МАГІСТЕРСЬКА РОБОТА

МР. ІПм - 12.00.00.000 ПЗ

Група ІПм-24-1

Грищук Артем

2025

Івано-Франківський національний технічний університет нафти і газу

Факультет інформаційних технологій

Кафедра інженерії програмного забезпечення

Грищук Артем Іванович

(прізвище, ім'я, по батькові)

УДК 004.9

(індекс)

МАГІСТЕРСЬКА РОБОТА

Моделі та методи обробки мовних складових в системах

комп'ютерного зору

(назва роботи)

Інженерія програмного забезпечення

(назва освітньої програми)

121 - Інженерія програмного забезпечення

(шифр і назва спеціальності)

Грищук А.І.

(підпис, ініціали та прізвище здобувача освітнього ступеня)

Науковий керівник Михайлюк Ірина Романівна, к.п.н., доцент

(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Допущено до захисту

Завідувач кафедри

доц. Бандура В.В.

(посада) (підпис) (дата) (ініціали та прізвище)

Нормоконтроль

доц. Вовк Р.Б.

(посада) (підпис) (дата) (ініціали та прізвище)

Робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

Івано-Франківськ – 2025

Івано-Франківський національний технічний університет нафти і газу

Факультет інформаційних технологій

Кафедра інженерії програмного забезпечення

Освітній рівень магістр

Спеціальність 121 – Інженерія програмного забезпечення

ЗАТВЕРДЖУЮ:
Зав. кафедрою **ІПЗ**
доц. В.В. Бандура
“ 04 ” вересня 2025 р.

ЗАВДАННЯ

НА МАГІСТЕРСЬКУ РОБОТУ СТУДЕНТУ

Грищуку Артему Івановичу

(прізвище, ім’я, по-батькові)

1. Тема магістерської роботи “ Моделі та методи обробки мовних складових в системах комп'ютерного зору”

керівник проекту (роботи) Михайлюк Ірина Романівна, к.п.н., доцент

затверджені наказом закладу вищої освіти від “ 05 ” листопада 2025 р. № 695/7

2. Строк подання студентом проекту (роботи) 15 грудня 2025 р.

3. Вихідні дані до проекту (роботи) Теоретичні концепції та формальні моделі побудови та функціонування програмних технологій обробки мови користувача

4. Зміст розрахунково - пояснювальної записки(перелік питань, які потрібно розробити)

1. Аналіз предметної області та методологій обробки мовлення для систем компютерного зору
2. Моделі та методи нейронних мереж для процесів обробки мовних складових
3. Застосування фреймворку для екстракції та алгоритм роботи системи ідентифікації мовлення
4. Імплементація моделей та методів обробки мовних складових в системах комп'ютерного зору

5. Перелік графічного матеріалу (з точним зазначенням обов’язкових креслень)

1. Спектрограма перетворення Фур'є (рис. 1.1)
2. Основні етапи процесу отримання MFCC (рис. 1.2)
3. Візуалізація Mel Scale (рис. 1.3)
4. Візуалізація трикутної віконної функції (рис. 1.4)
5. Представлення MFCC фрази «Привіт», вимовленої двічі (рис. 1.5)

6. Консультанти розділів проекту (роботи)

Розділ	Консультант	Підпис, дата
Перевірка на плагіат	доц., к.т.н. Вовк Р.Б.	

7. Дата видачі завдання 04 вересня 2025 р.

Керівник _____
(підпис)

Завдання прийняв до виконання_____
(підпис)

КАЛЕНДАРНИЙ ПЛАН

№ п/п	Назви етапів магістерської роботи	Строк виконання етапів роботи	Примітка
1	Підбір і вивчення літератури по темі магістерської роботи	15.09.2025	виконано
2	Аналіз предметної області та методологій обробки мовлення для систем компютерного зору	29.09.2025	виконано
3	Моделі та методи нейронних мереж для процесів обробки мовних складових	15.10.2025	виконано
4	Застосування фреймворку для екстракції та алгоритм роботи системи ідентифікації мовлення	08.11.2025	виконано
5	Імплементація моделей та методів обробки мовних складових в системах комп'ютерного зору	20.11.2025	виконано
6	Затвердження пояснювальної записки роботи завідувачем кафедри	15.12.2025	виконано

Студент – магістр _____
(підпис)

Керівник роботи _____
(підпис)

АНОТАЦІЯ

Магістерська робота: 76 с., 21 рис., 2 табл., 41 джерело.

Тема: Моделі та методи обробки мовних складових в системах комп'ютерного зору

Мета магістерської роботи - розробка, дослідження та реалізація моделей і методів обробки мовних складових для систем комп’ютерного зору, які забезпечують формування мультимодального представлення ознак об’єкту мовлення.

Об’єкт дослідження - процеси обробки мовних сигналів у системах комп’ютерного зору, які реалізують взаємозв’язок між аудіо- та візуальними модальностями.

Предмет дослідження - моделі, методи та алгоритми обробки мовних складових, що використовуються для формування, аналізу та синтезу візуальних образів на основі глибоких нейронних мереж.

Результати дослідження

В роботі розроблено методологію мультимодальної генерації зображення обличчя за голосом, яка базується на глибоких нейронних мережах та принципах кросмодального навчання.

Висновок

Розроблено архітектурне рішення на основі фреймворку Vec2Face, яке дозволяє здійснювати генерацію обличчя мовця за голосом з високим рівнем схожості. Отримано результати щодо узгодження латентних ознак різних модальностей у єдиному навчальному середовищі, що підвищує точність ідентифікації.

ОБРОБКА МОВЛЕННЯ; КОМП’ЮТЕРНИЙ ЗІР; ГЛИБОКЕ НАВЧАННЯ; НЕЙРОННІ МЕРЕЖІ; МУЛЬТИМОДАЛЬНА ГЕНЕРАЦІЯ; ІДЕНТИФІКАЦІЯ МОВЦЯ; СИНТЕЗ ОБЛИЧЧЯ ЗА ГОЛОСОМ; КРОСМОДАЛЬНЕ НАВЧАННЯ; ФРЕЙМВОРК.

ABSTRACT

Master Thesis: 76 pp., 21 fig., 2 tab., 41 sources.

Topic: Models and methods of processing speech components in computer vision systems

The purpose of the master's thesis is the development, research and implementation of models and methods of processing speech components for a computer vision system that provide the formation of a multimodal representation of a speech object feature.

The object of research is the processes of processing speech signals in computer vision systems that implement the relationship between audio and visual modalities.

The subject of research is models, methods and algorithms for processing speech components that are used to form, analyze and synthesize visual images based on deep neural networks.

Research results

The work has developed a methodology for multimodal image generation by voice, which is based on deep neural networks and the principles of cross-modal learning.

Conclusion

An architectural solution has been developed based on the Vec2Face framework, which allows generating a speaker's face by voice with a high level of similarity. Results were obtained on the coordination of latent features of different modalities in a single learning environment with identification accuracy.

SPEECH PROCESSING; COMPUTER VISION; DEEP LEARNING; NEURAL NETWORKS; MULTIMODAL GENERATION; SPEAKER IDENTIFICATION; FACE SYNTHESIS BY VOICE; CROSS-MODAL LEARNING; FRAMEWORK.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

CNN - Convolutional Neural Network - згорткова нейронна мережа

MFCC - Mel-Frequency Cepstral Coefficients - Мель-частотні Кепстральні Коефіцієнти

ReLU - Rectified Linear Unit – функція активації

GAN - Generative Adversarial Network - генеративно-змагальна мережа

VAD - Voice Activity - виявлення активності мовлення

GIK - Graphical User Interface - графічний інтерфейс користувача

ВСТУП

Актуальність теми.

У сучасному світі розвиток технологій штучного інтелекту та глибинного навчання зумовив появу нових напрямів інтеграції різних типів інформаційних сигналів — зокрема мовних і візуальних. Системи комп'ютерного зору, які раніше базувалися виключно на аналізі зображень або відеопотоків, дедалі частіше поєднуються з компонентами обробки природної мови, що відкриває нові можливості для створення мультимодальних систем сприйняття, ідентифікації та синтезу образів.

Важливою складовою таких систем є ефективне представлення та інтерпретація мовних ознак, які можуть містити не лише лінгвістичну, але й паралінгвістичну інформацію про особу мовця — його стать, вік, емоційний стан, а також індивідуальні акустичні характеристики. Використання методів глибокого навчання дозволяє моделювати складні нелінійні залежності між акустичними параметрами голосу та візуальними особливостями обличчя, що є основою для побудови систем Voice-to-Face синтезу, біометричної ідентифікації та безконтактного контролю доступу.

Розробка таких моделей передбачає вирішення низки наукових і технічних проблем, пов'язаних з перетворенням акустичних сигналів у числові ознаки, їх адаптацією до вимог нейронних мереж, узгодженням міждоменних просторово-часових представлень та формуванням спільного латентного простору для аудіо- та візуальних модальностей. Вирішення цих завдань має не лише теоретичне, а й прикладне значення для підвищення точності, надійності та універсальності систем комп'ютерного зору нового покоління.

Актуальність теми дослідження зумовлена стрімким розвитком мультимодальних систем штучного інтелекту, які поєднують можливості комп'ютерного зору, обробки мовлення та нейронних мереж. Традиційні системи розпізнавання або синтезу працюють лише з одним типом даних, що

обмежує їх ефективність у реальних умовах. Поєднання аудіо- та візуальних каналів інформації дозволяє створювати більш стійкі моделі, здатні компенсувати втрату даних однієї модальності за рахунок іншої, забезпечуючи таким чином підвищену надійність розпізнавання та ідентифікації.

Особливої ваги тема набуває в умовах зростання ролі біометричних технологій, систем безпеки, автоматизованого контролю особистості, а також у розробці інтерфейсів «людина–машина». Моделі, здатні відновлювати або передбачати візуальні характеристики людини за голосом, мають значний потенціал для використання у криміналістиці, соціальній інженерії, медичних застосуваннях та персоналізованих сервісах.

Наукова новизна роботи полягає в системному підході до розробки та аналізу моделей обробки мовних складових, які поєднують принципи глибинного навчання, методи екстракції акустичних ознак та алгоритми мультимодальної генерації. Практична значущість дослідження визначається можливістю його інтеграції у сучасні системи комп’ютерного зору, що потребують розширеного сенсорного сприйняття.

