НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

ФАКУЛЬТЕТ ПРИКЛАДНОЇ МАТЕМАТИКИ

КАФЕДРА СИСТЕМНОГО ПРОГРАМУВАННЯ І
СПЕЦІАЛІЗОВАНИХ КОМП'ЮТЕРНИХ СИСТЕМ

«На правах рукопису»
УДК _____

«До захисту допущено»

Завідувач кафедри СПСКС

_____ В.П.Тарасенко_
(підпис) (ініціали, прізвище)

"___"_____2018р.

Магістерська дисертація

на здобуття ступеня магістра

зі спеціальності 123 Комп'ютерна інженерія
Системне програмування

на тему: Метод та алгоритми оцінювання емоційного стану людини на основі аналізу голосових сигналів

Виконав: студент II курсу, групи _КВ-72мп_____
(шифр групи)

Яблонський Сергій Вікторович_____ _____
(прізвище, ім'я, по батькові) (підпис)

Науковий керівник д.т.н., проф. Яценко В.О. _____ _____
(посада, науковий ступінь, вчене звання, прізвище та ініціали) (підпис)

Рецензент_____ _____
(посада, науковий ступінь, вчене звання, науковий ступінь, прізвище та ініціали) (підпис)

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.
Студент _____
(підпис)

Київ – 2018 року

Національний технічний університет України

«Київський політехнічний інститут імені Ігоря Сікорського»

Факультет прикладної математики

Кафедра системного програмування і спеціалізованих комп'ютерних систем

Рівень вищої освіти – другий (магістерський) за освітньо-професійною програмою

Спеціальність (спеціалізація) – 123 «Комп'ютерна інженерія» («Системне програмування»)

ЗАТВЕРДЖУЮ

Завідувач кафедри СПСКС

_____ В.П. Тарасенко

«___»_____2018 р.

ЗАВДАННЯ
на магістерську дисертацію студенту

Яблонському Сергію Вікторовичу

1. Тема дисертації «Метод та алгоритми оцінювання емоційного стану людини на основі аналізу голосових сигналів», науковий керівник дисертації Яценко Віталій Олексійович, д.т.н., професор, затверджені наказом по університету від «30» жовтня 2018 р. №4030-с

2. Термін подання студентом дисертації «7» грудня 2018 р.

3. Об'єкт дослідження: процеси формалізації та математичної обробки числових даних, які характеризують розпізнавання емоції людини за її голосом.

4. Предмет дослідження: математична модель, яка характеризує розпізнавання емоції людини з визначеної множини емоцій.

5. Перелік завдань, які потрібно розробити:
   - Розгляд реалізації розробки методу оцінювання емоц. стану на основі голосу;
   - Опис оптимізованого методу оцінювання емоц. стану на основі голосу;
   - Аналіз існуючих методів оцінювання емоц. стану людини на основі голосу;
   - обґрунтувати вибір критеріїв оптимізації математичної моделі впливу техногенних факторів;
   - Створення структури програмного забезпечення
   - запропонувати методику комп'ютерного оцінювання емоційного стану;

6. Орієнтовний перелік графічного (ілюстративного) матеріалу:
   - теоретичні аспекти побудови та оптимізації математичної моделі;
   - демонстраційні таблиці голосових даних;

7. Орієнтовний перелік публікацій:

- Тези доповіді “Аналіз методів машинного навчання штучних нейронних мереж для розпізнавання емоцій за голосом”
- Тези доповіді “Методи глибинного машинного навчання для діагностики емоції людини за голосом”

8. Консультанти розділів дисертації

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв

9. Дата видачі завдання «04» жовтня 2017 р.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1.	Ґрунтовне ознайомлення з предметною галуззю	17.10.2017	
2.	Визначення структури магістерської дисертації; вивчення літератури, пошук додаткової літератури, патентний пошук	04.12.2017	
3.	Робота над першим розділом магістерської дисертації; проведення наукового дослідження	15.02.2018	
4.	Проведення наукового дослідження; робота над другим розділом магістерської дисертації; розроблення програмного забезпечення	05.04.2018	
5.	Проведення наукового дослідження; робота над статтею за результатами наукового дослідження	15.05.2018	
6.	Проведення наукового дослідження; робота над третім розділом магістерської дисертації	15.06.2018	
7.	Завершення роботи над основною частиною магістерської дисертації; підготовка ілюстративного матеріалу; підготовка матеріалів доповіді на конференції ПМК-2018	05.11.2018	
8.	Оформлення текстової і графічної частини магістерської дисертації	04.12.2018	

Студент С.В. Яблонський

Науковий керівник дисертації В.О. Яценко

РЕФЕРАТ

Актуальність теми. Автоматизоване розпізнавання емоційних станів на сьогодні є невирішеною проблемою, внаслідок того, що людські емоції зазвичай зовні слабо виражені й швидко змінюються. Прояв емоції людини може бути зафіксований зняттям показів датчиків фізичного стану (тиску, температури поверхні тіла та органів, електромагнітної активності мозку), але переважна більшість таких характеристик можуть бути отримані у безпосередньому контакті з людиною, що робить неможливим застосування характеристик на практиці.

Об'єктом дослідження є процеси формалізації та математичної обробки числових даних, які характеризують розпізнавання емоції людини за її голосом.

Предметом дослідження є математична модель, яка характеризує розпізнавання емоції людини з визначеної множини емоцій.

Мета роботи: розробка власного методу, здатного розпізнавати наявність певної емоції (із визначеної множини) у людини за її голосом.

Методи дослідження. В роботі використовуються методи математичного моделювання, методи оптимізації, синергетичні методи.

Наукова новизна роботи полягає в наступному:

1. Розроблено математичну модель розпізнавання певної емоції (із визначеної множини) у людини за спектрограмою її голосу.

2. Вперше запропоновано синергетичний метод розпізнавання емоцій людини за спектрограмою її голосу, що дозволило враховувати особливості аналізу спектрограми.

Практична цінність отриманих в роботі результатів полягає в тому, що запропоновані методи та засоби дають змогу з точністю 90% розпізнати певну емоцію у людини за спектрограмою її голосу.

Ідентифікація людини може підвищити ймовірність розпізнавання емоційних станів і стати подальшим розвитком системи

Апробація роботи. Основні положення і результати роботи були представлені та обговорювались на II науковій конференції магістрантів та аспірантів «Прикладна математика та комп'ютинг» ПМК-2018(Київ, 14-16 листопада 2018р.) та на V Міжнародній науково-технічній Internet-конференції «Сучасні методи, інформаційне програмне та технічне забезпечення систем керування організаційно-технічними та технологічними комплексами»(Київ, 22 листопада 2018р.).

Структура та обсяг роботи. Магістерська дисертація складається з вступу, чотирьох розділів, висновків та додатків.

У вступі надано загальну характеристику роботи, виконано оцінку сучасного стану проблеми, обґрунтовано актуальність напрямку досліджень, сформульовано мету і задачі досліджень, показано наукову новизну отриманих результатів і практичну цінність роботи, наведено відомості про апробацію результатів і їх впровадження.

У першому розділі розглянуто основні акустичні властивості голосу як індикатори депресії які можуть бути використані в медицині.

У другому розділі проведено аналіз існуючих методів розпізнавання емоції за голосом, та обрано метод машинного навчання.

У третьому розділі досліджено проблеми, пов'язані з аналізом і обробкою спектрограм; запропоновано програмну реалізацію алгоритму конвертації масивів звукових файлів в спектрограми.

У четвертому розділі проведено аналіз результатів машинного навчання, та класифікації емоцій за голосовими сигналами.

У висновках проаналізовано отримані результати роботи.

Магістерська дисертація виконана на 81 аркушах, містить 2 додатки та посилання на список використаних літературних джерел з _ найменувань. У роботі наведено _ рисунків та 24 таблиць.

Ключові слова: метод оптимізації, математична модель, синергетичний метод.

ABSTRACT

Actuality of theme. The automated recognition of emotional states today is an unresolved problem, as human emotions are usually outwardly weakly expressed and rapidly changing. The man's emotional manifestation can be recorded by removing the signals of the physical state sensors (pressure, body surface temperature and organs, electromagnetic activity of the brain), but the vast majority of such characteristics can be obtained in direct contact with the person, which makes it impossible to apply the characteristics in practice.

The object of research is the processes of formalization and mathematical processing of numerical data that characterize the recognition of human emotions in her voice.

The subject of the study is a mathematical model that characterizes the recognition of emotions of a person from a certain set of emotions.

The purpose of the work: the development of an own method, capable to recognize the presence of a certain emotion (from a certain set) of a person in her voice.

Research methods. Methods of mathematical modeling, methods of optimization, synergetic methods are used in this work.

The scientific novelty of the work is as follows:

1. A mathematical model of the recognition of a certain emotion (from a certain set) of a person is developed based on the spectrograph of her voice.

2. For the first time, a synergistic method for recognizing human emotions according to the spectrogram of its voice was proposed, which allowed taking into account the peculiarities of spectrograph analysis.

The practical value of the results obtained in the work consists in the fact that the proposed methods and facilities give the ability to accurately detect a certain emotion in a person by the spectrogram of her voice with an accuracy of 90%. Identification of a person can increase the likelihood of recognition of emotional states and become a further development of the system

Test work. The main provisions and results of the work were presented and discussed at the II scientific conference of masters and postgraduates "Applied Mathematics and Computer", PMK-2018 (Kyiv, November 14-16, 2018) and at the V International Scientific and Technical Internet Conference "Modern Methods , informational software and technical support of control systems for organizational, technological and technological complexes "(Kyiv, November 22, 2018).

Structure and scope of work. The master's dissertation consists of an introduction, four sections, conclusions and appendices.

The introduction provides a general description of the work, an assessment of the current state of the problem is carried out, the relevance of the direction is substantiated researches, the purpose and tasks of researches are formulated, the scientific novelty of the received results and practical value of work is shown, information on the testing of the results and their implementation is given.

The first chapter examines the basic acoustic properties of the voice as indicators of depression that can be used in medicine.

In the second section an analysis of the existing methods of emotional recognition by voice is conducted, and the method of machine learning is chosen.

The third section deals with the problems associated with the analysis and processing of spectrograms; The program implementation of the algorithm for converting sound file arrays into spectrographs is proposed.

In the fourth section, the analysis of the results of machine learning, and the classification of emotions by voice signals.

The conclusions are analyzed the results of work.

The master's dissertation is executed on _ sheets, contains _ applications and a link to the list of used literary sources from _ names. The work presents _ drawings and _ tables.

Keywords: optimization method, mathematical model, synergetic method.

РЕФЕРАТ

Актуальность темы. Автоматизированное распознавание эмоциональных состояний на сегодня является нерешенной проблемой, вследствие того, что человеческие эмоции обычно внешне слабо выражены и быстро меняются. Проявление эмоции человека может быть зафиксирован снятием показаний датчиков физического состояния (давления, температуры поверхности тела и органов, электромагнитной активности мозга), но подавляющее большинство таких характеристик могут быть получены в непосредственном контакте с человеком, что делает невозможным применение характеристик на практике.

Объектом исследования являются процессы формализации и математической обработки числовых данных, характеризующих распознавания эмоции человека по его голосу.

Предметом исследования является математическая модель, характеризующая распознавания эмоции человека с определенной множества эмоций.

Цель работы: разработка собственного метода, способного распознавать наличие определенной эмоции (с определенной множества) у человека по его голосу.

Методы исследования. В работе используются методы математического моделирования, методы оптимизации, синергетические методы.

Научная новизна работы заключается в следующем:

1. Разработана математическая модель распознавания определенной эмоции (с определенной множества) у человека за спектрограммой ее голоса.

2. Впервые предложен синергетический метод распознавания эмоций человека по спектрограммой ее голоса, что позволило учитывать особенности анализа спектрограммы.

Практическая ценность полученных в работе результатов заключается в том, что предложенные методы и средства позволяют с точностью 90% распознать определенную эмоцию у человека за спектрограммой ее голоса. Идентификация человека может повысить вероятность распознавания эмоциональных состояний и стать дальнейшим развитием системы

Апробация работы. Основные положения и результаты работы были представлены и обсуждались на II научной конференции магистрантов и аспирантов «Прикладная математика и компьютинг» ПМК-2018 (Киев, 14-16 ноября 2018р.) И на V Международной научно-технической Internet-конференции «Современные методы , информационное программное и техническое обеспечение систем управления организационно-техническими и технологическими комплексами» (Киев, 22 ноября 2018р.).

Структура и объем работы. Магистерская диссертация состоит из введения, четырёх глав, заключения и приложений.

Во введении дано общая характеристика работы, выполнена оценка современного состояния проблемы, обоснована актуальность направления исследований, сформулированы цели и задачи исследований, показано научную новизну полученных результатов и практическую ценность работы, приведены сведения об апробации результатов и их внедрение.

В первом разделе рассмотрены основные акустические свойства голоса как индикаторы депрессии которые могут быть использованы в медицине.

Во второй главе проведен анализ существующих методов распознавания эмоции по голосу, и выбран метод машинного обучения.

В третьем разделе исследованы проблемы, связанные с анализом и обработкой спектрограмм; предложено программную реализацию алгоритма конвертации массивов звуковых файлов в спектрограммы.

В четвертом разделе проведен анализ результатов машинного обучения, и классификации эмоций по голосовым сигналами.

В выводах проанализированы полученные результаты работы.

Магистерская диссертация выполнена на _ листах, содержит _ приложений и ссылки на список использованных литературных источников из _ наименований. В работе приведены _ рисунков и _ таблиц.

Ключевые слова: метод оптимизации, математическая модель, синергетический метод.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ, ПОЗНАЧЕНЬ, ТЕРМІНІВ

Персептрон (perceptron) - це однорівневе з'єднання нейронів Маккаллоу-Пітта, які називаються одношаровими мережами прямого доступу

Згорткова нейронна мережа (recurrent neural network)- це клас глибоких штучних нейронних мереж прямого розподілу, який найчастіше використовувався для аналізу візуальних образів.

Дискретизація по часу – це процес обчислення миттєвих значень аналового сигналу, що перетворюється, з деяким кроком у часі, який називають крок дискретизації

RNN – згорткова нейромережа

LTAS – метод довгострокового середнього спектру

ШНМ – штучна нейронна мережа

KSGD – календарний каскад стохастичного градієнтного спуску

БШП – багатошаровий персептрон

ReLU – шар зрізаних лінійних вузлів

SVM – метод опорних векторів

PPV – позитивна прогностична цінність

NPV – негативна прогностична цінність

CBR – постійна швидкість передачі даних

VBR – змінна швидкість передачі даних

ABR – середня швидкість передачі даних

AdaGrad – адаптивний градієнтний алгоритм

RMSProp – розповсюдження кореневого середнього квадрата

ВСТУП

Усі сучасні операційні системи мають вбудованого асистента, який може розпізнавати голос, виконувати прості голосові команди та спілкуватися з користувачем. На момент написання дипломної роботи, жоден такий асистент не розпізнає емоції людини за голосом.

Найрозповсюдженіший сьогодні метод розпізнавання голосу, об’єктів на зображенні, емоцій людини за зображенням – навчання штучної нейронної мережі. Даний метод полягає в тому що ми «навчаємо» комп’ютер знаходити результат на певних зразках, після чого комп’ютер намагається знайти результат на інших вхідних даних.

За рахунок використання синергетичного підходу та штучної нейронної мережі, можна розробити програму яка досить точно розпізнає емоції людини за голосом.

Дана робота розглядає саме метод та алгоритми оцінювання емоційного стану людини на основі аналізу голосових сигналів. Показаний в роботі метод в подальшому можливо реалізувати як одну складову вбудованого в операційну систему асистента.

1. АКУСТИЧНІ ВЛАСТИВОСТІ ГОЛОСУ ЯК ІНДИКАТОРИ ДЕПРЕСІЇ ТА ЇХ МОЖЛИВОСТІ ВИКОРИСТАННЯ В МЕДИЦИНІ

1.1 Проблема розпізнання емоцій людини за голосом

Акустичні властивості голосу були визначені як можливі ознаки депресії, і є докази що деякі вокальні параметри можуть бути використані щоб об’єктивно розрізняти голос людини у депресії або схильність до самогубства. Багато різних науковців проводили дослідження для аналізу та порівняння мовної акустики окремих чоловічих та жіночих зразків, що складаються з нормальних осіб та осіб, що несуть діагнози депресії та високого ризику, короткочасної суїцидальності. За оцінками, до одної з восьми осіб потребує лікування депресивних захворювань в своєму житті. Поширеність депресії – четверта в світі серйозна загроза для здоров’я, і як очікується, стабільно зростатиме при старінні світового населення. Подібні тенденції підкреслюють, що депресія та самогубство – серйозні та пов’язані з цим проблеми зі здоров’ям заслуговують на аналогічну увагу та ресурси, які зараз присвячені для виявлення, лікування та викорінення таких загроз для здоров’я.

Переважаюча проблема профілактики самогубств – це визначення ступені ризику суїциду в окремих пацієнтів. Досвідчені лікарі в даний час використовують історію, психологічні тестування, звіти від пацієнтів, звіти від інших людей та інші методи оцінки ризику. Поточна техніка оцінювання емоційного стану споживає значні обсяги часу, коли необхідна термінова оцінка ризику. Деякі з найбільш корисних інструментів та тих що використовуються для виявлення депресії, не можуть виявити схильність пацієнта до суїциду. Комп’ютерна інтерпретація важливих даних з усіх джерел включаючи досвідченого лікаря є нещодавнім діагностичним прогресом, але зараз не є широко доступним.

