {'0': 'нейтральна',
                    '1': 'спокій',
                    '2': 'радість',
                    '3': 'смуток',
                    '4': 'гнів',
                    '5': 'страх',
                    '6': 'відраза',
                    '7': 'здивування'}

        for a, b in emotions.items():
            if int(a) == my_prediction:
                emotion = b
        return emotion


if __name__ == '__main__':
predictions = MyPrediction('greta.wav')  # гнів
```

```
predictions.loaded_model.summary()
predictions.predict_now()
predictions = MyPrediction('10-16-07-29-82-30-63.wav')  # відраза
predictions.predict_now()



from keras.layers import Dense
from keras.layers import Conv1D
from keras.layers import Flatten
from keras.layers import Dropout
from keras.layers import Activation
from keras.models import Sequential
from sklearn.model_selection import train_test_split



def cnn_training(Xs, ys) -> None:
    X_training, X_testing, y_training, y_testing = train_test_split(Xs, ys, test_size=0.33, random_state=42)

    x_train_model = np.expand_dims(X_training, axis=2)
    x_test_model = np.expand_dims(X_testing, axis=2)

    my_cnn_model = Sequential()
    my_cnn_model.add(Conv1D(64, 5, padding='same', input_shape=(40, 1)))
    my_cnn_model.add(Activation('relu'))
    my_cnn_model.add(Dropout(0.2))
    my_cnn_model.add(Flatten())
    my_cnn_model.add(Dense(8))
    my_cnn_model.add(Activation('softmax'))

    print(my_cnn_model.summary)
    my_cnn_model.compile(loss='sparse_categorical_crossentropy', optimizer='rmsprop', metrics=['accuracy'])

    my_history = my_cnn_model.fit(x_train_model, y_training, batch_size=16, epochs=50,
validation_data=(x_test_model, y_testing))

    plt.plot(my_history.history['loss'])
    plt.plot(my_history.history['val_loss'])
    plt.title('Похибка')
    plt.ylabel('Похибка')
    plt.xlabel('Епохи')
    plt.legend(['Навчальна вибірка', 'Тестова вибірка'], loc='upper right')
    plt.savefig('loss.png')
```

```
    plt.close()

    plt.plot(my_history.history['accuracy'])
    plt.plot(my_history.history['val_accuracy'])
    plt.title('Точність розпізнавання')
    plt.ylabel('Точність')
    plt.xlabel('Епохи')
    plt.legend(['Навчальна вибірка', 'Тестова вибірка'], loc='lower right')
    plt.savefig('accuracy.png')