Метою магістерської роботи є розробка, дослідження та реалізація моделей і методів обробки мовних складових для систем комп’ютерного зору, які забезпечують формування мультимодального представлення ознак об’єкту мовлення.

Об’єктом дослідження є процеси обробки мовних сигналів у системах комп’ютерного зору, які реалізують взаємозв’язок між аудіо- та візуальними модальностями.

Предметом дослідження є моделі, методи та алгоритми обробки мовних складових, що використовуються для формування, аналізу та синтезу візуальних образів на основі глибоких нейронних мереж.

Завдання дослідження

Для досягнення поставленої мети в роботі визначено такі основні завдання:

- Проаналізувати сучасні підходи та методології обробки мовлення в контексті систем комп’ютерного зору.

- Дослідити методи екстракції акустичних ознак.

- Розглянути архітектури нейронних мереж, що застосовуються для аналізу аудіосигналів і побудови мультимодальних моделей.

- Розробити архітектурне рішення та алгоритм інтеграції мовних ознак у систему комп’ютерного зору.

- Створити реалізацію фреймворку для ідентифікації мовця та синтезу зображення обличчя на основі голосового сигналу.

- провести експериментальні дослідження та оцінити ефективність запропонованих моделей і методів.

Методи дослідження

У роботі використано комплекс теоретичних і прикладних методів:

- аналітичні методи — для узагальнення наукових джерел, формалізації процесів обробки мовлення та систематизації підходів до його аналізу;

- методи цифрової обробки сигналів — для перетворення мовного сигналу у частотно-часові представлення;

- методи машинного та глибокого навчання — для побудови і навчання моделей нейронних мереж;

- оптимізаційні алгоритми (Adam, RMSProp, SGD) — для підвищення точності та швидкості навчання;

- експериментальні методи — для тестування працездатності та оцінювання якості синтезованих візуальних образів.

Наукова новизна отриманих результатів

Запропоновано узагальнену модель обробки мовних складових у системах комп’ютерного зору, що базується на глибинних нейронних мережах та забезпечує кросмодальну трансформацію ознак між аудіо- та візуальними просторами.

Удосконалено підхід до представлення мовного сигналу шляхом інтеграції класичних коефіцієнтів MFCC з глибокими вбудовуваннями

акустичних ознак, що підвищує стійкість системи до шуму та варіацій голосу.

Практичне застосування результатів

Розроблені моделі та методи можуть бути використані у системах біометричної ідентифікації для розпізнавання особи за голосом та відновлення її візуального образу та у системах безпеки та контролю доступу, де потрібно підтвердження особи на основі мультимодальних даних.

Структура магістерської роботи. Робота складається зі вступу, трьох розділів та висновків. Загальний обсяг роботи становить 76 сторінок, і містить 21 рисунок, 2 таблиці, список використаних джерел із 54 найменувань.

РОЗДІЛ 1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА МЕТОДОЛОГІЙ ОБРОБКИ МОВЛЕННЯ ДЛЯ СИСТЕМ КОМПЮТЕРНОГО ЗОРУ

1.1. Методологія ідентифікації мовця на основі глибокого навчання

У даній роботі досліджується застосування методів глибокого навчання (ГН) для вирішення задач ідентифікації мовця та реконструкції обличчя за голосовою мовою. Актуальність дослідження зумовлена значним прогресом, продемонстрованим ГН в аналізі мовленнєвих сигналів, зокрема, у виявленні інформативних ознак. Архітектури ГН, зокрема згорткові нейронні мережі (ЗНМ), забезпечують ефективну екстракцію специфічних особливостей (фічеров) із часових та частотних представлень хвильових форм, що дозволяє формувати високоінформативні та щільно закодовані представлення даних (dense feature representations). У цьому контексті, наше дослідження пропонує та верифікує новий підхід до відтворення візуального образу (обличчя) мовця на основі акустичного сигналу, а також до його ідентифікації.

1.1.1. Огляд та використовувані компоненти методології

Спочатку в роботі проводиться систематичний огляд фундаментальних концепцій обробки мовлення (Speech Processing) та пов'язаних із ними прикладних систем. Далі представлено розроблені нами новітні методи обробки мовленнєвих сигналів, що базуються на архітектурах ГН.

Ключовим елементом роботи є фреймворк для ідентифікації мовця та синтезу обличчя (Face Synthesis) на основі голосових даних. Цей фреймворк являє собою каскад послідовно з'єднаних нейронних мереж, кожна з яких виконує критично важливу функцію в процесі перетворення "аудіо-в-зображення" (Audio-to-Image). Структура фреймворку включає такі ключові компоненти:

- Мережа аудіо-вбудовування (Audio Embedding Network) - відповідає за вилучення інваріантних, біометрично значущих векторів ознак

(embeddings) з акустичного сигналу, що кодують унікальні характеристики голосу мовця.

- Мережа кодування (Encoding Network) - трансформує отримані аудіо-вбудовування в проміжне латентне представлення, яке містить інформацію, необхідну для подальшої візуалізації.

- Мережа генерації обличчя (Face Generation Network): Використовує латентне представлення для синтезу передбачуваного зображення обличчя (predicted facial image) мовця. Для цього можуть бути використані генеративно-змагальні мережі (GANs) або варіаційні автокодувальники (VAEs).

1.1.2. Очікувані результати

Експериментальна частина дослідження підтверджує, що акустичний сигнал містить кореляційні ознаки, які можуть бути асоційовані з унікальними антропометричними та візуальними характеристиками обличчя. Отримані результати демонструють, що розроблена глибока нейронна мережа (ГНМ) здатна реконструювати обличчя мовця з високою ступеню достовірності, ґрунтуючись виключно на його голосовому записі. Крім того, наголошується на доцільності інтеграції розробленої мережі розпізнавання мовця (Speaker Recognition Network) з графічним інтерфейсом користувача (ГІК). Така інтеграція забезпечує візуалізацію внутрішніх даних та результатів роботи мережі, підвищуючи інтерпретованість системи та її практичну цінність у біометричних застосуваннях.

1.1.3. Аналіз сучасних підходів та методологій обробки мовлення в контексті систем комп'ютерного зору

Аналіз сучасних підходів та методологій обробки мовлення (ОМ, або NLP для тексту) в контексті систем комп'ютерного зору (КЗ) зосереджується на мультимодальному навчанні (Multimodal Learning). Це сфера, де

інформація з аудіо- чи мовленнєвого каналу використовується для покращення або умовного керування візуальними завданнями, і навпаки.

Сучасна інтеграція ОМ та КЗ реалізується через три основні напрямки: аудіо-візуальне навчання вбудовувань, умовна генерація та мультимодальне розуміння сцени.

1. Аудіо-візуальне навчання вбудовувань (Audio-Visual Embedding Learning)

Цей підхід має на меті знайти спільний латентний простір, де аудіо та візуальні дані, що відповідають одній сутності (наприклад, одній особі), розташовані близько.

Контрастивне навчання (Contrastive Learning) - мережа навчається мінімізувати відстань між вбудовуваннями позитивних пар (аудіо та відео, що відповідають одному мовцю/об'єкту) та максимізувати відстань між негативними парами (аудіо та відео від різних сутностей).

Приклади застосування:

- Верифікація мовця (Speaker Verification) - визначення того, чи належать голос і обличчя одній і тій же особі.

- Асоціація звуку та зображення (Sound-Image Association) - зіставлення звуків навколишнього середовища з об'єктами, які їх генерують (наприклад, гавкіт з собакою).

2. Умовна генерація (Conditional Generation)

Тут мовлення виступає як умова або керуючий сигнал для візуального синтезу чи модифікації.

Генерація обличчя-по-голосу (Voice-to-Face Synthesis) - застосування GAN (Conditional GANs). Вектор голосового вбудовування (як у SincNet або X-Vector) використовується для керування генератором, змушуючи його синтезувати зображення обличчя, яке відповідає ідентичності мовця.

Анімація обличчя (Face Animation) - використання аудіо для синтезу реалістичних рухів губ (lip synchronization) та міміки. Це критично для створення реалістичних "цифрових аватарів" або дубляжу відео.

3. Мультимодальне розуміння сцени (Multimodal Scene Understanding)

Мовлення та текст (як форма ОМ) використовуються для надання контексту та підвищення точності розуміння візуальної сцени.

Візуальна навігація, керована мовою (Language-Guided Navigation) - використання природної мови ("Поверни ліворуч біля червоної машини") для керування агентом (роботом, БПЛА) у віртуальному або фізичному середовищі, вимагаючи об'єднання КЗ для ідентифікації об'єктів та ОМ для розуміння інструкцій.

Обробка питань-відповідей на основі зображень (Visual Question Answering, VQA) - система відповідає на текстові питання про вміст зображення. Хоча це переважно NLP для тексту, воно вимагає складного взаємозв'язку між мовними вбудовуваннями та регіональними візуальними ознаками.

1.2. Методика щільності акустичних даних та перспективи обробки мовлення

Аудіодані характеризуються високою інформаційною щільністю, що містить значний обсяг невикористаної інформації, критично важливої для аналітичних систем. Обробка мовлення (Speech Processing) є глибоко дослідницькою та імплементованою дисципліною, яка займає значне місце в більшості сучасних технологічних рішень. Останні роки ознаменувалися інтеграцією базових функцій обробки мовлення у комерційні продукти, як-от домашні асистенти, системи перетворення голосу в текст (Voice-to-Text) та автоматичні генератори субтитрів.

Однак, більш складні напрямки обробки мовлення, зокрема аудіо-автентифікація користувача (Audio User-Authentication) та діаризація мовців (Speaker Diarization), лише починають досягати рівня точності та надійності, необхідного для їх широкого комерційного впровадження (Consumer-friendly). Діаризація мовців визначається як процес сегментації та

кластеризації зразків мовлення різних осіб в межах єдиного аудіозапису. Комбінація цих акустичних методик із обробкою зображень (Image Processing) суттєво підвищує їхню придатність для використання кінцевим споживачем. Таким чином, сфера обробки мовлення зберігає значний потенціал для подальшого зростання та дослідження нових аспектів.

1.2.1. Основні застосування ота методології обробки мовлення

У науково-дослідній сфері аналізу мовлення традиційно виділяють три фундаментальні та унікальні задачі: розпізнавання мовця (Speaker Recognition), розпізнавання мовлення (Speech Recognition) та синтез мовця (Speaker Synthesis). Незважаючи на те, що кожна з цих проблем вимагає розробки специфічних рішень, вони використовують спільні методології для інтерпретації мовленнєвих сигналів та екстракції інформативних ознак.