Медичні працівники потребують покращених фізіологічних показників, щоб доповнити їх методи дослідження і оцінити ризик суїциду. Зараз досліджуються вимірювання та моніторинг коливань біохімічних маркерів депресії та самогубства. Проте нейробіологічне походження депресії та самогубства ще не з'ясовано, що не дозволяє використовувати цю інформацію для прогнозування ризику. Альтернативою є підхід до проблеми з точки зору системи. В основі такого підходу лежить теорія про те, що функціонування системи та поведінка безпосередньо чи опосередковано впливають на біохімічні зміни, що виникають у станах депресії та самогубства. Було запропоновано, що, якщо зміна рівня системи може бути кількісно визначена та підтверджена співвідношенням із психологічним станом, то вони можуть бути використані для виявлення та моніторингу депресії та оцінки суїцидального ризику. Крім того, шляхом моніторингу таких системних змін у людей, дослідники зможуть зрозуміти нейробіологічні механізми депресії та самогубства.

Зростає наукове підтвердження того, що психомоторні розлади є найранішими та найбільш послідовними показниками розладів настрою. Було показано, що психомоторні симптоми передбачають позитивну відповідь на трициклічні антидепресанти та інгібітори зворотного захоплення серотоніну (SSRI). Встановлено, що між нормальними та депресивними групами людей спостерігаються відмінності у ваговій руховій активності, рух тіла, мовлення та тривалість моторного відгуку. Метою даного дослідження було подальше вивчення акустичних властивостей голосу як психомоторного симптома депресії та суїцидальності, а також перевірити його силу як диференційований діагностичний параметр емоцій людини. Деякі вокальні параметри раніше були визначені як можливі сигнали до депресії, і є дані, що ці параметри можуть використовуватися далі, щоб диференціювати депресію та

самогубство. В ідеалі, діагностичний прилад, що базується на мовній акустиці, міг би дати фахівцям кількісну фізіологічну міру короткотермінових суїцидальних станів високого ризику, що, таким чином, додасть до клінічних орієнтирів рішення, що стосуються госпіталізації та випуску пацієнтів. Крім того, ця кількісна міра може дати лікарям і терапевтам покращену метрику, за допомогою якої вони зможуть оцінити ефективність різних способів зменшення суїцидальності, такі як когнітивні, психофармакологічні та електросудомні терапії, як це було зроблено при депресії. Протягом багатьох десятиліть мова була зосереджена на значних дослідженнях, пов'язаних з психомоторними порушеннями, пов'язаними з психічними розладами. Депресивна мова часто характеризується як монотонна, рівноголосна, млява. Ці перцептивні властивості були пов'язані з акустичними коливаннями, що включали фундаментальні частоти, амплітудну модуляцію (АМ), структуру форманта, розподіл потужності, частоту та тривалість паузи. Попередні акустичні аналізи депресивної мови були виконані з метою кількісної оцінки цих перцептивних якостей. Прориви були зроблені в цій області, і дискримінаційні вправи, що включали нормальну, депресивну та шизофренічну промову, були успішно впроваджені з використанням ознак, що описують, АМ, структуру формант та розподіл потужності.

Середня, дисперсія, контур та куртоз були визначені як ключова статистика, корельована з депресією. Було також продемонстровано, що діапазон гучності і середня інтенсивність мовлення також корелює з депресивним настроєм. Формантичні дослідження, які дають важливе уявлення про поведінку вокальних шляхів при мовленнєвому виробництві та артикуляції, показали, що центральні частоти другого та третього формантів людини зменшуються в періоди депресії. Крім того, значне збільшення частоти формантів вимірювалось після покращення

індивідуума з депресивного епізоду. Розподіл потужності мови також визначався для подібного зсуву до та після лікування депресивних захворювань. Дослідження показали, що мова сильно депресивних індивідів містить більше енергії на частотах вище 500 Гц. Після обробки, зміни потужності вимірювалися на нижчих частотах.

Підтвердженням багатьох дослідників про необхідність здійснення інтегрованих багатовимірних аналізів на природній мові непрофесійних депресивних осіб було мотивація першого дослідження, представленого в цій роботі. Класифікатори, що використовують комбінації функцій вокального параметру, були перевірені та порівняні. Попередні дослідження депресії переважно використовували окремі статистичні заходи (тобто середні або дисперсійні), зібрані з одного голосового параметра для виконання завдань дискримінації. Очікуваний покращений показник чутливості, стабільності та точності класифікатора був отриманий в результаті впровадження цього інтегрованого підходу. Основна мета цього дослідження полягала в тому, щоб забезпечити неушкоджені заходи впливу депресії на мову. В даний час невідомо, чи побічні ефекти психоактивних ліків маскують психологічні наслідки для мови, тому ідеальним експериментальним сценарієм є той, який уникає використання суб'єктів які вживали дані препарати.

Завдання другого дослідження полягали у вивченні акустичних властивостей депресивної мови за допомогою багатовимірних аналітичних методів та проведення аналогічного, але першого разу, аналізу суїцидальної речі з високим ризиком. Кінцевою метою було визначити порівняльний аналіз, якщо існують значні акустичні відмінності між пробами мови, використовуваними для представлення цих груп. У цьому дослідженні визначено зв'язок між акустичними властивостями депресивної та суїцидальної мови з високим ступенем ризику, щоб

визначити, чи можна використовувати відмінності для розрізнення груп та класифікації психічного стану осіб.

1.2 Голосові характеристики депресії у жінок

Аудіозаписи з 38 пацієнток та десять жіночих терапевтів були вилучені з комплекту даних Vanderbilt II, існуючої стрічкової бібліотеки, для складання класів, що представляють собою контрольні, дистимічні та великі депресивні популяції. Всі піддослідні були від 27 до 60 років. Пацієнти в дослідженні Vanderbilt II були відібрані від осіб, які прийши за рекламою про недорогу психотерапією. На момент запису ці пацієнти були фізично здоровими і вважали, що вони не потребують альтернативного психологічного лікування, такого як медикаментозна терапія чи госпіталізація. Діагнози, засновані на третьому випуску діагностичного статистичного посібника (DSM), були призначені спеціальним лікарем в інтерв'ю, використовуючи комп'ютерну версію діагностичного інтерв'ю Національного інституту психічного здоров'я (NIMH). Пацієнти також були проаналізовані на інші психологічні, психотичні та неврологічні розлади.

У поточному дослідженні були включені лише ті пацієнти, які відповідали критеріям DSM-IV для дистимічного розладу, великого депресивного епізоду або основного депресивного розладу. У таблиці 1.1 підсумовуються діагнози DSM-IV пацієнтів та відповідні міжнародні класифікації захворювань, дев'яте видання, Клінічна модифікація (ICD-9-CM).

Таблиця 1.1 – DSM-IV діагнози і ICD-9-CM коди жінок з депресією

ICD-9-CM код	Діагноз DSM-IV	Кількість пацієнтів

300.40	Дистимічний розлад	17
296.22	Велика депресія, єдиний епізод (помірний)	1
296.32	Велика депресія, рецидив (помірний)	17
296.33	Велика депресія, рецидив (важка, без психотичних особливостей)	3

Контрольний клас складався з ліцензованих психологів та психіатрів, які пройшли навчання в психодіамінальній терапії для дослідження Vanderbilt II. Сегменти мовлення десяти жіночих терапевтів були випадково вибрані з записів або в сесійному діалозі пацієнт-терапевт, або після сесійних дисциплін і оцінок психотерапевта. І, нарешті, не було жодної пацієнтки з проблемами слуху, мови, мовних розладів або хронічних респіраторних захворювань.

Записи, використані в цьому дослідженні, були відібрані з сеансів терапії та дисциплін після сесії, що відбулися під час третього візиту пацієнта до офісу дослідження Vanderbilt II. Пацієнти зустрічалися з терапевтами щотижня протягом 25-ти тижнів, а також були зроблені аудіо та відеозаписи під час кожного з цих сеансів психотерапії. Зразки аудіо, використовувані в дослідженні, були зібрані з аудіозапису. Аудіозаписи раннього сеансу в програмі лікування були відібрані для того, щоб забезпечити вимірювання характеристик вироблення мовлення дистимії та основної депресії. Очікувалось, що психологічний стан суб'єктів покращиться протягом курсу лікування, а мета була виміряти параметри голосу до покращення. Записи першого та другого візиту офісу не використовувались, оскільки було висунуто гіпотезу, що стрес пацієнтів та психотерапевтів буде більш високим під час цих початкових візитів, особливо внаслідок їх усвідомлення запису аудіо та відеозаписів.

1.3 Голосові характеристики депресії у чоловіків

Аудіозаписи 24 нормальних пацієнтів, 21 пацієнт з депресією та 22 пацієнти із підвищеним ризиком самогубства були зібрані з існуючих баз даних для складання класів, що представляють кожну з цих груп. Всі пацієнти, які використовувались у дослідженні, були білими чоловіками у віці від 25 до 65 років. Клас контролю складався з сегментів мовлення з 24 терапевтів, довільно вибраних із записів діалогу пацієнта і психотерапевта. Ці терапевти були ліцензованими психологами та психіатрами, які навчалися в психодіамінальній теорії та пізнавальної психотерапії.

Аудіозаписи з 21 пацієнтом з депресією були зібрані з дослідження Vanderbilt II (6 пацієнтів), а також з дослідження, яке порівнювало ефекти когнітивної та фармакотерапії на депресію (15 пацієнтів). Непсихотичні, небіполярні депресивні амбулаторні пацієнти були основою дослідження щодо ефектів когнітивної терапії та фармакотерапії від депресії. Ці пацієнти відповідали наступним критеріям прийому:

1) основний депресивний розлад, як визначено в діагностичних критеріях дослідження (RDC);

2) показник депресії Бек 20 або більше;

3) шкала рейтингів Гамільтона (HRSD, 17-елементна версія) для оцінки депресії 14 або більше.

Критерії виключення були накладені на екран для інших психологічних, психотичних та медичних станів, які могли б перешкоджати цілям дослідження. Пацієнти, які відповідали всім вимогам до дослідження, були випадково віднесені до одного з чотирьох умов:

1) фармакотерапія без продовження;

2) фармакотерапія плюс продовження;

3) когнітивна терапія;

4) комбінована когнітивно-фармакотерапія.

Пацієнтів, призначених для когнітивної терапії (окремо або комбіновано), спостерігали максимум 20 сеансів кожні 50 хв протягом 12 тижнів. Пацієнти, призначені для фармакотерапії, отримували імірамін гідрохлорид протягом 12 тижнів. Рівень дозування починається з 75 мг / добу для кожного пацієнта і до кінця третього тижня збільшується до максимум 300 мг / добу. Рівні дозування проводилися постійно після третього тижня лікування.

Записи контрольних пацієнтів та сильно депресивних пацієнтів були відібрані з сеансів терапії та дисциплін після сесії, що відбулися під час третього візиту пацієнта до офісу Vanderbilt II або другого офіційного візиту до дослідження фармакотерапії. Аудіозаписи перших сеансів у програмах лікування були відібрані для того, щоб гарантувати, що голосові ефекти від основної депресії вимірюються більше, ніж ефекти або побічні ефекти, які приносять лікування. Як і в дослідженні голосових властивостей депресії у жінок, записи з першого офіційного візиту не використовувались, оскільки було висунуто гіпотезу, що стрес пацієнта та психотерапевта буде більш високим під час цього першого візиту.

Аудіозаписи з 22 пацієнтів із високим ризиком самогубства були отримані з базі даних мовлення предметів, що використовувалися Меріліном Сілверманом у своєму початковому дослідженні. Пацієнт із високим ризиком визначався як той, хто покінчив життя самогубством, спробував вчинити самогубство і зазнав невдачі, або пацієнт з фіксацією думки та методів самогубства. Зразки мов, що використовувались для представлення цього класу, були вилучені із терапії, телефонних переговорів між пацієнтами та психіатрами, а також записів про самогубство. Всі обрані записи були зроблені протягом декількох днів або декількох тижнів після самогубства пацієнта чи спроби самогубства.

Таблиця 1.2 – DSM-IV діагнози і ICD-9-CM коди чоловіків з великою депресією

ICD-9-CM код	Діагноз DSM-IV	Кількість пацієнтів
296.32	Велика депресія, рецидив (помірний)	20
296.33	Велика депресія, рецидив (важка, без психотичних особливостей)	1

Таблиця 1.3 – DSM-IV діагнози і ICD-9-CM коди чоловіків з великим ризиком самогубства

ICD-9-CM код	Діагноз DSM-IV	Кількість пацієнтів
296.33	Велика депресія, рецидив (важка, без психотичних особливостей)	17
296.34	Велика депресія, рецидив (важка, з психотичними симптомами)	2

У поточному дослідженні були включені лише пацієнти, які відповідали критеріям DSM-IV для основного депресивного епізоду чи основного депресивного розладу (Таблиці 1.2 та 1.3). Три пацієнти в дослідженні зробили самогубство перед діагнозом DSM-IV. І, нарешті, не було жодного пацієнта з проблемами слуху, мови, мовних розладів або хронічних респіраторних захворювань.

1.4 Запис та попередня обробка

Записи контрольних пацієнтів, депресивних пацієнтів та пацієнтів схильних до самогубства були отримані з попередніх досліджень, не орієнтованих на мовну акустику. Як наслідок, технічні характеристики обладнання для запису на магнітну плівку, використовувані для отримання зразків мови, невідомі для всіх досліджуваних суб’єктів. Неописаний комерційний магнітофон використовувався для запису пацієнтів та терапевтів першого досліду. Відомо, що для опису всіх учасників дослідження використовувались однакові процедури інтерв'ю та фізичне середовище. У другому дослідженні було значно більше варіацій у реєстраційному обладнанні та навколишньому середовищі, з яких були вилучені зразки суїцидальної мови з високим ступенем ризику. Даний набір високого ризику складався з репрезентативних зразків з 17 офісних візитів, використовуючи комерційний магнітофон Sony TC110 (з внутрішнім мікрофоном), три записи про самогубство та дві телефонні розмови. Реєстраційне обладнання, яке використовується для отримання телефонних розмов та повідомлень про самогубство, невідоме. Процедури співбесіди та фізичне середовище для реєстрації суб'єктів контролю та основних депресивних пацієнтів чоловіків були ідентичні тим, що застосовувались для жінок у першому досліді. Щоб компенсувати можливі відмінності в рівні запису між суб'єктами, всі мовні сигнали виділяються(тобто віднімають значення сигналу від сигналу і ділять різницю за стандартним відхиленням) і нормалізуються перед аналізом. Слід зазначити, що всі записи, використані в цьому дослідженні, вважалися високоякісними, і не було жодних ознак деградації обладнання. Єдині джерела шуму, включаючи фоновий шум, були зовнішніми для записуючого обладнання, і ефекти цих джерел були вилучені за допомогою редактора MicroSound (Micro Technology Unlimited, Raliegh, NC) після оцифровки голосових сигналів.

Приблизно 2хв та 30с невідредагованої мови були випадково витягнуті з сеансу терапії або після сеансу для представлення кожного пацієнта і терапевта. Усі магнітофони оцифровувались за допомогою 16-розрядного аналого-цифрового перетворювача. Частота дискретизації становила 10кГц з фільтром згладжування (тобто низькочастотний 5кГц), що точно відповідає частоті дискретизації. Форми оцифрованого мовлення були імпортовані в редактор MicroSound, де сторонні фонові шуми та голоси, крім голосу суб'єкта, були вилучені. Беззвучні паузи, що перевищують 0,5 с, також були видалені, щоб отримати запис безперервної мови. Розділ редагованих мовних сигналів на приблизно 20 сегментів завершено перед обробкою. Точки сегментації були обрані при нульових переходах або на початку пауз у мові. Ця процедура була прийнята для мінімізації введення підроблених частотних ефектів, що виникають внаслідок різких переходів у редагованому мовному сигналі.

1.5 Акустичні міри голосу та властивості артикуляції

1) Фундаментальний аналіз частот: видобуток тону та статистичний аналіз проводились за допомогою програмного забезпечення N!Power, розробленого компанією Signal Technology, Inc (Санта-Барбара, Каліфорнія). Алгоритм видобутку тону N!Power є розширенням алгоритму цептрального аналізу Юанга та Маркеля, розробленим Ноллом, в якому використовується процедура стеження за цептральним піком, щоб визначити, чи звучить канал голосу або неголосований. Наступні шість статистичних даних були розраховані для кожного 20-ти секундного сегмента мовлення: діапазону, дисперсії, середньої, скрегітності, куртозності та коефіцієнта варіації.

2) Аналіз амплітудної модуляції: алгоритм усереднення середніх квадратів (rms) був написаний у MATLAB (версія 4.2.c.1, MathWorks, Inc.,

Natick, MA) для збору статистичних даних, що описують характеристики АМ. Цей алгоритм був розроблений для обчислення та зберігання тих же шести статистичних даних, які використовуються в дослідженні.

3) Аналіз Форманту: у MATLAB була розроблена модель авторегресії (AP) двенадцатого порядку з використанням лінійного інтелектуального кодування (LPC) для розрахунку перших трьох частотних частот форманту та пропускної здатності (позначених і FBW, відповідно) записаної мови кожного члена класу. Алгоритм розподілив кожну представницьку вибірку учасника на декілька кадрів на 15 мс, а потім розраховував коефіцієнти LPC кожного мовного кадра. Частоти та смуги пропускання кожного формату мовлення кожного мовного кадра визначались шляхом прийняття коріння поліноми предиктора або всеполюсної моделі голосового тракту, отриманого з коефіцієнтів ЛПК. Нарешті, форманти частоти та пропускна здатність, розраховані з усіх кадрів, були усереднені часом для одержання одного вектора форманта для кожного члена класу. Форванти співвідношення були розраховані таким чином, щоб можна було дослідити зв'язок між частотами форманта.

Метод довгострокового середнього спектру (LTAS) не використовувався для обчислення частот форванту та пропускної здатності, оскільки обсяг мовлення, проаналізований в цих дослідженнях, зробив його обчислення дорогим та непрактичним. Слід зазначити, що підхід LTAS забезпечує більш точне відображення властивостей форманта, ніж підхід LPC. Підхід LPC був використаний для оцінки частоти та пропускної здатності форманта на основі його перевіреної корисності в області голосової науки та його обчислювальної ефективності.