def extract_mfccs(files_path):
    lst = []
    for subdirectories, directories, files_dir in os.walk(files_path):
        for my_file in files_dir:
            xs, my_rate = librosa.load(os.path.join(subdirectories, my_file), res_type='kaiser_fast')
            list_mfcc = np.mean(librosa.feature.mfcc(y=xs, sr=my_rate, n_mfcc=40).T, axis=0)
            my_file = int(my_file[7:8]) - 1
            features = list_mfcc, my_file
            lst.append(features)
```



Додаток Б. Презентаційні матеріали

Об’єкт дослідження

Процес розпізнавання емоцій людини за голосом з використанням технологій штучного інтелекту

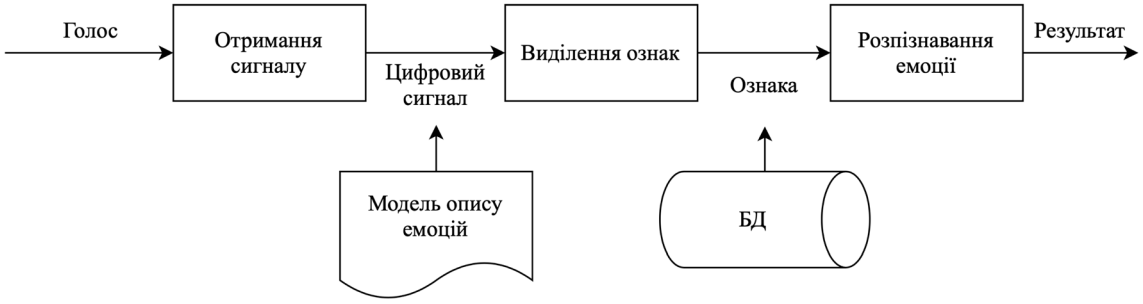
Мета дослідження

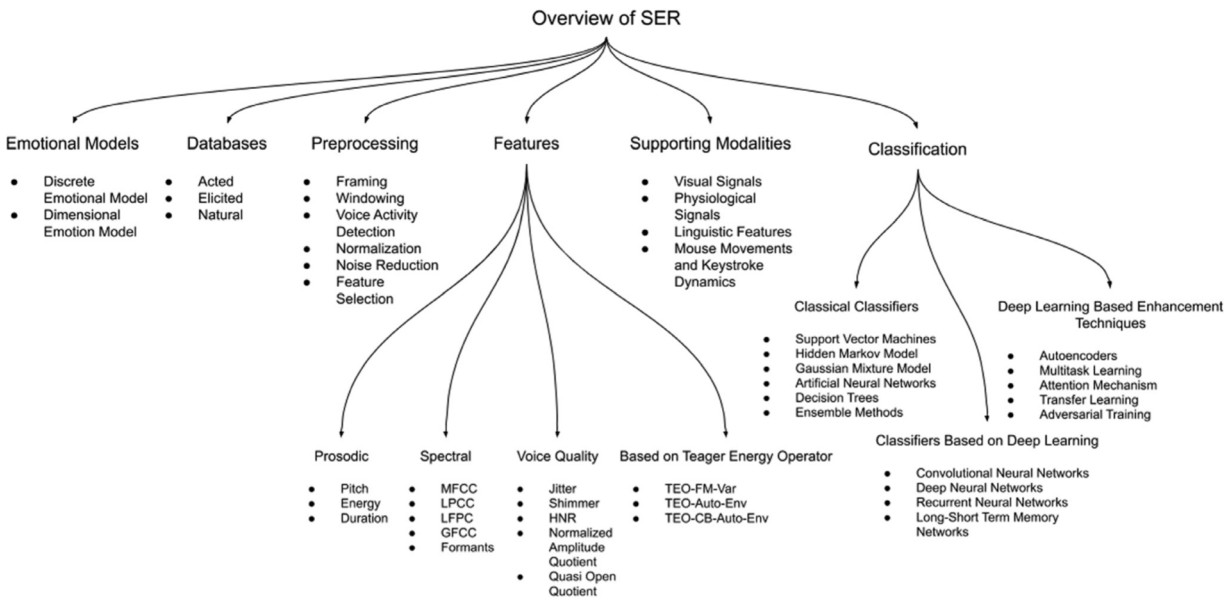