Розпізнавання Мовця (Speaker Recognition)

Розпізнавання мовця є задачею порівняння класів і полягає у ідентифікації особи мовця на основі унікальних характеристик, закодованих у голосовому фрагменті. Основні зусилля у цій галузі спрямовані на створення дискримінаційних представлень (unique representations) зразків мовлення та навчання моделей для порівняння з екстрагованими акустичними ознаками. Розпізнавання мовця поділяється на:

- верифікація мовця (Speaker Verification) - перевірка відповідності голосового зразка заявленій особі (задача "один-до-одного").

- ідентифікація мовця (Speaker Identification) - визначення особи невідомого мовця серед наявного набору даних (задача "один-до-багатьох").

Розпізнавання Мовлення (Speech Recognition)

Розпізнавання мовлення охоплює аналіз мовленнєвого сигналу з метою інтерпретації слів, що вимовляються, та їх перетворення у текстове представлення. Ці методи вже глибоко інтегровані в повсякденне життя через такі технології, як домашні асистенти та системи Speech-to-Text.

Профілювання голосу (Voice Profiling) — це спеціалізований напрямок досліджень, що має на меті інференцію фізіологічних та антропометричних атрибутів (таких як стать, вік, етнічна приналежність та особливості обличчя) виключно на основі акустичного сигналу. Існує значний обсяг наукової літератури, який демонструє потенційні, хоча й складні кореляції між цими атрибутами та параметрами голосу. Відомо, що такі фактори, як гендер, вік та фізичні параметри вокального тракту, що формують особливості обличчя, впливають на характеристики людського голосу. Проте, оскільки не всі фактори, пов'язані з обличчям, виявляються в голосі, завдання реконструкції обличчя мовця залишається складним науковим викликом. Людська здатність ментально зіставляти та ідентифікувати голоси та обличчя підтверджує можливість існування глибоких мультимодальних зв'язків, які активно досліджуються.

Синтез мовця (Speaker Synthesis)

Синтез мовця спрямований на вирішення проблеми відтворення вокальних характеристик, виразів та індивідуального тембру конкретної особи на основі екстрагованих ознак. Обмежувальним фактором у цій сфері часто є тривалість та повнота охоплення наданого аудіокліпу. Ідеальний аудіофрагмент повинен мати достатню довжину (наприклад, повноцінне речення) для захоплення максимальної кількості інтонаційних варіацій, необхідних для точної деривації (виведення) характеристик особи мовця.

1.3. Фундаментальні аспекти акустичного сигналу та його представлення

1.3.1. Природа звуку та частотний аналіз

У фізичній акустиці звук визначається як поширення змін тиску (або механічних коливань) у середовищі, які сприймаються сенсорними системами. Ці коливання фіксуються вимірювальними пристроями, зокрема мікрофонами, і є основою для аналізу частотних характеристик сигналу.

Для декомпозиції складного сигналу застосовується перетворення Фур'є (Fourier Transform, FT) — лінійне інтегральне перетворення, яке розкладає вхідний часовий сигнал на його конститутивні частотні компоненти (складові). Цей метод також надає інформацію про амплітуду (магнітуду) кожної частоти в спектрі. Завдяки FT, будь-який сигнал може бути представлений як сума синусоїдальних хвиль.

Канонічне визначення безперервного перетворення Фур'є для функції *g(t)* має вигляд:

$$\mathcal{F}\{g(t)\} = G(f) = \int_{-\infty}^{\infty} g(t) e^{-2\pi i f t} dt$$

у цьому виразі *f* позначає частоту, а *t* — час.

У галузі аналізу сигналів перетворення Фур'є є ключовим інструментом для:

- компресії даних - зберігаючи лише значущі частотні компоненти.

- ізоляції шуму - ідентифікуючи та фільтруючи небажані частотні діапазони.

- ідентифікації частотних патернів.

FT використовує фундаментальний принцип, що будь-яка складна хвильова форма формується як суперпозиція простіших гармонічних функцій (синусоїд та косинусоїд).

З оберненого перетворення Фур'є (Inverse Fourier Transform, IFT) можна відновити вихідний часовий сигнал виключно з його частотного спектру (G(f)). IFT є особливо корисним для реконструкції сигналів, які були модифіковані у частотній області (наприклад, після фільтрації).

1.3.2. Візуальне представлення розподілу за допомогою спектрограми

Спектрограма — це візуальне представлення розподілу амплітуди (енергії) сигналу як функції часу та частоти. Вона відображає потужність частотних компонентів (отриманих, наприклад, через періодограму або спектр потужності) вздовж часової осі.

Спектрограми перетворюють числові дані про частотні патерни на двовимірний візуальний образ, що є більш зручним для сприйняття людиною та обробки зображень. Це перетворення дозволяє застосовувати такі методи, як згорткові нейронні мережі (ЗНМ) та інші алгоритми навчання на основі зображень, оскільки задача аналізу мовлення може бути ефективно зведена до задачі класифікації зображень (Image Classification). Рисунок 1.1 це приклад візуалізації спектрограми.

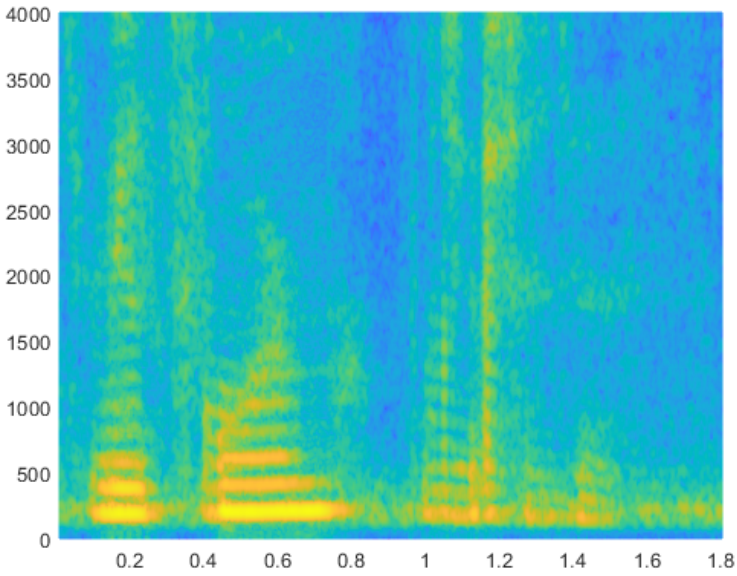

Рис. 1.1. Спектрограма перетворення Фур'є

Однак, використання спектрограм має суттєве обмеження: фазова інформація (phase information) вихідного сигналу втрачається в процесі побудови спектрограми. Ця втрата унеможливлює точну реконструкцію оригінального часового сигналу, ґрунтуючись лише на даних спектрограми (амплітуді та частоті).

У подальших розділах буде розглянуто роль перетворення Фур'є як критичного підготовчого етапу для обчислення Мел-частотних кепстральних коефіцієнтів (MFCCs), які є стандартним та високоефективним набором ознак для аналізу мовленнєвих сигналів людини.

1.4. Особливості методу Mel-Frequency Cepstral Coefficients (MFCC)

1.4.1. Теоретичні основи MFCC

Mel-Frequency Cepstral Coefficients (MFCCs) – це один з найбільш поширених методів представлення звукового сигналу у вигляді ознак (features), які використовуються для автоматичного розпізнавання мовлення, музики та інших аудіоаналітичних задач.

MFCC моделюють те, як людське вухо сприймає звук, а не те, як сигнал виглядає фізично. Вони ґрунтуються на мел-шкалі частот, яка відображає нелінійність людського слуху:

- на низьких частотах ми розрізняємо невеликі зміни частоти,

- на високих – слух менш чутливий.

Мел-частотний кепстр (Mel-frequency cepstrum) є репрезентацією короткострокового спектра потужності звукового сигналу. Мел-частотні кепстральні коефіцієнти (MFCCs) — це набір коефіцієнтів, які формують цей кепстр і є стандартним набором ознак у системах розпізнавання мовлення та ідентифікації мовця.

Розглянемо процес як утворюються MFCC:

1. Розбиття сигналу на фрейми – короткі інтервали (наприклад, 20–40 мс).

2. Віконна функція (наприклад, Hamming window), щоб уникнути різких переходів між фреймами.

3. Перетворення Фур'є (FFT) – для отримання спектру сигналу.

4. Мел-фільтрбанк – набір трикутних фільтрів, розташованих відповідно до мел-шкали, щоб підкреслити важливі для слуху ділянки спектра.

5. Логарифм енергії кожного фільтра – наближає нелінійність слухового сприйняття.

6. Дискретне косинусне перетворення (DCT) – перетворює логарифмічні енергетичні коефіцієнти у компактне представлення (cepstral coefficients).

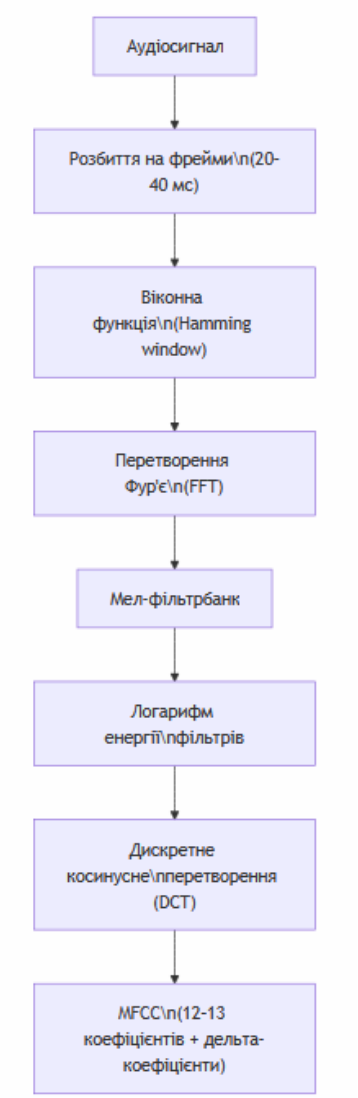

Рис. 1.2. Основні етапи процесу отримання MFCC

В результаті зазвичай отримують 12–13 основних MFCC для кожного фрейму + коефіцієнт енергії. Для кращого відображення динаміки додають дельта-коефіцієнти (зміна ознак у часі).