4) Аналіз потужності спектральної щільності: для розрахунку розподілу потужності мови в діапазоні частот від 0 Гц до 2000 Гц був застосований класичний метод оцінки PSD. Для оцінки спектра та

значення пікової потужності у вказаному частотному діапазоні був використаний алгоритм, написаний у MATLAB, який реалізував метод Уелша з неперекриваючими вікнами Хеммінга. Спектри були обчислені за допомогою 40-мс вільної мови, частоти дискретизації 10 кГц для оцифрування сигналів, швидких перетворень Фур'є 1024-точкових (FFT) та 100-точкових вікон Хеммінга.

1.6 Порівняльний статистичний аналіз особливостей класу

Алгоритми кожного з чотирьох алгоритмів акустичних (тобто. F0, AM, Formant, PSD) коли вони виконуються на зразках мовлення, що представляють кожного пацієнта та терапевта, генерують вихідну матрицю функцій (тобто статистика). Кожна вихідна матриця містила N рядків та M стовпців (матриця), де N була кількість 20-ти сегментів, отриманих з кожного репрезентативного запису мовлення, і M – число функцій, які використовуються для характеристики аналізованого акустичного параметра. Тому існував набір із чотирьох матриць параметрів, пов'язаних з кожним пацієнтом і терапевтом – F_0 матрицею, матрицею AM, матрицею Форманту та матрицею PSD. Далі, середня матриця кожного параметра була розрахована для отримання середнього вектора характеристик. Математично, контрольні, дистимічні, великі депресивні та суїцидальні класи були визначені таким чином, що мають чотири параметра вектора довжини M на клас-елемент-один вектор для кожного з чотирьох аналізованих вокальних параметрів. Вектори містять M середніх функцій, розрахованих з 20-х сегментів, створених з аудіозапису члена класу. Вектори параметрів, що представляють кожен клас, були імпортовані в систему аналізу онлайнового шаблону (PcOLPARS, PAR Government Systems, La Jolla, CA) та статистичний пакет SYSTAT (SPSS Inc., Chicago, IL) для аналізу функцій та дискримінаційного аналізу.

Проекційні аналізи та квантові кількісні ділянки були використані для перевірки припущення, що три класи даних були звичайно розподілені. У PcOLPARS були використані алгоритми проекції координат і власне значення, щоб переконатися, що кожен набір даних виявляє еліптичний унімодальний розподіл, характерний для багатовимірних нормальних розподілів. Квантові кількісні ділянки були виконані в SYSTAT для перевірки граничної норми кожної одноразової функції.

Для жіночих та чоловічих досліджень, попарно (тобто контрольно-дистимічні, контрольні-депресивні, великі депресивні суїцидальні і т. д.) статистичні аналізи проводились окремо по кожному з чотирьох наборів векторів вокальних параметрів, щоб визначити, які функції забезпечують найкраща дискримінація між кожною парою класів. Особливості, визначені як найкращі парні дискримінатори з кожного параметра голосів, були об'єднані для розробки "оптимального" класифікатора. Ця сама процедура була повторена для випадку, коли всі класи порівнювалися одночасно. Класифікаційні оцінки та показники чутливості, специфічності, позитивної прогностичної цінності (PPV) та негативної прогностичної цінності (NPV) були розраховані для можливості оцінки ефективності класифікатора.

Порівняльний статистичний аналіз включав розрахунок та порівняння матриць коваріації класів, порівняння характеристик класових голосових параметрів за допомогою аналізу дисперсії (ANOVA) з корекцією Бонферроні та одностороннього багатомірного аналізу дисперсії (MANOVA) та застосування лінійних або квадратичних дискримінаторів використовуючи метод "утримання". 95% довірчий інтервал був використаний у всіх статистичних аналізах. Матриці групової коваріації були розраховані в SYSTAT і порівняні, використовуючи криву для кваліфікації для рівності. Лінійні дискримінантні аналізи застосовувалися

до класових пар з однаковими матрицями коваріації. В тих випадках, коли матриці групової коваріації були значно відмінні, використовували квадратичні дискримінанти. Весь дискримінантний аналіз проводився в SYSTAT. У обох випадках дискримінаційного аналізу для компенсації розмірів невеликих (N<30) класів використовувався метод дискретимантного аналізу «утримання» або «Джекнайф». Лінійний дискримінантний аналіз з використанням автоматичного кроку назад використовувався в поєднанні з методом дискримінантного рангу аналізу функцій для визначення оптимального підмножини ознак для дискримінації класів з однаковими матрицями коваріації.

Метод дискримінантного рангу аналізу функцій використовувався в pcOLPARS для визначення найкращих класових дискримінаторів за допомогою прогностичної дійсності. Функції, вибрані в якості найкращих класових дискримінаторів за допомогою цього методу, порівнювалися з тими, що вибираються методом зворотного кроку лінійного дискримінантного аналізу, щоб визначити, яка підмножина функцій з кожного з чотирьох вокальних параметрів повинна використовуватися для інтегрованого класифікатора.

1.7 Результати

1. Study #1—Vocal Characteristics of Depression in Unmedicated Women

Результати аналізу попарного та все класу дискримінантних аналізів, виконаних на жіночих популяціях дослідження, наведені у таблицях 1.4 та 1.5. У таблиці 1.4 наведені аналізовані класи, найкращі ознаки розпізнавання та сукупна оцінка класифікації, отримані для кожного аналізу з використанням зазначених розрізнювальних ознак . У таблиці 1.5

викладені експлуатаційні характеристики кожної дискримінантної функції з значеннями чутливості, специфічності, PPV та NPV.

Акустичні особливості, що характеризують динаміку та AM, в цілому були неефективними дискримінаторами. AM була єдиною функцією з цих двох наборів функцій для демонстрації будь-якої дискримінуючої сили, про що свідчать результати контрольно-дистимного попарного аналізу. Дистимічні пацієнти були погано (66%) диференційовані від основних депресивних пацієнтів на основі мовної акустики. Однак дистимічна і велика депресивна мова ефективно диференціювалася з контрольної мови переважно на основі функцій форманта та PSD. Дистимічна мова характеризувалася значним звуженням FBW у порівнянні з контрольною мови. У виступі сильно депресивних пацієнтів виявлені підвищені F_1 та PSD, а також зниження PSD_1, FBW_2 та FBW_3. Незважаючи на те, що вимірюване збільшення значень F_1 було значним між контролем та основними депресивними класами, воно не виявлялося як корисна дискримінаційна ознака. Результати парного дослідження депресивного контрольного мазка вказують на структуру спектрального сплеску з великою депресією. Відсоток загальної потужності в піддіапазоні від 0 до 500 Гц зменшувався(PSD_1), тоді як частка потужності у вищих піддіапазонах збільшувалася. Загалом, показники ефективності, наведені в таблицях IV та V, показують, що функції форманта та PSD забезпечують чудову дискримінацію між контролем та основними депресивними групами. 95% довірчі інтервали для жіночого форманта та статистики PSD представлені у таблицях 1.6 і 1.7, відповідно.

Одночасний статистичний аналіз контрольних, дистимічних та основних депресивних груп виявив суттєві відмінності тільки в особливостях, отриманих від дослідження форманта. , FBW, FBW та FBW

були визначені як такі, що істотно відрізняються серед цих класів. Аналіз функцій з використанням передбачуваної дійсності та методу дискримінантного рангу, який визначив функції FBW та FBW, що володіють найбільшою дискримінаційною силою. Середня шкала коефіцієнта класифікації 58% була отримана з використанням методу "утримання" з квадратичним класифікатором. Класифікатор був майже таким же ефективним у класифікації суб'єктів контролю (70%) як сильно депресивних пацієнтів (71%). Проте класифікатор погано працював при призначенні дистимічних пацієнтів до свого класу (35%). Фактично, неправильно класифіковані дистимічні пацієнти були майже рівномірно розподілені між контрольним та основними депресивними класами.

Таблиця 1.4 – Сумарний аналіз класифікації жіночих пацієнтів

Проаналізовані класи	Найкращі розпізнавальні особливості	Точність, %
Контрольний/Дистимічний	RMS AM, FBW_2	78
Контрольний/Великий депресивний	FBW_2, FBW_3, PSD_1, PSD_2	94
Дистимічний/Великий депресивний	FBW_3	66
Контрольний/Дистимічний/Великий депресивний	FBW_2, FBW_3	58

Таблиця 1.5 – Чуттєвість, специфічність, позитивне передбачуване значення(PPV) та негативне передбачуване значення(NPV) для жіночих пацієнтів

Проаналізовані класи	Чуттєвість	Специфічність	PPV	NPV

Контрольний/Дистимічний	0.66	0.85	0.80	0.75
Контрольний/Великий депресивний	0.90	0.95	0.90	0.95
Дистимічний/Великий депресивний	0.63	0.68	0.59	0.71
Контрольний/Дистимічний/ Великий депресивний	0.50	0.62	0.70	0.55

Таблиця 1.6 – Статистика формант для жінок

	Контрольний	Дистимічний	Великий депресивний
F_1	(298.2, 373.8)	(323.3, 421.7)	(389.5, 454.5)
F_2	(1056.6, 1231.4)	(1082.6, 1203.4)	(1161.7, 1252.3)
F_3	(1984.9, 2089.1)	(1987.2, 2052.8)	(2005.6, 2048.4)
FBW_1	(201.23, 286.8)	(226.7, 293.3)	(266.2, 315.8)
FBW_2	(605.6, 646.5)	(532.6, 593.4)	(546.8, 591.2)
FBW_3	(619.3, 688.7)	(583.6, 646.4)	(560.0, 596.0)

Таблиця 1.7 – Статистика PSD для жінок

	Контрольний	Дистимічний	Великий депресивний
Пікова сила	(18.7, 27.3)	(21.1, 24.9)	(17.9, 22.1)
Пікове місце, Гц	(220.9, 351.1)	(270.5, 403.6)	(318.0, 466.0)

Сила в PSD_1, %	(0.75, 0.89)	(0.67, 0.83)	(0.62, 0.78)
Сила в PSD_2, %	(0.09, 0.15)	(0.13, 0.21)	(0.14, 0.20)
Сила в PSD_3, %	(0.005, 0.05)	(0.03, 0.05)	(0.05, 0.09)
Сила в PSD_4, %	(0.005, 0.05)	(0.02, 0.06)	(0.03, 0.09)

Зразки голосу, зібрані з сильно депресивних пацієнтів та із високим ризиком самогубства, характеризувалися зменшенням діапазону смуги пропускання другого форматування (FBW_2) та підвищеними частотами форманта (F_1, F_2, F_3) та смугою пропускання третьої форми (FBW_3). Унікальні класові акустичні властивості також виникли для кожної з цих груп. AM діапазон та AM відхилення були значно підвищені тільки у сильно депресивній мові. Суїцидальна мова продемонструвала значний зсув у силі від нижчих до високих частот. 95% довірчі інтервали для статистики AM, форманта та PSD для чоловіків представлені у таблицях 1.8 і 1.9, відповідно.

Сильно депресивна та суїцидальна мова були ефективно розпізнані та відділені від контрольної мови на основі функцій форманта та PSD. Класифікатори, розроблені з цих дискримінаційних ознак (як зазначено у другому рядку таблиці 1.10), були більш ефективними в класифікації контрольних суб'єктів (88%), ніж сильно депресивних (76%) або пацієнтів з високим ризиком самогубства (73%), відповідно.

Визначено діапазон середньоквадратичного діапазону, коефіцієнт варіації AM та PSD_3 як оптимальні дискримінатори сильно депресивної та суїцидальної мови. Квадратичний класифікатор, розроблений за цими

ознаками, був значно ефективніший для класифікації сильно депресивної мови (86%), ніж суїцидальних зразків (77%).

АМ відхилення, F_1, F_3, і PSD_2 з'явилися в якості найкращих дискримінаційних особливостей для три класу проблеми. Інтегрована квадратична дискримінантна функція з використанням цих функцій характеризувалася здатністю класифікувати зразки з усіх класів з майже рівною ефективністю (контроль 75%, основний депресивний 71%, самогубний ризик 75%). Таблиця 1.11 підсумовує результати попарного і загального класу дискримінантних аналізів. Таблиця 1.12 підсумовує характеристики продуктивності кожної дискримінантної функції з десятками чутливості, специфічності, PPV та NPV. Як показано в таблицях 1.11 та 1.12, функції, вибрані для попарних досліджень, забезпечували дуже подібну та послідовну класифікаційну ефективність між групами. Помірне зниження продуктивності класифікатора спостерігалось лише в комплексному або загальному аналізі.

Таблиця 1.8 –

	Контрольний	Дистимічний	Великий депресивний
F_1	(298.2, 373.8)	(323.3, 421.7)	(389.5, 454.5)
F_2	(1056.6, 1231.4)	(1082.6, 1203.4)	(1161.7, 1252.3)
F_3	(1984.9, 2089.1)	(1987.2, 2052.8)	(2005.6, 2048.4)
FBW_1	(201.23, 286.8)	(226.7, 293.3)	(266.2, 315.8)
FBW_2	(605.6, 646.5)	(532.6, 593.4)	(546.8, 591.2)
FBW_3	(619.3, 688.7)	(583.6, 646.4)	(560.0, 596.0)

Таблиця 1.9 –

Проаналізовані класи	Найкращі	Точні

	розпізнавальні особливості	сть, %
Контрольний/Дистимічний	RMS AM, FBW_2	78
Контрольний/Великий депресивний	FBW_2, FBW_3, PSD_1, PSD_2	94
Дистимічний/Великий депресивний	FBW_3	66
Контрольний/Дистимічний/Великий депресивний	FBW_2, FBW_3	58

Таблиця 1.10 –

	Контрольний	Дистимічний	Великий депресивний
Пікова сила	(18.7, 27.3)	(21.1, 24.9)	(17.9, 22.1)
Пікове місце, Гц	(220.9, 351.1)	(270.5, 403.6)	(318.0, 466.0)
Сила в PSD_1, %	(0.75, 0.89)	(0.67, 0.83)	(0.62, 0.78)
Сила в PSD_2, %	(0.09, 0.15)	(0.13, 0.21)	(0.14, 0.20)
Сила в PSD_3, %	(0.005, 0.05)	(0.03, 0.05)	(0.05, 0.09)
Сила в PSD_4, %	(0.005, 0.05)	(0.02, 0.06)	(0.03, 0.09)

Таблиця 1.11 –

Проаналізовані класи	Найкращі розпізнавальні особливості	Точність, %
Контрольний/Дистимічний	RMS AM, FBW_2	78

Контрольний/Великий депресивний	FBW_2, FBW_3, PSD_1, PSD_2	94
Дистимічний/Великий депресивний	FBW_3	66
Контрольний/Дистимічний/Великий депресивний	FBW_2, FBW_3	58

Таблиця 1.12 –

Проаналізовані класи	Чуттєвість	Специфічність	PPV	NPV
Контрольний/Дистимічний	0.66	0.85	0.80	0.75
Контрольний/Великий депресивний	0.90	0.95	0.90	0.95
Дистимічний/Великий депресивний	0.63	0.68	0.59	0.71
Контрольний/Дистимічний/Великий депресивний	0.50	0.62	0.70	0.55

2. АНАЛІЗ МЕТОДІВ РОЗПІЗНАВАННЯ ЕМОЦІЇ ЗА ГОЛОСОМ

2.1 Ідентифікація голосу

Проблема ідентифікації емоції людини за її голосом дуже схожа на ідентифікацію самої людини за її голосом, оскільки загальна схема роботи системи розпізнавання буде виглядати однаково:

1) Спектрограма голосу людини
2) Виявлення ознак
3) Класифікація
4) Порівняння з порогом

Основні методи розпізнавання:

- Розпізнавання на основі ознак
- Нейромережевий підхід до розпізнавання

Визнання на підставі ознак свідчить про те, що кожен об'єкт може характеризуватися вектором ознак і набором значень атрибутів. Цей метод дуже складний, оскільки важко правильно описати, яка емоція і як її перетворити на чисельне значення. Цей метод також класифікується як текстово залежний метод розпізнавання голосу.

Нейромережевий підхід. Штучна нейронна мережа(ШНМ) – сукупність нейронів, які виконують операції на даних, які мають властивості класифікації. Така система використовує поняття машинного навчання, тобто система навчається виконувати певні задачі, такі як розпізнавання певних образів на зображеннях. Машинне навчання у штучних нейронних мережах можна розглядати як процес налаштування архітектури мережі та ваг зв’язків з наявних вихідних даних, і з кожною ітерацією функціонування такої мережі поліпшується.

2.2 Машинне навчання

Загалом існує три види машинного навчання: з вчителем, без вчителя, змішаний.

Машинне навчання ШНМ з вчителем означає, що для кожного вектору на вході з множини навчальних даних існує значення вихідного(цільового) вектору. Мережеві ваги встановлюються, якщо для кожного вхідного вектора вони не отримують прийнятного рівня відхилення вхідного вектора від цілі.

Навчання нейронної мережі без вчителя складається з вхідних векторів і не вимагає знань правильних відповідей на вхідних векторах. Таким чином, нейронна мережа знаходить подібні зображення між зразками в системі даних і розподіляє його на категорії.

Змішане навчання нейронної мережі полягає в тому, що частина ваг налаштовується за допомогою навчання з учителем, а інша за допомогою самонавчання.

Для визначення якості роботи нейронної мережі використовують функцію втрат. За таку функцію зазвичай обирають:

- Евклідова відстань
- Середньоквадратична похибка
- Функція кросентропії

Мережа вважається навченою, якщо функція втрат набуває мінімального значення. Основною ідеєю цього методу є поширення сигналів помилок від виходів мережі до його входів у напрямку зворотного прямого поширення сигналів в нормальному режимі роботи нейронної мережі. Процедура повернення помилки може бути застосована кілька разів, щоб поширювати градієнти на всіх рівнях, від виходу (результат прямого проходження нейронної мережі) та вхідних даних до мережі. У процесі навчання нейронної мережі вага зв'язків між нейронами регулюється на основі методу градієнтного спуску. На практиці

модифікація цього методу зазвичай використовується, коли процедура зниження градієнта застосовується для вивчення прикладів. Цей підхід називається методом стохастичного градієнта, що значно прискорює час навчання нейронної мережі.