Розробити інтелектуальний модуль розпізнавання емоцій людини за голосом з використанням методів глибинного навчання

Новизна роботи

- Поєднання згорткових нейронних мереж, мел-частотних кепстральних коефіцієнтів та наборів даних RAVDESS і TESS
- За результатами програмної реалізації модуля розпізнавання емоцій за голосом збільшено точність валідації до 0.8

Система розпізнавання емоцій за голосом





***Джерело:** Mehmet Berkehan Akçay, Kaya Oğuz, Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, Speech Communication, Volume 116, 2020.*

Моделі емоційних станів

- Дискретна модель: радість, смуток, страх, гнів, відраза, здивування
- Багатовимірна модель: привабливість, збудження, контроль, влада, стрес, ...

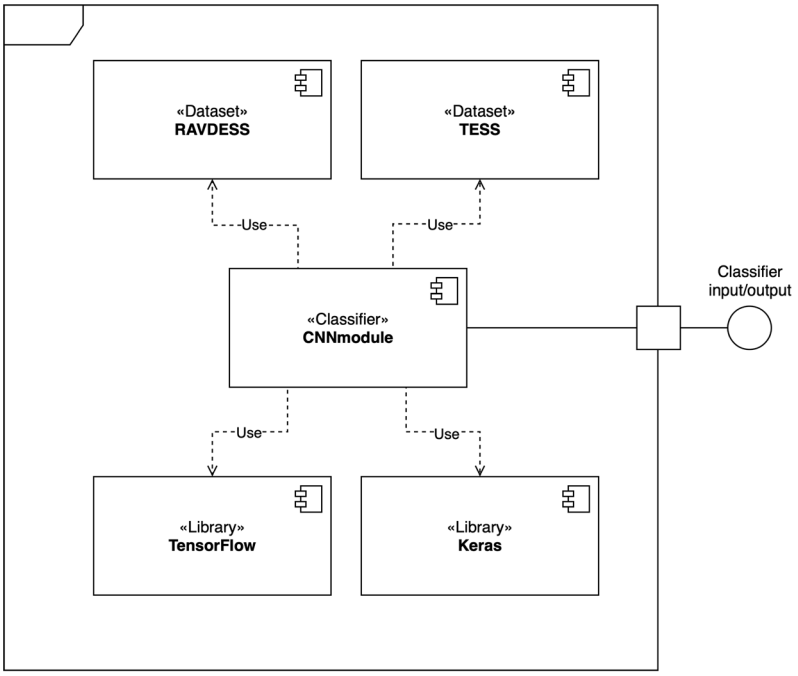
Набори даних
TESS

- 2 актори (жінки)
- 26 та 64 роки
- Англійська мова
- 8 емоційних станів
- 2800 аудіофайлів
- 200 слів

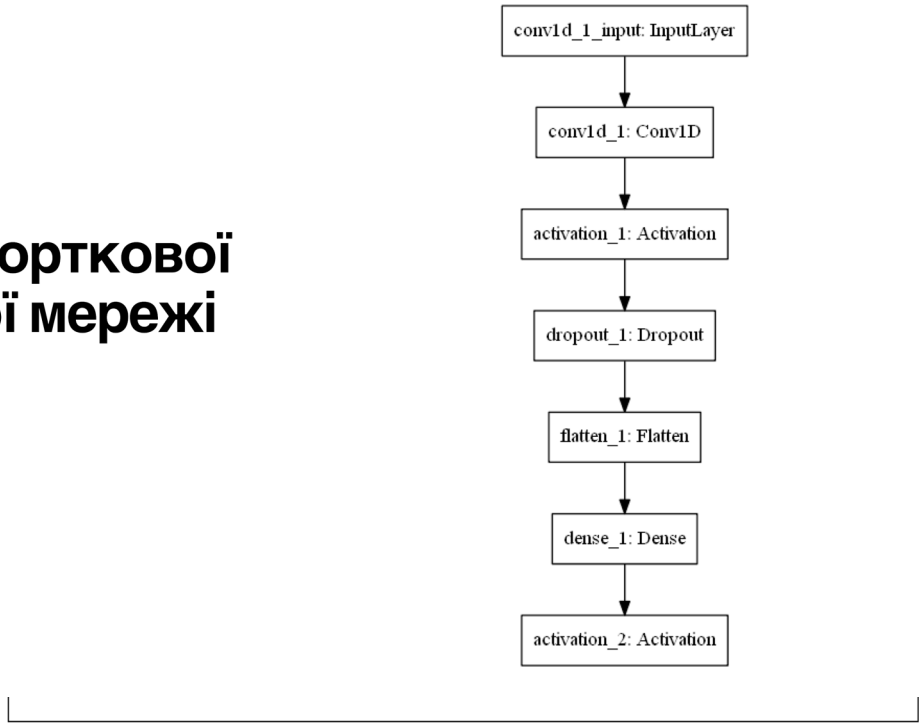
Класифікатори

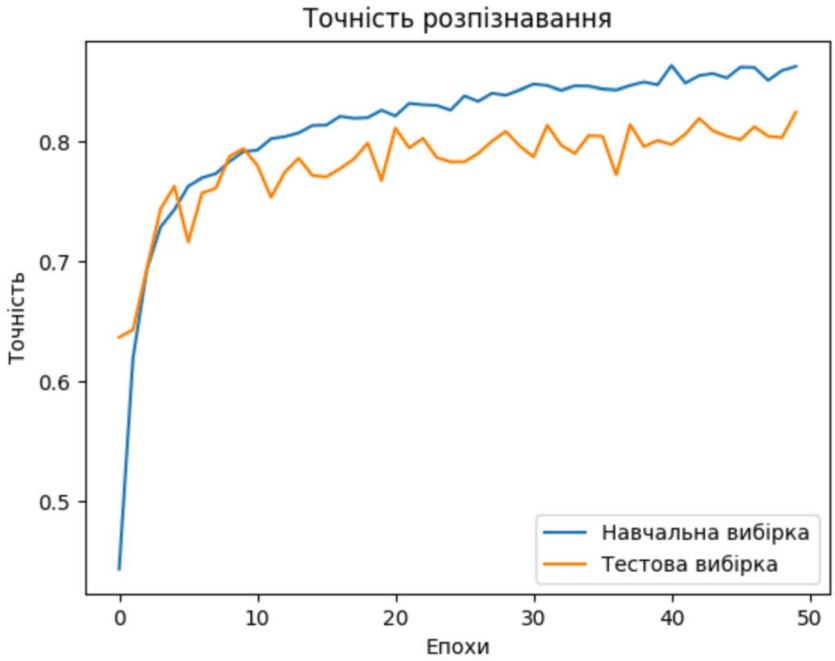
- Традиційні класифікатори (прихована марковська модель, гаусівська змішана модель, метод опорних векторів)
- Класифікатори на основі глибинного навчання (згорткові нейронні мережі, рекурентні нейронні мережі)
- Методи машинного навчання для покращення класифікації (автокодувальники, механізм уваги)

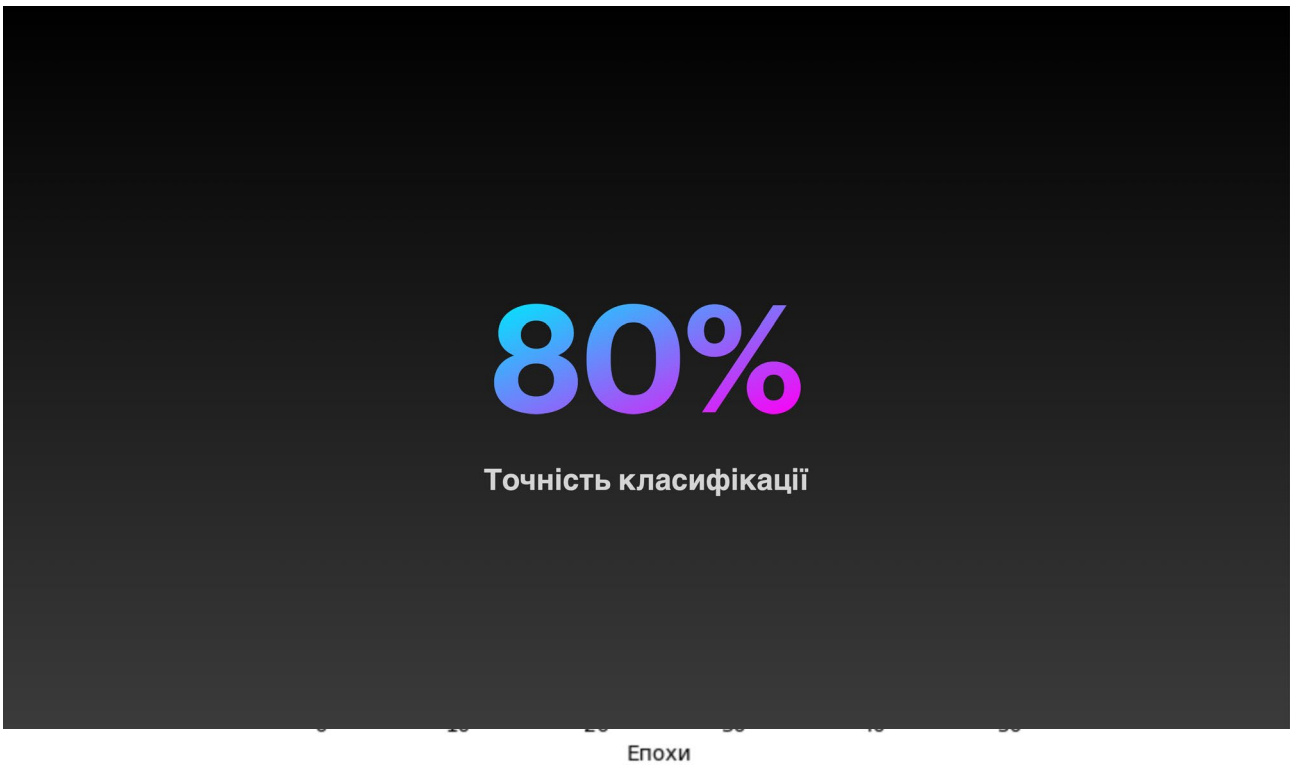
Діаграма компонентів

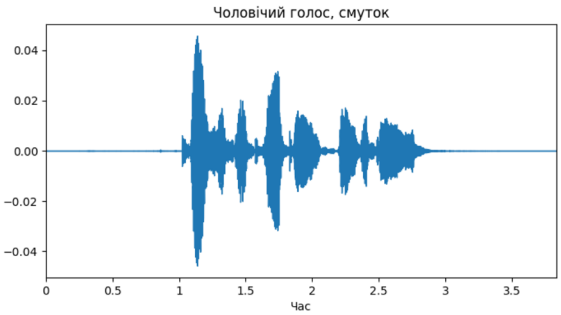


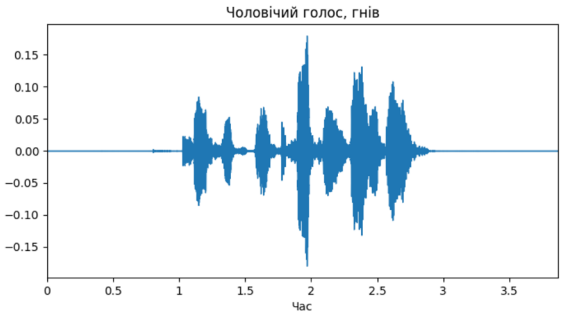
Модель згорткової нейронної мережі




Точність розпізнавання
Точність
0.8
0.7
0.6
0.5
0
10
20
30
40
50
Епохи
Навчальна вибірка
Тестова вибірка


80%
Точність класифікації
Епохи


Чоловічий голос, смуток
0.04
0.02
0.00
−0.02
−0.04
0
0.5
1
1.5
2
2.5
3
3.5
Час


Чоловічий голос, гнів
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
0
0.5
1
1.5
2
2.5
3
3.5
Час

Висновки

- Проведено аналіз методів та підходів до вирішення задачі розпізнавання емоцій за голосом
- Модифіковано існуючі методи, розроблено інтелектуальний модуль, досягнута точність валідації 0.8
- Виконана оцінка ефективності отриманих результатів дослідження
- Запропоновані сценарії подальшого використання дослідження

Дякую за увагу!