Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Кібербезпеки
(повна назва)

Кафедра Інформаційно-мережної інженерії
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіті другий (магістерський)

Розпізнавання емоцій людини за її мовою та зображенням

(тема)

Виконав:
здобувач 2 року навчання,
групи ІМІм-24-2
Зінченко Є.А.
(прізвище, ініціали)

Спеціальність 172 Електронні комунікації та радіотехніка
(код і повна назва спеціальності)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Інформаційно-мережна інженерія
(повна назва освітньої програми)

Керівник: доц. Омельченко С.В.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ Москалець М.В.
(підпис) (прізвище, ініціали)

2026 р.

Не містить відомостей, заборонених до відкритого публікування

Студент _____ / Зінченко Є.А. /
(підпис) (прізвище та ініціали)

Керівник _____ / Омельченко С.В. /
(підпис) (прізвище та ініціали)

Харківський національний університет радіоелектроніки

Факультет Кібербезпеки

Кафедра Інформаційно-мережної інженерії

Рівень вищої освіти другий (магістерський)

Спеціальність 172 Електронні комунікації та радіотехніка

(код і повна назва)

Тип програми освітньо-наукова

Освітня програма Інформаційно-мережна інженерія

(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
« » травня 2026 р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Зінченко Євгенію Аркадійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Розпізнавання емоцій людини за її мовою та зображенням

затверджена наказом по університету від« 23» березня 2026 р. № 232 Ст

2. Термін подання студентом роботи до екзаменаційної комісії 26 травня 2026 р.

3. Вхідні дані до роботи Виконати огляд методів розпізнавання емоційного стану людини. Розглянути особливості оцінювання мовленевих та візуальних ознак. Розглянути методи класифікації. Дослідити системи розпізнавання емоцій людини.

4. Перелік питань, що потрібно опрацювати у роботі Вступ

1 . Огляд методів розпізнавання емоційного стану людини

2. Оцінювання мовленевих ознак

3. Оцінювання візуальних ознак

4. Вибір та класифікація ознак

5 Дослідження систем розпізнавання емоцій

Висновки

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри) слайди презентації в форматі Power Point (назва та мета роботи, методи, алгоритм, висновки)

6. Консультанти розділів роботи (п.6 включається до завдання за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата
Основна частина	доцент Омельченко С.В.		

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Вступ	23.03	виконано
2	Огляд методів розпізнавання емоційного стану людини	23.03-05.04	виконано
3	Оцінювання мовленевих ознак	06.04-12.04	виконано
4	Оцінювання візуальних ознак	13.04-20.04	виконано
5	Вибір та класифікація ознак	21.04-05.05	виконано
6	Дослідження систем розпізнавання емоцій	06.05-22.05	виконано
7	Висновки	16.05-25.05	
8	Оформлення пояснювальної записки	25.05	виконано

Дата видачі завдання 28 березня 2026 р.

Здобувач _____
(підпис)

Керівник роботи _____ доц. Омельченко С.В.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка: 76 с., 19 рис., 34 джерел, 1 додатка.

Об'єкт дослідження – методи розпізнавання емоційного стану людини, які поєднують акустичну та візуальну інформацію.

Мета роботи – дослідження методів розпізнавання емоційного стану людини.

Досліджуються методів розпізнавання емоцій на основі полігаусівської моделі, методу К -найближчих сусідів, нейронних мереж та лінійного дискримінантного аналізу.

Результати показують, що комбінація аудіо та візуальної інформації досягає кращої загальної точності, ніж будь-яка з них окремо.

РОЗПІЗНАВАННЯ ЕМОЦІЙ, ПОЛІГАУСІВСЬКА МОДЕЛЬ, НЕЙРОННІ МЕРЕЖІ, ФІЛЬТРИ ГАБОРА, БАНК ФІЛЬТРІВ.

THE ABSTRACT

Explanatory note: 76 p., 19 fig., 18 sources, 1 app.

The object of the study is methods for recognizing human emotional states that combine acoustic and visual information.

The purpose of the work is to study methods for recognizing human emotional states.

Emotion recognition methods based on the polygaussian model, the K-nearest neighbors method, neural networks, and linear discriminant analysis are studied.

The results show that the combination of audio and visual information achieves better overall accuracy than either of them separately.

EMOTIONS RECOGNITION, POLYGAUSSIAN MODEL, NEURAL NETWORKS, GABOR FILTERS, FILTER BANK.

ЗМІСТ

ПЕРЕЛІК СКОРОЧЕНЬ

PCA – (Principal Component Analysis) – аналіз головних компонентів ;

FACS – (Facial Action Coding System) — система кодування мімічних рухів);

SVM – (Support Vector Machine) -Метод опорних векторів;

LP – (Linear Programming) – лінійне програмування.

ВСТУП

Оскільки комп'ютери стали невід'ємною частиною нашого життя, виникла потреба в більш природному інтерфейсі зв'язку між людьми та машинами.

Щоб досягти цієї мети, комп'ютер повинен бути здатним сприймати поточну ситуацію та реагувати по-різному залежно від цього сприйняття. Щоб зробити взаємодію людини з комп'ютером більш природною та зручною, було б корисно надати комп'ютерам здатність розпізнавати емоційний стан так само, як і людина.

Емоційний стан людини є критично важливою галуззю досліджень у мультимедійній спільноті. Вона відіграє вирішальну роль у таких галузях, як проектування біокомп'ютерів, безпека та спостереження, а також навчання людей з особливими потребами. Для розпізнання емоційного стану людини використовують вираз облич, особливості інтонацій мовлення, жестів, моделювання тіла та розпізнавання рухів. Повна система розпізнання емоційного стану людини інтегрує всі аспекти цих компонентів.

Основними характеристиками людського спілкування є множинність та мультимодальність каналів зв'язку. Прикладами каналів людського спілкування є слухові канали, які передають мову та голосову інтонацію, та візуальні канали, які передають міміку та жести. Модальність – це відчуття, яке використовується для сприйняття сигналів. Прикладами модальностей є зір, слух та дотик. У реальному особистому спілкуванні активується багато різних каналів та модальностей, і таким чином комунікація є дуже гнучкою та надійною. Саме так також інтелектуальна система повинна бути розроблена для забезпечення надійної, природної, ефективної та результативної взаємодії людини та машини.

У сфері людсько-контактної ідентифікації мовлення міміка є основними завданнями системи розпізнавання емоцій. Вони вважаються двома основними показниками афективного стану людини, і тому відіграють дуже важливу роль у розпізнаванні емоцій. Тому необхідно досліджувати методи, за допомогою яких комп'ютер може розпізнавати людські емоції з аудіовізуальної інформації. Такі методи можуть сприяти комунікації між людьми та комп'ютерами, а також таким застосуванням, як навчальне середовище, розваги, обслуговування клієнтів та освітнє програмне забезпечення.

Емоції предметом пильного інтересу як східної, так і західної філософії ще з часів Лао-цзи (шосте століття до н. е.) на сході та Сократа (470-399 рр. до н. е.) на заході, і найсучасніші роздуми про емоції в психології можна пов'язати з тією чи іншою західною філософською традицією. Однак багато хто вважає, що початок сучасного наукового дослідження природи емоцій покладається на дослідження Чарльзом Дарвіном емоційного вираження у тварин і людей.

Дослідження людських емоцій у психології та нейрофізіології швидко розширилися за останні роки. Цей швидкий розвиток частково зумовлений досягненнями в технології візуалізації, а також новим інтересом до ідеї про те, що певні окремі емоції можуть обслуговуватися окремими системами мозку. Ця остання теоретична позиція лежить в основі ідеї про те, що вибрані набори так званих «базових» емоцій складають основу людських емоцій. Усі емоції обробляються ланцюгом взаємопов'язаних структур мозку, відомим як лімбічна система.

Сучасні дослідження емоцій у психології показують, що певні емоції пов'язані з різними сигналами обличчя, і що вони були спільними для культур усього світу. Ці основні емоції можуть кодуватися частково різними системами мозку. Наприклад, було виявлено, що мигдалина відіграє значну роль у розпізнаванні мімічних та голосових виразів страху. Широке дослідження вимірів емоцій показує, що принаймні шість емоцій.

Ці емоції універсальні. Кілька інших емоцій та багато їх комбінацій були досліджені, але залишаються непідтвердженими як розрізнювані. Набір із шести основних емоцій: щастя, смуток, гнів, страх, здивування та огида.

Для покращення інтелектуальної взаємодії між комутаторами за останні кілька років було докладено багато зусиль до машинного розпізнавання людських емоцій. Більшість цих робіт зосереджені або лише на мовленні, або лише на виразі обличчя. Порівняно мало існуючих робіт поєднують ці дві модальності в одну єдину систему для аналізу емоцій. Однак, деякі емоції є домінантними за аудіо, тоді як інші – за візуальними. Коли одна модальність не спрацьовує або недостатньо ефективна для визначення певної емоції, інша модальність може допомогти виконати розпізнавання. Інтеграція аудіо та візуальних даних надасть більше інформації про емоційний стан людини. Взаємодоповнюючий зв'язок цих двох модальностей щодо різних емоцій допоможе досягти вищої точності розпізнавання.

Через те, що більшість існуючих систем обробляють аудіо та візуальну інформацію окремо, не існує стандартної бази даних для аудіовізуального розпізнавання людських емоцій. Більшість робіт у дослідницькій спільноті використовують базу даних, яку вони зібрали самостійно. Більшість систем обмежені базою даних лише однієї мови. Однак спосіб, яким люди передають емоційний стан, може відрізнятися залежно від їхнього культурного походження, мови та акценту. Ефективна система розпізнавання емоцій повинна бути здатною адаптуватися до цих аспектів. Існують притаманні їм риси, незалежні від цих умов. Метою дослідження є розробка ефективної системи для розпізнавання людських емоцій.

Система розпізнавання емоцій для розпізнавання афективного стану людини з аудіовізуальної інформації повинна тестуватись на базі даних,що незалежні від мови, мовця та контексту. Аудіо та візуальні ознаки оцінюються з емоційних даних окремо для представлення вокальних та мімічних характеристик людей при різних емоціях. Відбір ознак необхідний для того щоб виявити значущі ознаки, одночасно зменшуючи розмірність простору ознак. Для досягнення найвищої точності розпізнавання необхідно порівнювати продуктивність різних алгоритмів класифікації. Крім того, оскільки різні емоції можуть мати різні значущі ознаки, а ознаки для розрізнення комбінацій різних емоцій також можуть бути різними, необхідно використовувати схему з кількома класифікаторами для аналізу цих сценаріїв.

Для проведення дослідження розпізнавання людських емоцій потрібна відеобаза даних, яка може достовірно передати емоційний стан людини. Продуктивність системи розпізнавання емоцій значною мірою залежить від якості бази даних емоцій, на якій вона навчається. Працюючи з мовленням та мімікою, необхідно приділяти особливу увагу тому, щоб конкретна емоція, що озвучується та виражається, була правильною. Для запису емоційних даних з достатньо високою точністю може використовуватись цифрова відеокамера.

1 ОГЛЯД МЕТОДІВ РОЗПІЗНАВАННЯ ЕМОЦІЙНОГО СТАНУ ЛЮДИНИ

Необхідність розуміння емоційного стану людини стало великим викликом, який включає багатопрофільні дослідження від психології до обробки сигналів. Нещодавні досягнення в аналізі мовлення та зображень, а також розпізнаванні образів відкривають можливості автоматичного виявлення та класифікації емоційних аудіовізуальних сигналів. Застосування розпізнавання емоцій можна передбачити в широкій галузі людського комп'ютерного інтелекту.

На якість аналізу емоцій значно впливають такі вхідні дані, як голос, міміка, мова тіла та розуміння семантичного значення слів тощо. Однак загальновизнано, що мовлення та міміка є двома основними модальностями в людському спілкуванні, а отже, і для більшості застосувань взаємодії людини з комп'ютером. Мова тіла, яку люди виражають у різних емоціях, сильно залежить від культурного походження, особистості статі, віку тощо. Важко змоделювати емоційну поведінку людини в загальному вигляді. Хоча розуміння семантичного значення слів певною мовою може допомогти розпізнати емоції, два речення можуть мати однакове лексичне значення, але дуже різну емоційну інформацію. Хоча більшість досліджень використовують окремо мовлення або міміку, але деякі системи поєднують обидві ці модальності.

1.1 Огляд методів розпізнавання емоцій на основі мовлення

Більшість існуючих систем розпізнавання емоцій на основі мовлення складаються з трьох компонентів: виділення ознак, вибір ознак та класифікація (рис. 1.1). Основною проблемою тут є виділення ознак, які можуть достовірно відображати вокальні характеристики людського мовлення в різних емоційних станах. Існує два основних типи ознак: просодичні ознаки та фонетичні ознаки.

Просодичні ознаки в основному пов'язані з ритмічними аспектами мовлення, тоді як фонетичні ознаки зосереджені на лінгвістичних аспектах мовлення. Після виділення ознак мовлення вони застосовуються алгоритми вибору ознак для вибору значущих ознак, тоді як класифікатори

використовуються для зіставлення поточних ознак з ознаками, отриманими для відповідних емоційних станів на етапі навчання.

Рисунок 1.1 - Архітектура системи розпізнавання мовлення

Коуї та Дуглас-Коуї [1] розширили погляд на розпізнавання емоцій, зазначивши, що мовленнєві особливості, які раніше асоціювалися з емоціями, також можуть бути пов'язані з такими порушеннями, як шизофренія та глухота. Їхні дослідження більше зосереджувалися на інтенсивності та ритмі, а також досліджували важливість глухих звуків.

Амір та Рон [2] повідомили про метод ідентифікації емоцій на основі аналізу сигналу через ковзні вікна. Було визначено набір базових емоцій, і для кожної емоції обчислювалася точка відліку. У кожен момент часу обчислювалася відстань вимірюваного набору параметрів від точки відліку, яка використовувалася для обчислення нечіткого індексу належності для кожної емоції.

Сато та Морісіма [3] продемонстрували використання нейронних мереж для аналізу та синтезу емоцій у мовленні. У їхній статті також обговорювалося, як емоції кодуються в мовленні. Автори класифікували емоції як «гнів», «відраза», «щастя», «сум» та «нейтральність». Вони коротко обговорили шість параметрів мовлення, які були проаналізовані, та зробили деякі емпіричні спостереження щодо зв'язку між цими параметрами та певними емоціями. Звукові зразки, використані в експерименті, були записані професійним актором, проте група людей-випробуваних могла правильно розпізнати його емоцію лише у 70% випадків.

Делларт та інші [4] дослідили кілька методів статистичного розпізнавання образів для класифікації висловлювань відповідно до їхнього емоційного змісту. Різні методи вибору ознак, включаючи перспективний перший вибір (PFS), прямий вибір (FS) та голосування більшості спеціалістів, порівнюються для покращення продуктивності системи. Як категорії емоцій використовується

набір із чотирьох емоцій: щастя, смуток, гнів та страх. Застосована база даних емоційного мовлення містить 1000 зразків мовлення від 5 суб'єктів людського суспільства.

Ця система досягає рівня розпізнавання до 79,5%. Однак у своїх заключних зауваженнях автори зазначили, що результати залежать від мовця, і що методи класифікації повинні бути перевірені на повністю заповнених базах даних.

Ніколсон та ін. [5] запропонували незалежну від мовця та контексту систему для розпізнавання емоцій у мовленні за допомогою нейронних мереж. У статті розглянуто як просодичні, так і фонетичні ознаки. Досліджено набір із восьми емоцій, а саме: радість, дражнення, страх, смуток, огида, гнів, здивування та нейтральність. Для тестування системи було використано велику базу даних із загалом 100 мовців, 50 чоловіків та 50 жінок, які використовують слова зі збалансованими фонемами. Ця база даних забезпечує незалежність експериментів від мовця та контексту. Однак, для отримання найкращої підмножини ознак не застосовувалися жодні методи вибору ознак, і коефіцієнт розпізнавання становив лише близько 50%.

У статті К. М. Лі, С. Нараянана та Р. П'єраччіні [6] діалоги людини і машини, що записано з комерційного застосунку, класифікується за двома основними емоціями: негативними та позитивними. Ознаки, що використовуються класифікаторами - це частоти основного тону та енергії мовного сигналу. Для зменшення розмірності ознак та максимізації точності класифікації було застосовано аналіз головних компонентів (PCA).