Інші методи оптимізації: метод найменших квадратів (алгоритм Левенберга-Маркварта та алгоритм Ньютона-Гаусса), квазіньютонівські методи, спосіб градієнта сполучення та ін. У 1986 році Ромельхарт запропонував метод моментів, який запам'ятовує зміну ваги для кожної ітерації і приймає враховуючи це при подальшій регуляції ваг нейронної мережі. На відміну від методу стохастичного градієнта, підхід намагається зберегти однаковий напрямок руху під час регулювання вагових факторів, запобігаючи коливання. Найсучаснішими методами оптимізації є AdaGrad (адаптивний градієнтний алгоритм), RMSProp (розповсюдження кореневого середнього квадрата), календарний каскад стохастичного градієнтного спуску (KSGD) та Адам (адаптивний момент оцінювання). Ці методи - модифікації методу стохастичного градієнта та прогнозування змін швидкості навчання під час навчання нейронної мережі

Вхідні дані для такої нейронної мережі є спектрографами людини. Важливо вибрати модель для навчання нейронних мереж. Для роботи з голосом людини краще вибрати модель глибокого машинного навчання, тоді машина сама виділятиме ознаки голосу, оскільки завдання класифікації емоцій для голосу є складним завданням для людини. Однак цей підхід також є недоліком, оскільки тоді ми не можемо знати, за яким принципом автомобіль розпізнає ознаки голосу. Для машинного навчання важливою є використання великих зразків навчального матеріалу, оскільки неадекватні рівні навчання відбуваються тоді, коли замість підведення підсумків отриманої інформації модель просто запам'ятовує це.

У практичних завданнях навчання з вчителем найчастіше використовуються нейронні мережі прямого розподілу, такі як багатошаровий персептрон (рис 2.1). Переходячи з одного шару на інший, приховані нейрони обчислюють зважену суму входів до них з попереднього шару і застосовують нелінійну функцію - функцію активації до результату. Важливим критерієм для функції активації є його диференціація. Найбільш відомі функції активації - гіперболічна тангенс і сигмоїд

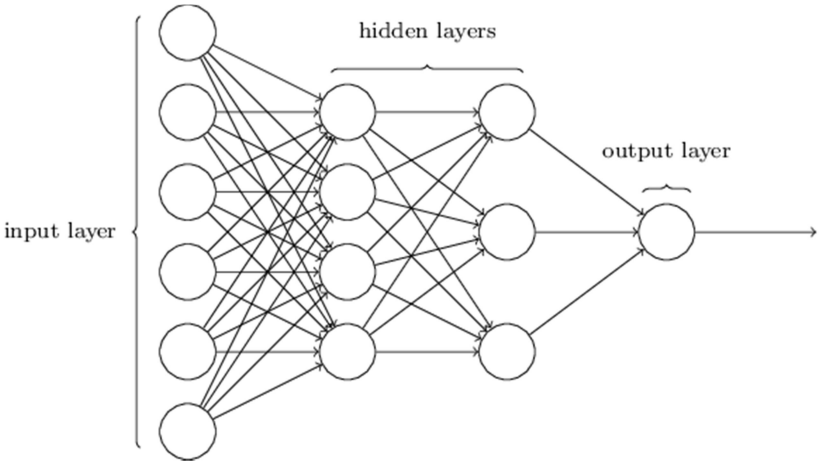

Рисунок 2.1 – Архітектура багатошарового перцептрона

Згідно з теоремою про загальну наближеність, нейронна мережа з одним прихованим шаром може наближати будь-яку безперервну функцію багатьох змінних з будь-якою точністю. Тому для прискорення вивчення мережі більшість дослідників використовують архітектуру з одним прихованим шаром.

2.3 Глибинне навчання.

Глибинне навчання є полем машинного навчання на основі набору алгоритмів, які намагаються моделювати абстракції високого рівня в даних, з використанням глибиною граф з безліччю шарів, побудованих з декількох лінійних або нелінійних перетворень.

Глибинне навчання є частиною більш широкої сім'ї методів навчання, заснованих на вивченні даних. Спостереження (наприклад, зображення) може бути представлено багатьма способами, такими як вектор значень яскравості для пікселів або абстрактним шляхом, як множина країв, областей певної форми тощо. Деякі подання краще, ніж інші, у спрощенні навчального завдання (наприклад, розпізнавання обличчя або вираз обличчя). Однією з обіцянок глибокого навчання є заміна ознак ручної роботи за допомогою ефективних алгоритмів автоматичного або напівавтоматичного навчання особливостей та ієрархічного розподілу знаків.

Алгоритми навчання глибини базуються на розподілених представленнях. Припущення, що лежить в основі розподілених уявлень, полягає в тому, що спостережувані дані генеруються взаємодією факторів, організованих на рівні. Поглиблене навчання додає припущення, що ці рівні факторів відповідають різним рівням абстракції або побудови. Змінні кількості та шари можуть бути використані для забезпечення різних ступенів абстракції.

Глибинне навчання використовує цю ідею ієрархічних пояснювальних факторів, де поняття нижчого рівня вивчають абстрактні поняття найвищого рівня. Ці архітектури часто будуються за допомогою шару жадібного методу. Глибинне навчання дозволяє розгадати ці абстракції та захоплювати функції, корисні для навчання.

Для завдань керованого вивчення методів глибокого вивчення уникати особливостей проектування, перетворення даних у компактні проміжні уявлення в аналогічні основні компоненти та виведення шаруватих структур, що усувають надмірність у презентації.

2.4 Моделювання нейронів

У спробі імітувати певні здібності мозку Уоррен МакКаллох та Уолтер Піттс встановили у 1943 році спрощену модель біологічного нейрона, що називається моделлю МакКаллох-Піттса, яка складається з декількох входів і одного виходу з центральним процесором (ЦП).

На рис. 1.2 відображено модель нейрона, яка описується:

$$y = f(\sum_{i=1}^{N} w_i x_i - v_t)$$

де xi = вхідні сигнали, i = 1, 2, 3, ..., N;

wi = синаптичні ваги;

vt = поріг або зсув;

f (*) = функція активації або функція стиснення або елемент який обробляє;

y = вихідний сигнал нейрона.

Використання порогу vt означає, що функція активації f (*) зміщена. МакКаллох та Піттс не давали ніякого способу, в якому вузол або нейрон могли самостійно відрегулювати або адаптувати свої синаптичні шкали в навчальному процесі. У 1949 році Хебб запропонував просту математичну формулу, здатну адаптивно змінювати вага нейронів пропорційно активності між перед і постсинаптичними нейронами:

$$\Delta w_i(n) = \mu y(n) x_i(n)$$

де μ- позитивна постійна швидкість навчання за весь час n.

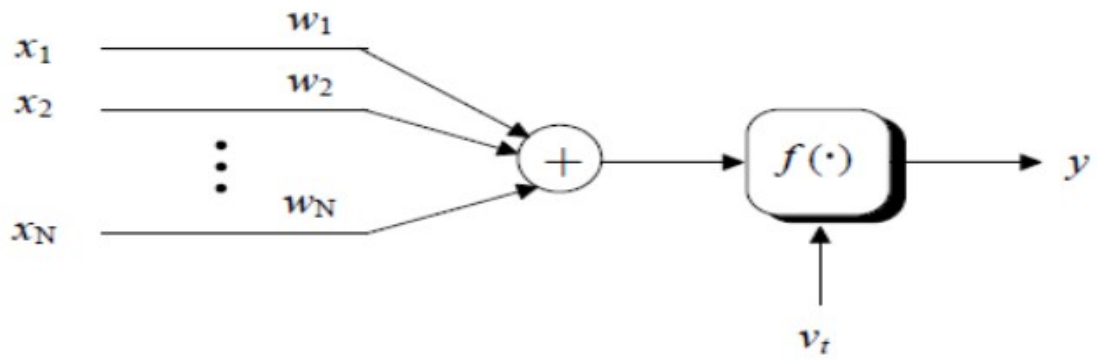

Рисунок 2.2 – Модель нейронів Маккаллох – Піттс.

2.5 Персептрон

У 1958 році Розенблат демонстрував деякі практичні застосування, використовуючи персептрон. Персептрон - це однорівневе з'єднання нейронів Маккаллоу-Пітта, які називаються одношаровими мережами прямого доступу. Мережа здатна лінійно розділяти вхідні вектори на структуру класу гіперструктур. Лінійна асоціативна пам'ять є прикладом одношарової нейронної мережі. У цій програмі мережа з'єднує (вектор) з вхідним шаблоном (вектор), і інформація зберігається в мережі за допомогою модифікацій синаптичних масштабів мережі.

Рис. 2.3 ілюструє персептрон, який описується:

$$y_i = f\left(\sum_{j=1}^{N} w_{ij}x_j - v_t\right)$$

де i = 1, 2, ..., M (вихідні вузли), j = 1, 2, ..., N (входи).

Розенблатт розробив правила навчання, що базуються на шкалах, коректуються пропорційно помилкам між вихідними нейронами та бажаними результатами (ціль). Вага припадає:

$$\Delta w_{ij}(n) = \mu[d_i(n) - y_i(n)]x_j(n)$$

де i = 1, 2, ..., M (виходи), j = 1, 2, ..., N (входи), a di - бажаний вихід вузла i у часі n.

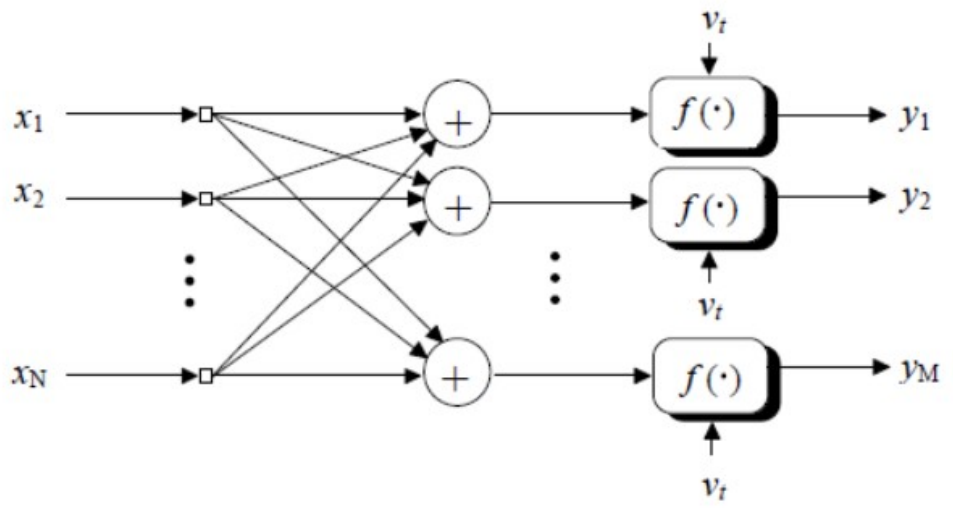

Рисунок 2.3 – Персептрон з одним шаром (одношарова мережа фідів)

2.6 Багатошарові персептрони

Для того, щоб отримати статистику вищого замовлення, таку як реалізація простої XOR або логічної XNOR функції (без попереднього процесора, який часто використовується в одношаровому персептроні), Мінський і Паперт були запропоновані в 1969 році нелінійним багатошаровим персептрони або багатошарові мережі з прямим з'єднанням, які математично показали, що існують принципові обмеження на те, що можна розрахувати для одного персептрона. Багатошарові персептрони (БШП) являють собою один або декілька прихованих шарів, обчислювальні вузли яких відповідно називають прихованими нейронами. Функція прихованих нейронів полягає у втручанні між зовнішнім входом і виходом мережі. Рис 2.4. Представляє тришаровий багатошаровий персептрон з одним прихованим шаром і виходом. Вихідні вузли в вхідному шарі мережі складаються з N елементів шаблону, які складаються з вхідних сигналів, застосованих до K-нейронів у другому шарі або першого прихованого шару (l = 1). Вихідні сигнали M нейрони - це кінцевий шар (l = L) мережі, що є загальною мережною відповіддю на шаблон, що надається вихідними вузлами. Просто правильний вибір

кількості прихованих вузлів можна спочатку розрахувати для кращого узагальнення.

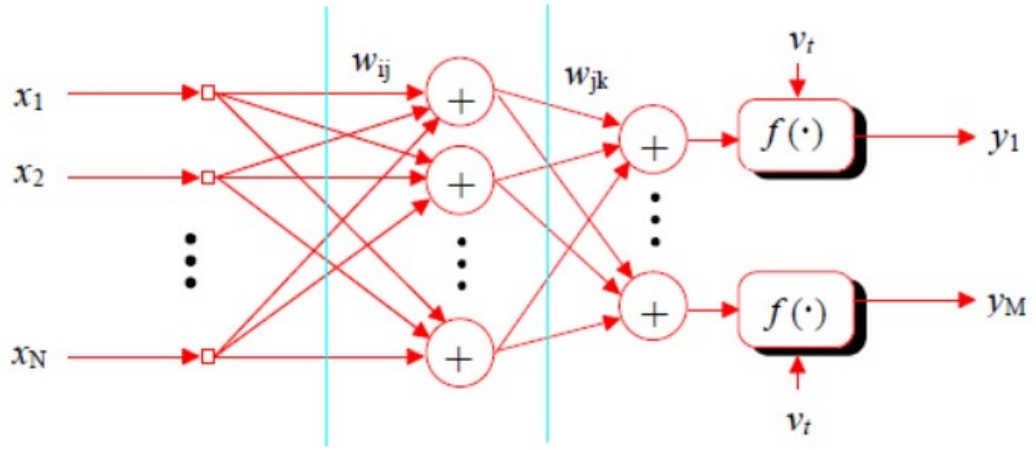

Рисунок 2.4 – Тришарові багатошарові персептрони з одним прихованим шаром і виходом

1) Вхідний шар джерела N;

2) Прихований шар прихованих вузлів K;

3) Вихідний рівень.

Всі три архітектури нейронної мережі, описані до цього часу, використовують функцію активації f (), яка визначається як вихід нейрона з точки зору рівня активності на його вході (від -1 до 1 або від 0 до 1). У табл. 1.1 наведено основні типи функцій активації.

Найбільш практичними функціями активації є сигмоїд і гіперболічний тангенс функції. Це тому, що вони мають похідні.

Таблиця 1.1 – Загальні функції активації

Назва	Визначення
Лінійна	$f(x) = kx$
Крокова(зазвичай: b = 1, d =0, x_k = 0)	$f(x) = \begin{cases} \beta, \text{якщо } x \geq x_k \\ \delta, \text{якщо } x < x_k \end{cases}$
Лінійної зміни	$f(x) = \begin{cases} \rho, \text{якщо } x \geq \rho \\ x, \text{якщо } \lvert x \rvert < \rho \\ -\rho, \text{якщо } x \leq -\rho \end{cases}$

Сигмоїд	$f(x) = \frac{1}{1+e^{-\alpha x}}, \alpha > 0$
Гіперболічний тангенс	$f(x) = \tanh(\gamma x) = \frac{1-e^{-2x}}{1+e^{-2x}}, \gamma > 0$
Гауса	$f(x) = \frac{1}{\sqrt{2\pi\sigma}} exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$

2.7 Згорткова нейронна мережа

Згорткова нейронна мережа - це клас глибоких штучних нейронних мереж прямого розподілу, який найчастіше використовувався для аналізу візуальних образів. NNM використовують різноманітні багатошарові перчітрони, призначені для використання мінімальної кількості попередньої обробки.

NNM складається з вхідних і вихідних шарів, а також декількох прихованих шарів. Приховані шари NNM зазвичай складаються з рухомих шарів, шарів агрегації, повністю об'єднаних шарів і шарів нормалізації. Цей процес описується в нейронних мережах як згортка за домовленістю. З математичної точки зору, це скоріше взаємна кореляція, ніж згортка. Це стосується лише показників в матриці, а отже, і ваги, на якій знаходиться індекс.

2.8Порівняння шарів

2.8.1 Згорткові шари

Згорткові шари застосовують до входу операцію згортки, передаючи результат до наступного шару. Згортка імітує реакцію окремого нейрону на зоровий стимул. Кожен згортковий нейрон обробляє дані лише для свого рецептивного поля.

Хоч повноз'єднані нейронні мережі прямого поширення й можливо застосовувати як для навчання ознак, так і для класифікування даних,

застосування цієї архітектури до зображень є непрактичним. Було би необхідним дуже велике число нейронів, навіть у поверхневій (протилежній до глибинної) архітектурі, через дуже великі розміри входу, пов'язані з зображеннями, де кожен піксель є відповідною змінною. Наприклад, повноз'єднаний шар для (маленького) зображення розміром 100 × 100 має 10 000 ваг. Операція згортки дає змогу розв'язати цю проблему, оскільки вона зменшує кількість вільних параметрів, дозволяючи мережі бути глибшою за меншої кількості параметрів. Наприклад, незалежно від розміру зображення, області замощування розміру 5 × 5, кожна з одними й тими ж спільними вагами, вимагають лише 25 вільних параметрів. Таким чином, це розв'язує 32 проблему зникання або вибуху градієнтів у тренуванні традиційних багатошарових нейронних мереж з багатьма шарами за допомогою зворотного поширення.

Розмір ємності виходу згорткового шару контролюють три гіперпараметри:

• Глибина ємності виходу контролює кількість нейронів шару, що з'єднуються з однією й тією ж областю вхідної ємності. Ці нейрони вчаться активуватися для різних ознак входу. Наприклад, якщо перший згортковий шар бере як вхід сире зображення, то різні нейрони вздовж виміру глибини можуть активуватися в присутності різних орієнтованих контурів, або плям кольору.