Процес отримання MFCCs включає послідовну низку операцій, що відображають як фізичні властивості сигналу, так і психоакустичні особливості людського слуху:

1. Частотний Аналіз (Перетворення Фур'є).

Початковий сигнал піддається перетворенню Фур'є (зазвичай Швидкому перетворенню Фур'є, FFT) для отримання його спектра потужності. На цьому етапі частоти спрощуються та розділяються для подальшої обробки та відображення.

2. Відображення на Mel-Scale Mapping.

Отримані частоти зі спектра потужності відображаються на Мел-шкалу (Mel Scale). Мел-шкала є нелінійною функцією, що моделює залежність між сприйнятою людиною висотою звуку (пітчем) і фізичною частотою (Гц). Вона є лінійною на низьких частотах і логарифмічною на високих, що відображає чутливість слуху. На рисунку 1.3 візуально зображено шкалу Мела.

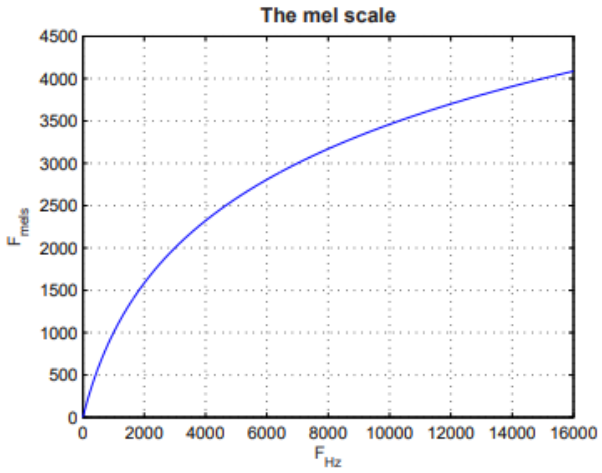

Рис. 1.3. Візуалізація Mel Scale

Найбільш поширеною та загальновизнаною функцією для перетворення частоти f (у Гц) в Мел-частоту m є формула:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right)$$

3. Застосування трикутних віконних функцій (Triangular Windowing).

Мел-частоти фільтруються за допомогою набору перекривних трикутних віконних функцій (Mel-filter bank). Віконні функції є фільтрами

амплітуди, які зважують (згладжують) спектральні значення, що потрапляють у визначений частотний діапазон, виконуючи функцію зменшення спектральних спотворень (tapering).

Трикутна віконна функція зазвичай має вигляд рівняння і візуальну форму, що подана на рисунку 1.4:

$$w[n] = 1 - \left| \frac{n - n_c}{L/2} \right|, \quad 0 \leq n \leq N$$

де n_c — центр вікна, L — його ширина, а N — кількість вибірок.

Це забезпечує візуальне представлення у формі рівнобедреного трикутника над N частотними вибірками.

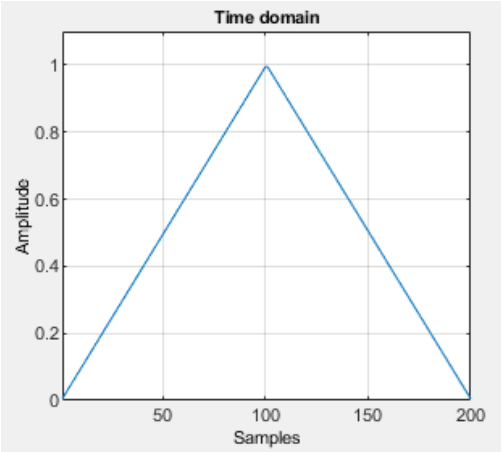

Рис. 1.4. Візуалізація трикутної віконної функції

4. Логарифмування та дискретне косинусне перетворення (DCT).

Після фільтрації застосовується логарифмічна функція до потужностей Мел-частот. Логарифмування знижує динамічний діапазон і імітує нелінійну чутливість слухового апарату. Наступним кроком є застосування дискретного косинусного перетворення (DCT) до логарифмічних потужностей. DCT ефективно декорелює коефіцієнти і концентрує значущу інформацію у перших кількох коефіцієнтах, які й становлять MFCCs.

MFCCs є амплітудами спектра, отриманого після DCT.

1.4.2. Обмеження та застосування MFCC

Методика MFCC є високо перцептивно-орієнтованою (perceptually engineered), що не гарантує оптимальності отриманих представлень для всіх завдань обробки мовлення.

Основне відоме обмеження MFCCs полягає у їхній низькій стійкості до шуму. У середовищах із високим рівнем шуму їхня ефективність суттєво знижується. Це вимагає обов'язкового застосування нормалізації та методів шумозаглушення (de-noising) для мінімізації інтерференції. Сучасні дослідження включають розробку спеціалізованих нейронних мереж для адаптивного видалення шуму із вхідного сигналу перед екстракцією ознак.

Як ілюстрація, MFCC-спектрограма являє собою двовимірне зображення, де горизонтальна вісь представляє час, а вертикальна вісь — унікальні MFCC-коефіцієнти. Кожен часовий сегмент (зразки) формує стовпець, а кожен рядок відповідає певному коефіцієнту.

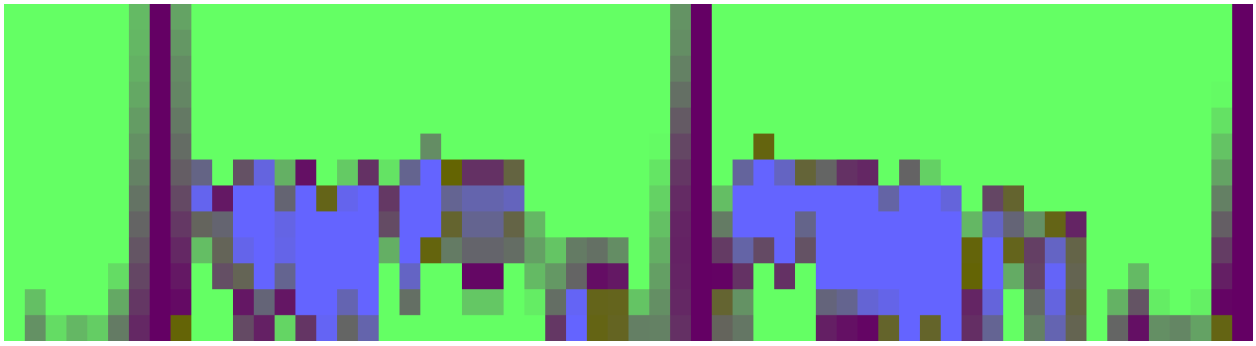

Рис. 1.5. Представлення MFCC фрази «Привіт», вимовленої двічі

У програмній реалізації, як показано на прикладі (рис. 1.5), MFCC-ознаки можуть бути вилучені в реальному часі. Наприклад, використання функцій, подібних до navigator.mediaDevices.getUserMedia(...) та Meyda.createMeydaAnalyzer(...), дозволяє захопити та проаналізувати аудіопотік. Відображення MFCCs часто активується лише тоді, коли

середньоквадратичне значення (RMS) вхідного аудіосигналу перевищує встановлений поріг, що забезпечує аналіз лише активного мовлення.

Висновки до розділу

У першому розділі здійснено комплексний аналіз теоретичних і методологічних основ обробки мовних сигналів у контексті систем комп’ютерного зору. Розглянуто сучасні підходи до ідентифікації мовця, що базуються на методах глибокого навчання, та окреслено ключові компоненти архітектур, які забезпечують високу точність розпізнавання. Особливу увагу приділено методології побудови моделей на основі нейронних мереж, здатних до самоадаптації та узагальнення акустичних закономірностей. Окрему увагу приділено методу Mel-Frequency Cepstral Coefficients (MFCC), який залишається одним з базових інструментів у задачах обробки мовлення. Досліджено його математичну основу, переваги у відображенні психоакустичних характеристик сприйняття звуку, а також недоліки, пов’язані з обмеженою здатністю до моделювання складних акустичних контекстів.

РОЗДІЛ 2. МОДЕЛІ ТА МЕТОДИ НЕЙРОННИХ МЕРЕЖ ДЛЯ ПРОЦЕСІВ ОБРОБКИ МОВНИХ СКЛАДОВИХ

2.1. Концептуальні основи нейронних мереж у глибокому навчанні

2.1.1. Архітектура та функціональні елементи нейронних мереж

Нейронна мережа (НМ) є конгломератом взаємопов'язаних вузлів (nodes) або перцептронів, що імітує біологічну структуру мозку, де нейрони та синаптичні зв'язки забезпечують обчислення та прийняття рішень. Кожен вузол має настроювані ваги (adjustable weights), які модифікуються в процесі навчання.

Перцептрон являє собою базовий елемент НМ, що складається з вхідних зв'язків, активаційної функції та виходу (рис. 2.1).

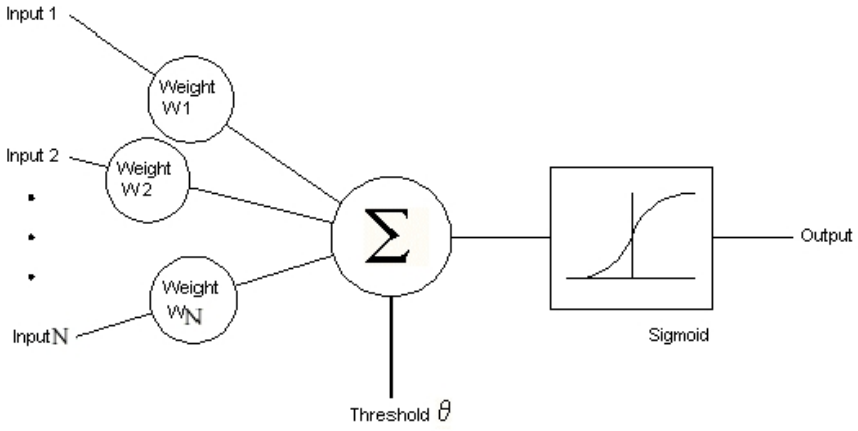

Рис. 2.1. Схема базової структури вузла

Функція активації (Activation Function) — це, як правило, диференційовна функція, яка визначає вихідне значення відповідного вузла для заданого входу. Модифікація ваг відбувається за допомогою алгоритму градієнтного спуску (gradient descent), що використовує градієнт функції втрат. Функція активації зазвичай є диференційованою функцією, яка визначає вихід кожного відповідного вузла для заданого входу за допомогою

алгоритму градієнтного спуску. Зазвичай вони мають форму Sigmoid, Softmax, TanH, як показано на рисунку 2.2.