У статті Т. Л. Неве та Ф. С. Вей [7] розпізнавач базується на дискретній прихованій марковській моделі (HMM), та коефіцієнтах потужності короткочасного мовлення за частотою Mel. Досліджуються шість основних емоцій: гнів, неприязнь, страх, щастя, смуток та здивування. Універсальна кодова книга побудована на основі емоцій, що спостерігаються для кожного експерименту. База даних складається з 90 емоційних висловлювань від двох мовців. Система забезпечує середню точність 72,22% та 60% відповідно для двох мовців.

Квон та ін. [8] обрали групу просодичних та фонетичних ознак як вхідні дані для методу прямого відбору (FS) та зворотного виключення (BE), що використовуються для відбору ознак. Порівнюються різні алгоритми класифікації, включаючи метод опорних векторів (SVM), дискримінантний

аналіз (DA) та HMM. Однак ця система досягає лише 42,3% для розпізнавання емоцій для 5-ти емоцій.

Шуллер та ін. [9] пропонують підхід до поєднання акустичних ознак та мовної інформації для автоматичного розпізнавання емоцій в автомобільному середовищі. Отримані акустичні ознаки представлені статистикою висоти тону та енергії, тоді як мовна модель є стандартною прихованою марковською моделлю автоматичного розпізнавання мовлення (ASR). Для об'єднання двох типів мовних ознак задіяна багатошарова нейронна мережа сприйняття (MLP).

Верверідіс та ін. [10] використовують 87 мовленнєвих ознак з 500 емоційних висловлювань, зібраних від чотирьох суб'єктів. Послідовний прямий відбір використовується для виявлення значущих ознак з перехресною перевіркою як критерієм. Потім до вибраних ознак застосовується аналіз головних компонентів для подальшого зменшення розмірності. Для класифікації висловлювань на п'ять класів використовуються два класифікатори. Ця система досягає загальної точності лише 51,6%. Автори також надають підказки щодо того, що систему можна покращити, звівши класифікацію до проблеми двох класів, оскільки ознаки, які можуть розділяти два класи, можуть відрізнятися від тих, що розділяють п'ять класів.

1.2 Огляд методів розпізнавання емоцій за аналізом виразу обличчя

Щоб розпізнати людські емоції за зображенням обличчя або послідовністю зображень, спочатку слід визначити область обличчя. Потім інформацію про вираз обличчя можна витягти зі спостережуваного зображення обличчя або послідовності зображень. У випадку статичної картинки, вилучення інформації про вираз обличчя означає локалізацію обличчя та його рис на зображенні. У випадку послідовності зображень обличчя це означає відстеження руху обличчя та його рис у послідовності зображень. Риси обличчя зазвичай є помітними компонентами обличчя, такими як рот, очі, ніс, брови та підборіддя. Однак, ми також можемо використовувати риси моделі обличчя для представлення обличчя. Такий тип моделі може представляти обличчя як єдине ціле (цілісне представлення), як набір рис (аналітичне представлення) або як комбінацію цих двох (гібридний підхід). Застосоване представлення обличчя та тип вхідних зображень визначають вибір механізмів вилучення інформації про вираз обличчя [11].

Після того, як обличчя та його зовнішній вигляд були сприйняті, наступним кроком є визначення виразу обличчя, який воно передає. Це передбачає класифікацію виразів обличчя за набором заздалегідь визначених категорій, з якими ми хочемо мати справу. Хоча категоризація людських емоцій є досить складним питанням, більшість досліджень з аналізу виразів обличчя використовують емоційну категоризацію виразів обличчя Екмана [12], в якій визначено шість основних емоцій, а саме: щастя, смуток, здивування, страх, гнів та огида. На рисунку 2.2 зображено архітектуру системи аналізу виразів обличчя.

На рисунку 1.2 показано приклади емоційних даних.

Рисунок 1.2- Відеокліпи виразів обличчя

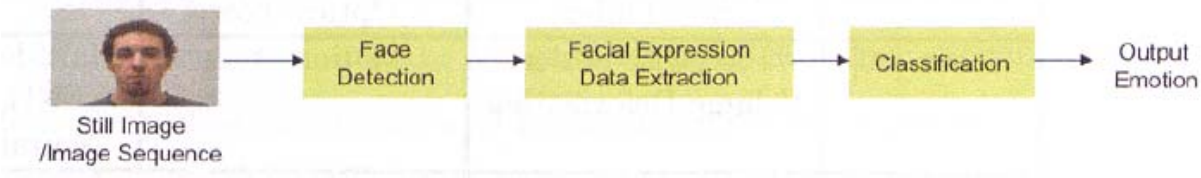

Рисунок 1.3- Архітектура системи аналізу міміки

Система кодування мімічних рухів (FACS), запропонована П. Екманом та В. В. Фрізеном [11] є одним із найпопулярніших представлень міміки. Воно

базується на переліку всіх одиниць дії (AU) обличчя, які викликають рухи обличчя. У FACS є загалом 46 AU, які пояснюють зміни у виразі обличчя. Поєднання цих одиниць дії призводить до великого набору можливих виразів обличчя. Навчений FACS-кодер розкладає спостережуваний вираз на конкретні AU, які викликали цей вираз. FACS кодується з відео, і код забезпечує точну специфікацію динаміки (тривалість, час початку та зміщення) руху обличчя, а також морфологію (конкретні дії обличчя, що відбуваються). Ця модель FACS широко використовується в дослідженнях розпізнавання виразів обличчя [13-15]. Відстежуючи риси обличчя та вимірюючи кількість рухів обличчя, можна класифікувати різні вирази обличчя.

Робота Екмана надихнула багатьох дослідників комп'ютерного зору на проведення аналізу виразів обличчя за допомогою обробки зображень та відео. Самал та Айєнгар [16] надають огляд ранніх робіт у 1992 році, тоді як Пантік та Роткранц дають вичерпний огляд робіт кінця 1990-х років [17]. У цьому розділі ми розглянемо

Ханеда та ін. [18] використовують експертну систему для класифікації виразів обличчя. Площа обличчя витягується на основі зображень елементів Hue та Quasi-chroma у модифікованій кольоровій системі HSV та інформації про рух об'єкта у відеокадрах. Витягнута область обличчя потім нормалізується за допомогою центральних точок очей, кінчика носа та двох кутових точок рота. Потім кількість ознак ока, рота та інших ділянок обличчя витягується за допомогою різних методів. Автоматична генерація знань досягається шляхом автоматичної побудови функцій належності. На основі побудованих функцій належності та шляхом об'єднання всієї інформації будуються набори правил для розпізнавання виразів обличчя. Ця система протестована на 4410 відеокадрах трьох осіб та досягає точності в середньому понад 90%.

Лайонс та ін. [27] використовують набір багатомасштабних, багатоорієнтаційних фільтрів Кабора для попереднього перетворення зображень. Потім сітка автоматично реєструється з обличчям за допомогою еластичної стрічки.

Метод зіставлення графів. Коефіцієнти Габора, вибіркові на сітці, об'єднуються в один вектор як ознаки. Для зменшення розмірності простору ознак застосовується аналіз головних компонентів (PCA). Вони тестують свою систему з базою даних із 193 зображень, зроблених 9 японськими жінками, і

досягають 75% точності класифікації виразів за допомогою лінійного дискримінантного аналізу (LDA).

Пантік та Роткранц [20] пропонують автоматичну систему розпізнавання рухів обличчя, що використовує статичне зображення з подвійним ракурсом. Обличчя розпізнається за допомогою методу сегментації вододілу з маркерами, в якому маркери витягуються на основі колірної моделі HSV. Для просторового відбору контуру профілю та контурів компонентів обличчя, таких як очі та рот, використовується багатодетекторний підхід до локалізації рис обличчя. Вони повідомляють про середній рівень розпізнавання 86% шляхом класифікації рухів обличчя в групу з 32 окремих рухів мімічних м'язів, що відбуваються разом або в поєднанні, використовуючи міркування на основі правил.

Сілва та Хуей [21] визначають положення ока та губ за допомогою низькочастотної фільтрації та

Методи виявлення країв. Для обчислення векторів локального руху рис обличчя застосовано методи підрахунку країв та кореляції зображень з оптичним потоком. За допомогою нейронної мережі було досягнуто середнього коефіцієнта розпізнавання емоцій 60%.

Коен та ін. [22] представляють та тестують різні класифікатори для розпізнавання виразів обличчя людини з відеопослідовностей. Використовується алгоритм відстеження обличчя під назвою Piecewise Bezier Volume Deformation tracker (PBVD), і 12 Motion-Units (MU) витягуються як основні ознаки для класифікації. Вони представляють багаторівневий HMM-класифікатор для динамічної класифікації, в якому також враховується часова інформація. Досліджуються два типи баєсівських мережевих класифікаторів: наївний баєсівський (NB) та деревоподібний наївний баєсівський (TAN) та нейронна мережа для виконання класифікації на одному кадрі. Незалежний від особи експеримент з використанням власної бази даних показує, що TAN-класифікатор забезпечує найкращий коефіцієнт правильного розпізнавання 66,53%.

Ма та Хорасані [23] використовують двовимірне (2-D) дискретне косинусне перетворення (DOT) для всього зображення обличчя як детектор ознак, а конструктивну одношарову нейронну мережу прямого зв'язку як класифікатор виразів обличчя. База даних містить зображення облич, на яких у сцені є лише обличчя. Експериментальні результати показують, що

конструктивна нейронна мережа перевершує метод векторного зіставлення та нейронну мережу.

Кім та Б'єн [24] використовують адаптивний метод визначення порогу на основі кольорової гістограми в колірному просторі на основі PCA для виявлення області обличчя. З компонентів обличчя витягуються п'ять ознак для представлення виразу обличчя, де компоненти обличчя витягуються за допомогою підходу «від крупних до дрібних деталей». Вони також застосували алгоритм на основі гістограми для вибору ознак. Ця система досягає точності 94,3% завдяки використанню нечіткої нейронної мережі для класифікації.

1.3 Розпізнавання емоцій з інтеграцією аудіо та візуальної інформації

Сільва та ін. [25] повідомляють про результати сприйняття людських емоцій людьми. Вони виявили, що деякі емоції (сум, страх) є сильно домінантними на слуху, деякі емоції (щастя, здивування та гнів) є сильно домінантними на візуальному рівні, тоді як деякі (огида) є змішано домінантними. Це означає, що інтеграція аудіо та візуальної інформації потенційно перевершить будь-яку з них окремо.

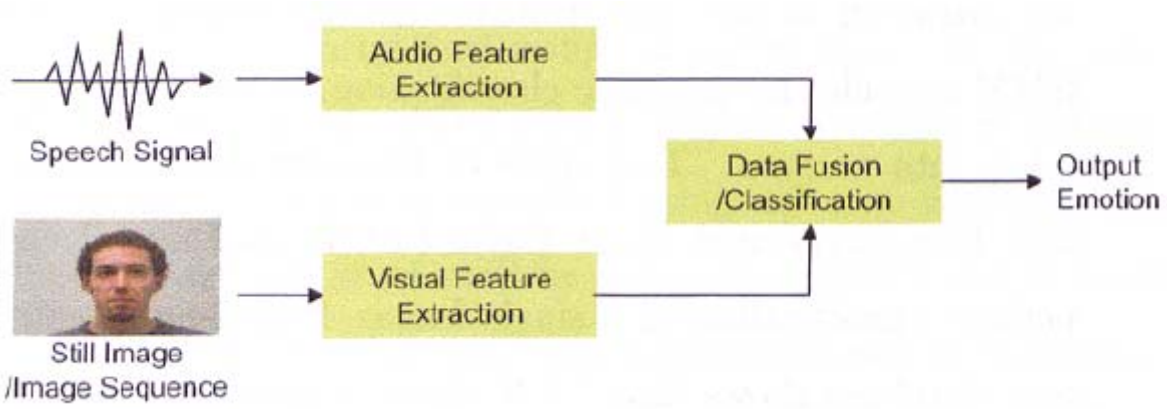

Рисунок 1.4- Архітектура бімодального розпізнавання емоцій

Архітектуру аудіовізуальної системи розпізнавання емоцій зображено на рисунку 1.4. Аудіо та візуальні ознаки витягуються з мови та зображень відповідно. Ці два набори даних потім об'єднуються в класифікатор для розпізнавання емоцій.

Сонг та ін. [26] пропонують аудіовізуальну систему розпізнавання емоцій, засновану на прихованій марковській моделі. Витягнуті аудіоознаки – це ознаки

висоти тону та енергії. Рух брів, повік та щоки витягуються як ознаки виразу, тоді як рух губ та щелепи – як ознаки візуального мовлення. Витягнуті трипотокові аудіовізуальні ознаки об'єднуються в потрійну HMM для класифікації. Цю систему було протестовано на базі даних із 684 зразків, і заявлена загальна точність становить 85%.

Сілва [27] окремо створюють аудіо- та відеосистеми. В аудіосистемі висота тону визначається як ознаки, а для класифікації використовується метод найближчих сусідів. У відеосистемі вони відстежують рух країв губ, куточків рота та брів за допомогою алгоритму оптичного потоку, тоді як прихована марковська модель навчається як класифікатор. Для об'єднання результатів класифікації аудіо та відео використовується система на основі правил. Результати розпізнавання показують, що бімодальний підхід перевершує окремі системи.

Чен та ін. [28] використовують статистику контуру тангажу, енергетичної обвідної та їх похідні для представлення характеристик емоційного мовлення, тоді як візуальна інформація отримується шляхом відстеження положення брів, підняття щік та відкриття рота. Аудіо- та відеоознаки просто об'єднуються в один вектор для класифікації. Вони порівняли два методи класифікації: метод найближчого середнього і гаусівська модель. Завдяки використанню перехресної перевірки з виключенням одного елемента, бімодальний підхід досягає 97,2%, що значно покращує коефіцієнт розпізнавання порівняно з аудіо (75%) та відео (69,4%).

Го та ін. [29] також демонструють перевагу бімодального розпізнавання емоцій над одномодальністю. У своїй системі вони застосовують аналіз головних компонентів (PCA) для класифікації ознак міміки, які витягуються за допомогою багатороздільного аналізу на основі дискретного вейвлет-перетворення (DWT). Для розпізнавання емоцій з мовного сигналу пропонується схема багаторівневого прийняття рішень для незалежного виконання класифікації на кожному вейвлет-піддіапазоні. Нарешті, методи розпізнавання обличчя та мовлення об'єднуються за допомогою схеми на основі правил, що забезпечує кращу продуктивність.

Таким чином продемонстровано, що аудіовізуальний підхід працює краще, ніж методи, засновані на одній модальності [26-29], завдяки тому, що передається більше інформації, а також сприяє взаємодоповнюючий зв'язок цих двох модальностей. Загалом, бімодальна система розпізнавання емоцій

складається з кількох компонентів: виділення аудіоознак, виділення візуальних ознак, вибір ознак та класифікація. У цьому розділі ми окремо розглянемо ці аспекти.

1.4 Огляд методів оцінювання аудіоознак

У галузі обробки аудіо існують два основні типи ознак: просодична ознака та фонетична ознака. Просодичні ознаки стосуються більш музичних аспектів мовлення, таких як висхідні або спадні тони та акценти або наголоси. Фонетичні ознаки стосуються типів звуків, що беруть участь у мовленні, таких як голосні та приголосні, а також їх вимова. Розуміння мовлення традиційно використовує лише фонетичні ознаки. У розпізнаванні людських емоцій на основі мовлення більшість робіт зосереджені на просодичних ознаках, які можна представити як статистичні характеристики висоти тону, енергії, нахилу та темпу мовлення тощо. Однак, як показано в [5-9], фонетичні характеристики, які відображають лінгвістичний фактор мовлення, також можуть сприяти розпізнаванню емоцій, і тому їх також слід враховувати.

Основним обмеженням існуючих систем є те, що сфера їхніх досліджень обмежена однією мовою. Однак спосіб, у який люди виражають свій емоційний стан у мові, може відрізнятися залежно від їхнього культурного походження, мови, акценту тощо. Для більш загального застосування система розпізнавання емоцій повинна бути здатною розпізнавати афективний стан людини незалежно від цих аспектів.

1.5 Огляд методів оцінювання візуальних ознак

Успішне та точне виявлення людського обличчя на сцені може значно зменшити шум та негативний вплив фону, і тому є дуже важливим для аналізу виразів обличчя. Впровадження схеми виявлення обличчя сильно залежить від типу вхідних зображень [30]. Більшість попередніх робіт використовують інформацію про краї [21] або колір [18,20] для виявлення обличчя. У певних програмах взаємодії людини з комп'ютером, якщо обличчя розглядається фронтально, а фон не дуже складний, інформація про колір є ефективним інструментом для ідентифікації областей обличчя та певних рис обличчя. Метод

сегментації на основі кольору має перевагу простоти та швидкості, що дуже важливо для реального застосування.