• Крок контролює те, як стовпчики глибини розподіляються за просторовими вимірами (шириною та висотою). Коли кроком є 1, ми рухаємо фільтри на один піксель за раз. Це веде до сильного перекриття рецептивних полів між стовпчиками, а також до великих ємностей виходу. Коли ми робимо крок 2 (або, рідше, 3 чи більше), то фільтри, просуваючись, перестрибують на 2 пікселі за раз. Рецептивні поля

перекриваються менше, й отримувана в результаті ємність виходу має менші просторові розміри.

• Іноді зручно доповнювати вхід нулями по краях вхідної ємності. Розмір цього доповнення є третім гіперпараметром. Доповнення забезпечує контроль над просторовим розміром ємності виходу. Зокрема, іноді бажано точно зберігати просторовий розмір вхідної ємності.

2.8.2 Агрегувальні шари.

Іншим важливим поняттям ШНМ є агрегування (англ. pooling), яке є різновидом нелінійного зниження дискретизації. Існує декілька нелінійних функцій для реалізації агрегування, серед яких найпоширенішою є максимізаційне агрегування (англ. max pooling). Воно розділяє вхідне зображення на набір прямокутників без перекриттів, і для кожної такої 33 підобласті виводить її максимум. Ідея полягає в тому, що точне положення ознаки не так важливе, як її грубе положення відносно інших ознак. Агрегувальний шар слугує поступовому скороченню просторового розміру представлення для зменшення кількості параметрів та об'єму обчислень у мережі, і відтак також для контролю перенавчання. В архітектурі ЗНМ є звичним періодично вставляти агрегувальний шар між послідовними згортковими шарами. Операція агрегування забезпечує ще один різновид інваріантності відносно паралельного перенесення.

Агрегувальний шар діє незалежно на кожен зріз глибини входу, і зменшує його просторовий розмір. Найпоширенішим видом є агрегувальний шар із фільтрами розміру 2×2 (рис. 1.5.), що застосовуються з кроком 2, який знижує дискретизацію кожного зрізу глибини входу в 2 рази як за шириною, так і за висотою, відкидаючи 75 % збуджень. В цьому випадку кожна операція взяття максимуму діє над 4 числами. Розмір за глибиною залишається незмінним.

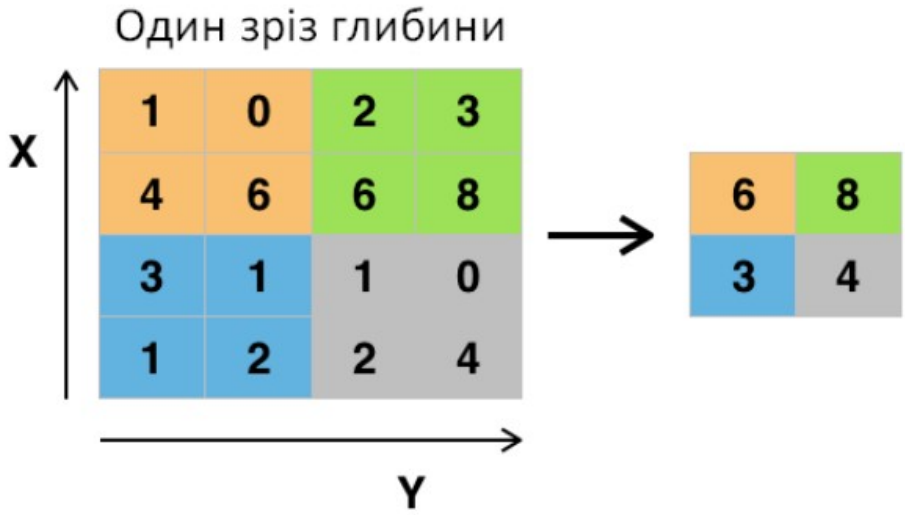

Рисунок 2.5 – Максимізаційне агрегування

2.8.3 Шар зрізаних лінійних вузлів (ReLU)

ReLU є абревіатурою від англ. Rectified Linear Units, зрізаних лінійних вузлів. Цей шар застосовує ненасичувальну передавальну функцію $f(x) = \max(0, x)$, посилює нелінійні властивості функції ухвалення рішення і мережі в цілому, не зачіпаючи рецептивних полів згорткового шару. Для посилення нелінійності застосовуються й інші функції, наприклад, насичувальні гіперболічний тангенс $f(x) = \tanh(x)$, $f(x) = |\tanh(x)|$, та сигмоїдна функція $f(x) = (1 + e^{-x})^{-1}$. ReLU часто віддають перевагу перед іншими функціями, оскільки він тренує нейронну мережу в декілька разів швидше без значної розплати точністю узагальнення.

2.8.4 Повноз'єднаний шар

Насамкінець, після кількох згорткових та максимізаційно агрегувальних шарів, високорівневі міркування в нейронній мережі здійснюються повноз'єднаними шарами (англ. fully connected layers). Нейрони у повноз'єднаному шарі мають з'єднання з усіма збудженнями попереднього шару, як це можна бачити у звичайних нейронних мережах.

Їхні збудження відтак може бути обчислювано матричним множенням, за яким слідує зсув упередженості.

2.8.5 Шар втрат

Шар втрат визначає, як тренування штрафує відхилення між передбаченими та справжніми мітками, і є, як правило, завершальним шаром. Для різних завдань у ньому можуть використовувати різні функції втрат. Нормалізовані експоненційні втрати (англ. softmax) застосовуються для передбачення єдиного класу з K взаємно виключних класів. Сигмоїдні перехресно-ентропійні втрати застосовуються для передбачення K незалежних значень імовірності в проміжку [0,1]. Евклідові втрати застосовуються для регресії до дійснозначних міток.

2.9 Проблема вибору моделі даних

Віртуально всі алгоритми виявлення аномалій створюють модель нормального патерна даних, нормальної поведінки системи, і потім обчислюють «ступінь аномальності» певної точки даних (набору точок) на основі відхилення від цього патерна. Наприклад, модель даних може бути статистичною регресійною моделлю, або моделлю близькості. Ці моделі мають різне тлумачення «нормальної поведінки» даних. Ступінь аномальності досліджуваної (нової) точки даних оцінюють обчисленням подібності точки даних і моделі. В деяких випадках модель може бути задана алгоритмічно. Наприклад, алгоритм виявлення викидів на основі методу найближчих сусідів моделює дані в термінах розподілу відстані поміж nнайближчих сусідів. В цьому випадку, викиди знаходяться на великій відстані від більшості даних

Зрозуміло, що вибір моделі даних є важливим завданням. Некоректно вибрана модель даних може бути причиною незадовільних

результатів роботи алгоритму. Наприклад, модель лінійної регресії може давати результати низької якості, якщо дані, на яких побудована ця модель, кластеризовані випадковим чином.

Правильний вибір моделі даних вимагає добре розуміння предметної галузі. Наприклад, регресійні моделі добре працюють у методах виявлення викидів на даних розподілених лінійно (рис. 1.1). Для розподілу поданому на рис. 1.2, більше підходять кластерні моделі.

Також важливою складовою є правильний вибір складності (узагальнення) моделі. Складна модель, з високим рівнем узагальнення, з надто великою кількістю параметрів буде «перенавчатись», і знайде спосіб «пристосуватися» до викидів. Простіша модель, побудована з хорошим розумінням даних, швидше за все, призведе до кращих результатів. 12 Надмірно спрощена модель, погано «допасована» до даних, ймовірно прийме нормальні патерни за викиди.

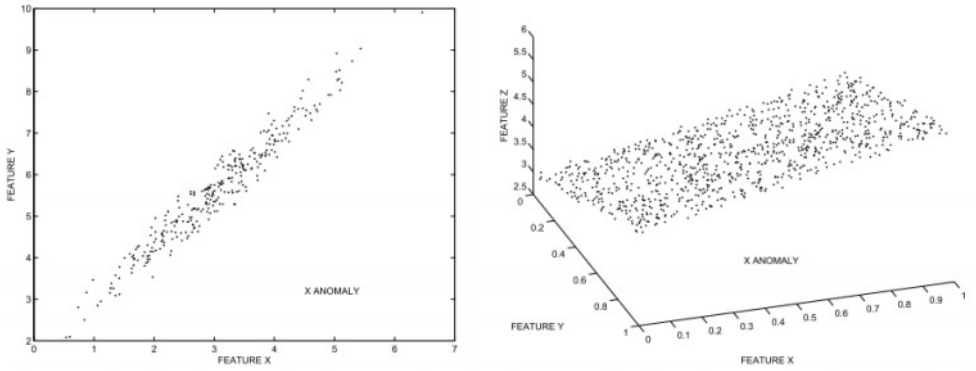

Рисунок 2.6 - Приклад даних з лінійним розподілом

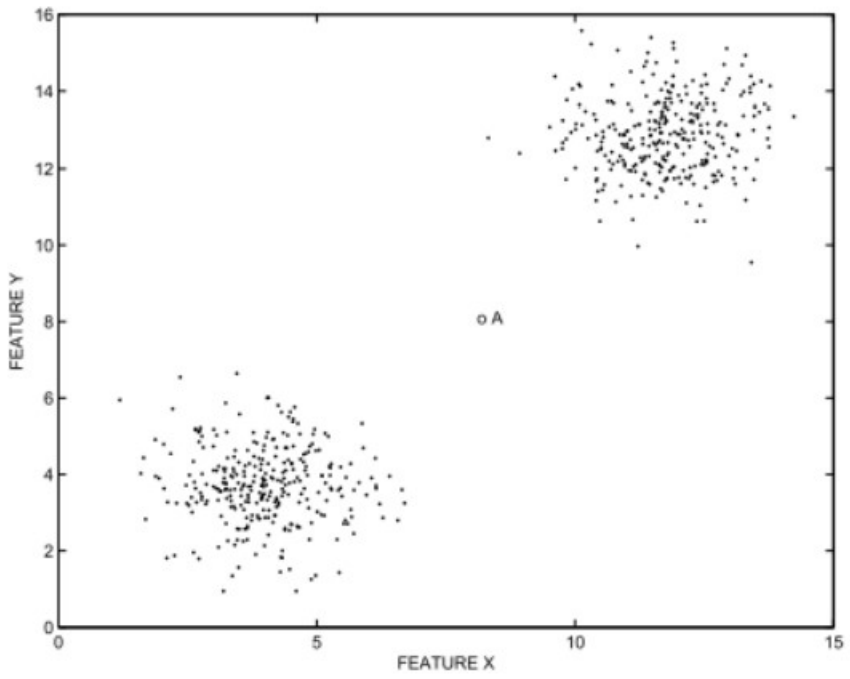

Рисунок 2.7 – Приклад даних, на яких аномалій можуть визначатись кластерним аналізом

2.10 Метод опорних векторів

Метод опорних векторів (SVM) - це один із найпотужніших методів машинного навчання. Незважаючи на те, що вона розроблялася головним чином для класифікаційних завдань, її модифікації успішно використовуються в проблемах виявлення аномалій. Метод підтримки векторів є дискримінаційним класифікатором: він "тримає" межу між кластерами даних. максимальна гіперплощадна межа.

Основними перевагами методу є:

- висока ефективність на даних великих розмірностей

- залишається ефективним навіть у випадках коли розмірність даних перевищує кількість записів в навчаючій вибірці

- ефективне використання оперативної пам’яті так як метод застосовує підмножину навчаючих точок в вирішуючій функції (опорні вектори)

- універсальність: у вирішуючій функції можуть використовуватись різноманітні функції ядра (будуть розглянуті нижче).

Недоліками методу опорних векторів є:

- класифікатор не дає ймовірнісних оцінок, проте в разі необхідності ці оцінки можуть бути розраховані методом крос-валідації (хоча цей процес не є ефективним в термінах обчислювальної складності).

Як зазначалося вище, метод векторних векторів побудований гіперплоскою (або набором гіперплощин) у великорозмірному просторі. Інтуїтивно зрозуміло, що найкращий розподіл досягається за допомогою гіперплоски, яка проходить на найбільшій відстані до найближчих точок даних будь-якого класу. Ця відстань називається функціональною, а в загальному випадку більша функціональна маржа забезпечує меншу генералізаційну похибку класифікатора

Ми формалізуємо проблему класифікації даних, встановлених методом референтних векторів. Точка даних буде розглядатися як p-мірний вектор. Завдання класифікації полягає в розподілі заданого набору точок (p-1) -мерної гіперплощини. Існує багато різних варіантів гіперпор, які можуть відокремити однакові дані. Варіант найкращої гіперплощини - це той, який пропонує найбільший інтервал поділу. Якщо така площина існує, вона називається максимальною гіперплоскою поля.

2.11 Статистичні та ймовірнісні моделі

Хоча ці моделі також не є предметом дослідження, їх огляд є важливим для розуміння багатьох методів машинного навчання. В цих методах дані моделюють у формі розподілу ймовірностей з параметрами які навчаються. Наприклад, модель суміші функцій Гауса – породжувальна модель яка характеризує дані в формі генеративного процесу що містить суміш гаусовських кластерів. Параметри цього розподілу навчаються з

використанням EM-алгоритму (Expectation-Maximization) на колекції даних. Результатом роботи алгоритму за цим методом є ймовірність приналежності точки даних певному кластеру, а також густина розподілу «допасування» точки даних до моделі (fit). Це забезпезчує природній спосіб виявлення викидів, як точок з найнижчим рівнем «допасування».

- Ключове припущення: Нормальні точки даних розміщені в регіонах статистичого розподілу з високими значеннями ймовірностей, тоді коли аномалії «лежать» в регіонах статистичного розподілу з низькими значеннями щільності ймовірності.
- Загальний підхід: Оцінити статистичний розподіл вихідних даних, зробити оцінку чи тестові дані належать до цього розподілу чи ні.
  - Якщо дане спостереження лежить на відстані більшій за 3 стандартні відхилення від середнього значення вибірки, вважаємо його за аномальним.
  - На аномальних даних наступний вираз приймає нетипово великі значення:

$$T^2 = \frac{n}{n+1}(X - \bar{X})' S^{-1} (X - \bar{X})$$

Основною перевагою імовірнісних моделей є можливість їх застосування до будь-якого типу даних (або змішаного типу даних), для яких існує відповідна генеруюча модель. Наприклад, для категоріальних (нечислових) даних дискретний розподіл Бернуллі може застосовуватися до кожного компонента суміші. Оскільки розглянуті моделі працюють з ймовірністю, проблема нормалізації (розглянута нижче) вже була врахована.

Основним недоліком імовірнісних моделей є спроба "адаптувати" збір даних до певного типу розподілу, що часто неприйнятно для типу досліджуваних даних. Коли кількість параметрів зростає, дослідники стикаються з проблемою перепідготовки - модель нормальної моделі може

додати і викиди. Також параметри цих моделей важко інтерпретувати з точки зору галузі - це усуває виконання однієї з завдань аналізу аномалій - виявлення причин ненормальної поведінки.

Як типова проблема класифікації, набір емоційних звукових сигналів є важливим для побудови моделей розпізнавання. Більшість існуючих звукосховищ пов'язані з такими темами, як розпізнавання мови, ідентифікація динаміків, класифікація жанру музики тощо. Спеціально розроблені набори даних повинні бути побудовані таким чином, щоб відповідати проблемі емоційного розпізнавання відповідно до характеристик емоційних звукових сигналів.

Можливі ресурси та існуючі набори емоційної мови вводяться відповідно у наступних підрозділах.

2.12 Таксономія набору даних про емоційну мову

Метод мовлення для вивчення емоцій повинен відображати різні типи емоцій унікальними моделями або конфігураціями акустичних настанов для надійного спілкування з основними емоційними станами динаміка. Існує досить велика кількість досліджень, присвячених способу отримання емоційних мовних тіл, які, як правило, можна розділити на три основні категорії: природний вокальний вираз, індукований емоційний вираз та моделювання емоційного виразу.

Природний вокальний вираз фіксується в природних емоційних станах різних типів. Він має дуже високу екологічну ефективність, яка вимірює ступінь, до якої фактичні характеристики відносяться до базової сили (згідно з термінологією Брансвік). Крім того, він більш точно відображає акустичні функції та інші клавіші смайлика на мові. Вираз природного вокалу можна отримати з реального життя або з телевізійної програми, наприклад, ток-шоу або шоу ігорної взаємодії. Незважаючи на

високу екологічну обгрунтованість, природне вираз вокалу має значні недоліки. Дійсно, сегмент з очевидними емоційними підказками у зразках природних голосів, як правило, страждає від низької якості запису, що призводить до труднощів у визначенні точного характеру основних емоцій. Незважаючи на те, що це викликає зацікавленість у характері вокальних емоцій, тренувальні схеми з незрозумілими емоційними станами можуть перешкоджати тренуванню алгоритму розпізнавання, послаблюючи акустичну кореляцію між звуковими характеристиками та емоційними станами.

ндуковані емоції обумовлені вживанням психоактивних лікарських засобів або деякими особливими обставинами, такими як індукція стресу через складні завдання, які необхідно виконувати під час тиску, подання емоційних плівок або слайдів, або способи зображення. Психолог завжди говорить про мову, але цей метод не може забезпечити бажану емоційну мову, тому що люди не завжди мають однакову реакцію на одне і те ж роздратування.