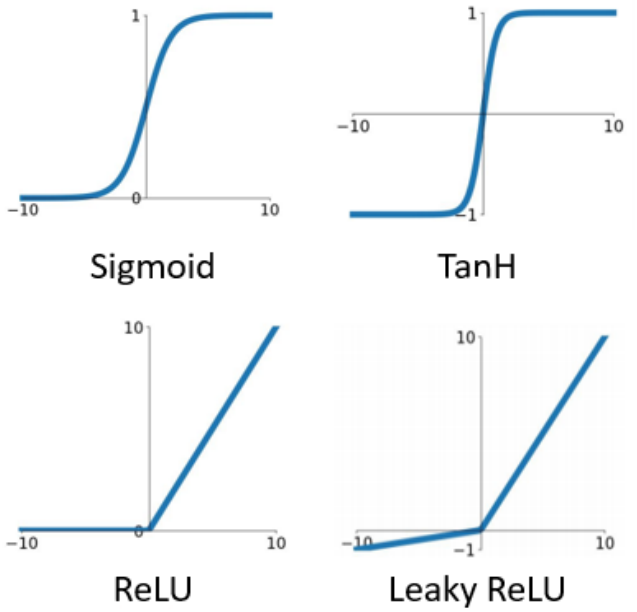

Рис. 2.2. Графічні представлення поширених функцій активації

Типові активаційні функції включають:

Сигмоїда (Sigmoid) - генерує вихід у діапазоні [0,1], що корисно для інтерпретації як імовірності. Основні недоліки: проблема зникаючого градієнта (vanishing gradient), ненульова центрованість виходу та повільна збіжність.

- Гіперболічний тангенс (TanH) - вихід у діапазоні [−1,1]. Зазвичай перевершує Сигмоїду через центрованість виходу навколо нуля, але також страждає від проблеми зникаючого градієнта.

- ReLU (Rectified Linear Unit) - функція max(0,x). Забезпечує значне прискорення збіжності (за оцінками, до шести разів швидше, ніж TanH) завдяки своїй простоті. Не страждає від проблеми зникаючого градієнта у позитивній області. Критична проблема: "вмираючі ReLU" (Dying ReLU), коли вага вузла зміщується до стану, в якому він ніколи більше не активується.

- Leaky ReLU - усуває проблему "вмираючих ReLU" шляхом введення невеликого від'ємного нахилу для від'ємних входів (max(ax,x), де a — мале число, наприклад 0.01). Це гарантує, що вузол не "помре", оскільки його вага завжди може бути оновлена.

2.1.2. Парадигми навчання та регуляризація

Нейронні мережі зазвичай класифікуються за методами навчання:

1. Навчання з учителем (Supervised Learning).

Модель тренується на розмічених даних, де кожен вхідний зразок має відповідний правильний вихід. Усі запропоновані в цьому дослідженні роботи в області аналізу мовця відносяться до цієї парадигми.

2. Навчання без учителя (Unsupervised Learning).

Модель отримує лише вхідні дані без міток правильних виходів і повинна самостійно виявляти приховані структури чи патерни.

Dropout (Викидання) — це техніка регуляризації, застосовувана до шарів під час тренування. Вона полягає у випадковому тимчасовому вимкненні вузлів шару з імовірністю p. Цей механізм запобігає перенавчанню (overfitting) моделі та унеможливлює спів-адаптацію вузлів до конкретних вхідних даних, сприяючи формуванню більш стійких та узагальнюючих ознак у мережі.

Розглянемо принцип роботи Dropout:

1. Механізм дикидання

Під час кожного етапу тренування (прямого та зворотного поширення) Dropout випадковим чином тимчасово виключає (встановлює на нуль) певну частку (зазвичай від 10% до 50%) нейронів (разом з їхніми вхідними та вихідними зв'язками) у певному шарі.

Якщо частка виключення (dropout rate) встановлена на p=0.5 (50%), це означає, що під час кожної ітерації кожен нейрон у цьому шарі має ймовірність p бути вимкненим.

2. Створення "розрідженої" мережі

Кожна ітерація тренування використовує випадково вибраний "розріджений" варіант (sub-network) оригінальної мережі. Таким чином, замість навчання однієї великої мережі, Dropout ефективно навчає ансамбль (множину) менших, незалежних мереж, які спільно використовують ваги.

3. Запобігання коадаптації

Основна перевага Dropout полягає у запобіганні коадаптації (co-adaptation). Це явище, коли певні нейрони в сусідніх шарах починають сильно залежати один від одного для виконання конкретного завдання. Випадкове вимкнення нейронів змушує будь-який нейрон бути більш надійним і менш залежним від присутності конкретних інших нейронів, оскільки його "сусіди" можуть бути відсутніми на наступному кроці.

Таблиця 2.1.

Фази застосування

Фаза	**Дія Dropout**	**Призначення**
Тренування	Активний. Випадково вимикає нейрони згідно з вірогідністю p.	Регуляризація, запобігання перенавчанню, навчання надійних ознак.
Тестування/Висновок	Неактивний. Всі нейрони присутні.	Використовувати повну, навчену потужність мережі для точного прогнозування.

Оскільки під час тренування нейрони присутні лише з ймовірністю (1−p), їхні вихідні значення в середньому менші. Якби під час тестування використовувалася вся мережа без коригування, вихідні значення були б значно більшими (відбулося б зміщення).

Щоб компенсувати це, під час тестування ваги нейронів у шарі, де застосовувався Dropout, масштабуються (множаться) на коефіцієнт (1−p). Наприклад, якщо p=0.5, ваги множаться на 0.5.

У сучасних реалізаціях, таких як PyTorch або TensorFlow, часто використовується інвертований Dropout (inverted dropout), де масштабування

на 1/(1−p) відбувається вже на етапі тренування, що дозволяє залишити мережу без змін на етапі тестування.

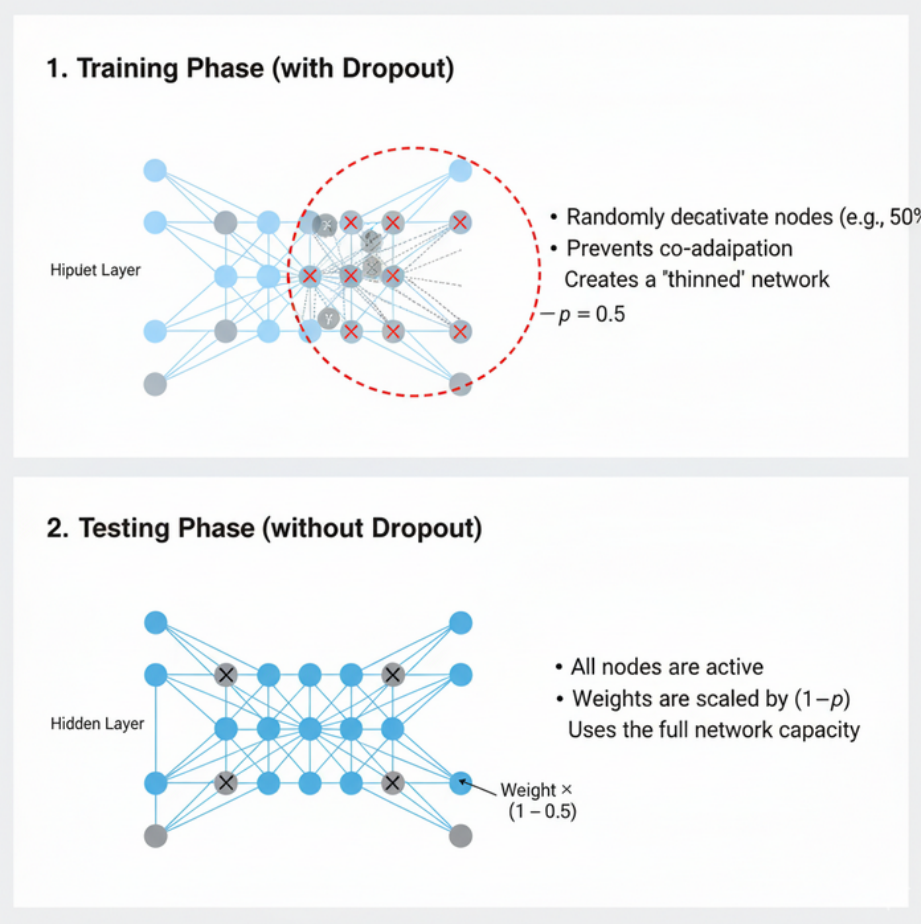

Рис. 2.3. Графічна інтерпретація техніки Dropout

Техніка Dropout візуально демонструє, як випадкове тимчасове видалення нейронів змінює структуру мережі під час тренування.

1. Повна мережа (навчання та тестування)

Перше зображення на рис. 2.3 показує повну нейронну мережу (Full Network) у стандартному стані. Кожен нейрон (коло) у прихованих шарах (Hidden Layers) з'єднаний з усіма нейронами в попередньому та наступному шарах.

Навчання: Без Dropout мережа схильна до коадаптації.

Тестування: Використовується повна структура.

2. Мережа з Dropout (лише під час навчання)

Друге зображення на рис. 2.3 ілюструє застосування Dropout. Під час кожного кроку тренування певна частка нейронів у прихованих шарах випадково викидається (тимчасово вимикається).

Викидання: Нейрони, які були вимкнені, зазвичай закреслені або зникають, а їхні вхідні та вихідні зв'язки видаляються на цьому кроці.

Ефект: Це створює розріджену підмережу (Thinned Sub-network). Мережа стає меншою, і вона вимушена знаходити більш надійні та незалежні ознаки для передачі інформації.

3. Компенсація ваг (тестування)

Хоча під час тестування Dropout не застосовується, і всі нейрони присутні (як у пункті 1), ваги нейронів, до яких застосовувався Dropout, мають бути масштабовані (помножені на коефіцієнт $1-p$).