Після виявлення області обличчя, людські емоції можна розпізнати, досліджуючи рухи точок, що належать до рис обличчя, таких як рот, око, брова, а потім аналізуючи взаємозв'язки між цими рухами [14,22,26,28]. Цей метод, заснований на послідовності зображень, може досягти вищої точності, але обчислювальна складність є високою. Вилучення ключового кадру з послідовності зображень для представлення характеристик усього відео є варіантом досягнення компромісу між точністю та обчислювальною складністю. Аналіз розпізнавання емоцій зі статичних зображень [18-20] також демонструє ефективність розпізнавання на основі нерухомих зображень.

Щодо вилучення рис обличчя зі статичних зображень, як холістичного [19], так і аналітичного [18] підходи вивчалися в минулому. Цілісний підхід розглядає людське обличчя як єдине ціле, тоді як аналітичний підхід намагається проаналізувати основні риси людського обличчя, такі як очі, брови та рот. Хоча вираз обличчя людини може певною мірою відображатися кількома його компонентами, частина інформації також втрачається. Цілісний підхід потенційно краще відобразить характеристики людського обличчя при різних емоціях.

Вибір ознак є дуже важливим кроком у задачах розпізнавання образів. Зазвичай з джерела даних витягується велика кількість ознак. Але повне використання цих ознак може не досягти найкращих результатів. Великий простір ознак також спричиняє проблему прокляття розмірності, і система стає непридатною. Як зменшити розмір простору ознак без втрати занадто великої кількості інформації та отримати кращі результати, надихає на дослідження у сфері вибору ознак.

У минулому досліджувалися різні алгоритми зменшення розмірності та вибору ознак [31-32]. Критерієм вибору алгоритму вибору ознак є досягнення компромісу між ефективністю та обчислювальною складністю. Субоптимальні методи, такі як прямий відбір та зворотне виключення, є хорошим компромісом між цими двома аспектами, і тому часто використовуються в цій галузі [4, 8,10].

1.5 Огляд методів класифікації

Алгоритми класифікації можна розділити на дві групи: контрольовані та неконтрольовані, тоді як контрольовані методи можна додатково розділити на лінійні та нелінійні класифікатори. Вибір алгоритму класифікації залежить від проблеми. Для розпізнавання людських емоцій ми намагаємося класифікувати людські емоції на кілька заздалегідь визначених базових емоцій, навчаючи машину за допомогою зразків, відомих класам. Це контрольований випадок.

Зазвичай ми не маємо повного уявлення про дані, які будуть застосовані до класифікації, і тому потрібно порівнювати різні алгоритми класифікації. Різні

У літературі використовуються такі алгоритми, як k-найближчих сусідів, дискримінантний аналіз, нейронна мережа, прихована марковська модель, метод опорних векторів, баєсівський класифікатор. В аудіовізуальному розпізнаванні емоцій деякі роботи використовують схему на основі правил для виконання об'єднання аудіо- та візуальної класифікації [27, 29]. Цей метод на основі правил зазвичай базується на апріорних знаннях про аудіо- та відеофактори в системі.

У більшості попередніх робіт для класифікації багатокласових задач застосовувався один класифікатор. Однак, суттєві ознаки, вибрані в двокласовій задачі, можуть відрізнятися від тих, що вибрані в багатокласовому сценарії [10]. Окремий клас також має різні ознаки, які відрізняють його від інших класів. У цій роботі спочатку порівнюємо продуктивність деяких алгоритмів класифікації. На основі експериментальних результатів створюється багатокласифікаторна схема для аналізу цих сценаріїв. Цей метод декомпозиції допомагає досягти вищої точності розпізнавання.

2 ОЦІНЮВАННЯ МОВЛЕНЕВИХ ОЗНАК

Люди здатні розпізнавати емоції інших людей, слухаючи їхні голоси. Хоча в усьому світі використовуються різні мови та акценти.

Спосіб, яким люди виражають свої емоції мовою, залежить від їхнього культурного походження, особистості, віку, статі тощо. У більшості випадків ми можемо сприймати почуття інших людей. Для більш природних застосувань людської інтелектуальної взаємодії нам потрібно надати комп'ютерам ті ж можливості, що й людині. Як один з основних показників афективного стану людини, мова відіграє важливу роль у машинному розпізнаванні людських емоцій.

Для створення більш загальної системи розпізнавання емоцій вилучення ознак, які можуть дійсно представляти універсальні характеристики запланованої емоції, є справжнім викликом. Гарною еталонною моделлю є система людського слуху. У попередніх роботах досліджувалося кілька різних типів ознак. Оскільки просодія вважається основним показником емоційного стану мовця [40], більшість досліджень використовують просодічні ознаки. Однак, кепстральний коефіцієнт Mel-частоти (MFCC) та формантна частота також широко використовуються в розпізнаванні мовлення та деяких інших програмах обробки мовлення. Оскільки наша мета — імітувати сприйняття емоцій людиною та визначити можливі ознаки, які можуть передавати основні емоції в мовленні незалежно від мови, мовця та контексту, ми також досліджуємо ці два типи ознак.

На рисунку 2.1 показано блок-схему системи вилучення аудіоознак, що використовується в цій роботі. Вона складається з п'яти компонентів. Етап попередньої обробки виконує шум зменшення та усунення паузи. Потім попередньо оброблений сигнал проходить через процес віконної сегментації для сегментації вихідних сигналів на короткі часові мовні кадри. Просодичні, MFCC та формантні частотні характеристики потім витягуються окремо на основі короткочасного спектрального аналізу. У цьому розділі ми детально обговоримо кожен із компонентів.

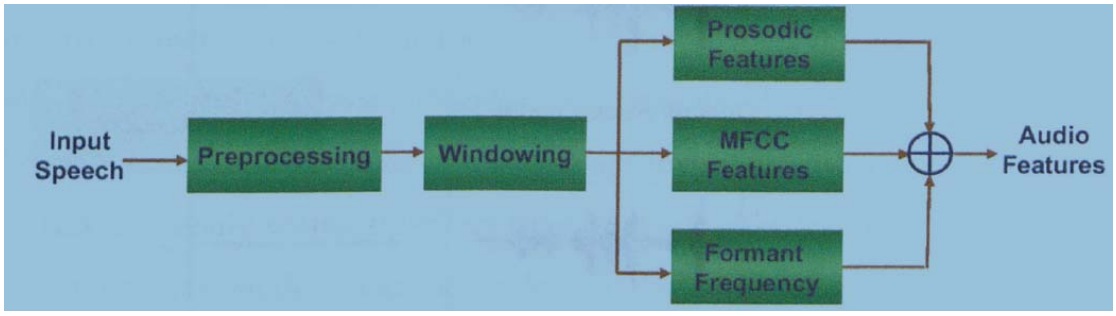

Рисунок 2.1- Система вилучення аудіоознак

2.2 Попередня обробка

Емоційні дані записуються у відносно спокійному середовищі. Однак «шипіння» записуючого пристрою та фоновий шум можуть суттєво впливати на вилучення ознак з мовного сигналу, а отже, на продуктивність системи розпізнавання. Крім того, передній та задній фронти, що утворюються в процесі запису, не надають жодної корисної інформації та є повністю зайвими. Щоб зменшити вплив шуму та тиші на мовленнєве висловлювання, ми виконуємо зменшення шуму та усунення переднього та заднього фронтів на етапі попередньої обробки. На рисунку 3.2 показано процес попередньої обробки.

Наявні «шипіння» записуючого пристрою та шум зменшуються шляхом порогового регулювання з застосуванням вейвлет-коефіцієнтів [33]. Традиційні методи низькочастотної фільтрації, які є лінійно-інваріантними в часі, можуть розмити різкі особливості сигналу, і іноді важко відокремити шум від сигналу там, де їхні спектри Фур'є перетинаються. Вейвлет-метод має перевагу в ефективному зменшенні шуму без спотворення вихідного сигналу.

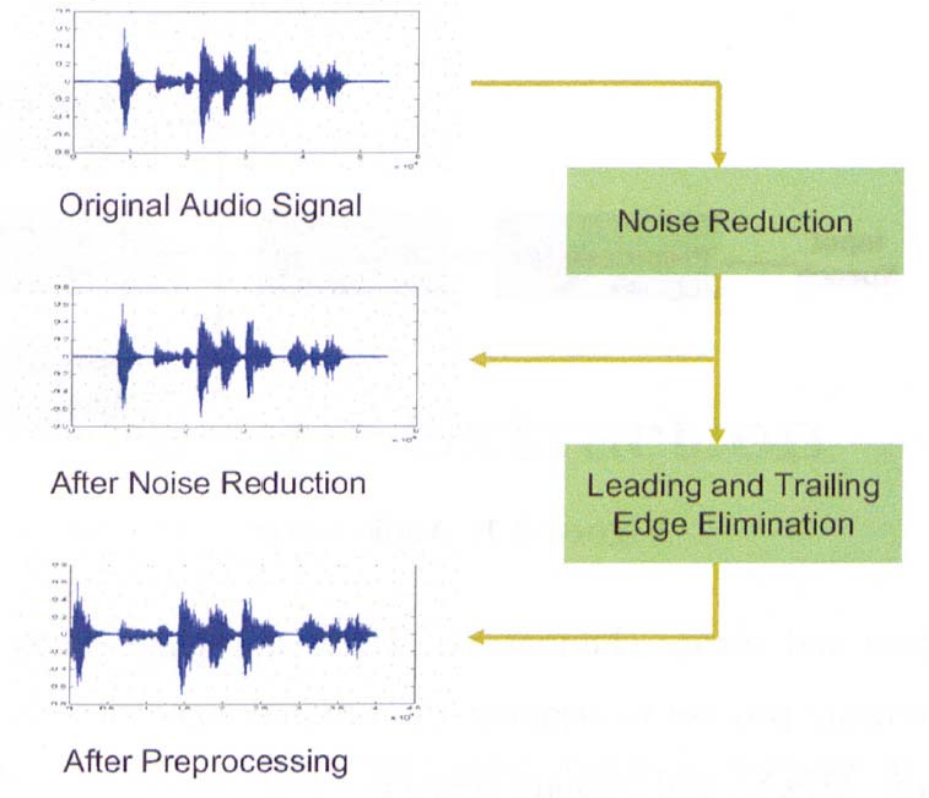

Рисунок 2.2- Послідовність етапу попередньої обробки

Для вейвлетів, які є нелінійними функціями, амплітуда, а не розташування, спектрів Фур'є відрізняється від амплітуди шуму. Це дозволяє використовувати порогове регулювання коефіцієнтів вейвлету для видалення шуму. Якщо енергія сигналу зосереджена в невеликій кількості коефіцієнтів вейвлету, значення будуть великими порівняно з шумом, енергія якого розподілена на велику кількість коефіцієнтів. Ці локалізуючі властивості вейвлет-перетворення дозволяють дуже ефективно фільтрувати шум із сигналу. У той час як лінійні методи жертвують придушенням шуму для розширення особливостей сигналу, зменшення шуму за допомогою вейвлетів дозволяє залишатися незмінними особливостям вихідного сигналу.

Щоб виключити періоди мовчання, які не впливають на людські емоції, ми виконуємо усунення переднього та заднього фронтів сигналу зі зниженим шумом. Ми оцінюємо максимальну амплітуду шуму за заздалегідь визначений короткий проміжок часу. Поріг розраховується шляхом додавання запасу шуму до максимальної амплітуди. Значення запасу оцінюється на основі експериментів. Потім ми виключаємо передній та задній фронти, видаляючи ті

частини, де амплітуда нижче порогу. У цій роботі заздалегідь визначений період мовчання становить 250 мс, а як значення запасу вибрано значення 0,01.

2.3 Застосування вікон

Щоб виділити ознаки з емоційного мовного сигналу, ми виконуємо спектральний аналіз. Метод спектрального аналізу є надійним лише тоді, коли сигнал стаціонарний, тобто статистичні характеристики сигналу незмінні відносно часу. Для мовлення це справедливо лише протягом коротких часових інтервалів артикуляційної стабільності, протягом яких можна виконати короткочасний аналіз шляхом розбиття сигналу x(n) на послідовність віконних послідовностей,які називаються кадрами. Ці мовленнєві кадри потім можна обробляти окремо.

Тоді

$$x_t(n) = w(n)x'_t(n), \quad n = 0,..., N-1, \;\; t = 0,...T-1, \tag{2.1}$$

де w(n) – імпульсна характеристика вікна, N – розмір вікна, T – кількість кадрів, а $x'_t(n)$ – кадр до застосування віконної функції. Прийнятим вікном є вікно Хеммінга (рис. 3.3), імпульсна характеристика якого w(n) є піднятим косинусоїдальним імпульсом

$$w(n) = 0{,}54 - 0{,}46\cos\left(\frac{2\pi n}{N-1}\right), \; n = 0,..., N-1 \tag{2.2}$$

Порівняно з прямокутною формою вікна, вікно Хеммінга має перевагу у зменшенні ефекту витоку [42]. Розмір вікна визначається шляхом врахування компромісу між часовою та частотною роздільною здатністю. Крім того, для згладжування переходу та усунення можливих проміжків між блоками зазвичай використовуються вікна, що перекриваються. Спектральний аналіз виконується для мовних кадрів з 512 точок з 50% перекриттям.

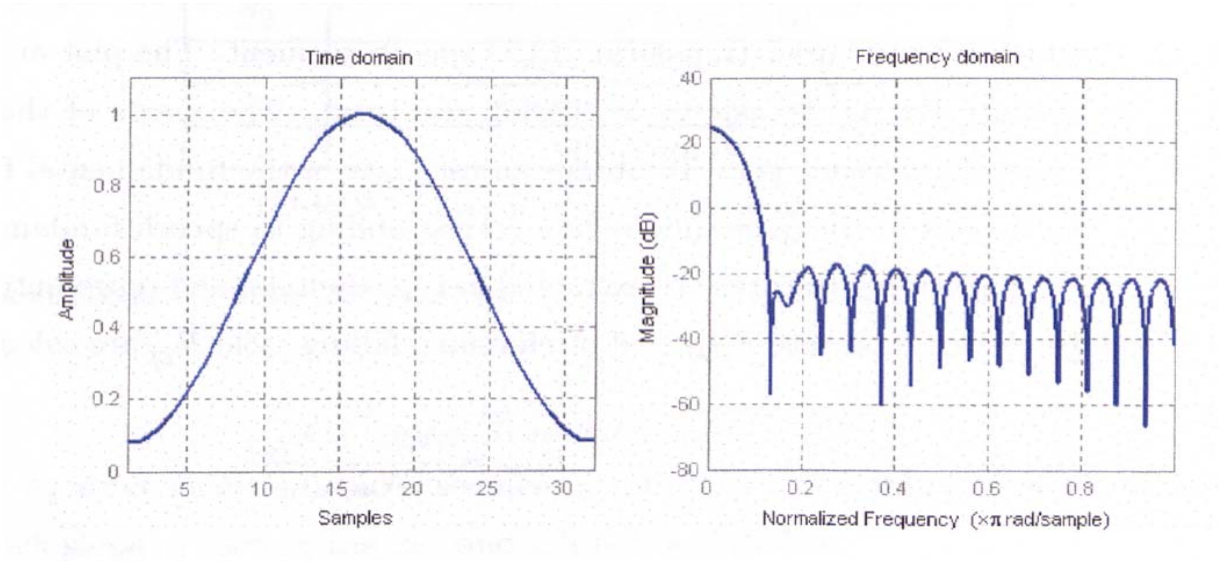

Рисунок 2.3- Вікно Хемінга (32 точки). Ліворуч: часова область, праворуч: частотна область

2.4 Просодичні особливості

Просодія головним чином пов'язана з ритмічними аспектами мовлення та зазвичай представлена статистикою та варіаціями основної частоти, інтенсивності, темпу мовлення тощо. Раціонально використовувати 25 просодичних ознак, перелічених у таблиці 2.1.