Третій спосіб отримати зразки мовлення - імітувати або зобразити емоційний вираз. Мова йде про акторів, у тому числі простих людей та професійних акторів, для голосового вираження певних емоцій. Ці зразки використовують цей вміст і дають деякі емоції. Емоції на зображеній мові мають більш типові вирази, ніж індуковані емоції, іноді навіть більш інтенсивні, ніж природні емоції. Проблема полягає в тому, що в зображених емоціях деякі очевидні підказки можуть бути надмірно підкреслені, тоді як деякі більш витончені натяки можна ігнорувати і не можуть точно відображати всі підказки в емоційному вокалі. Крім того, деякі міркування вказують на те, що показані емоції можуть спричинити щось із культурного тлі динаміки більше, ніж коли вони виникають у природному середовищі. Проте, стверджується, що всі виступи для

громадськості мають більш-менш сприйняття дій. Поки відображені емоції можуть бути визнані слухачами, вони відображають принаймні деякі емоційні моделі. Деякі дослідники віддають перевагу деяким дослідникам, тому що моделювання вокального образу емоцій має тенденцію додати набагато інтенсивніші, сильно контрольовані, прообразні вирази, ніж індуковані стани або навіть природні емоції.

2.13 Набори даних емоційної мови

Існує кілька зображених баз даних емоційного висловлювання на різних мовах. Згідно з, вокальні емоційні вирази можуть бути принаймні значною мірою обумовлені універсальними психофізіологічними механізмами, оскільки судді з різних культур, що говорять на різних мовах, визнають виражені емоції з набагато кращою точністю, ніж шанс. Цей пункт також може бути підтриманий, який показав, що навіть маленькі діти, які ще не розмовляють, можуть також визнати емоційні підказки з мови дорослих. У роботі Чжу незалежна від мови машина визнання людських емоцій у мові також реалізується з тілом емоційної мови з різних предметів та різних мов для розробки та тестування можливостей системи, що підтвердило роботу з людськими випробуваннями. Тому мови, що використовуються в базах даних емоційних мов, не мають значного впливу на вивчення вокальних емоцій.

Кілька баз даних були побудовані за різними даними

Висновки

Використання ШНМ цілком задовольняє вимоги до вирішення задачі класифікації, але показники якості та точності залежать від правильної конфігурації нейронної мережи, чіткості поділу вхідних даних на класи та наявності взаємозв’язку між параметрами ціх даних. З іншого боку

спектрограма – найбільш характеризоване та зрозуміле для глибинного навчання відображення звукових хвиль. Форматування цієї спектограми, ширина вибірки даних, та використання необхідної функції віконного трансформування Фур'є ще підлягають дослідженню з метою вдосконалення методики класифікації.

3. ЕКСПЕРЕМЕНТАЛЬНІ ДАНІ ТА МЕТОДИ ЇХ ОБРОБКИ

3.1 Цифрове відображення аудіо сигналу

Аудіо сигнал – являє собою коливання, тобто це залежність амплітуди коливання певної звукової хвилі від часу коливання. Процес перетворення аналового звукового сигналу в дискретний, називають аналогово-цифрове перетворення або оцифрування.

Дискретизація по часу – це процес обчислення миттєвих значень аналового сигналу, що перетворюється, з деяким кроком у часі, який називають крок дискретизації (рис. 3.1).

Рисунок 3.1 – Дискретизація сигналу

Частота дискретизації (частота вибірки, частота семплювання) – це кількість замірів величини сигналу, які здійснюються за одну секунду. З цього випливає, що для того, щоб збільшити точність перетворення сигналу, потрібно зменшити крок дискпетизації, оскільки значення амплітуди буде частіше реєструватися. Теорема Шеннона (з англ. Shannon) підтверджує цю думку (в нашій літературі частіше можна зустріти цю теорему під назвою «теорема Котельникова»). Згідно з цією теоремою, для того щоб аналоговий сигнал міг бути оцифрованим, а потім відновленим,

необхідно і достатньо, щоб частота дискретизації, була більшою або дорівнювала б подвоєній верхній частоті аналового сигналу. Тобто, аналоговий звуковий сигнал, частота найвищого складового спектра якого дорівнює F_m, точно описується послідовністю оцифрованих значень амплітуди сигналу, при виконанні для частоти дискретизації F_d умови: $F_d \geq 2F_m$. Оскільки діапазон частот які людина здатна почути становить 0-22кГц, це означає що частота дискретизації сигналу повинна дорівнювати не менше ніж 44кГц, для того щоб дискретизований сигнал містив інформацію на всьому діапазоні чутних для людини частот.

Не зважаючи на це, записати значення амплітуди сигналу які ми виміряли в чисельній формі, не достатньо для дискретизації сигналу. Оскільки тут основною проблемою дискретизації є неможливість записати виміряні значення сигналу з ідеальною точністю.

3.2 Стиснення з втратами

До втрат деякої аудіо інформації призводить кодування з втратами, через що таке кодування називається «з втратами». Таке кодування призводить до того, що транскодований сигнал під час відтворення звучить як оригінал, але фактично перестає бути ідентичним йому. В основі більшості методів кодування з втратою лежить використання психоакустичних властивостей слухової системи людини, а також різні трюки, пов'язані з повторним кількісним і попереднім поширенням сигналу. У частоті в процесі стиснення звукові дані аналізуються датчиком з метою виявлення різних частин звуку, які можна знехтувати. Маскуються частоти, нечутні та слабочутними деталями звуку - все це можна принести в жертву, щоб досягти більш високого значення коефіцієнта стиснення. Якщо звук важливий лише для читабельності (наприклад, у телефоні, де наявність частот вище 19,4 кГц не є необхідною), аудіоінформація в

процесі кодування зазнає серйозного "спрощення", що разом із використанням "розумного " та успішних "жадібні" алгоритмів стиснення даних дозволяють досягти найвищих показників стиснення (1:50 і вище). Якщо якість звуку вимагає більших вимог (наприклад, у портативних та домашніх аудіопристроях), звукові матеріали є більш ніжним кодуванням. Слід зазначити, що ступінь агресивності кодера по відношенню до частин звуку може бути скоригована (ця здатність, однак, залежить від конкретної реалізації). У середньому, сучасні програвачі, навіть при такому високому ступені стиснення, як 1:10, можуть забезпечити відмінний звук, якість якого середній прослухувач середнього обладнання оцінюється рівною якості звучання оригінальних звукових даних.

3.3 MP3 файл

MP3 - формат файлу для зберігання звукової інформації. Він використовує алгоритм стиснення втрат, призначений для значного зменшення розміру даних, необхідних для відтворення запису, і забезпечення якості відтворення звуку, що точно відповідає оригіналу (на думку більшості слухачів), але при значному втраті якості при прослуховуванні до високоякісної аудіосистеми.

Файл MP3 має стандартний формат 384, 576 або 1152 зразкових кадрів (в залежності від версії та рівня MPEG), і всі фрейми мають асоційовану інформацію заголовка (32 біти) та додаткову інформацію (9, 17 або 32 байти залежно від версії MPEG та стерео / моно).

Рівень 3 MPEG-1 визначає кілька бітрейтів: 32, 40, 48, 56, 64, 80, 96, 112, 128, 144, 160, 192, 224, 256 та 320 кбіт / с, а також доступну частоту дискретизації 32, 44,1 та 48 кГц. Частота дискретизації 44,1 кГц практично завжди використовується, оскільки вона також використовується для аудіо CD, головного джерела для створення MP3-файлів. Інтернет використовує

велику кількість бітових частот. 128 Кбіт/с є найпоширенішим явищем, оскільки він забезпечує адекватну якість звуку в відносно невеликому просторі. 192 Кбіт/с часто використовуються тими, хто помічає артефакти при низьких швидкостях передачі. Оскільки пропускна здатність Інтернету та простір на жорсткому диску стають більш доступними, файли на повільній швидкості 128 Кбіт/с повільно замінюються більш високими швидкостями передачі даних, наприклад, 192 Кбіт/с, деякі з яких кодуються до файлу MP3 з максимальною швидкістю 320 Кбіт/с.

Навпаки, нескомпресований звук, що зберігається на компакт-диску, має бітову швидкість 1411,2 Кбіт/с (16 біт/зразка × 44100 вібрацій/сек × 2 каналів / 1000 біт/кілобайт).

В стандартах MPEG-2 та (неофіційних) стандартів MPEG-2.5 були доступні додаткові бітові та частоти дискретизації: швидкість передачі 8, 16, 24 та 144 Кбіт/с та частоти дискретизації 8, 11.025, 12, 16, 22.05 та 24 кГц

Нестандартні швидкості передачі даних до 640 Кбіт / с можуть бути досягнуті за допомогою кодувача LAME та вільного формату, хоча ці файли можуть відтворювати декілька MP3-програвачів. Відповідно до стандарту ISO дешифратори повинні мати змогу декодувати потоки до 320 кбіт/с.

Оскільки формат MP3 підтримує двоканальне кодування (стерео), існує 3 режими:

- Стерео - двоканальне кодування, в якому канали кодуються незалежно один від одного. Таким чином, задана бітна швидкість ділиться на два канали. Наприклад, якщо задана бітна швидкість становить 192 кбіт / с, то для кожного каналу вона буде дорівнювати лише 96 кбіт / с.
- Моно - одноканальне кодування. Різниця між каналами буде повністю стерта, оскільки два канали змішуються в один, вони закодовані та

відтворюються на обох стереосистемних каналах. Єдина перевага цього режиму може бути лише оригінальною якістю в порівнянні зі стереорежимом з тим самим бітрейтом, оскільки один канал має в два рази більше бітів, ніж стереорежим. Але відмінності між каналами, які ви не почуєте, тому що канал тут лише один.

- Спільне стерео (Joint Stereo) - оптимальний спосіб подвійного каналу кодування, в якому ліва і права канали перетворюються на їх суму та різницю. Для більшості аудіофайлів канал з різницею набагато тихіший, ніж канал із сумою, тому більшу частину розміщується більша частина бітрейту. Таким чином, якість вихідного файлу відрізняється у найкращій стороні стереорежиму з однаковим бітрейтом, особливо при низькому. Вважається, що цей режим не підходить для аудіо стереоматеріалів, в якому два канали суб'єктивно відтворюють абсолютно інший матеріал, оскільки він стирає відмінності між каналами. Це помилкова ідея, адже насправді кодек MP3 працює на частотах, і певні частоти в більшості випадків перетинаються в обох каналах, тобто однакові дані все ще присутні, а різні - закодовані окремо. Цей метод двоканального кодування особливо ефективний при використанні змінних бітрейтів.

CBR означає постійну швидкість передавання, тобто постійну швидкість передавання, яка встановлюється користувачем і не змінюється під час кодування продукту, тому кожна секунда роботи відповідає такої ж кількості закодованих бітів даних. Насправді цей режим кодування не є оптимальним, оскільки він не підходить для більшості динамічних музичних творів із бітрейтом нижче 256 Кбіт/с.

VBR - це змінна швидкість, тобто змінна швидкість передавання даних або змінна швидкість передачі даних, яка динамічно змінює програму кодування під час кодування, залежно від насиченості закодованого звукового матеріалу та визначеної користувачем якості

кодування. Цей метод кодування MP3 є найбільш прогресивним і все ще розвивається і покращується, оскільки аудіозміст різних насичувань може бути закодований певною якістю, що зазвичай вище, ніж при налаштуванні середнього значення методу CBR. Крім того, розмір файлу зменшується через фрагменти, які не потребують високої бітрейту. Єдиним недоліком цього методу кодування є повна нездатність прогнозувати розмір вихідного файлу.

ABR - це середня швидкість передачі даних, тобто середня швидкість передачі даних, яка є гібридом VBR та CBR: швидкість передачі даних в кбіт / с встановлюється користувачем, і програма змінює її, постійно коригуючи її за певною швидкістю передачі даних. Таким чином, кодер не може встановлювати максимальні та мінімальні значення бітрейту, оскільки ризик не вписується в заданий користувачем бітрейт. Це очевидний недолік цього методу, оскільки це впливає на якість вихідного файлу, що буде трохи краще, ніж використання ЦБР, але набагато гірше, ніж використання VBR. З іншого боку, цей метод дозволяє отримати найменший бітрейт (може бути будь-яке число від 8 до 320, проти одноразового 16-значного методу CBR) і обчислити розмір вихідного файлу.

3.4 Структура MP3

Файл MP3 складається з декількох кадрів, а вони у свою чергу складаються з заголовку та блоку даних, як показано на рис. 3.2. Дана послідовність фрагментів називається потоком елементів. Фрагменти є залежними елементами ("резерв байтів"), і саме тому не можуть бути вилучені довільно. Блок даних MP3 файлу містить стиснуту аудіоінформацію у формі частот та амплітуд. Заголовок складається з маркера, який визначає правильний фрагмент MP3 і біта який вказує, що

MPEG використовується, і два біти показують використання шару 3; Іншими словами, він визначає MPEG-1 Audio Layer 3 або MP3. Наступні значення можуть відрізнятися залежно від типу файлу. Стандарт ISO/IEC 11172-3 визначає діапазон значень для кожного розділу заголовка разом із загальною специфікацією. Більшість MP3-файлів зараз містять метадані ID3, які передують або слідують фрагменту MP3.

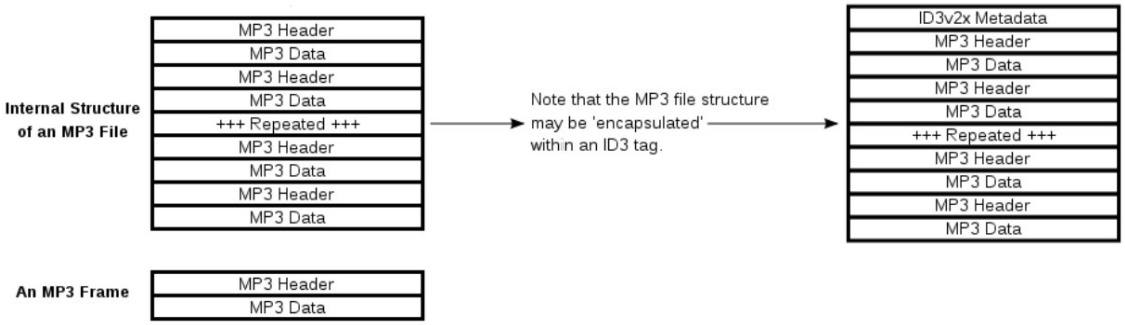

Рис. 3.2. Структура MP3 файлу

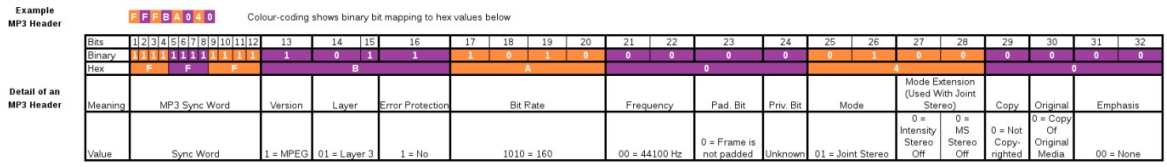

Рис. 3.3. Структура фрейму

Теги — мітки у межах Mp3-файла (на початку і в кінці). У них може бути записана інформація про авторство, альбом, році випуску і інша інформація про трек. У пізніших версіях тегів можливе зберігання обкладинок альбомів і тексти пісні. Існують різні версії тегів.

ID3 — формат метаданих. Найчастіше використовується разом з аудіоформатом MP3. Містить дані про назву трека, альбому, ім'я виконавця тощо. Ця інформація може використовуватись, наприклад, програвачами мультимедіа для відображення відомостей про трек чи автоматичного сортування. Назва «ID3» — це скорочення «Identify a MP3». Існує дві несумісних версії ID3: ID3v1 та ID3v2. Хоч ID3

розроблявся для MP3, розробникам немає перешкод вбудовувати його також і в інші формати.

Після створення формату MP3 виникла проблема зберігання метаданих. MP3 цього не реалізовував. У 1996 році Еріку Кемпу прийшла ідея як вирішити цю проблему: додати маленький шматочок даних до файлу. Стандарт, тепер відомий як ID3v1 швидко став стандартом де-факто зберігання метаданих у MP3-файлі.

У відповідь на критику стандарту ID3v1 у 1998 році був розроблений стандарт ID3v2. Він мало схожий на першу версію.

Існує три версії ID3v2: ID3v2.2 — перша публічна версія ID3v2. Використовує 3-символьний ідентифікатор кадру замість 4-символьного. Наприклад, TT2, а не TIT2. Тепер вважається застарілим. ID3v2.3 має 4-символьні ідентифікатори кадру. Кадр може мати кілька значень, розділених символом «/». На даний час, це найпопулярніший стандарт. ID3v2.4 — найновіша версія стандарту, представлена у листопаді 2000. Дозволяє зберігати текст у кодуванні UTF-8, замість UTF-16. Для розділення значень використовується нульовий байт, тому символ «/» можна вільно використовувати в тексті. Також додана можливість розміщати тег у кінці файлу, як у версії ID3v1.

3.5 Візуалізація блоку даних

Інформація з блоків даних MP3 файлу витягується в масив значень амплітуд у часі. У чистому вигляді цю інформацію дуже складно обробити, а тим більше характеризувати, тому потрібен спосіб представити набір даних інакше. Для нашої задачі буде зручним відобразити наявний масив у вигляді спектрограми (рис. 3.4.).

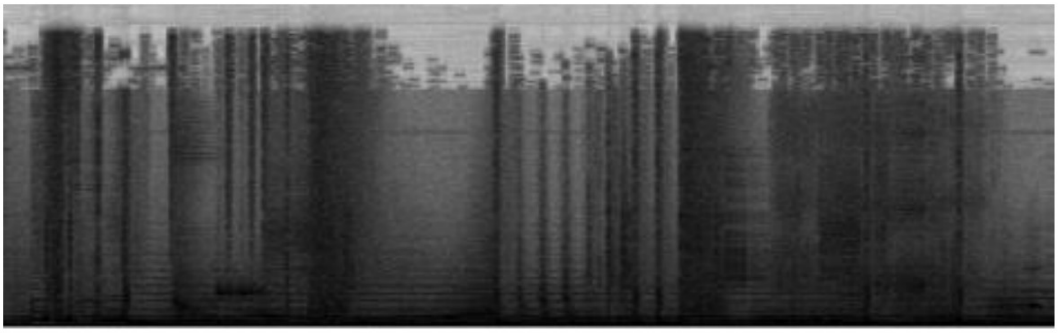

Рис. 3.4. Відрізок спектрограми аудіо файлу

Це зображення показує залежність спектральної щільності потужності сигналу від часу. Найбільш поширеним уявленням спектрограми є двовимірна діаграма: на горизонтальній осі представлено час, по вертикальній осі - частота; третій вимір із зазначенням амплітуди на певній частоті в конкретний момент часу представлено інтенсивністю або кольором кожної точки зображення. Спектрограма зазвичай створюється одним з двох способів: апроксимується, як набір фільтрів, отриманих із серії смугових фільтрів (це був єдиний спосіб до появи сучасних методів цифрової обробки сигналів), або розраховується за сигналом часу, використовуючи віконне перетворення Фур'є. Ці два способи фактично утворюють різні квадратичні частотно-часові розподілу, але еквівалентні при деяких умовах.