Наприклад, якщо ймовірність викидання $p=0.5$ (50%), тоді під час тестування вихід кожного нейрона множиться на 0.5 (або, що частіше, його ваги були помножені на 2, тобто на $1/(1-p)$, під час тренування, як у випадку інвертованого Dropout).

На етапі тестування ми повертаємося до повної мережі, але з адаптованими вагами, які враховують, що нейрони були відсутні 50% часу тренування.

2.1.3. Структурні компоненти та оптимізація

Повністю з'єднаний шар (Fully Connected Layer) — це шар, у якому кожен вузол має зв'язок із усіма вузлами попереднього шару. Це дозволяє інтегрувати всі ознаки попереднього шару в кожну наступну одиницю, що може підвищити продуктивність.

У цьому шарі виконується основна математична операція: зважена сума входів з попереднього шару, до якої додається зсув (bias), а потім застосовується активаційна функція. Однак така архітектура є обчислювально витратною при масштабуванні. На рисунку 2.4 зображено, як може виглядати проста повністю зв'язана мережа.

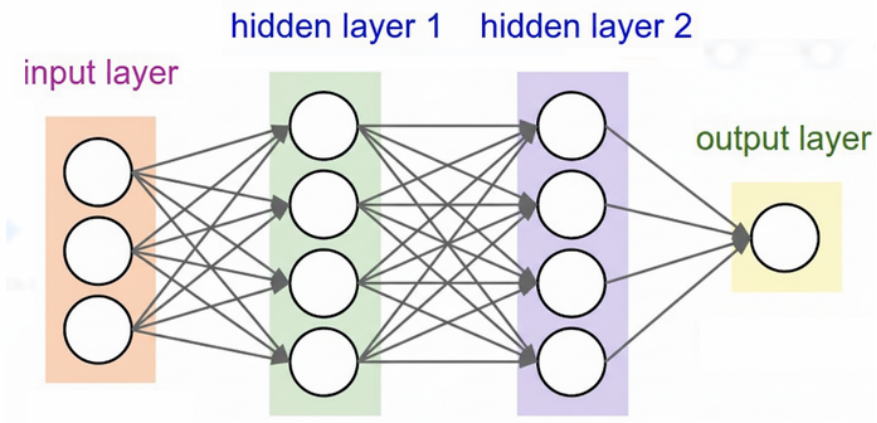

Рис. 2.4. Приклад структури FC-шарів

FC-шари є універсальними, але в сучасних глибоких нейронних мережах їх часто використовують у кінці архітектури, після того, як більш ефективні шари (наприклад, згорткові) вилучили первинні ознаки.

Функція втрат (Loss Function) використовується для оцінки ефективності виходу НМ (у випадку навчання з учителем) шляхом порівняння передбаченого результату з фактичним (еталонним) значенням та обчисленням показника помилки.

Зворотне поширення помилки (Back-propagation) — це алгоритм, який використовується для обчислення градієнта функції втрат щодо ваг мережі. На основі цього градієнта функція оптимізації (наприклад, Stochastic Gradient Descent, Adam) коригує ваги НМ, щоб мінімізувати помилку та наблизити модель до кращого оптимуму в просторі параметрів.

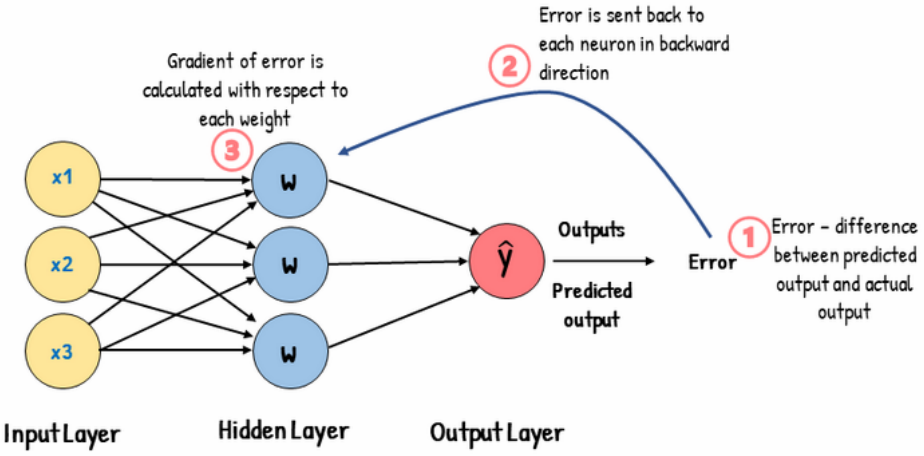

Рис. 2.5. Алгоритм зворотного поширення помилки

Навчання зазвичай відбувається ітеративно у пакетах (batches) протягом заданої кількості епох (epochs). Кожна ітерація включає тестування, обчислення втрат, зворотне поширення та оновлення ваг. Нейронні мережі, побудовані лише на цих базових елементах, називаються прямопоширеними мережами (Feed-Forward Networks).

2.2. Принципи глибокого навчання та згорткові нейронні мережі

2.2.1. Особливості глибокого навчання

Глибоке навчання (Deep Learning, DL) є широкою категорією архітектур нейронних мереж, які характеризуються наявністю великої кількості шарів (глибини), що забезпечують ієрархічне перетворення сирих вхідних даних у високорівневі ознаки. Основна мета DL полягає у рафінуванні (refining) вхідної інформації для виконання складних операцій, таких як класифікація або регресія.

Традиційні прямопоширені мережі (Feed-Forward Networks), які покладаються переважно на повністю з'єднані шари (Fully Connected Layers, FC), ефективні для задач з низьковимірними або одновимірними векторними даними (наприклад, лінійна регресія). Однак, вони погано масштабуються при роботі зі складними багатовимірними даними, зокрема із зображеннями. Наприклад, кольорове зображення (RGB) складається з трьох матриць, і подання його як єдиного одновимірного вектора для FC-шару призводить до експоненційного зростання кількості ваг на кожному вузлі, що робить традиційні НМ обчислювально неефективними для таких проблем.

2.2.2. Згорткові нейронні мережі (CNN)

Згорткова нейронна мережа (Convolutional Neural Network, CNN) — це спеціалізована форма глибокого навчання, яка оптимально підходить для екстракції та обробки ознак у багатовимірних даних, таких як зображення та спектрограми.

Ключова відмінність CNN від традиційних мереж полягає у послідовності шарів екстракції ознак — згорткових (Convolutional) та пулінгових (Pooling) — які передують стандартній FC-мережі. CNN здатна захоплювати просторові та часові кореляції шляхом збереження структурної цілісності вхідних даних, на відміну від прямопоширених мереж, які сплющують (flattening) вхідні дані в одновимірний вектор. Наявність принаймні одного згорткового шару є необхідною умовою для класифікації мережі як згорткової.

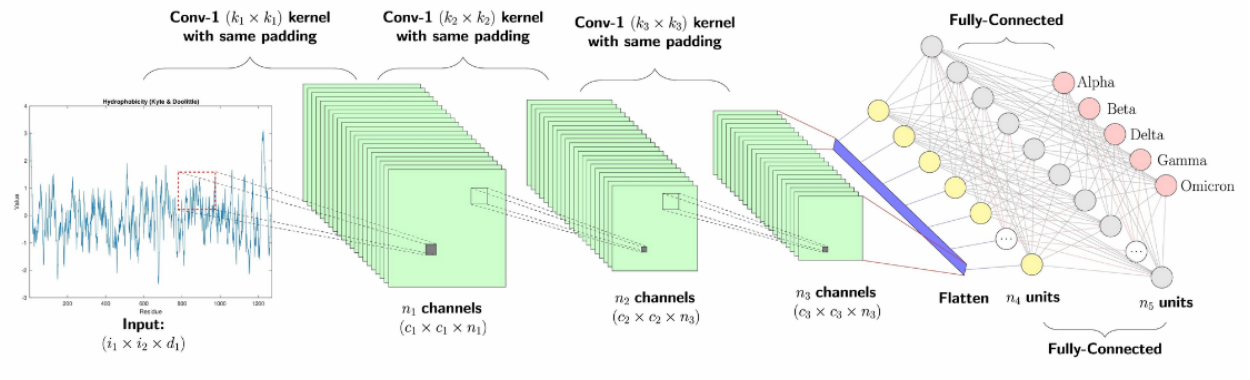

Рис. 2.6. Архітектура згорткової нейронної мережі

Згортковий шар (Convolutional Layer) виконує операцію згортки (convolution), яка математично є ковзним скалярним добутком (sliding dot product) між двома матрицями:

1. Ядро (Kernel) або фільтр - невелика матриця навчальних параметрів (ваг). Ядро має менші розміри у площині (x та y), але таку ж глибину (z, кількість каналів), як і вхідна область.

2. Рецептивне поле (Receptive Field) - обмежена область вхідної матриці, що обробляється в даний момент.

Операція згортки полягає у множенні ядра на відповідну область рецептивного поля. Ядро сканує вхідне зображення, зміщуючись на визначену відстань (крок, stride), і генерує меншу матрицю згорнутих ознак (convolved features), яка є закодованим представленням обробленого регіону.

Ядро, таким чином, діє як фільтр, що виявляє специфічні патерни (наприклад, краї або текстури).

Пулінгові шари (Pooling Layer) відповідають за зменшення просторового розміру (downsampling) вихідних даних мережі. Пулінг сприяє зниженню дисперсії між вхідними даними, згладжуючи дрібні деталі та поглинаючи шум, що підвищує інваріантність моделі до невеликих зсувів або деформацій.

Макс-пулінг (Max Pooling) вибирає максимальне значення з кожного регіону. Зазвичай демонструє кращу продуктивність, оскільки ефективніше мінімізує вплив шуму.

Середній пулінг (Average Pooling) обчислює середнє значення елементів у регіоні.

Після декількох послідовних операцій згортки та пулінгу результуюча матриця ознак сплющується (перетворюється на одновимірний вектор) і передається на вхід до повністю з'єднаної мережі. Ця фінальна частина функціонує як традиційна прямопоширена мережа, виконуючи фінальну класифікацію або операцію виведення ознак на основі високоабстрагованого представлення.

2.3. Архітектурне рішення та дизайн мереж для реконструкції обличчя та ідентифікації мовця

У цьому розділі представлено опис двох основних компонентів запропонованої архітектури: мережі вбудовування (Embedding Network) для екстракції акустичних ознак, необхідних для генерації обличчя, та дизайну користувацького інтерфейсу (GUI) для ідентифікації мовця.