Висота тону оцінюється на основі Фур'є-аналізу логарифмічного амплітудного спектру сигналу [43] [44]. Взявши логарифм частотного спектру, низькочастотні компоненти спектра перебільшуються. Якщо логарифмічний амплітудний спектр містить багато регулярно розташованих гармонік, то Фур'є-аналіз спектру покаже пік, що відповідає інтервалу між гармоніками: тобто основній частоті. Цей метод продемонстровано на рисунку 3.4. Верхній графік - це форма хвилі сигналу в часовій області. Графік посередині показує спектр, який є перетворенням Фур'є сегмента мовлення. Графік внизу - це дискретне перетворення Фур'є логарифмічного спектру. Вісь x нижнього графіка має одиниці частоти [44]. Щоб отримати оцінку основної частоти, ми шукаємо пік в області частоти, що відповідає основним частотам мовлення. Енергетичні характеристики витягуються в часовій області та представлені в децибелах (дБ).

Таблиця 2.1-Список вилучених просодичних ознак

Індекс	Опис функцій
1	Середній тон
2	Медіана кроку
3	Стандартне відхилення висоти тону
4	Макс. висота тону
5	Діапазон висоти звуку
6	Коефіцієнт зміни висоти тону
7	Співвідношення висоти/спаду висоти тону
8	Зростаючий максимальний нахил
9	Падаючий кут нахилу Макс.
10	Зростаючий кут нахилу. Середнє значення
11	Середнє значення падіння нахилу
12	Максимальний діапазон зростання висоти тону
13	Максимальний діапазон падіння висоти тону
14	Середнє значення діапазону зростання висоти тону
15	Середнє значення діапазону падіння висоти тону
16	Загальний середній нахил тангажу
17	Загальне стандартне відхилення нахилу тангажу
18	Загальний середній нахил тангажу
19	Середнє значення енергії (дБ)
20	Медіана енергії (дБ)
21	Стандартне відхилення енергії (дБ)
22	Макс. енергія (дБ)
23	Діапазон енергії 23 (дБ)
24	Середня тривалість паузи
25	Швидкість виступів

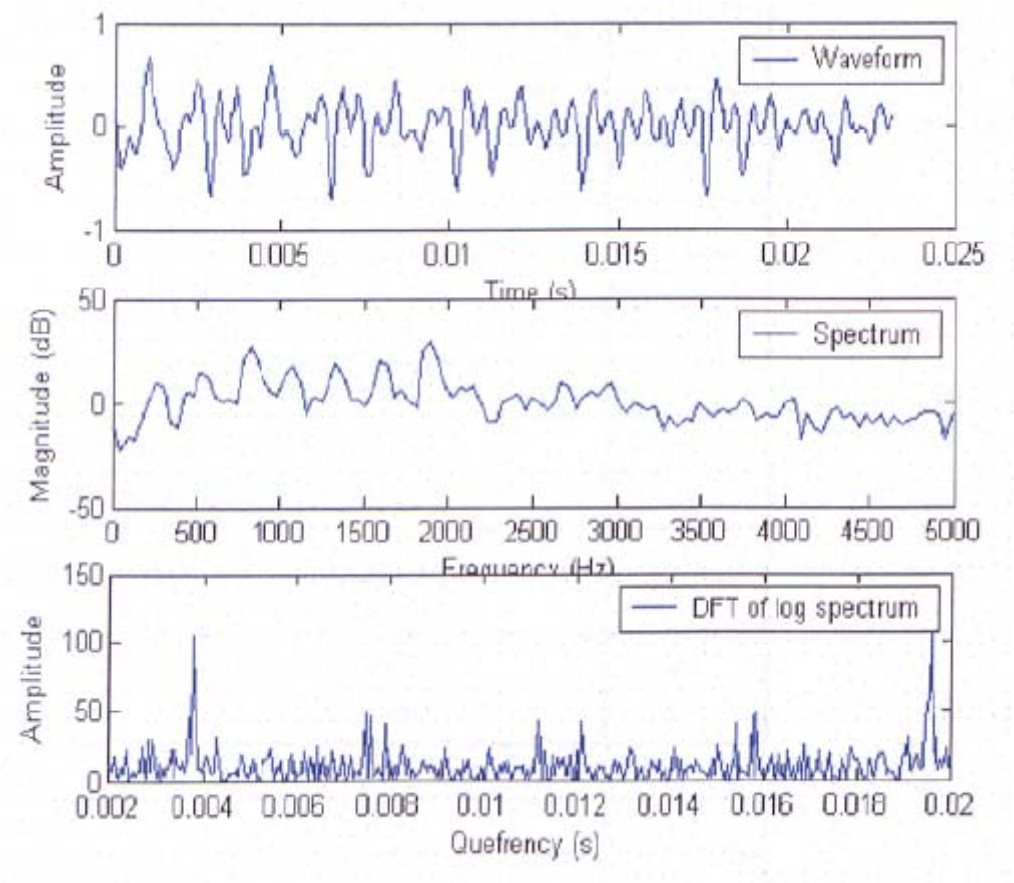

Рисунок 2.4-Оцінка висоти тону. Вгорі: форма хвилі в часовій області, Посередині: спектр амплітуди, Внизу: ДПФ логарифмічного спектру амплітуди

Паузи не використовуються для розрахунку інших параметрів, тому вони відкидаються. Діапазон висоти тону (амплітуди) визначається шляхом сканування кривої, знаходження максимального та мінімального висоти тону (амплітуди) та обчислення різниці.

Середні значення та стандартне відхилення обчислюються відповідно за формулою:

$$\bar{x} = \sum_{i=1}^{N} x_i p(x_i)$$

$$\sigma = \sqrt{\sum_{i=0}^{N} (x_i - \bar{x})^2 p(x_i)}, \qquad (2.3)$$

де N — діапазон x_i, ймовірність появи значень x_i обчислюється за допомогою нормалізованої гістограми.

Медіанні значення потім обчислюються шляхом знаходження значення x_i такого, що:

$$\sum_i p(x < x_i) \approx 0{,}5..$$

Мел-частотний кепстральний коефіцієнт (MFCC), мабуть, є найпопулярнішим рішенням у галузі розпізнавання мовлення, ідентифікації тощо. Метою MFCC є імітація поведінки людського вуха шляхом застосування кепстрального аналізу. Оскільки наша мета — визначити можливі акустичні особливості, які можуть сприяти розпізнаванню людських емоцій, ми також досліджуємо цей тип ознак.

Розрахунок MFCC для спекельного сигналу складається з попередньої обробки, віконної обробки, а потім перетворення Фур'є, Mel-масштабування та оберненого косинусного перетворення для кожного часового кадру. Перед оберненим перетворенням величина спектру робиться логарифмічною. Ця логарифмічна шкала є характеристикою людського слухового апарату. На рисунку 3.5 показано послідовність дій процедури вилучення ознак MFCC.

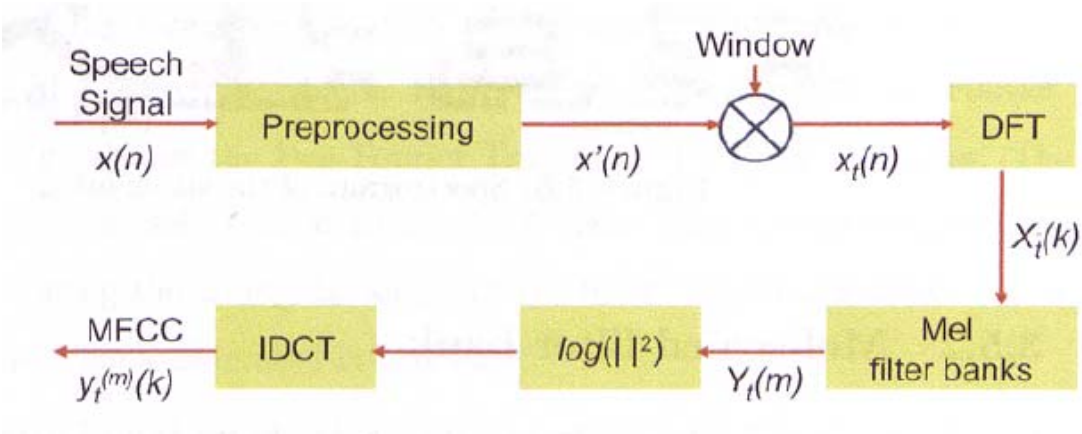

Рисунок 2.5- Потік обробки MFCC

2.5 Спектральний аналіз

Тут також застосовується процес попередньої обробки та взяття вікон. Для виконання спектрального аналізу мовні сигнали необхідно перетворити в частотну область. Це робиться за допомогою дискретного перетворення Фур'є.

На рисунку 2.6 показано спектрограми та пов'язані з ними форми хвиль шести емоцій, сформованих одним із суб'єктів експерименту. На спектрограмі час представлений вздовж горизонтальної осі, тоді як частота відкладена вздовж вертикальної осі.

Можна помітити, що кожен клас емоцій демонструє різні закономірності.

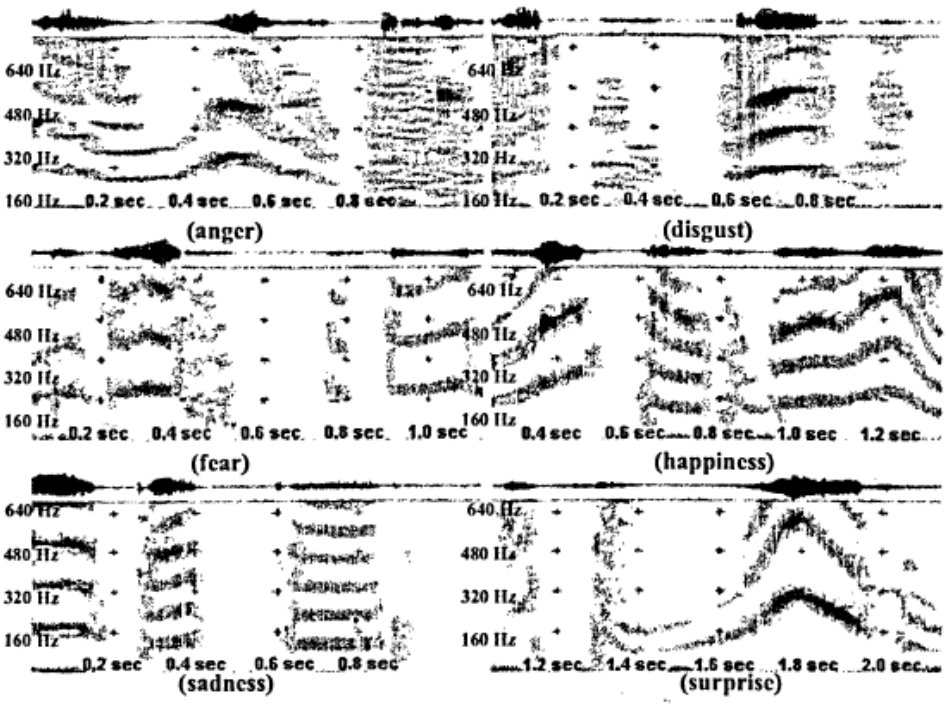

Рисунок 2.6- Спектрограма шести емоцій

2.6 Банк фільтрів з масштабуванням за шкалою Мел

Щоб спростити спектр без значної втрати даних, сигнал, перетворений Фур'є, зазвичай пропускається через набір смугових фільтрів, які належним чином інтегрують спектр у визначених діапазонах частот. У мовному сигналі більшість важливої та корисної інформації розташована в нижній смузі частот. Шкала Мела є найширше використовуваною перцептивною шкалою, яка призначена для захоплення та підкреслення інформації в низькочастотній смузі. Банк фільтрів зазвичай складається з трикутних фільтрів з перекриттям частот, так що центральна частота фільтра відповідає верхній частоті попереднього фільтра та нижчій частоті наступного фільтра. Центральна частота кожного банку фільтрів Мела рівномірно розташована до 1 кГц і дотримується логарифмічної шкали після 1 кГц. Крім того, щоб підкреслити низькочастотні

компоненти, величина фільтра зазвичай встановлюється на 1 у низькочастотній смузі, зменшуючи її зі збільшенням частоти.

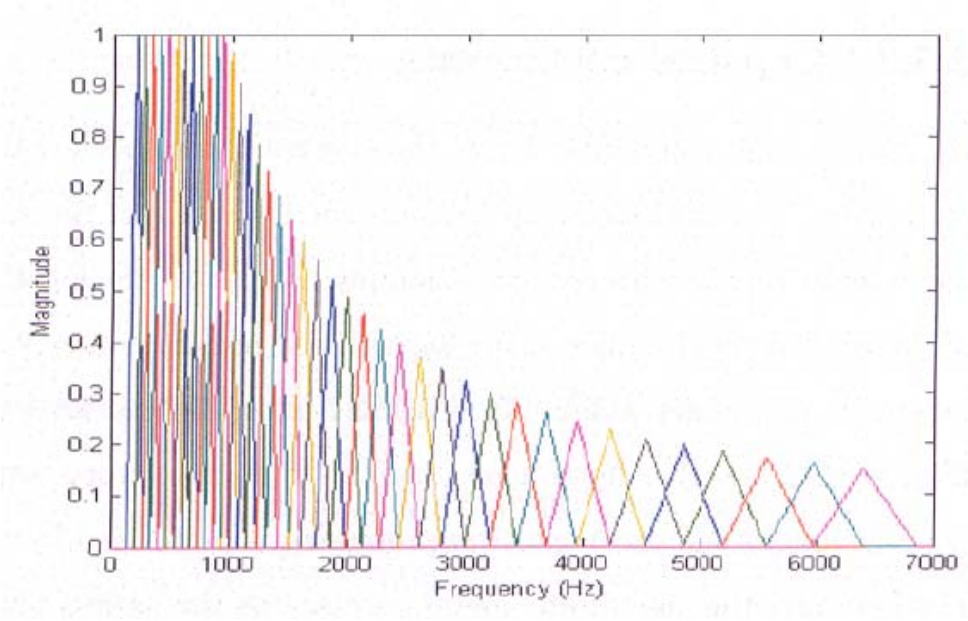

Рисунок 2.7- Проектування банку фільтрів з масштабуванням за Мелом

Зазвичай діапазон частоти, що охоплюється банком фільтрів, лежить від 20 Гц до половини частоти дискретизації сигналу. На рисунку 3.7 показано діаграму ідеального банку фільтрів з масштабуванням за Мелом. Однак, коли ми обчислюємо перетворення Фур'є, ми зазвичай виконуємо швидке перетворення Фур'є (ШПФ) за допомогою комп'ютера. Роздільна здатність за частотою тісно пов'язана з розміром ШПФ. Це призводить до обчислювальної похибки в апроксимації трикутної форми фільтра. На рисунку 3.8 показано розташування банку фільтрів, згенерованого в цій роботі.

Частота («частота дискретизації/розмір FFT»)

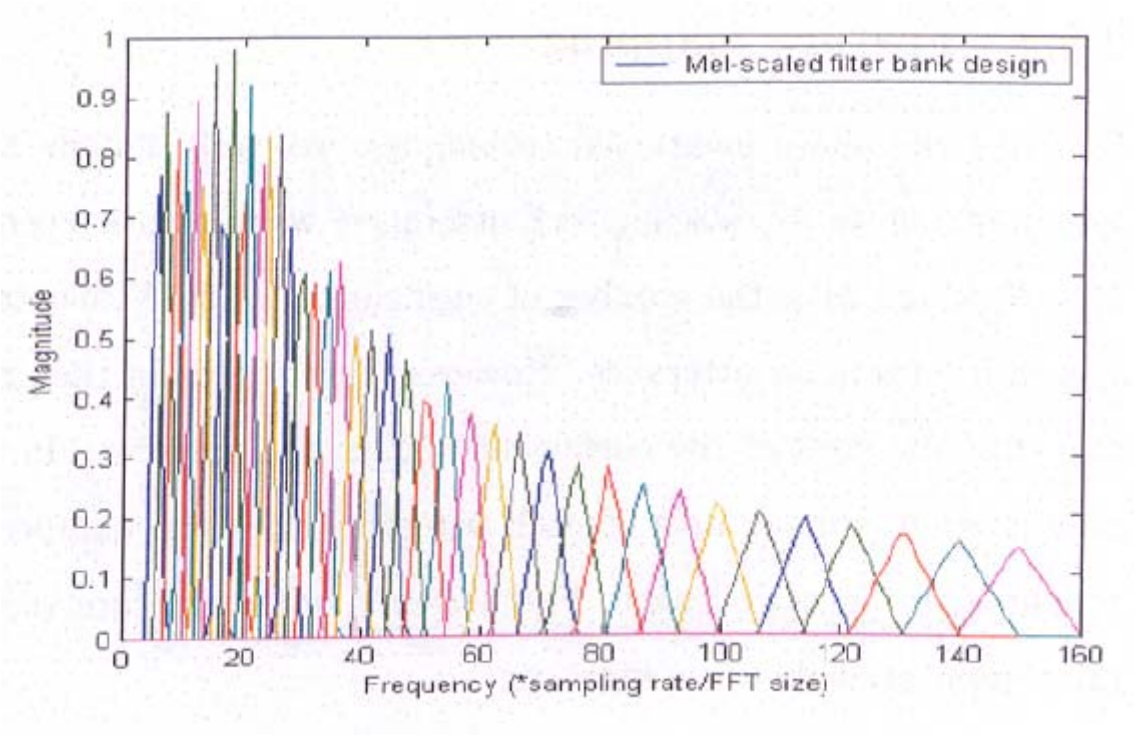

Рисунок 2.8- Банк фільтрів масштабованого за шкалою Mel

2.7 Кепстральні коефіцієнти

Використовуючи банк фільтрів з масштабуванням Mel, спектр згладжується подібно до людського вуха. Наступним кроком є обчислення логарифма квадрата величини коефіцієнтів. Це зводиться до простого обчислення логарифма величини коефіцієнтів завдяки алгебраїчній властивості логарифмування, яка повертає логарифм степеня до множення на коефіцієнт масштабування. Взявши логарифм коефіцієнтів фільтра, можна змоделювати характеристики слухової системи людини, оскільки обробка величини та логарифмування також виконується людським вухом. Крім того, операція з величиною відкидає непотрібну фазову інформацію, тоді як логарифм виконує динамічне стиснення, що робить вилучення ознак менш чутливим до змін динаміки.