Для створення спектограмми сигнал розбивається на частини, які, як правило, перекриваються, і потім проводиться перетворення Фур'є, щоб розрахувати величину частотного спектра для кожної частини. Кожна частина відповідає вертикальній лінії на зображенні - значення амплітуди в залежності 23 від частоти в кожен момент часу. Спектри або терміни їх виконання розташовуються поруч на зображенні. 3.6. Віконне перетворення Фур'є Перетворення Фур'є - операція, що зіставляє однієї функції дійсної змінної іншу функцію дійсної змінної. Формула розрахунку перетворення відображена на рис. 3.5. Отримана нова функція

описує коефіцієнти («амплітуди») при розкладанні вихідної функції на елементарні складові - гармонійні коливання з різними частотами:

$$\hat{f}(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-ix\omega} dx \quad (3.1)$$

Віконне перетворення Фур'є – це різновид перетворення Фур'є, яка визначається формулою, де $W(\tau - t)$— деяка віконна функція:

$$F(t, \omega) = \int_{-\infty}^{\infty} f(\tau) W(\tau - t) e^{-i\omega\tau} d\tau \quad (3.2)$$

Існує безліч математичних формул візуально поліпшують частотний спектр на розриві кордонів вікна. Для цього застосовуються такі перетворення:

- прямокутне (рис 3.7.)
- фрагмент синусоїди
- синус в кубі
- перетворення Гаусса
- перетворення Хеннінга (рис. 3.8.)
- перетворення Хеммінга (рис. 3.9.)
- перетворення Розенфілда
- перетворення Блекмана-Харріса (рис. 3.10.)

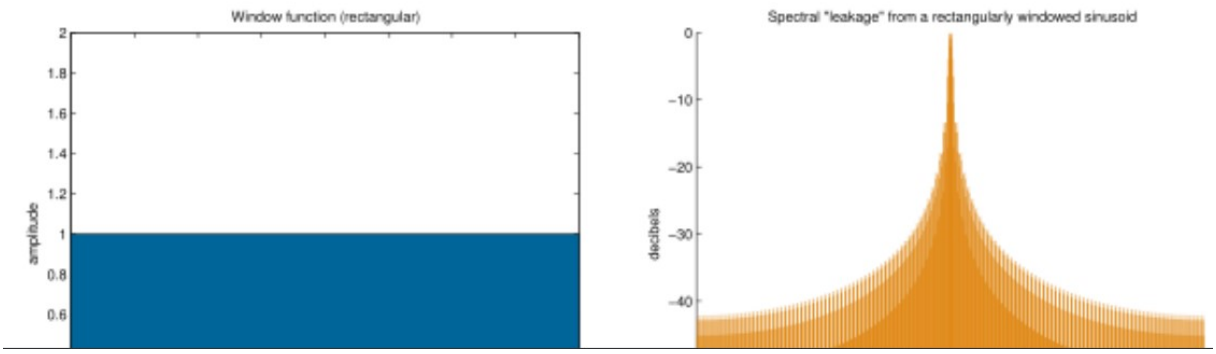

Рисунок 3.7 – Прямокутне перетворення

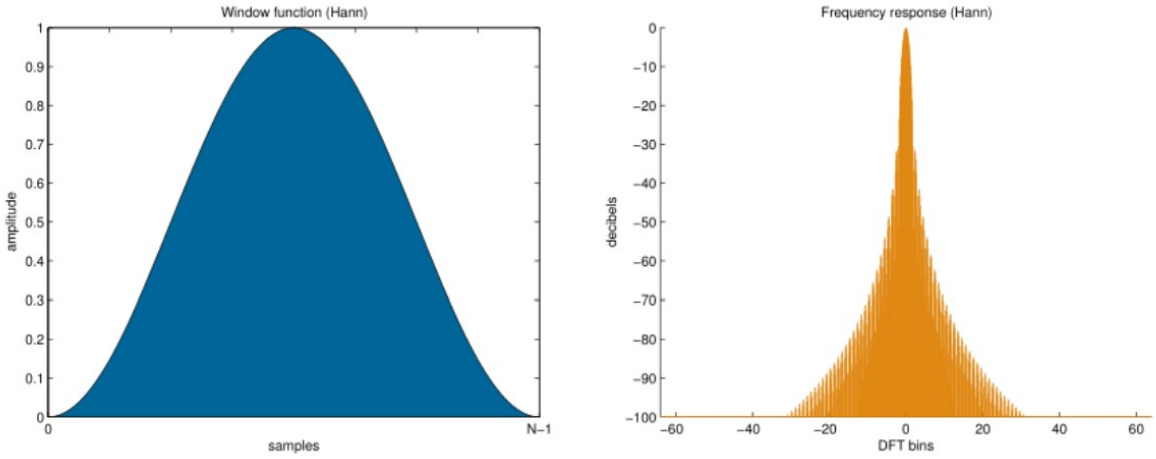

Рис. 3.8 – Перетворення Хеннінга

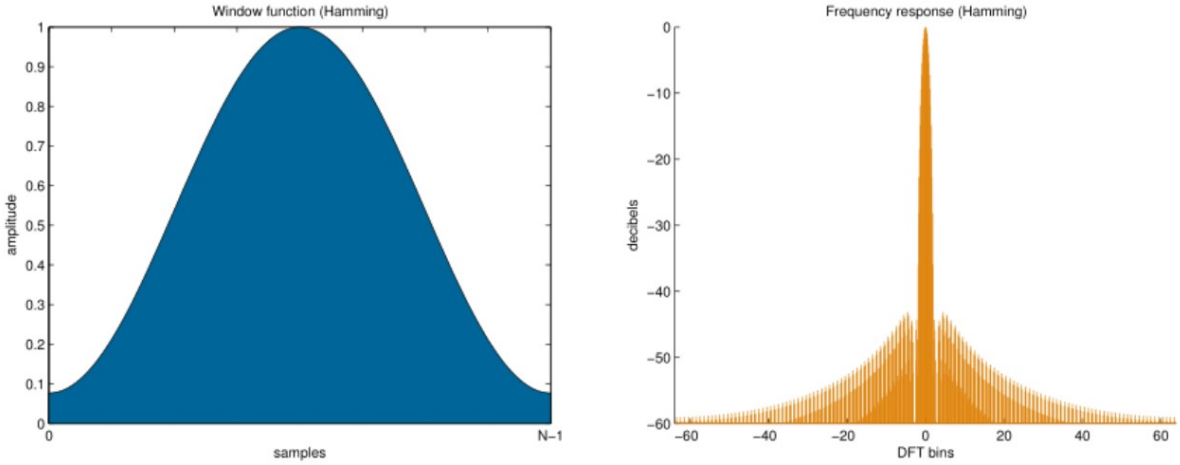

Рис. 3.9 – Перетворення Хеммінга

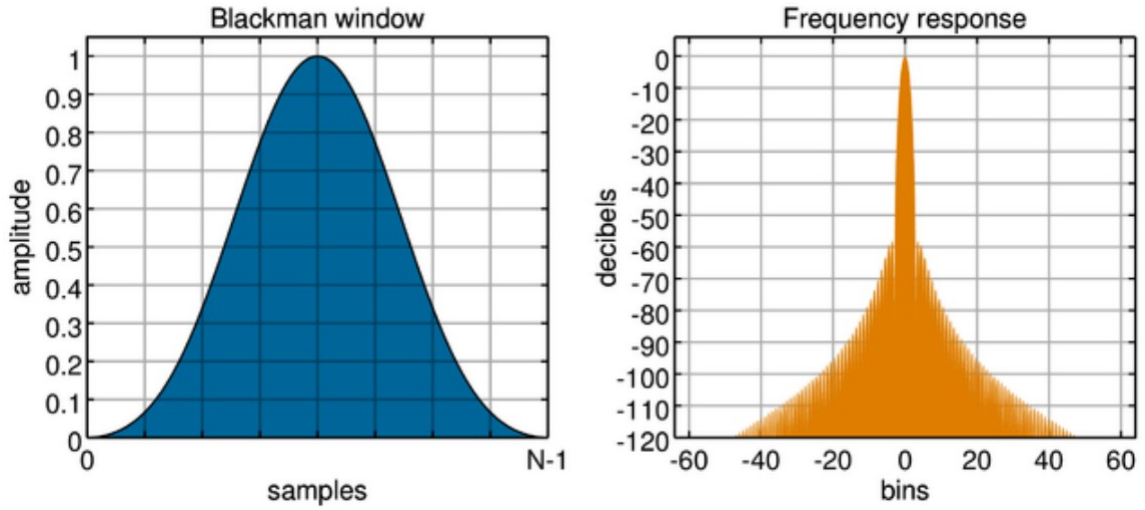

Рис. 3.10 – Перетворення Блекмана

За допомогою цих перетворень ми зможемо отримати спектрограми аудіо файлів, за якими будемо навчати програму класифікувати їх.

3.6 Задача класифікації

Задачи класифікації зустрічаються дуже часто в різних областях діяльності людини. Як і задача регресійного аналізу, задача класифікації вирішується з метою подальшого прогнозування змінної відгуку (номери класу). Передбачається, що вже є якась кількість n об'єктів, для кожного з яких відомий деякий набір з m ознак (чинників) і номер класу, до якого цей об'єкт належить, тобто сирі дані, які використовуються для вирішення задачі класифікації, мають вигляд:

Таблиця 3.1

Номер спостереження	Значення факторів			Значення змінної відгуку(номеру класу)
1	$x_{1,1}$	⋯	$x_{1,1}$	y_1
⋯	⋯	⋯	⋯	⋯
i	$x_{1,1}$	⋯	$x_{1,1}$	y_1
⋯	⋯	⋯	⋯	⋯
n	$x_{1,1}$	⋯	$x_{1,1}$	y_n

Тут значення змінної відгуку - номер класу, якому належить об'єкт, тобто $yi \in \{1, \dots, K\}$ для всіх $i = 1, \dots, n$ (2.3) K – (відома) кількість класів. Нейронні мережі є передовим способом вирішення завдань класифікації завдяки точності визначення і швидкості роботи. Для нашої задачі необхідна конфігурація мережі, що навчається з учителем.

3.7 Перетворення аудіо даних

Класична частота дискретизації становить 44100 Гц - на кожну секунду звуку записано 44100 значень і в два рази більше для стерео. Це означає, що 3- хвилинна стереофонічна пісня містить 7 938 000 зразків. Для нас це дуже великий обсяг даних, відповідно, потрібно зменшити його до більш керованого рівня, щоб ми мали можливість виробляти над нам будь-які операції. Для початку можна відкинути стереоканал, оскільки він містить надлишкову інформацію.

Ми скористаємося перетворенням Фур'є для відображення наших аудіо даних у вигляді частотної області. Це більш просте і компактне представлення даних, які ми будемо експортувати у вигляді спектрограми. В результаті цього ми отримаємо PNG-файл, який містить еволюцію всіх частот нашої пісні в часі.

Частота дискретизації 44100 Гц, про яку ми говорили раніше, дозволяє нам відновлювати частоти до 22050 Гц (за теоремою Котельникова [12]), але тепер, коли частоти витягнуті, ми можемо використовувати набагато більше значення роздільної здатності. Встановимо точність відображення на спектрограмі 50 38 пікселів в секунду (20 мс на піксель). Цього більш ніж достатньо, щоб бути впевненим у використанні всієї необхідної нам інформації.

На рис. 3.11. зображений відрізок голосу після процесу перетвореня (зразок 12.8s).

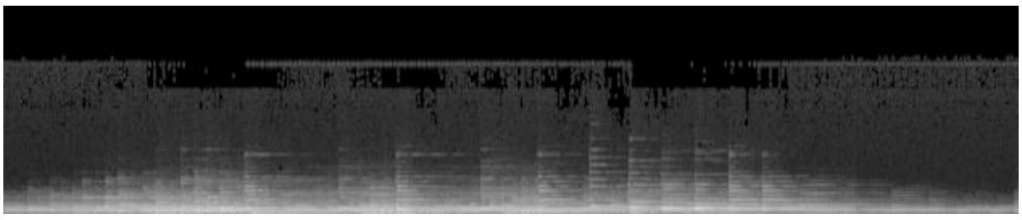

Рис. 3.11. Спектрограма витягу з голосу

Масштабована амплітуда частоти зображена у відтінках сірого, де білий є максимальним, а чорний - мінімальним. Я використовував

спектрограму зі 128 частотними рівнями, тому що в ній міститься вся необхідна інформація про звук - ми можемо з легкістю відрізнити різні частоти.

На етапі анализу найбільш зручної нейронної мережі встав вибір між рекурентною та згортковою. Для першої необхідно було б «годувати» кожен стовпчик зображення окремо по порядку, але людина здатна визначити емоцію за декілька секунд, тому цей варіант віявився не найкращим.

Для прискорення роботи згорткової моделі створему фрагменти фіксованої довжини спектрограми і розглянемо їх як незалежні зразки, що представляють емоції. Для цього ми можемо використовувати квадратні скибочки, скоротивши спектрограму до 128x128 пікселів (рис. 3.12.).

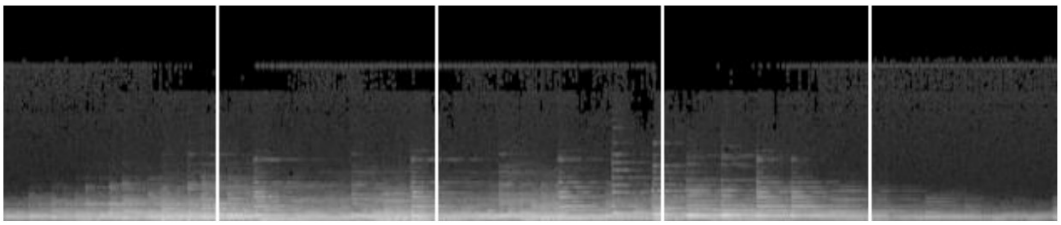

Рис. 3.12. «Нарізана» спектрограма

3.8 Загальний план вирішення задачі класифікації бібліотеки аудіо файлів

Застосування будь якої нейронної мережі у чистому вигляді не дасть ніяких результатів, бо ця мережа нічого не знає про те, що саме вона повинна класифікувати та за якими ознаками. Тому спочатку треба підготовити дані для тренування, та декілька з них виділити для тестування навченої мережі. При цьому кожен аудіо файл обов'язково повинен бути поміченим відповідним класом завдяки розташуванні у відповідній директорії. Після цього треба буде сконфігорувати нейронну мережу за згортковою моделлю враховуючи кількість пікселів сформованих спектрограм для вірного порядку та кількості необхідних

шарів нейронів. Останнім кроком ми натренуємо модель глибинного навчання та запустемо тестувальні дані для прогнозування жанрів, до яких вони ймовірніше відносяться.

Висновки

Використання моделі глибинного навчання у сукупності з перетворенням вхідного аудіо сигналу на спектрограму частот повністю вирішило проблему класифікації великого масиву аудіо даних. Але й ця система не позбавлена недоліків, таких як відносно повільна процедура Фур'є перетворень та групування отриманих частотних спектрів до єдиної спектрограми; точності класифікації все одно залежна від кількості вхідної вибірки.

4. АНАЛІЗ РЕЗУЛЬТАТІВ

4.1 Структура розробленого програмного продукту

Для демонстрації процесу розпізнавання емоцій людини за її голосом, було розроблено та реалізовано програмний продукт мовою Python 3.6 та використано відкриту програмну бібліотеку для машинного навчання TensorFlow. В программі використовується машинне навчання

Ефективність даного підходу експериментується як на Берлінському наборі даних, так і на наборах DES. Нижче вперше представленно набори даних Berlin і DES. Потім експериментальні представлені результати.

Дана робота перевіряє два набори даних, доступних для громадськості, зокрема набір даних в Берілі та набір даних DES. Ці два набори даних будуть представлені більш детально в наступних підрозділах.

4.1.1 Берлінський набір даних

Берлінська база даних емоційної мови розроблена професором Сендлмейєром та його колегами в кафедрі комунікації, Інституту мовлення та зв'язку, Берлінського технічного університету. Ця база даних містить зразки промов від п'яти акторів та п'яти актрис, десять різних слів німецької мови з семи видів емоцій: гнів, нудьга, відраза, страх, щастя, смуток і нейтральність. У цій базі даних існує всього 535 речових зразків, в яких 302 випуски мов є жіночими голосами та 233 зразки чоловічого голосу. Довжина зразків мовлення коливається від 3 секунд до 8 секунд, а частота дискретизації становить 16 кГц.