2.3.1. Мережа вбудовування акустичних ознак (Embedding Network)

Аналіз мовленнєвих акустичних даних неминуче стикається із проблемою зашумленості (noise pollution), де важливі особливості мовця

маскуються сторонніми звуками. Подібно до того, як людський слуховий апарат природно фільтрує шум, згорткові шари (Convolutional Layers) у нейронних мережах використовуються для емуляції фільтрації та акцентування інформативних ознак.

Для забезпечення оптимальної продуктивності екстрактора ознак наша реалізація використовує архітектуру з п'ятьма послідовними згортковими шарами. Це дозволяє ієрархічно розширювати та рафінувати дані, вбудовані у вхідний вектор MFCC (Mel-Frequency Cepstral Coefficients).

Між згортковими шарами інтегровані такі компоненти:

- усереднюючий пулінг (Average Pooling) - використовується після кожного згорткового шару. Цей метод допомагає стабілізувати ознаки, запобігаючи надмірному акценту на екстремальних значеннях, та сприяє рівномірному розподілу шуму у векторі ознак.

- Пакетна нормалізація (Batch Normalization) - застосовується для прискорення швидкості навчання та поліпшення потоку градієнта між шарами, що підвищує стабільність тренування.

- Функція активації ReLU - використовується завдяки її властивості швидкої збіжності та простоті обчислень.

Завдяки такій комбінації методів, модель генерує тонко налаштовані та високодискримінативні вбудовування акустичних даних мовця.

2.3.2. Представлення потоку даних

Потік даних запропонованої моделі починається з вихідного аудіофайлу (хвильової форми).

Попередня Обробка (Pre-processing)

- Нормалізація та сегментація. Для видалення сторонніх даних і забезпечення консистентності, застосовується нормалізація для згладжування амплітудних піків. Також використовуються порогові значення амплітуди для виділення активного мовлення шляхом видалення сегментів тиші.

- Форматування. Через варіативність часової тривалості кліпів, аудіосигнали поділяються на рівні часові фрагменти та трансформуються у часово-інваріантний формат.

- Зниження частоти дискретизації (Downsampling). Хоча більшість сучасних аудіофайлів мають високі частоти дискретизації (наприклад, 44100 вибірок на секунду), використання зниженої частоти, наприклад, 16000 вибірок, є компромісним рішенням. При цьому зберігається життєво важлива інформація, необхідна для ідентифікації, водночас суттєво зменшується час обробки.

Екстракція MFCC

Для компресії сигналу та збереження критичної інформації застосовується екстракція MFCCs. MFCCs є оптимальним вибором для екстракції ознак, пов'язаних з людиною, оскільки їхня спеціалізована функція відображення (Мел-шкала) найбільш точно емулює сприйняття людським вухом частотного діапазону. Це робить їх ідеальними для встановлення кореляції між фізичними характеристиками та вокальними ознаками.

Недоліком є те, що MFCCs є статичними за своєю природою і недостатньо стійкі до викидів (outliers), які погано узгоджуються з Мел-шкалою.

Отриманий вбудований MFCC-вектор подається в мережу вбудовування. Мережа конулюється (згортається) у вектори більшої довжини, щоб виявити глибшу інформацію з максимальних значень вихідних MFCCs. Фінальний вихідний шар мережі редукує розмір вектора до початкового розміру вхідних ознак, який потім слугує входом для мережі перетворення голосу в обличчя.

2.3.3. Мережа відображення ознак (Feature Mapping Network)

Мережа кодування (Encoding Network) використовується для перетворення вектора акустичних ознак у вектор ознак обличчя. Для цього завдання використовується повністю з'єднана прямопоширена мережа (Fully

Connected Feed-Forward Network). Використання FC-шарів є доцільним, оскільки вони дозволяють всім вхідним ознакам впливати на вихід кожного вузла, що підвищує дискримінативну здатність мережі.

Вхідний шар відповідає розміру акустичних ознак, а вихідний — розміру вектора ознак обличчя. Приховані шари мають конусоподібну структуру зростання за кількістю вузлів (наприклад, [128,256,512]). Ця структура призначена для підвищення складності взаємодії ознак і формування кращих зв'язків між вокальними та візуальними ознаками.

Для активації та оптимізації використовується Leaky ReLU для прискорення збіжності та запобігання деактивації вузлів. Для тренування обрано оптимізатор Adam завдяки його високій обчислювальній ефективності та швидкій збіжності, особливо при роботі з великими наборами ознак.

Для тренування використовується комбінована функція втрат, що поєднує кутову втрату (angular loss) і SoftMax-втрату:

$$\text{Loss} = (\text{angular\_loss} \times 1.0) + (\text{soft\_max\_loss} \times 0.5)$$

Цей модифікований підхід до функції втрат спрямований на посилення дискримінативності ознак та покращення їх представлення після кожної ітерації оновлення ваг.

2.3.4. Фреймворк Vec2Face

Фреймворк Vec2Face є кінцевим етапом. У ньому вектор ознак голосу, згенерований мережею кодування, проходить через лінійні шари, які розширюють його до вектора ознак обличчя. Це перетворення відбувається на основі навчених ваг, які акцентують ті характеристики голосового вектора, що корелюють із візуальними ознаками обличчя. Отриманий вектор

ознак обличчя потім використовується для генерації фінального зображення обличчя.

Фреймворк Vec2Face (є ключовим компонентом архітектури, призначеним для трансформації високоабстрагованого акустичного представлення (вектора голосу) у візуальне представлення (вектор обличчя), яке потім використовується для синтезу зображення обличчя мовця.

Vec2Face являє собою кінцевий модуль синтезу, який функціонує як декодер у загальній схемі "голос-в-обличчя", базуючись на попередньому етапі — мережі Кодування Ознак (Feature Mapping Network).

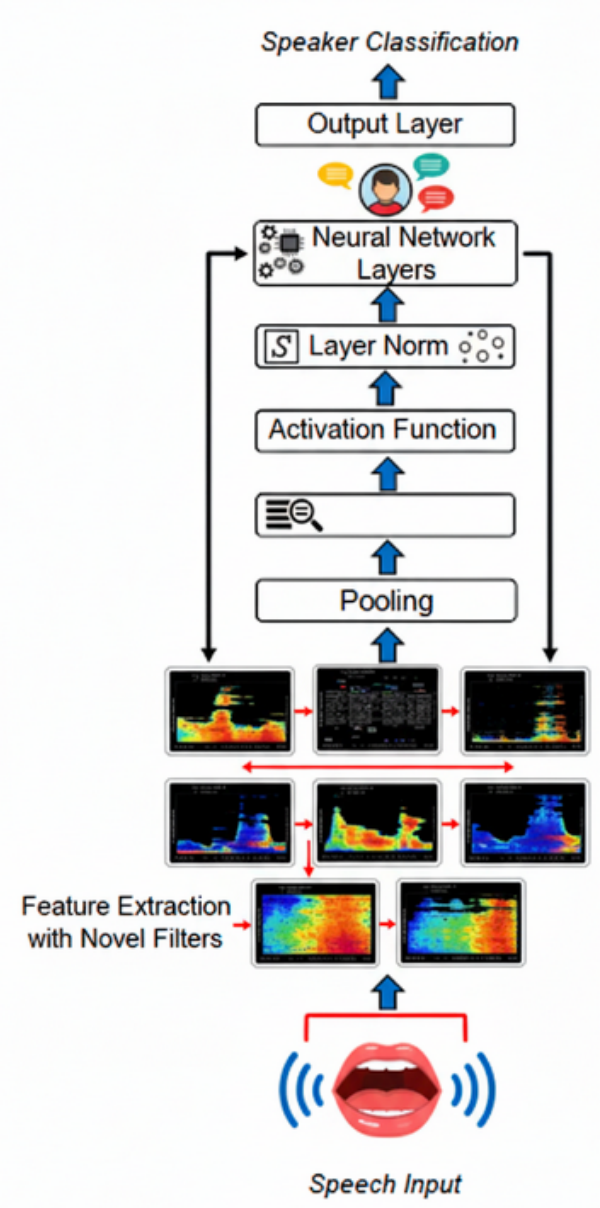

Рис. 2.7. Алгоритмічна схема роботи фреймворку

Фреймворк Vec2Face, по суті, є мережею відображення (Mapping Network), яка отримує на вхід вектор, що містить дискримінативні акустичні ознаки мовця, та генерує на виході вектор, який містить параметричні ознаки обличчя.

1. Вхідний Вектор (Voice Feature Vector)

На вхід Vec2Face подається тонко налаштований, високоабстрагований вектор ознак голосу (Voice Feature Vector), що є виходом мережі кодування (Feature Mapping Network). Цей вектор вже пройшов процес очищення, згортки та відображення, і, відповідно до гіпотези дослідження, містить ключові кореляційні ознаки, які пов'язані з фізичними характеристиками мовця (такими як стать, вік, форма голови тощо).

2. Лінійні Шари (Linear Layers)

Vec2Face зазвичай складається з одного або декількох повністю з'єднаних (Fully Connected, FC) шарів (також званих лінійними шарами). Ці шари виконують основну роботу з трансформації простору ознак:

- Розширення простору ознак: Основна функція полягає у розширенні (upscaling) вхідного вектора голосу до вектора ознак обличчя (який зазвичай має більшу розмірність для кодування складної візуальної інформації).

- Використання навчених ваг: Трансформація керується навченими вагами мережі. Ці ваги були оптимізовані під час тренування, щоб акцентувати (підвищувати значущість) ті компоненти вхідного акустичного вектора, які найсильніше корелюють із цільовими візуальними характеристиками.

3. Вихідний Вектор (Face Feature Vector)

Виходом фреймворку є вектор ознак обличчя (Face Feature Vector). Це параметричне представлення обличчя. Залежно від фінальної архітектури, цей вектор може кодувати різні аспекти:

- латентний простір обличчя: Вектор може бути точкою у латентному просторі попередньо навченої генеративної моделі (наприклад, StyleGAN або VAE/GAN, навченої на великому наборі облич).

- параметри моделі: Вектор може кодувати параметри певної моделі обличчя (наприклад, 3DMM-коефіцієнти для форми та текстури обличчя).

Сам по собі Vec2Face не генерує зображення; він генерує вектор. Для створення фінального зображення обличчя вихідний вектор Vec2Face повинен бути переданий у Генеративний Модуль (Generative Module).