MFCC – це обернене дискретне косинусне перетворення логарифма величини вихідного сигналу банку фільтрів:

$$y_t^{(m)}(k) = \sum \log\{|Y_t(m)|\} \cdot \cos\left(k\left(m - \frac{1}{2}\right)\frac{\pi}{M}\right) \quad (2.4)$$

де m – індекс фільтра, а M – загальна кількість фільтрів.

2.8 Відображення ознак

Використовуючи вищезгадані методи, ми можемо обчислити MFCC (коефіцієнти кадрів збереження) для кожного мовленнєвого висловлювання. Для кожного мовленнєвого висловлювання ми обчислюємо матрицю коефіцієнтів розміром MX, де M – кількість коефіцієнтів, а N – загальну кількість мовленнєвих кадрів у висловлюванні. Однак довжини висловлювань різняться, а отже, і розміри матриці коефіцієнтів різні. Для полегшення класифікації ознаки кожного висловлювання, що відображаються в просторі ознак, повинні мати однакову довжину. Крім того, при векторі ознак високої розмірності обчислювальні витрати високі.

Зазвичай, у розпізнаванні мовлення загальна кількість коефіцієнтів, що використовуються, становить від дев'яти до тринадцяти. Це пояснюється тим, що більша частина енергії сигналу стискається в перші кілька коефіцієнтів через властивості косинусного перетворення. У цій роботі ми беремо перші тринадцять коефіцієнтів як корисні дані. Потім ми обчислюємо середнє значення, медіану, стандартне відхилення, максимум і мінімум кожного порядку m=l,2,...,13 як витягнуті ознаки. Таким чином, ми витягли загалом 65 ознак MFCC.

2.9 Характеристики формантної частоти

Формантні частоти – це властивості системи голосового тракту. Їх потрібно виводити з мовного сигналу, а не вимірювати. Спектральна форма збудження голосового тракту сильно впливає на спостережувану обвідну, і тому ми не можемо гарантувати, що всі резонанси голосового тракту

спричинятимуть піки у спостережуваній спектральній обвідній, а також що всі піки у спектральній обвідній спричинені резонансами голосового тракту.

Оцінка формантної частоти базується на моделюванні мовного сигналу так, ніби він генерується певним типом джерела та фільтра [43]. Щоб знайти найкращу систему узгодження, ми використовуємо метод лінійного прогнозування. Лінійне прогнозування моделює сигнал так, ніби він генерується сигналом мінімальної енергії, що проходить через чисто рекурсивний фільтр з нескінченною частотною характеристикою (HR). На рисунку 3.9 показано графік частотної характеристики фільтра за допомогою лінійного прогнозного кодування (LPC).

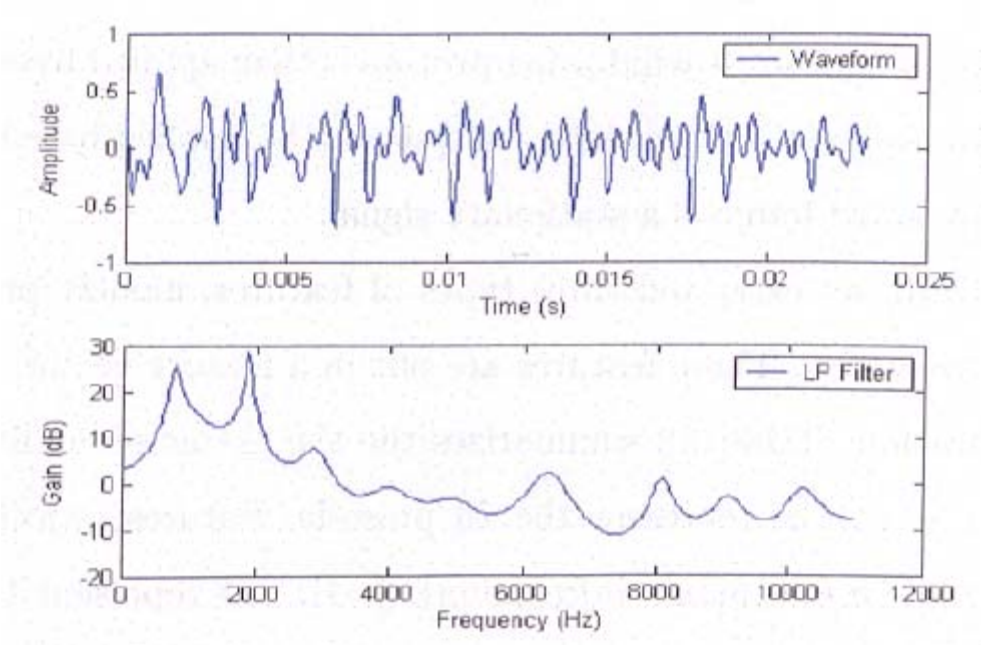

Рисунок 2.9-Оцінка формантної частоти. Вгорі: форма хвилі в часовій області, Внизу: частотна характеристика HR-фільтра з використанням LPC

Щоб знайти формантні частоти з фільтра, нам потрібно знайти розташування резонансів, з яких складається фільтр. Це можна зробити, розглядаючи коефіцієнти фільтра як поліном та розв'язуючи задачу щодо коренів полінома. Щоб зробити розмір ознак однорідним та досягти компромісу між ефективністю імітації системи голосового тракту та розмірністю простору ознак, ми беремо середнє значення, медіану, стандартне відхилення, максимум та мінімум перших трьох формантних частот як витягнуті ознаки.

Таким чином, продуктивність системи розпізнавання людських емоцій значною мірою залежить від точності та ефективності вилучених ознак. Як один з основних аспектів запропонованої нами системи, точна ідентифікація мовленнєвих ознак має вирішальний вплив на швидкість розпізнавання всієї системи. Оскільки ми не маємо попередніх знань про значення окремих мовленнєвих ознак, ефективності вилучення ознак можна краще досягти, вилучивши велику кількість різних типів ознак та виконавши вибір ознак.

Таким чином, представлено методи, які використовувались для вилучення аудіоознак з мовленнєвого висловлювання, щоб відобразити емоційне мовлення на відповідний простір ознак. Щоб зменшити негативний вплив шуму та непотрібної тиші, спочатку виконуємо попередню обробку. Потім застосовується процес відвікнування для сегментації мовлення на короткі часові кадри, за допомогою якого можна застосувати спектральний аналіз, виходячи з припущення, що сегментований кадр є стаціонарним сигналом.

Загалом, ми виділили три типи ознак, а саме: просодичний, MFCC та формантну частоту. Ці ознаки поміщені у вектор ознак як аудіоознаки для класифікації.

У таблиці 2.2 підсумовано вилучені аудіоознаки, де p{i), i = 1,...,25 представляють 25 просодичних ознак.

Таблиця 2.2- Просодичні ознаки

Ознаки	Розмірність	Позначки
Просодична	25	[p(I), i =1,...25]
MFCC	65	[mfcc (j, mean), mfcc (j, median), mfcc (j,std), mfcc (j, ymax),mfcc(j,min), j=l,...,13]
Форманти	15	[ff (k,mean),ff (k, median), ff (k,std), ff(k,max),ff(k,min), k=l,2,3]
Overall Audio	105	[p(l),...,p(25),mfcc(1,mean),...$_t$ mfcc(13, min),ff(1, mean),. ff(3,min]

3 ОЦІНЮВАННЯ ВІЗУАЛЬНИХ ОЗНАК

Вираз обличчя є ще одним важливим фактором розпізнавання людських емоцій. Обличчя є одним з основних каналів виведення, за допомогою яких людина виражає емоції.

стан. У минулому досліджувалися різні методи аналізу виразів обличчя, які можна умовно розділити на дві групи. Одна полягає у трактуванні людського обличчя як цілісного цілого, а інша – у представленні обличчя за допомогою визначних компонентів, таких як рот, очі, ніс, брови та підборіддя. Аналіз компонентів обличчя призведе до втрати певної інформації, тому ми виконуємо аналіз обличчя, розглядаючи факт у глобальному сценарії.

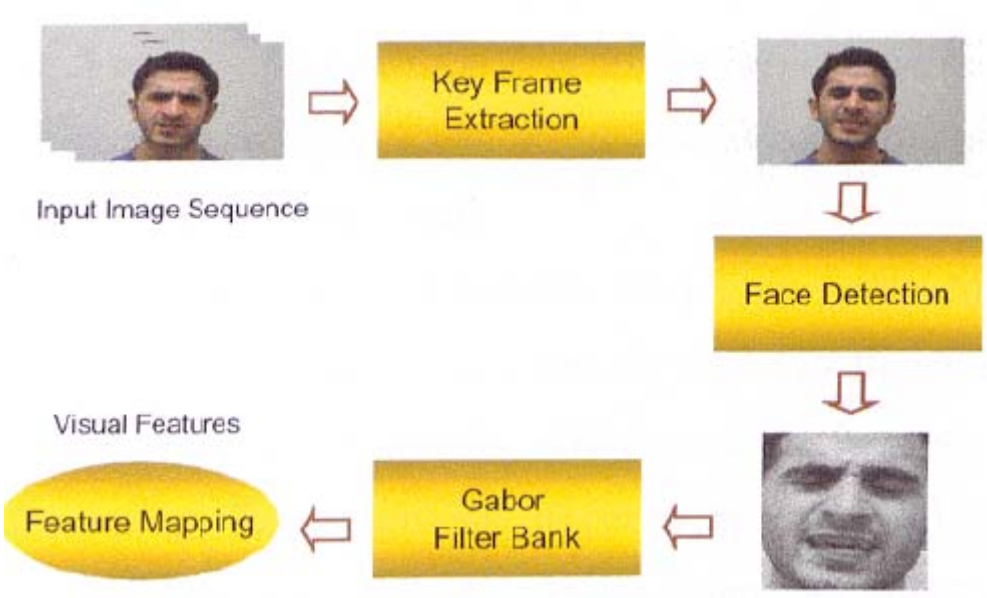

Рисунок 3.1- Послідовність вилучення візуальних ознак

На рисунку 3.1 показано, як відбувається вилучення візуальних ознак. Щоб пришвидшити час обробки, ми використовуємо ключовий кадр для представлення емоційного стану експериментального суб'єкта у відеокліпі. Ключовий кадр - це кадр, у якому відповідна мова має найвищу амплітуду. Потім для виявлення обличчя на фоні використовується схема розпізнавання обличчя на основі колірної моделі HSV. Вирази обличчя представлені вейвлет-ознаками Габора.

3.2 Розпізнавання обличчя

У минулому вивчалися різні підходи до розпізнавання облич. Прикладами цих підходів є аналіз головних компонентів, аналіз кольору шкіри, перетворення Хафа тощо. Серед цих різних методів багатообіцяючим виявився колірний аналіз. Однією з головних переваг колірного аналізу є те, що швидкість обробки інформації про колір набагато вища, ніж у інших методів. Ця характеристика дуже важлива для взаємодії людини з комп'ютером у реальному часі.

Розглянемо схему розпізнавання облич на основі колірної моделі HSV. У колірній моделі HSV компоненти H та S містять хроматичну інформацію, а V представляє інформацію про яскравість. Колірна модель HSV точно відповідає людській інтуїції щодо кольору.

Відтінок – це характеристика кольору, яка вимірюється кутом навколо вертикальної осі. Він виглядає як колірне коло, яке позначається червоним, жовтим, зеленим, блакитним, синім, пурпуровим тощо. Насиченість використовується для опису того, наскільки чистим є колір або скільки білого додано до нього. Яскравість – це міра відносної яскравості.

Компоненти RGB зображення можна перетворити на колірний простір HSV за допомогою такої формули:

$$\mathrm{H} = \{\mathrm{H}_1 \qquad if B \le G; \qquad 360^\circ - H_1 \qquad if B > G\},$$

$$H_1 = \cos^{-1}\left\{\frac{0.5[(R-G)+(R-B)]}{\sqrt{(R-G)(R-G)+(R-B)(G-B)}}\right\}, \qquad (3.1)$$

$$S = \frac{\max(R,G,B) - \min(R,G,B)}{\max(R,G,B)},$$

$$V = \frac{\max(R,G,B)}{256}$$

Використовуємо метод апроксимації планарної обвідної [46] для апроксимації кольору шкіри людини.

У методі планарної обвідної піксель вважається пікселем шкіри, якщо колір пікселя задовольняє наступні дві умови:

$$S \geq Yh_S;\ V \geq Th_V;\ S \leq -H - 0.1V + 110;\ H \leq -0.4V + 75,$$

$$if\quad H \geq 0,\quad S \leq 0{,}08(100 - V)H + 0.5V \qquad else\quad S \leq 0.5H + 35,$$

де Th_S та Th_υ встановлені на 10 та 40 відповідно [34].

Після застосування сегментації шкіри в результаті неминуче спостерігаються деякі ділянки, що не належать до шкіри, такі як невеликі ізольовані плями та вузькі смужки, оскільки їхній колір потрапляє в колірний простір шкіри. Збереження цих штучних ділянок шкіри призведе до негативних наслідків для останньої обробки. Ми застосовуємо морфологічну операцію для реалізації процедури очищення. Спочатку виконується операція закриття, щоб з'єднати вузькі проміжки між ділянками шкіри, а потім застосовується операція відкриття, щоб видалити невеликі ізольовані цибулини та розділити області, з'єднані тонкими смужками. Нарешті, ми виконуємо операцію заповнення, щоб видалити чорні ізольовані отвори.

На рисунку 3.2 показано процедуру та операції застосованої схеми розпізнавання обличчя. Виявлена область обличчя відображається назад на вихідне зображення та нормалізується до зображення на рівні сірого розміром 128 x 128 як вхідний сигнал для банку фільтрів Габора.