Таблиця 4.1 – Текст набору даних Берліну

Код	Текст (німецькою)	Переклад українською
A01	Der Lappen liegt auf dem Eisschrank	Тканина лежить на холодильнику
A02	Das will sie am Mittwoch abgehen	Вона хоче піти у середу

A04	Heute abend könnte ich es ihm sagen	Я міг сказати йому цього вечора
A05	Das schwarze Stück Papier befindet sich da oben neben dem Holzstück.	Чорний шматок паперу знаходиться поруч із шматочком дерева.
A07	In sieben Stunden wird es soweit sein.	Він буде готовий через сім годин.
B01	Was sind denn das für Tüten, die da unter dem Tisch stehen?	Які сумки під столом?
B02	Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter.	Вони тільки що провели це, і тепер вони знову спускаються.
B03	An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.	На вихідних я завжди їхав додому та відвідав Агнес.
B09	Ich will das eben wegbringen und dann mit Karl was trinken gehen.	Я просто хочу зняти це, а потім поїхати напій з Карлом.
B10	Die wird auf dem Platz sein, wo wir sie immer hinlegen.	Вона опиниться на площі, де ми завжди її покладаємо.

Зразки з емоційним станом "відраза" в нашому експерименті ігноруються через невідповідність акустичних особливостей. Інші зразки нарізані на сегменти 2 секунди. Хвилин частин зразків, що тривають більше 1,5 секунд, також зберігаються. Числа сегментів, отриманих в кінцевому експерименті в даному наборі даних кожного з емоційних станів, наведені в таблиці 4.2.

Таблиця 4.2 – Числа сегментів для емоційних станів у наборі даних Берліна

	Гнів	Щастя	Страх	Нейтр. стан	Смуток	Нудьга

Жінки	87	57	39	49	74	65
Чоловіки	75	34	39	46	49	49

4.1.2 Набір даних DES

Даний набір даних DES записано для Центру особистого спілкування (КПК), Ольборгського університету, Данія, у рамках проекту VAESS (Голоси, ставлення та емоції в синтезі мовлення). Звукові файли були записані в моно, використовуючи 16-розрядну PCM під частотою дискретизації 20 КГц.

Чотири актори були використані для запису DES, перерахованих у таблиці. 4.3

Таблиця 4.3 – Стать і вік 4-х акторів для запису набору DES

Ініціали	Стать	Вік
DHC	Жінка	34
KLA	Жінка	52
JZB	Чоловік	38
HO	Чоловік	52

У цьому наборі даних розглядаються п'ять емоцій: нейтральний, сюрприз, щастя, смуток і гнів. Для кожної емоції кожен актор записав наступні сегменти:

- 2 простих слова
- 9 фраз
- 2 проходи вільного мовлення

Окремі слова та речення беруться як тестовий набір у нашому експерименті. Оскільки проходи значно перевищують слова та речення, а емоції, що містяться в цих уривках, не є настільки типовими, як у словах та

реченнях, вони в наших тестах ігноруються. Тексти та переклади наведено в Таблиці 4.4.

Таблиця 4.4 – Слова та фрази в DES

Слова	Текст	Переклад
1	Ja.	Так.
2	Nej.	Ні.
Sentences		
1	Du er en sød dreng.	Ти гарний хлопецт
2	Jeg er ikke sulten.	Я не голодний.
3	Jeg ved det heller ikke.	Я також не знаю.
4	Hvad er det?	Що це?
5	Hvor er du?	Де ти?
6	Hvor skal du hen?	Куди ти йдеш?
7	Kom med mig!	Пішли зі мною!
8	Kommer du her igen?	Це знову ти?
9	Jeg synes vi mangler nogle, som er lidt længere.	Я думаю нам потрібно щось довше

У даній роботі ці дві набори даних використовуються для експериментальної оцінки нашого підходу. Оскільки існує більше емоційних типів та більше акторів у наборі даних Берліна, ніж набір даних DES, повномасштабні експерименти ведуться за допомогою набору даних Берліна, і попередні результати були зазначені в доповіді дослідження

4.2 Результати базовані на берлінському наборі даних

У наших експериментах дані в кожному випадку діляться на 10 груп випадковим для перехресної валідації, і середній з цих 10 результатів

розглядається як кінцевий результат. У кожний час експерименту 50% зразків використовують як тренувальний набір, а інші 50% зразків використовуються як набір тестів. Оскільки існує лише вісім зразків "відрази" у чоловічих зразках, що набагато менше, ніж інші типи, а акустична особливість цієї емоції несумісна, цей тип опущений під час навчання та тестування. Висвітлено також вплив ґендерної інформації на точність класифікації емоцій. Для кожної класифікаційної схеми оцінюються та порівнюються три експериментальних параметри, що використовують відповідно лише зразки жіночого мовлення, зразки чоловічої мови та комбінацію всіх зразків (змішані зразки).

4.2.1 Гармонічні та Zipf функції проти частоти та функцій на основі енергії

Цей перший експеримент спрямований на вивчення внесків наших гармонічних та Zipf функцій для покращення точності класифікації емоцій, коли вони використовуються в доповнення до класичних частотних та енергетичних функцій. Для цього експерименту в схему класифікації не вносяться жодних інновацій, і ми використовуємо лише кілька відомих глобальних класифікаторів, кожен з яких використовує той самий набір функцій. Таким чином, створюються два набори експериментальних результатів. Перша містить результати, отримані глобальними класифікаторами, коли використовуються лише класичні частотні та енергетичні функції. Другий набір експериментальних результатів отримується, коли попередні класичні частотні та енергетичні функції розширені, а також включають функції гармонії та Zipf.

Класифікатор перевіряється з декількома конфігураціями параметрів, і зберігаються лише найкращі результати.

Таблиця 4.5 – Точність розпізнання емоцій на основі статі

Frequency and energy feature set (FES)			All features (FES+Harmonic +Zipf features)		
Female	Male	Mixed	Female	Male	Mixed
60.38±2.26	57.91±2.56	60.38±2.26	65.73±2.85	64.45±2.47	64.47±1.93

Найвищі рівні розпізнавання наведені в таблиці. 4-6 і табл. 4-7. Як ми бачимо з цих таблиць, запропоновані нами додаткові функції допомагають поліпшити принаймні 4 бали продуктивність, досягається всіма світовими класифікаторами, які отримують за частотними та енергетичними характеристиками, найкраще вдосконалення на чоловічих емоційних зразках із продуктивністю виграш з 6 очок. Наступний експеримент точно покаже актуальність наших гармонічних та Zipf функцій у процесі класифікації.

Розпізнавання емоцій було досліджено за допомогою трьох основних типів баз даних: емоцій, природних спонтанних емоцій та викликаних емоцій. Найкращі результати, як правило, отримуються з діючими базами емоцій, оскільки вони містять сильні емоційні вирази.

Таблиця. 4.6 – Змішана матриця глобального класифікатора з частотними та енергетичними характеристиками

Голос		Гнів	Щастя	Страх	Нейтр. стан	Смуток	Нудьга
Жіночий	Гнів	67.55	23.67	6.24	1.39	0.00	1.15
	Щастя	35.56	45.60	12.50	3.70	0.00	2.64
	Страх	14.87	23.85	37.18	12.31	4.87	6.92
	Нейтр. стан	0.20	1.22	3.47	61.43	3.27	30.41
	Смуток	0.00	0.00	0.27	8.02	86.96	4.76
	Нудьга	2.00	1.85	6.62	30.92	8.15	50.46
Чоловічий	Гнів	82.67	8.80	6.80	0.67	0.53	0.53
	Щастя	36.18	39.12	20.00	3.53	0.00	1.18
	Страх	12.82	10.26	55.38	11.54	6.92	3.08

	Нейтр. стан	1.52	3.26	5.00	48.91	10.87	30.43
	Смуток	0.20	0.82	2.86	10.61	61.22	24.29
	Нудьга	1.84	1.43	1.84	27.96	26.73	40.20
Змішані	Гнів	71.82	21.71	4.73	0.46	0.00	1.27
	Щастя	43.31	41.02	8.63	3.52	0.35	3.17
	Страх	13.85	25.90	38.72	6.92	7.44	7.18
	Нейтр. стан	1.84	1.22	3.27	57.14	6.73	29.8
	Смуток	0.00	0.41	1.63	5.84	80.16	11.96
	Нудьга	2.62	2.46	3.54	24.00	12.31	55.08

Таблиця 4.7 – Змішана матриця глобального класифікатора з усіма характеристиками

		Гнів	Щастя	Страх	Нейтр. Стан	Смуток	Нудьга
Жіночий	Гнів	73.44	21.71	2.66	0.46	0.12	1.62
	Щастя	38.03	50.53	6.51	1.94	2.11	0.88
	Страх	12.56	23.59	41.79	5.9	10.51	5.64
	Нейтр. стан	1.02	1.02	0.61	60.00	6.12	31.22
	Смуток	0.00	0.14	0.68	5.30	86.68	7.20
	Нудьга	1.69	1.23	2.00	22.62	8.77	63.69
Чоловічий	Гнів	84.93	9.33	5.33	0.27	0.00	0.13
	Щастя	30.88	46.18	17.65	2.94	0.88	1.47
	Страх	11.03	18.72	55.38	7.44	6.15	1.28
	Нейтр. стан	3.26	0.65	3.26	58.04	7.39	27.39
	Смуток	0.00	1.63	3.06	4.29	73.27	17.76
	Нудьга	1.22	0.20	1.22	22.65	24.49	50.20
Змішаний	Гнів	74.57	18.40	5.19	1.05	0.00	0.80
	Щастя	38.81	44.76	11.14	1.76	1.76	1.76
	Страх	10.53	15.4	56.48	5.52	8.99	3.08
	Нейтр. стан	0.95	1.48	2.63	62.7	5.69	26.55
	Смуток	0.00	0.49	2.04	4.89	79.97	12.62
	Нудьга	1.50	1.14	2.73	23.66	14.95	56.02

4.3. Результати базовані на наборі даних DES

Заохочені попередніми результатами на наборі даних Берліна, ми також оцінюємо ефективність наших нових функцій та нашу багатостадійну модель розміщення емоцій на основі демографічного моделювання на наборі даних DES. Нагадаємо, що у наборі даних DES існують лише п'ять станів емоції: гнів, щастя, нейтральність, смуток та сюрприз. Використовуючи перший розмір розпаду, а потім оціночні вимірювання в моделі розмірної емоції, як ми робили для нашої попередньої класифікації з шістьма емоціями, ми вивели наступну ієрархічну класифікаційну схему. Цей процес перш за все розрізняє всі стани емоцій, згідно з розмахом збудження, у дві великі класи емоцій, що збігаються з гнівом, щастям і сюрпризом, з одного боку, і нейтраллю і смутком, з іншого боку. Ці широкі класи емоцій поділяються ще трьома класифікаторами для отримання остаточних станів емоцій.

Для того, щоб порівняти цей результат з роботою Ververidis, в цьому експерименті застосовується таке ж співвідношення між навчанням та тестуванням, що встановлено як 90% та 10% з перехресною валідацією. Таблиця 4.8 підсумовує показники точності, а табл. 4.9 дає матрицю плутанини такої оцінки. Як ми бачимо, в нашій роботі досягається середня точність корекції класифікації 81%. Для порівняння, найкраща продуктивність в літературі, згідно з нашими знаннями на одному і тому ж наборі даних, становить 66% точності класифікації для чоловіків зразків Ververidis

Table 4.8 – Рівень точності на наборі DES (%)

Жіночі	Чоловічі	Змішані
85.14±2.02	87.02±1.44	81.22±1.27

Table 4.9. – Змішана матриця набору DES (%)

Жіночий	Predicted Actual	Гнів	Щастя	Нейтр. стан	Смуток	Подив
	Гнів	**76.86**	14.71	2.94	1.37	4.12
	Щастя	9.22	**86.08**	0	1.18	3.53
	Нейтр. стан	1.37	2.55	**85.88**	8.43	1.76
	Смуток	0	0.96	8.46	**89.04**	1.54
	Подив	4.81	4.81	1.67	1.11	**87.59**
Чоловічий	Гнів	**84.51**	5.49	2.16	2.35	5.49
	Щастя	4.63	**85.37**	3.15	0.37	6.48
	Нейтр. стан	4.91	3.27	**87.09**	3.64	1.09
	Смуток	0.37	0.74	6.85	**90.93**	1.11
	Подив	5.9	6.56	0.49	0	**87.05**
Змішаний	Гнів	73.43	13.14	2.84	3.63	6.96
	Щастя	6.86	**80.67**	1.62	1.62	9.24
	Нейтр. стан	3.68	3.49	**81.89**	8.87	2.08
	Смуток	0.38	0.94	8.21	**88.77**	1.7
	Подив	7.22	7.83	1.39	2.52	**81.04**

Матриці в крос-мовних тестах для чотирьох емоцій наведені в табл. 4.10. Правильні класифікаційні рівні для більшості випадків для чотирьох емоційних типів становлять близько 50%, за винятком надзвичайно низького рівня для гніву з жіночими зразками та дуже високою рівня для смутку з чоловічими зразками, коли набір даних Берліну використовується як навчальний набір, а набір даних DES використовується як набір тестів.

Table. 4.10 – Змішана матриця крос-мовного тесту (%)

	Training – Berlin, Testing – DES					Training – DES, Testing - Berlin			
	Predicted Actual	Гнів	Щастя	Нейтр.	Смуток	Гнів	Щастя	Нейтр.	Смуток

				стан				стан	
Female	Гнів	**15.69**	19.61	60.78	3.92	**52.87**	39.08	0.00	8.05
	Happiness	0.00	**41.18**	52.94	5.88	35.09	**54.39**	1.75	8.77
	Neutral	9.80	5.88	**74.51**	9.80	26.53	16.33	**40.82**	16.33
	Sadness	0.00	5.77	23.08	**71.15**	14.86	2.70	25.68	**56.76**
Male	Гнів	**56.86**	23.53	7.84	11.76	**52.00**	30.67	17.33	0.00
	Happiness	29.63	**48.15**	14.81	7.41	26.47	**55.88**	14.71	2.94
	Neutral	7.27	21.82	**49.09**	21.82	17.39	8.70	**47.83**	26.09
	Sadness	7.41	5.56	3.70	**83.33**	10.20	10.20	26.53	**53.06**

ВИСНОВКИ

В даній роботі ми запропонували, в доповнення до класичних частотних та енергетичних функцій, метод машинного навчання. Крім того, ми маємо запропонувати ієрархічну класифікаційну схему (схему DEC), яка використовує альтернативний вимір збудження та оцінки з розмірної емоційної моделі, що стосується нечіткого оточення деяких дискретних станів емоції, що мають подібні акустичні корелятиви. Спочатку експерименти на базі даних Берліна показують, що наш новий запропонований метод допомагає покращувати рівень розпізнавання емоцій, коли він використовуються у доповнення до класичних частотних та енергетичних функцій. Ефективність нашого підходу також була підтверджена на іншому загальнодоступному наборі даних, наборі даних DES.

Проте існують ще кілька питань, які потрібно розглянути в майбутньому.

По-перше, оскільки загальної згоди щодо кількості та видів дискретних емоцій немає, типи емоцій, які розглядаються на практиці, зазвичай залежать від застосування або набору даних. Наша схема спирається на інтуїтивне відображення стану дискретних емоцій у моделі розмірної емоції. У цій роботі це інтуїтивне відображення було зроблено вручну та емпірично. Явно потрібна схема автоматичного картографування, особливо коли кількість емоцій збільшується, і їх види різняться.

По-друге, оскільки емоції дуже суб'єктивні, а емоції кордону між замкнутими емоціями в розмірному просторі, як правило, не дуже чіткі, судження про емоційний стан, переданий висловом, може бути між деякими емоційними станами або навіть множинними відповідно до особи.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Hopfield J.J. Neural with graded response have collective computational properties like those of two-state neurons // Proceedings of the National Academy of Science of the USA, 1984.–vol.81.–P.3088-3092.
2. Dempster A.P., Laird N.M., Rubin D.B. Maximum Likelihood Estimation from Incomplete Data via the EM algorithm // Journal of the Royal Statistical Society B, 1977.– vol.39.–P.1-38
3. Bhat N.V., McAvoy T.J. Determining model structure for neural models by network pruning // Computers in Chemical Engineering, 1992.–vol.16.–P.271-281.
4. . Elsner L., Koltracht I., Neumann M. On the convergence asynchronous paracontractions with application to tomographic reconstruction from incomplete data // Linear Algebra and its Applications, 1990.–vol.130.–P.65-82.
5. Grossberg S. Contour enhancement, short-term memory, and consistencies in reverberating neural networks // Studies in Applied Mathematics, 1973.–L11.–P.213-257.
6. Kohonen T. Self-organized formation of topologically correct feature maps // Biological Cybernetics, 1982.–vol.43.–P.59- 69.
7. Yuille A.L., Kammen D.M., Cohen D.S. Quadrature and the development of orientation of selective cortical cells by Hebb rules // Bioligical Cybernetics, 1984.–vol.61.–P.183-194.
8. Ziehe A., Müller K.R. TDSEP – an efficient algorithm for blind separation using time structure // ICANN'98, Skovde, 1998.–P.675-680.
9. Szyld D.B., Jones M.T. Two-stage and multisplitting methods for the parallel solution of linear systems // SIAM J. Matrix Anal. Appl., 1992.–vol.2.–P.671-679.

10. Rosenblatt F. Two theorems of statistical separability in the perceptron. // Symposium on the Mechanization of Thought Processes, 1959.–P.421-456.
11. Nesterenko B.B., Novotarskiy M.A. Mathematical simulation for parallel asynchronous methods of boundary value problems of mathematical physics // 16th IMACS World Congress, 2000.–6 p.
12. Miller K.D., MacKay D.J.C. The Role of Constraints in Hebbian Learning // Neural Computation, 1994.–vol.6.– P.100-126.
13. O'Leary D.P., White R.E. Multi-splitting of matrices and parallel solution of linear systems // SIAM Journal of Algebraic and Discrete Mathematic, 1985.–vol.6.–P.630-640.
14. Oja E. A Simplified neuron model as a principal component analyzer // J. Math. Biol., 1982.–vol.15.–P.267-273,.
15. Oja E. Neural networks, principal components and subspaces // Int. J. Neural Systems, 1989.–vol.1.–P.61-68.