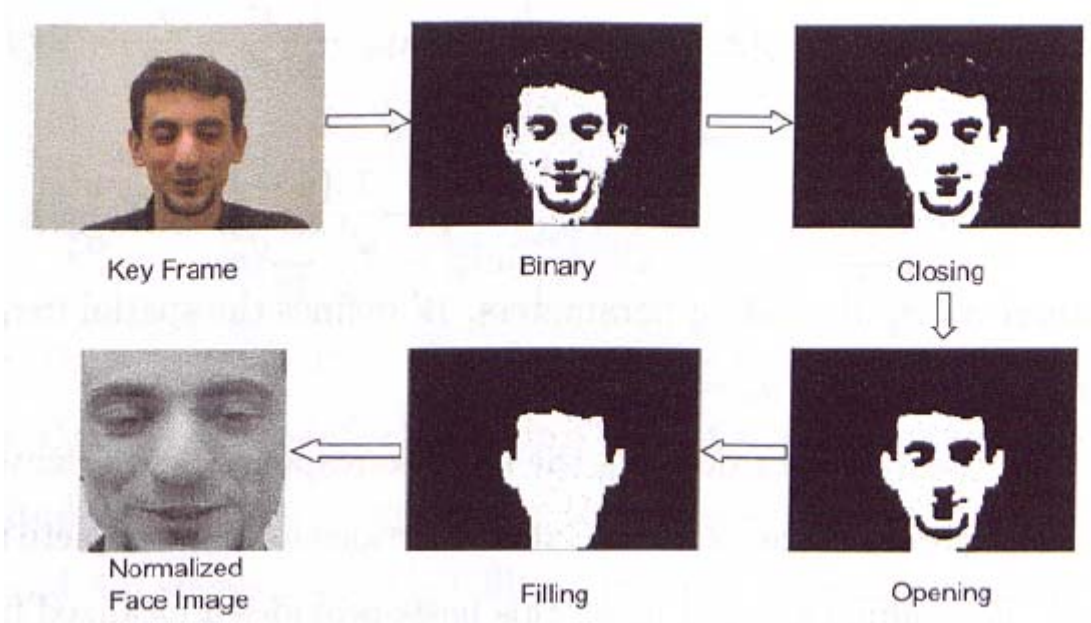

Рисунок 3.2- Процедура застосування схеми розпізнавання облич

3.3 Вейвлет-представлення Габора

Використання вейвлет-ознак Габора для представлення виразу обличчя було досліджено та показано як дуже ефективне в літературі [27]. Це дозволяє описувати просторову частотну структуру зображення, зберігаючи при цьому інформацію про просторові відносини. Банк фільтрів Габора реалізовано за допомогою алгоритму, запропонованого в [47].

3.4 Функції Габора

У просторовій області імпульсна характеристика фільтрів Габора є гаусовим ядром, модульованим синусоїдальною плоскою хвилею:

$$g(x,y) = s(x,y)w(x,y), \tag{3.2}$$

де s{x, y) — комплексна синусоїдальна функція, відома як носій, а w{x, y) — двовимірна функція гаусової форми, відома як обвідна.

Двовимірну функцію Габора g{x, y) та її перетворення Фур'є G{u, v) можна записати як:

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp[-\frac{1}{2}(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}) + 2\pi jWx],$$

$$G(u,v) = \exp\{-\frac{1}{2}(\frac{(u-W)^2}{\sigma_u^2}+\frac{v^2}{\sigma_v^2}]\}, \tag{3.3}$$

де σ_x, σ_y – параметри масштабування, W визначає просторову частоту синусоїдальної форми $\sigma_u = 1/2\pi\sigma_x$, $\sigma_v = 1/2\pi\sigma_y$.

У частотній області імпульсна характеристика є орієнтованою гаусовою функцією, центрованою на частоті несучої. Функції Габора утворюють повний, але неортогональний базисний набір. Розгортання сигналу за допомогою цього базису забезпечує локалізований частотний опис. Розглянуто клас самоподібних функцій, які називаються вейвлетами Габора. Нехай g{x, y) — материнський

вейвлет Габора, тоді цей самоподібний словник фільтрів можна отримати шляхом відповідних розширень та обертань g{x, y).

3.5 Розробка словника фільтрів Габора

Неортогональність вейвлетів Габора означає наявність надлишкової інформації у відфільтрованих зображеннях. Таким чином, зменшення цієї надлишковості необхідно враховувати при розробці словника фільтрів. Цю проблему можна вирішити, забезпечивши підтримку половини пікової величини відгуків фільтра в частотному спектрі, щоб вони торкалися один одного.

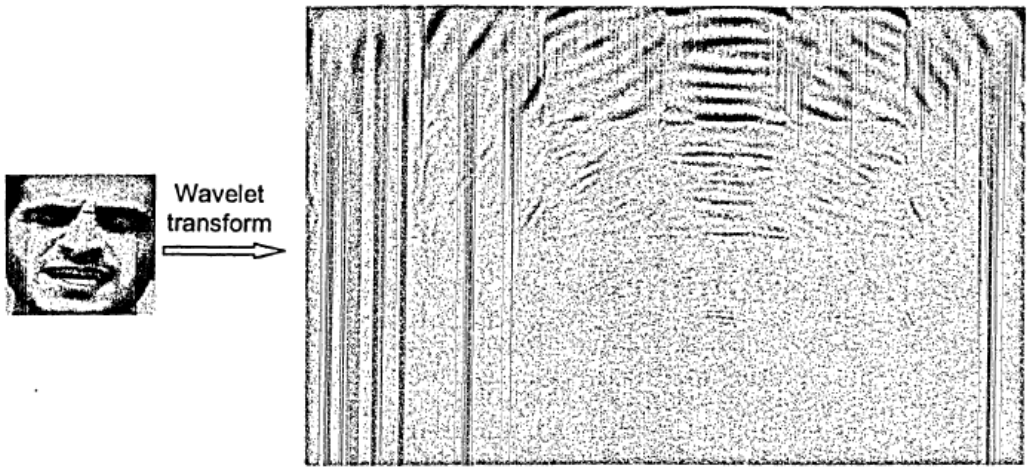

Рисунок 3.3- Приклад зображення, перетвореного за допомогою вейвлета Габора

3.6 Представлення ознак

У розділі 4.2 ми згадували, що зображення людського обличчя нормалізується до зображення рівнів сірого розміром 128 x 128. Якщо взяти це зображення як вхідне та позначити як I{x,y), вейвлет-перетворення Габора цього зображення можна визначити як:

$$W_{mn}(x,y) = \int I(x_1,y_1) g^{*}_{mn}(x-x_1, y-y_1) dx_1 dy_1, \qquad (3.4)$$

де * вказує на комплексно спряжене. У цьому випадку ми отримуємо матрицю з дуже великою коефіцієнтом для кожного зображення. Ми використовуємо загалом 4 x 6 = 24 фільтри Гобора, і таким чином розмір матриці становить 128 x 128 x 24 = 393216. З простором ознак такого великого розміру вартість обчислень дуже висока, і тому це не дуже підходить для системи, яка вимагає швидкої обробки. Ми беремо середнє значення umn та стандартне відхилення amn величини коефіцієнтів перетворення кожного фільтра для представлення ознак:

$$\mathrm{u}_{\mathrm{mn}} = \int\int |W_{mn}(x,y)| dxdy ,$$

$$\sigma_{\mathrm{mn}} = \int\int (|W_{mn}(x,y)| - u_{mn})^2 dxdy . \qquad (3.5)$$

Вектор ознак побудовано з використанням umn та amn як компонентів ознак. Використовано чотири масштаби S=4 та шість орієнтацій K=Q, що призводить до отримання вектора ознак з 48 вимірами:

$$f_{Gabor} = [u_{00} \quad \sigma_{00} \quad u_{01} \quad \sigma_{01} \quad ..... \quad u_{35} \quad \sigma_{35}] .$$

Для вилучення ознак для представлення візуальних характеристик людського обличчя в різних емоційних станах спочатку з послідовності зображень вилучається ключовий кадр. Для виявлення людського обличчя на фоні використовується схема розпізнавання обличчя, заснована на колірній моделі HSV. Потім виявлена область обличчя нормалізується до зображення рівнів сірого розміром 128 x 128. Ми використовуємо стратегію проектування словника фільтрів Габора для розробки банку фільтрів Габора з 4 орієнтаціями та 6 масштабами. Нормалізоване зображення обличчя пропускається через цей банк фільтрів, і обчислюються коефіцієнти. Нарешті, з виходу кожного фільтра обчислюються середнє значення та стандартне відхилення. Ці середні значення та стандартні відхилення розглядаються як візуальні ознаки. Загалом ми вилучили 48 візуальних ознак. Ці ознаки в поєднанні зі 105 аудіоознаками поміщаються у вектор ознак розміром 153 для останньої класифікації.

4 ВИБІР ТА КЛАСИФІКАЦІЯ ОЗНАК

Розпізнавання людських емоцій є, по суті, проблемою розпізнавання образів.

Щоб побудувати повноцінну систему розпізнавання людських емоцій, необхідно застосувати класифікатор для зіставлення витягнутих ознак з відповідним емоційним станом. Це включає процедуру машинного навчання, за допомогою якої комп'ютер вивчає характеристики кожного емоційного стану. Хороший класифікатор повинен успішно вивчати ці характеристики та правильно класифікувати нові вхідні дані. Однак, який класифікатор нам слід обрати, є великим викликом. Порівнював деякі популярні класифікатори, включаючи модель гаусової суміші (GMM), K-найближчих сусідів (K-NN), нейронні мережі (NN) та лінійний дискримінантний аналіз Фішера (FLDA).

Ми витягли загалом 153 ознаки з кожного зразка для представлення як аудіо, так і візуальної інформації. З такою великою розмірністю обчислювальна складність є високою. Більше того, не всі витягнуті ознаки можуть сприяти класифікації, деякі з ознак можуть навіть мати негативний ефект через можливу залежність між ознаками. Щоб зменшити розмірність простору ознак, зберігаючи при цьому точність розпізнавання, нам потрібно застосувати алгоритм вибору ознак. Також порівнювалась продуктивність двох алгоритмів вибору ознак, а саме: аналізу головних компонентів (PGA) та покрокового методу.

4.1 Вибір функцій

Продуктивність системи розпізнавання образів критично залежить від дискримінантної здатності ознак з точки зору розділення образів, що належать до різних класів у просторі ознак. Важливість вибору відповідної підмножини з вихідного набору ознак тісно пов'язана з проблемою "прокляття розмірності" у функціональній апроксимації, де точки вибірки даних стають дедалі рідшими зі збільшенням розмірності функціональної області, так що скінченна множина вибірок може бути недостатньою для характеристики вихідного відображення. Крім того, обчислювальні вимоги зазвичай більші для реалізації високорозмірного відображення. Щоб вирішити ці проблеми, ми зменшуємо

розмірність вхідної області, вибираючи відповідну підмножину ознак з вихідного набору.

Кінцевою метою вибору ознак є вибір кількості ознак з вилученого набору ознак, яка дає мінімальну похибку класифікації. Існує кілька популярних алгоритмів вибору ознак. Вичерпний пошук – це оптимальний алгоритм, який включає пошук усіх можливих комбінацій ознак. Хоча оптимальна підмножина ознак може бути гарантована, вона зазвичай не є вибором через високі обчислювальні витрати. Алгоритм вибору ознак за методом гілок та меж [38] може бути використаний для знаходження оптимальної підмножини ознак, але він вимагає, щоб критеріальна функція вибору ознак була монотонною, що означає, що значення критеріальної функції ніколи не може бути зменшено шляхом додавання нових ознак до підмножини ознак. Це точно не той випадок, коли розмір вибірки відносно малий. Метод обгортки– це ще один варіант вибору підмножини ознак. Однак він дуже повільний та обчислювально дорогий, оскільки весь набір даних необхідно аналізувати Для порівняльного дослідження також представлено популярний алгоритм скорочення простору ознак, аналіз головних компонентів (PCA).

4.2 Покроковий метод

Виконуємо вибір ознак за допомогою покрокового методу. Покроковий метод являє собою комбінацію процедур прямого та зворотного введення. Процедура прямого введення в список вводить змінні одну за одною, доки більше не можна буде ввести. Порядок введення змінних визначається їхньою значущістю в моделі. Процедура зворотного введення вводить усі змінні в список за один крок, потім видаляє незначущі змінні одну за одною, доки більше не можна буде видалити. Однак обидві ці дві процедури мають проблему вкладеності. Наприклад, вибрана 4-елементна підмножина не міститься у вибраній 5-елементній підмножині. Поєднання цих двох процедур може ефективно зменшити цей ефект вкладеності.

Покроковий метод починається з однієї ознаки та додає по одній ознакі кожного разу. Критерієм для визначення включення або виключення ознаки є відстань Махаланобіса. Відстань Махаланобіса – це міра того, наскільки значення випадку за незалежними змінними відрізняється від середнього значення всіх випадків. Велика відстань Махаланобіса ідентифікує випадок як

такий, що має екстремальні значення за однією або кількома незалежними змінними. Відстань Махаланобіса визначається як:

$$D_{mah} = (X_i - Y_i)^T S^{-1} (X_i - Y_i), \tag{4.1}$$

де X_i та Y_i – вектори ознак, S^{-1} – обернена коваріаційна матриця. Для кожного кроку одна ознака додається або видаляється з вибраної підмножини ознак, щоб максимізувати відстань Махаланобіса між класами.

4.3 Аналіз головних компонент

Аналіз головних компонентів – це популярний алгоритм зменшення розмірності в області перетворення. Перетворення призначене для представлення вихідних ознак за рахунок зменшення кількості перетворених ознак, зберігаючи при цьому більшу частину внутрішнього інформаційного змісту вихідних даних.

Щоб PCA працював належним чином, першим кроком є створення набору даних, середнє значення якого дорівнює нулю. Це можна зробити, віднявши середнє значення від кожного виміру даних. Відняте середнє значення є середнім значенням для кожного виміру і може бути розраховане як:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{4.2}$$

де X позначає всю множину зразків в одному вимірі, n – кількість зразків, Xi – компонента множини X.

Потім нам потрібно обчислити коваріаційну матрицю. Коваріація – це міра, яка дозволяє з'ясувати, наскільки розміри відрізняються від середнього значення один відносно одного.

Її можна обчислити за допомогою наступного рівняння:

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\cdot\left(Y_i - \overline{Y}\right)}{(n-1)} \tag{4.3}$$

Коваріація завжди вимірюється між двома вимірами, для m-вимірного набору даних потрібно обчислити різні значення коваріації. Щоб обчислити матрицю коваріації, нам потрібно обчислити всі можливі значення коваріації між усіма різними вимірами та помістити їх у матрицю. Визначення матриці коваріації для набору даних з m вимірами таке:

$$C^{m\times m} = [c_{i,j}], \qquad c_{i,j} = \operatorname{cov}(Dim_i, Dim_j) \tag{4.4}$$

де $C^{m\times m}$ — матриця з m рядками та m стовпцями. Наприклад, запис у рядку 2, стовпці 3 коваріаційної матриці — це значення коваріації, обчислене між 2-м та 3-м вимірами.

Оскільки коваріаційна матриця є квадратною, можемо обчислити власні вектори, кожен з відповідним власним значенням для цієї матриці. Ці власні вектори є ортогональними. Власний вектор з найбільшим власним значенням є першою головною компонентою набору даних. Потім ми можемо впорядкувати власні вектори за відповідними власними значеннями, від найбільшого до найменшого. Це дає компоненти в порядку значущості. Компоненти з меншою значущістю можна ігнорувати, і таким чином можна досягти зменшення розмірності.

Останній крок — формування нових векторів даних шляхом проектування вихідних даних на вектори головних компонент. Після того, як ми вибрали компоненти, які хочемо зберегти в наших даних, і сформували вектор ознак, помістивши власні вектори в матрицю з власними векторами у стовпцях, ми просто беремо транспонований вектор і множимо його ліворуч від транспонованого вихідного набору даних

$$X_{pca} = V_{eig} \times X_{original},$$

де V_{eig} — матриця з власними векторами у стовпцях, транспонованими таким чином, що власні вектори тепер розташовані в рядках, з найбільш значущим власним вектором зверху; $X_{original}$ – це транспоновані дані, скориговані за середнім значенням.

Новий набір даних X_{pca} зі зменшеною розмірністю може бути використаний для останньої класифікації.

4.4 Класифікація на основі моделі гаусової суміші

Щоб класифікувати витягнуті ознаки за різними людськими емоціями, нам потрібно вибрати класифікатор, який може належним чином моделювати дані та досягти кращої точності класифікації. Порівняна ефективність кількох популярних класифікаторів, включаючи модель гаусової суміші (GMM), K-найближчих сусідів (K-NN), нейронну мережу (NN) та лінійний дискримінантний аналіз Фішера (FLDA).

GMM широко використовується для ідентифікації, верифікації та розпізнавання мовлення мовця. GMM моделює функцію щільності ймовірності як зважену суму кількох різних гаусових функцій щільності $p_k(y)$

$$P_{GMM}(y) = \sum_{k=1}^{K} c_k p_k(y), \tag{4.5}$$

де c_k – ймовірності компонентів, а K – кількість компонентів. Використовується алгоритм максимізації очікувань (EM) для оцінки параметрів GMM. Параметри, включаючи ймовірність компонентів c_k, середнє значення μ_k та коваріаційну матрицю $\sigma_{k,j}$, оцінюються за допомогою наступних рівнянь:

$$c_k^{i+1} = \frac{1}{N} \sum_{n=1}^{N} \psi_k(n),$$

$$\mu_k^{i+1} = \frac{\sum_{n=1}^{N} \psi_k(n) y_n}{\sum_{n=1}^{N} \psi_k(n)},$$

$$\sigma_{k,j}^{i+1} = \frac{\sum_{n=1}^{N} \psi_k(n)(y_n - \mu_{k,j}^i)^2}{\sum_{n=1}^{N} \psi_k(n)}$$

$$\psi_k(n) = \frac{c_k^i p_k(y_n | \mu_k^i, \Sigma_k^i)}{\sum_{j=1}^{K} c_j^i p_j(y_n | \mu_j^i, \Sigma_j^i)}, \quad (4.6)$$

)

$\psi_k(n)$— апостеріорна ймовірність, N — розмір вектора ознак, а i — номер ітерації.

Для класифікації GMM зазвичай виконується за модульною архітектурою, яка передбачає навчання окремої GMM для кожного окремого класу. На рисунку 5.1 показано архітектуру застосованої схеми класифікації GMM.

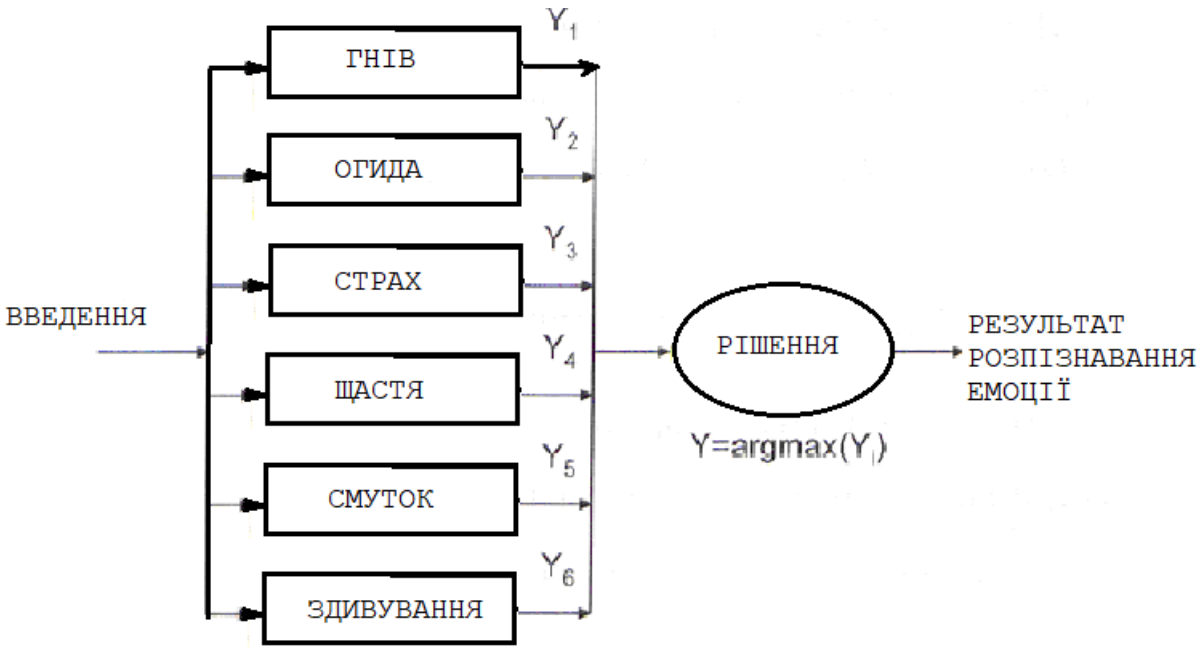

Рисунок 5.1- Архітектура застосованої схеми GMM

Вхідний сигнал позначається відповідною емоцією з максимальним вихідним сигналом:

$$H = \arg\max(Y_i) \tag{4.7}$$

де i =1, 2, ...,6.

4.5 Приняття рішень на основі методи K-найближчих сусідів

K-найближчих сусідів – це непараметричний метод класифікації [53].

Алгоритм k-найближчих сусідів полягає у визначенні підмножини k векторів, які лежать найближче до у відносно функції подібності до шаблону D(y,x).

Використовувана функція подібності до шаблону - це евклідова відстань, яка визначається як:

$$D(x, y) = \sqrt{\sum_{i=1}^{n} |y_i - x_i|^2} \ . \tag{4.8}$$

Якщо ми використовуємо Fq для позначення частоти кожної мітки класу в підмножині k найближчих сусідів, вхідний вектор можна класифікувати за таким правилом:

$$l_{out} = \arg\max\{F, \quad q = 1,2,..., p\} \tag{4.9}$$

4.6 Нейронна мережа

Штучні нейронні мережі [51] за останні кілька років зазнали вибухового інтересу та застосовуються в надзвичайно широкому діапазоні проблемних областей. На відміну від класичних методів статистичної класифікації, таких як байєсівський класифікатор, нейронній мережі не потрібні знання про розподіл ймовірностей. Вона може вивчити вільні параметри, ваги та зміщення шляхом навчання на прикладах. На рисунку 5.2 показано архітектуру використаної нейронної мережі.

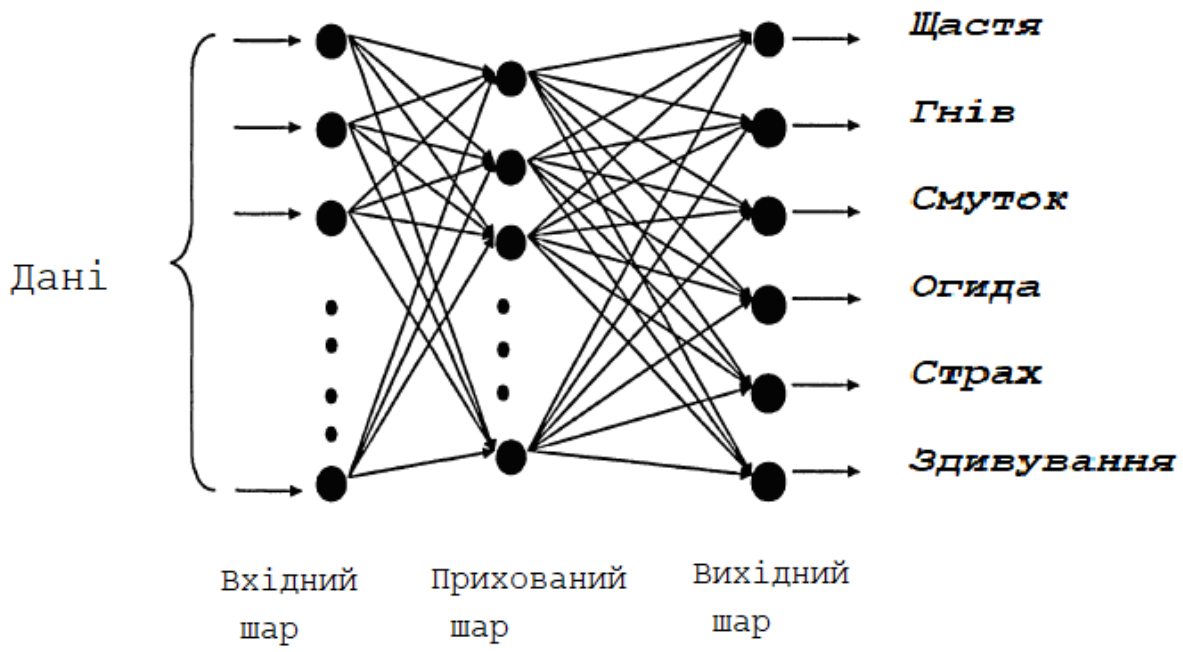

Рисунок 5.2- Архітектура застосованої нейронної мережі

Для навчання використовується алгоритм зворотного поширення. Кількість нейронів у вхідному шарі дорівнює розмірності вектора ознак. Кількість нейронів у прихованому гіперболічно-тангенційному сигмоподібному шарі Nhidden визначається наступним правилом:

$$N_{hidden} = \left| \sqrt{N_{input} N_{output}} \right|,$$

де Ninpui та Noutput позначають кількість нейронів у вхідному шарі та вихідному шарі відповідно.

Передавальна функція гіперболічного тангенса сигмоподібної осі визначається як:

$$\sigma(n) = \left(\frac{2}{1 + e^{-2n}} \right) - 1,$$

де a(n) – це вихідний сигнал нейрона, виражений у вигляді індукованого локального поля n.

Шість нейронів у вихідному лінійному шарі відповідають шести емоціям. У процесі навчання ми присвоюємо значення 1 вихідному нейрону, який

відповідає тій самій емоції, що й вхідний сигнал, та значення 0 усім іншим нейронам вихідного шару. У процесі моделювання новий вхід позначається як емоція з найбільшим вихідним сигналом.

4.7 Лінійний дискримінантний аналіз Фішера

Лінійний дискримінантний аналіз припускає, що дискримінантна функція g(x) є лінейно-рівневою функцією даних x. У випадку задачі c-класу дискримінантна функція визначається як [53]:

$$g_i(x) = w_i^T x + w_{i0}\,. \tag{4.10}$$

У LDA Фішера ваги w оцінюються шляхом максимізації критеріальної функції Фішера J f (w) :

$$J_F(w) = \frac{w^T S_b w}{w^T S_\omega w}, \tag{4.11}$$

де S_b – матриця розсіювання між класами, S_ω – матриця розсіювання всередині класу.

Якщо ми використовуємо Hi для представлення середнього значення класу i, h представляє середнє значення всього набору даних, тоді S_b та S_ω можна розрахувати як:

$$S_b = \sum_{i=1}^{C} \frac{n_i}{n}(\mu_i - \mu)(\mu_i - \mu)^T,$$

$$S_\omega = \sum_{i=1}^{C} \frac{n_i}{n} \sum_{x_k \in C_i} (x_k - \mu_i)(x_k - \mu_i)^T, \tag{4.12}$$

де C – кількість класів, Q позначає вибірки в класі i, n – загальна кількість вибірок, а n – кількість вибірок у класі i. При цьому $\frac{n_i}{n}$ представляє апріорну ймовірність.

Оскільки наша мета — знайти w, яке максимізує критерій можна вивести та знайти, розв'язавши узагальнену задачу на власні значення:

$$\mathrm{J_F}(w) = w^T S_b S_\omega^{-1} w.$$

Після обчислення вагових коефіцієнтів знаходимо постійне значення для кожної дискримінантної функції, яке максимізує роздільність між класами. Для класифікації, вхідні дані групуються в клас, який дає найбільше значення дискримінантної функції.

У таблиці 4.1 наведно застосовані алгоритми. Продуктивність алгоритму вибору та класифікації ознак тісно пов'язана з властивими характеристиками вилучених ознак. Аналіз та порівняння різних підходів допоможе нам отримати більше розуміння досліджуваної проблеми та досягти вищої точності розпізнавання.

Таблиця 4.1- Алгоритми вибору та класифікації ознак

	Алгоритм
Оцінювання ознак	Метод головних компонент, Покроковий метод
Класифікація	Модель гаусової суміші, Нейронна мережа, Метод K-найближчих сусідів, Лінійний дискримінантний аналіз Фішера

5 ДОСЛІДЖЕННЯ СИСТЕМ РОЗПІЗНАВАННЯ ЕМОЦІЙ

5.1 Особливості експериментальних досліджень

Система розпізнавання емоцій працює на основі результатів отриманих на основі методів вибору ознак, оцінювання ознак, та використання їх для класифікації емоцій. На рисунку 5.1 показано огляд запропонованої системи розпізнавання.

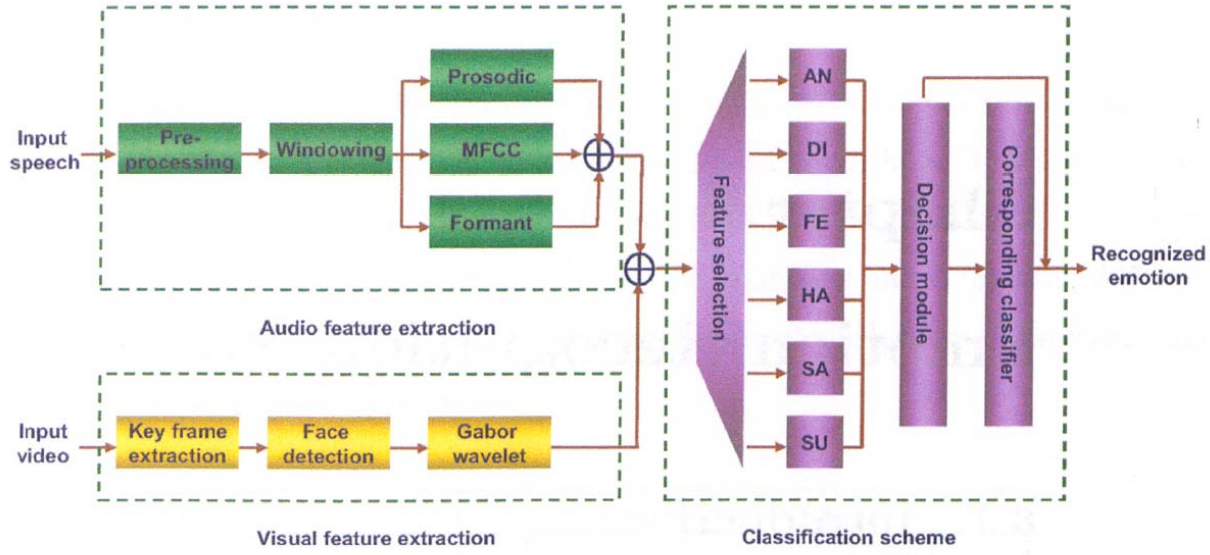

Рисунок 5.1- Система розпізнавання емоцій

Вхідна відеопослідовність пропускається через два різні канали для вилучення аудіо- та візуальних характеристик послідовності. Ці два потоки ознак потім об'єднуються в один потік, і застосовується процес вибору ознак для виявлення значущих ознак, одночасно зменшуючи розмірність, а отже, і обчислювальні витрати. Система навчається за допомогою даних відомого класу, а новий вхід може бути класифікований у відповідному класі за допомогою прийнятої схеми класифікації.

Проведені експерименти базувалися на відеозразках восьми учасників. Для навчання та тестування було використано загалом 500 зразків, кожен з яких містив одну з шести емоцій. З цих зразків 360 зразків (від шести учасників) було відібрано для навчання, а решта 140 (від двох інших учасників) – для тестування. Не було жодного перекриття між учасниками навчання та

тестування. У таблиці 5.1 детально наведено експериментальні дані щодо різних емоцій.

Спочатку експериментальні результати застосування різних алгоритмів класифікації, а потім порівнюємо різні алгоритми вибору ознак використовуючи класифікатор найкращої продуктивності. Потім вводиться схема з кількома класифікаторами, що базується на аналізі окремого класу та комбінацій різних класів. Нарешті, для порівняльного дослідження також представлені результати перехресної валідації за принципом «виключити один елемент».

Таблиця 5.1- Експериментальна структура бази даних емоцій

Емоції	Тренування	Тестування	Вього
Гнів	60	26	86
Огида	59	21	80
Страх	56	22	78
Щастя	62	25	87
Смуток	59	21	80
Здивування	64	25	89
Всього	360	140	500

5.2 Результати експериментальних досліджень алгоритмів розпізнавання емоцій на основі полігаусівської моделі

У першому експерименті виконали класифікацію за просодичними, аудіо, візуальними та аудіовізуальними ознаками окремо. Порівнювалась ефективність моделі гаусової суміші (GMM), K-найближчих сусідів (K-NN), нейронної мережі (NN) та лінійного дискримінантного аналізу Фішера (FLDA).

Класифікатор GMM реалізовано за модульною архітектурою. Для кожного окремого класу навчається окремий GMM. Параметри, включаючи ваги, середнє значення та стандартне відхилення кожного компонента, оцінюються за допомогою алгоритму EM. У наших експериментах, оскільки ми не маємо точного уявлення про розподіл даних, ми пробуємо діапазон значень k, щоб розподіл даних можна було змоделювати як суму k гаусових функцій. Застосовуючи схему класифікації GMM до 25 просодичних ознак, 105 аудіоознак, 48 візуальних ознак та 153 аудіовізуальних ознак, ми отримуємо експериментальні результати, показані в таблиці 6.2, де найкращий коефіцієнт розпізнавання кожної категорії виділено жирним шрифтом.

Таблиця 5.2 - Результати розпізнавання GMM

K	Просодичний	Аудіо	Візуальний	Аудіовізуальний
2	52.62%	58.57%	27.20%	63.27%
3	56.95%	62.11%	26.72%	72.37%
4	49.99%	64.72%	28.28%	58.56%
5	53.04%	57.51%	25.35%	62.24%
6	54.23%	61.65%	21.80%	60.69% •
7	60.60%	61.66%	27.93%	60.23%
8	50.97%	56.97%	21.68%	65.38%
9	51.53%	57.86%	26.72%	55.24%
10	51.18%	55.49%	24.38%	57.91%

5.3 Результати експериментальних досліджень алгоритмів розпізнавання емоцій на основі методу K - найближчих сусідів

У K-NN вхідний сигнал позначається класом, який має найбільшу кількість вибірок серед k найближчих сусідів. Популярним способом визначення значення k є використання перехресної перевірки з виключенням одного елемента. Однак значення k, вибране за допомогою цього методу, може бути не найкращим значенням k. У наших експериментах ми не встановлюємо k на фіксоване значення. Ми проводимо експерименти з десятьма значеннями k від одного до десяти, а результати експериментів наведено в таблиці 5.3, де найвищі результати виділено жирним шрифтом.

Таблиця 5.3-Результати розпізнавання K-NN

K	Просодичний	Аудіо	Візуальний	Аудіовізуальний
1	40.43%	46.43%	135.71%	57.14%
2	45.71%	50.71%	30.00%	50.71%
3	50.71%	55.71%	30.71%	57.14%
4	48.57%	55.71%	29.29%	57.14%
5	50.71%	58.57%	29.29%	57.14%
6	50.00%	55.71%	30.71%	60.00%
7	49.29%	61.43%	29.29%	62.86%
8	49.29%	62.14%	30.71%	57.86%
9	52.14%	61.43%	28.57%	59.29%
10	55.00%	59.29%	27.14%	60.00%

5.4 Результати експериментальних досліджень алгоритмів розпізнавання емоцій на основі нейронних мереж

Також досліджується тришарова нейронна мережа прямого зв'язку для класифікації. Кількість нейронів вхідного шару дорівнює розмірності вхідного набору ознак, тоді як вихідні нейрони відповідають шести класам емоцій. Для навчання мережі використовується алгоритм зворотного поширення. Новий вхід позначається як клас, який дає максимальне вихідне значення. У таблиці 5.4 наведено експериментальні результати застосування нейронної мережі до просодичних, аудіо, візуальних та аудіовізуальних ознак.

Таблиця 5.4-Результати розпізнавання нейронної мережі

Класифікатор	Просодичний	Аудіо	Візуальний (за зображенням)	Аудіовізуальна
Нейронна мережа	49.29%	51.43%	35%	56.43%

5.5 Результати експериментальних досліджень алгоритмів розпізнавання емоцій на основі лінійного дискримінантного аналізу Фішера

Застосований класифікатор FLDA має шість виходів, що відповідають шести емоціям, що нижче називається глобальним класифікатором. Вхідний сигнал позначено класом, який дає максимальне вихідне значення. У таблиці 5.5 наведено експериментальні результати.

Таблиця 5.5- Результати розпізнавання FLDA

Класифікатор	Просодичний	Аудіо	Візуальний	Аудіовізуальний
FLDA	65,71%	66,43%	49,29%	70,00%

5.6 Порівняння результатів розпізнавання

Порівняння результатів розпізнавання з використанням різних алгоритмів класифікації за просодичними, аудіо, візуальними та аудіовізуальними ознаками наведено в таблиці 5.6.

Таблиця 5.6 - Результати розпізнавання різних класифікаторів

	GMM	NN	K-NN	FLDA
Просодичний	60.60	49.29	55.00	65.71
Аудіо	64.72	51.43	62.14	66.43
Візуальний	28.28	35 00	35.71	40.20
Аудіовізуальний	65.38	56 43	62.88	70.00

Експериментальні результати показують, що комбінація аудіо та візуальної інформації працює краще, ніж будь-який з них окремо. Ми порівняли продуктивність GMM, K-NN, NN та FLDA. Очевидно, що FLDA перевершує інші класифікатори. GMM та K-NN – це статистичні методи, що базуються на оцінці функції щільності ймовірності. Нейронні мережі оцінюють оптимальні значення ваги та зміщення між різними шарами за допомогою процесу

навчання. Усім їм потрібен достатньо великий навчальний набір. У нашому випадку розмір навчального набору містить 3GO вибірок, а розмірність простору ознак становить 153. Висока розмірність простору ознак та розрідженість навчального набору обмежують точність оцінки, і таким чином знижують продуктивність.

Підтверджено, що просодічні ознаки є потужним показником емоційного стану людини в мовленні. За допомогою FLDA 25 просодичних ознак забезпечують точність розпізнавання 65,71%, що дуже близько до 105-вимірного набору аудіоознак, до якого також включені MFCC та формантні ознаки. Однак це також демонструє, що фонетичні ознаки також сприяють класифікації.

Можна також зазначити, що взаємодоповнюючий зв'язок аудіо та візуальної інформації покращує продуктивність системи. Таблиця 5.7, 5.8 та 5.9 детально опишіть матрицю плутанини застосування FLDA до аудіо, візуальних та аудіовізуальних характеристик (гнів, огида, страх, щастя, SA: смуток, здивування).

Таблиця 5.7- Матриця плутанини FLDA щодо аудіофункцій (усі у %)

	Виявлено					
Відомо	Гнів	Огида	Страх	Щастя	Смуток	Здивування
Гнів	88.46	0	0	3.85	0	7.69
Огида	4.76	61.90	28.57	4.76	0	0
Страх	0	18.18	59.09	4.55	9.09	9.09
Щастя	8.00	20.00	16.00	48.00	8.00	0
Смуток	0	4.76	28.57	9.52	57.14	0
Здивування	16.00	O	4.00	0	0	80.00

Таблиця 5.8- Матриця плутанини FLDA за візуальними ознаками (усі у %)

	Виявлено					
Відомо	Гнів	Огида	Страх	Щастя	Смуток	Здивування
Гнів	34.62	0	42.31	7.69	15.38	0
Огида	0	47.62	23.81	23.81	4.76	9.52
Страх	18.18	9.09	68.18	0	0	4.55
Щастя	4.00	16.00	4.00	52.00	8.00	16.00
Смуток	9.52	4.76	14.29	14.29	47.62	9.52
Здивування	16.00	8.00	16.00	8.00	4.00	48.00

Таблиця 5.9- Матриця перепутування з застосуванням аудіовізуальних функцій (усі у %)

	Виявлено					
Відомо	Гнів	Огида	Страх	Щастя	Смуток	Здивування
Гнів	76.92	0	3.85	7.69	0	11.54
Огида	0	76.19	19.05	4.76	0	0
Страх	4.55	9.09	77.27	4.55	0	4.55
Щастя	4.00	16.00	8.00	56.00	16.00	4.00
Смуток	0	9.52	23.81	4.76	61.90	0
Здивування	12.00	0	12.00	4.00	0	72.00

З таблиць 5.7 та 5.8 очевидно, що гнів, огиду, смуток та здивування краще класифікувати за допомогою аудіо, тоді як страх та щастя краще розпізнаються за допомогою візуальної інформації. Поєднання аудіо та візуальних ознак покращує точність розпізнавання огиди, страху, щастя та смутку. Однак точність класифікації гніву та здивування не така висока, як до інтеграції даних (таблиця 5.8). Це свідчить про те, що деякі ознаки можуть негативно впливати на класифікацію. Тому для вирішення цієї проблеми необхідний процес відбору ознак.

Результати показують, що комбінація аудіо та візуальної інформації досягає кращої загальної точності, ніж будь-яка з них окремо. Застосований покроковий метод ефективно зменшує розмірність простору ознак, одночасно досягаючи кращої точності розпізнавання. Запропонована адаптивна схема

мультикласифікатора забезпечує помітне покращення як загальної, так і індивідуальної точності розпізнавання класів.

ВИСНОВКИ

Розглянуто розпізнавання емоцій на основі полігаусівської моделі, методу К -найближчих сусідів, нейронних мереж та лінійного дискримінантного аналізу.

Запропонована система досягає дуже обнадійливих результатів завдяки використанню запропонованої схеми класифікації за мовленевою та візуальною інформацією. Систему було протестовано з використанням бази даних, в якій були зібрані зразки від різних суб'єктів.

Проводено аналіз емоцій на основі шести основних емоцій. Однак, необхідно відмітити, що емоційні стани людини не мають чіткої межі. У дослідженні вивчали голосові та мімічні вирази обличчя та запропоновано систему для розпізнавання цих виразів. Згідно з результатами експерименту, точність класифікації на основі візуальних ознак є близько 40 %. Це демонструє, що представлення візуальних ознак є недостатньо сильним.

Очевидно, що гнів, огиду, смуток та здивування краще класифікувати за допомогою аудіо, тоді як страх та щастя краще розпізнаються за допомогою візуальної інформації. Поєднання аудіо та візуальних ознак покращує точність розпізнавання огиди, страху, щастя та смутку.

Досліджено алгоритми розпізнавання емоційного стану людини, та показано переваги та ефективність поєднання мультимодальної інформації в розпізнаванні людських емоцій. Це відкриває потенціал та можливість аналізу інших проблем взаємодії людини з комп'ютером. Поєднання різних модальностей передає більше інформації про інтерфейс та допомогає підвищити продуктивність системи.

Для розпізнавання коли об'єднується ознаки аудіо та відео точність розпізнавання емоційного стану около 70% .

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. R. Cowie, E. Douglas-Cowie. Automatic statistical analysis of the signal and prosodic signs of emotion in speech.- Proceedings of 4th International Conferences on Spoken Language Processing, vol. 3, pp. 1989-1992, Philadelphia, USA,October 1996

2. N. Amir, S. Ron. Toward an automatic classification of emotions in speech.- Proceedings of 5th International Conference on Spoken Language Processing.- vol. 3, pp. 555-558, Sydney, Australia, November/December 1998.

3. J. Sato, S. Morishima. Emotion modeling in speech production using emotion space.- Proceedings of IEEE International Workshop on Robot and Human Communication.- pp. 472-477, Tsukaba, Japan, November 1996.

4. F. Dellaert, T. Polzin and A. Waibel. Recognizing emotion in speech. Proceedings of the 4th International Conferences on Spoken Language Processing, vol. 3.- pp. 1970-1973, Philadelphia, USA, October 1996.

5. J. Nicholson, K. Takabashi and R. Nakatsu. Emotion recognition in speech using neural networks.- Neural Computing and Applications, vol. 9.- pp. 290-296, 2000.

6. C. M. Lee, S. S. Narayanan, and R. Pieraccini. Classifying emotions in humanmachine spoken dialogs.- Proceedings of International Conference on Multimedia and Expo, vol. 1, pp. 737-740, Switzerland, August 2002.

7. T. L. Nwe, F. S. Wei and L. C. De Silva. Speech based emotion classification.- Proceedings of IEEE Region 10 Conference on Electrical and Electronics Technology, vol. 1, pp. 297-301, Singapore, August 2001.

8. O. Kwon, K. Chan, J. Hao, and T. Lee, Emotion recognition by speech signals.- Proceedings of European Conference on Speech Communication and Technology, pp. 125-128, Geneva, Switzerland, September 2003.

9. B. Schuller, G. Rand M. Lang, Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief

network architecture.- Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 577-580, Montreal, Canada, May 2004.

10. D. Ververidis, C. Kotropoulos, and I. Pitas, Automatic emotional speech classification.- Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 593-596, Montreal, Canada, May 2004.

11. P. Ekman and W. V. Friesen, Facial Action Coding System (FACS) .- Manual, Palo Alto: Consulting Psychologists Press, 1978.

12. P. Ekman, M. O. Sullivan. The role of context in interpreting facial expression: Comment on Russell and Fehr.- Journal of Experimental Psychology, General, vol. 117, pp. 86-88, 1987.

13. G. Donato, M. S. Barlett, J. C l Hager, P. Ekman and T. J. Sejnowski. Classifying facial actions.- IEEE Transactions on Pattern Analysis and Machine Intelligence,vol. 21, pp. 974-989, October 1999.

14. D. Yang, T. Kunihiro, H. Shimoda, and H. Yoshikawa. A study of real-time image processing method for treating human emotion by faeial expression.- Proceedings of IEEE International Conference on Systems, Man, and Cybernetics,vol. 2, pp. 360-364, Japan, October 1999.

15. M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski. Measuring facial expressions by computer image analysis.- Psychophysiology, vol. 36, pp. 253-263, 1999.

16. A. Samal and P. A. Iyengar. Automatic recognition and analysis of human faces and facial expressions: A survey.- Pattern Recognition, vol. 25, pp. 65-77, 1992.

17. M. Pantic and L. J. M. rothkrantz. Automatic analysis of facial expressions: The state of the art.- IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 1424-1445, 2000.

18. K. Haneda, T. Muraguchi and O. Nakamura. The recognition of faical expressions using expert system.- Proceedings of IEEE Canadian Conference on Electrical and Computer Engineering, vol. 2, pp. 1195-1198, Quebec, Canada, May 2003.

19. M. J. Lyons, J. Budynek, A. Plante, and S. Akamatsu. Classifying facial attributes using a 2-D Gabor wavelet representation and discriminant analysis.- Proceedings of the 4th International Conference on Automatic Face and Gesture Recognition, pp. 202-207, France, March 2000.

20. M. Pantic and L. J. M. rothkrantz. Facial action recognition for facial expression analysis from static face image.- IEEE Transactions on Systems, Man, and Cybernetics-Part B:Cybernetics, vol. 34, pp. 1449-1461, 2004.

21. L. C De Silva and S. C. Hui. Real-time facial feature extraction and emotion recognition.- Proceedings of 4th International Conference on Information, Communications and Signal Processing, vol. 3, pp. 1310-1314, Singapore, December.2003

22. I. Cohen, N. Sebe, Y. Sun, M. S. Lew and T. S. Huang. Evaluation of expression recognition techniques.- Proceedings of International Conference on Image and Video Retrieval, pp. 184-195, Urbana, IL, USA, July 2003.

23. L. Ma and K. Khorasani. Facial expression recognition using constructive feedforward neural networks.- IEEE Transactions on Systems, Man and Cybernetics-Part B:Cybernetics, vol. 34, pp. 1588-1595, 2004.

24. D. Kim and Z. Bien. Fuzzy neural networks(FNN)-based approach for personalized facial expression recognition with novel feature selection method.- Proceedings of IEEE International Conference on Fuzzy Systems, vol. 2, pp. 908-913, St.Louis, MO, USA, May 2003.

25. L. C. De Silva, T. Miyasato and R. Nakatsu. Facial emotion recognition using multi-modal information.- Proceedings of IEEE International Conference on Information, Communications and Signal Processing, vol. 1, pp. 397-401, Singapore, September 1997.

26. M. Song, C. Chen, and M. You. Audio-visual based emotion recognition using tripled hidden Markov model.- Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 5, pp. 877-880, Montreal, Canada, May 2004.

27. L. C. De Silva, P. C. Ng, Bimodal emotion recognition.- Proceedings of 4th IEEE International Conference on Automatic Face and Gesture Recognition, pp. 332 - 335, France, March 2000.

28. L. S. Chen, H. Tao, T. S. Huang, T. Miyasato, and R. Nakatsu, Emotion recognition from audiovisual information.- Proceedings of IEEE 2nd Workshop on Multimedia Signal Processing, pp. 83-88, California, USA, December 1998

29. H. Go, K. Kwak, D. Lee, and M. Chun. Emotion recognition from the facial image and speech signal.- Proceedings of SICE 2003 Annual Conference, vol. 3, pp. 2890-2895, Japan, August 2003.

30. M. H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey.- IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.24, pp. 34-58, January 2002.

31. P. M. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection.- IEEE Transactions on Computers, vol. 26, pp. 917-922, September 1977.

32. J. Kittler, Feature set algorithms. Pattern Recognition and Signal Processing.- pp. 41-60. Sijthoff and Noordhoff, Alphen aan den Rijn, Netherlands, 1978.

33 G. Bailly, C. Benoit, and T. R. Sawallis, Talking Machines: Theories, Models, and Designs.- Elsevier Science Publishers, Amsterdam: 1992.

34. C. Garcia and G. Tziritas, Face detection using quantized skin color regions merging and wavelet packet analysis.- IEEE Transactions on Multimedia, vol.1, pp 264-277, September 1999.