Вектор ознак обличчя діє як керуюча умова (controlling latent code) для генератора. Генеративний Модуль (наприклад, архітектура GAN-декодера) використовує цей вектор для синтезу високоякісного, фотореалістичного зображення, яке відображає закодовані риси обличчя.

Таким чином, Vec2Face виконує критичну роль моста між акустичною модальністю та візуальною модальністю, завершуючи процес перетворення:

2.4. Застосування фреймворку для екстракції ознак та алгоритм роботи системи ідентифікації мовлення

2.4.1. Фреймворк для екстракції ознак з хвильової форми

Фреймворк являє собою архітектуру глибокої нейронної мережі (ДНМ), спеціалізовану для виконання ідентифікації мовця безпосередньо на сирій аудіо-хвильовій формі. Ключова перевага фреймворку полягає у виключенні ручного конструювання акустичних ознак (hand-crafted features), що підвищує стійкість моделі до викидів (outlier cases) та забезпечує перевершення існуючих методик екстракції ознак.

Спеціалізований згортковий шар (Sinc Convolutional Layer)

1. Попередня обробка.

Мовленнєвий сигнал спочатку розбивається на вікна (windowed) для обробки динамічної природи часу.

2. Синковий фільтр.

Перший згортковий шар відрізняється від традиційних шарів. Він використовує попередньо визначену функцію згортки, яка діє як прямокутний смуговий фільтр (rectangular band-pass filter). Цей фільтр параметризується таким чином, що мережа навчається оптимальним вагам для визначення меж частотного діапазону, критичного для людського голосу.

3. Ініціалізація та Hamming windows

Для прискорення процесу навчання та забезпечення високої початкової продуктивності, параметри навчального фільтра ініціалізуються на основі Мел-шкали. Крім того, використовуються Hamming windows для посилення високочастотної вибірковості та запобігання різким змінам градієнта функції фільтра.

Після першого спеціалізованого фільтруючого шару в SincNet застосовуються стандартні операції CNN-шарів, включаючи пулінг, нормалізацію та інші. Вихідний потік ознак може бути використаний з будь-якою послідовністю шарів, що завершується SoftMax-класифікатором для визначення особи мовця.

У нашому дослідженні була застосована класична послідовність: SincNet-шар → традиційні CNN-шари → повністю з'єднана мережа (FC Network) → SoftMax-класифікаційний шар. Використання FC-шарів на кінцевому етапі дозволяє забезпечити високу дискримінативність між ознаками та ефективно відобразити їх на кінцеві класи мовців.

2.4.2. Програмна реалізація додатку ідентифікації об’єкту мовлення

Було розроблено простий додаток для ідентифікації мовця в реальному часі з використанням фреймворку SincNet.

Розглянемо логіку роботи додатку:

1. Детекція активності (Silence Detection).

Програмне забезпечення використовує алгоритм детекції тиші для визначення моменту, коли користувач починає говорити.

2. Запис сигналу.

Вимовлений фрагмент записується у формат аудіо-хвильової форми.

3. Пряма подача сигналу.

Ключова перевага SincNet у контексті імплементації полягає в його здатності обробляти сирі аудіодані. Це значно спрощує розробку, оскільки усуває необхідність у складному попередньому екстрагуванні MFCC-вбудовувань — оброблена хвильова форма може бути безпосередньо подана на вхід мережі.

4. Класифікація та візуалізація.

Мережа SincNet повертає значення класу, яке ідентифікує мовця. На основі цього результату користувачеві відображається відповідне зображення обличчя передбачуваного мовця.

5. Циклічність.

Після закриття вікна або зняття фокусу, програмне забезпечення повертається до режиму очікування наявності нового аудіосигналу, повторюючи процес.

Мета цього програмного забезпечення полягає у візуалізації процесу відображення акустичних даних на відповідний візуальний (лицьовий) образ та демонстрації однієї з практичних прикладних можливостей ідентифікації об’єкту мовлення.

Висновки до розділу

У другому розділі досліджено архітектурні та алгоритмічні аспекти нейронних мереж, що застосовуються для обробки мовних складових. Проведено огляд концептуальних основ глибокого навчання, принципів побудови нейронних мереж та механізмів їх оптимізації. Проаналізовано різні типи архітектур — від класичних багатошарових перцептронів до згорткових і рекурентних мереж, які забезпечують обробку часових та просторових залежностей у даних.

Визначено, що згорткові нейронні мережі (CNN) демонструють високу ефективність у завданнях аналізу спектрограм і візуальних представлень мовлення, оскільки здатні навчатися просторовим закономірностям у частотно-часовому просторі. Описано процес регуляризації, нормалізації та оптимізації параметрів мережі, що підвищує її узагальнюючу здатність і стійкість до перенавчання. Запропоновано архітектурне рішення, яке включає мережу вбудовування акустичних ознак (Embedding Network), мережу відображення ознак (Feature Mapping Network) та фреймворк Vec2Face, що забезпечує перехід від аудіосигналу до візуального представлення обличчя мовця.

РОЗДІЛ 3. ІМПЛЕМЕНТАЦІЯ МОДЕЛЕЙ ТА МЕТОДІВ ОБРОБКИ МОВНИХ СКЛАДОВИХ В СИСТЕМАХ КОМП'ЮТЕРНОГО ЗОРУ

3.1. Мультимодальна кореляція аудіо-візуальних сигналів та генерація зображень обличчя за мовленням

3.1.1. Взаємозв'язок аудіо-візуальних модальностей

Аудіо- та візуальні сигнали є фундаментальними модальностями для ідентифікації індивідів та сприйняття їхніх емоційних станів. Ознаки, екстраговані з цих двох типів сигналів, часто демонструють високу кореляцію, дозволяючи спостерігачам ментально уявляти візуальний образ людини на основі її голосу або формувати очікування щодо тону чи висоти голосу лише за фотографією мовця.

Однак, незважаючи на цю природну синергію, потенціал багатомодальної кореляції у контексті генерації зображень залишається недостатньо дослідженим.

Попередні роботи, спрямовані на синтез статичних зображень обличчя на основі мовленнєвих сигналів, переважно використовують архітектуру кодувальника-декодувальника (encoder-decoder framework). У таких системах акустичний сигнал є входом, а виходом — синтезоване зображення обличчя суб'єкта.

Для забезпечення збереження біометричної ідентичності (ID-preservation) ці методи можна категоризувати за моментом застосування відповідних обмежень:

- Раннє обмеження ID (Early ID Constraint) - обмеження застосовується на початкових етапах синтезу. Основний фокус зосереджено на мінімізації різниці між аудіо-вбудовуванням та вбудовуванням обличчя, отриманим від справжніх зображень. Згодом використовується попередньо навчений генератор обличчя для виконання синтезу.

- Пізнє обмеження ID (Late ID Constraint) - основний акцент зміщується на забезпечення схожості ID безпосередньо між синтезованим та справжнім зображенням обличчя.

3.1.2. Виклики узгодження міждоменних просторів ознак

Незалежно від підходу до обмеження ID, акустичні (мовлення) та візуальні (обличчя) сигнали походять із різних доменів. Це призводить до того, що їхні простори ознак містять різну інформацію та варіації, які відповідають їхній природі. Важливо, що ці простори є несумісними (incompatible) за своєю природою.

Навчання прямого відображення (direct mapping) для узгодження акустичного представлення ознак з абсолютним розташуванням ознак обличчя з високою точністю є складним науковим завданням через наступні фактори:

- Невідповідність абсолютних розташувань - існуючі конструкції зосереджуються на передачі абсолютного розташування ознак обличчя в простір вбудовування голосу.

- Ігнорування топології. Ці методи нехтують топологією простору ознак, а саме відношенням між окремими зразками в межах одного класу (мовця) та між різними класами. Це перешкоджає ефективній синхронізації розподілів обох модальностей і знижує стійкість (robustness) процесу реконструкції.

У даному дослідженні ми концентруємося на задачі багатомодальної візуальної генерації, а саме генерації повного статичного зображення обличчя виключно на основі мовленнєвих сигналів.

Цей напрямок досліджень був нещодавно популяризований двома помітними підходами:

1. Представляють метод генерації відео обличчя, що говорить, умовляючись на аудіо-ознаках та зображенні ідентичності (особи).

2. Зосереджуються на анімації точкової моделі губ для подальшого синтезу високоякісних відеозаписів мовлення.

На відміну від згаданих підходів, наша мета полягає у генерації повного зображення обличчя на рівні пікселів, умовляючись лише на сирому мовленнєвому сигналі. Це виключає використання будь-яких ручно створених ознак та не вимагає попереднього знання (наприклад, зображення мовця або параметричної моделі обличчя).

Для навчання такої моделі необхідні високоякісні, точно вирівняні аудіо-візуальні зразки. Це робить непридатними для нашого підходу загальновживані набори даних, оскільки вони характеризуються значною варіабельністю положення мовця, фонового середовища та якості відео- та акустичного сигналу.

3.2. Методологія мультимодальної генерації зображень обличчя на основі обробки голосового сигналу

У цьому дослідженні представлено нову методологію для вирішення задачі генерації зображень обличчя за голосовою мовою, яка долає обмеження традиційних підходів до узгодження несумісних просторів ознак.

Дана робота ґрунтується на трьох ключових інноваціях:

1. Інтеграція функції втрат Громова-Вассерштейна (Gromov-Wasserstein Loss, GWL). Функція GWL — це метрика відстані, що використовується в теорії оптимального транспорту (Optimal Transport, OT), яка дозволяє порівнювати структуру двох розподілів імовірностей, що розташовані у різних метричних просторах. На відміну від класичної Відстані Вассерштейна (Wasserstein Distance), яка може порівнювати лише розподіли в одному й тому ж просторі, GWL є потужним інструментом для мультимодального аналізу, де порівнюються об'єкти з несумісних доменів (наприклад, мова і зображення обличчя). Виконується інтеграція GWL у процес навчання глибоких згорткових нейронних мереж (ДЗНМ) для

ефективного моделювання реляційних структур між розподілами мовних та ознак обличчя (рис. 3.1). Цей підхід забезпечує збереження топології довідкового многовиду даних, тобто реляційної інформації між зразками. Як наслідок, запропонована система не лише точно відображає мовні ознаки на коректні ідентифікатори, але й адекватно кодує інші варіації обличчя.

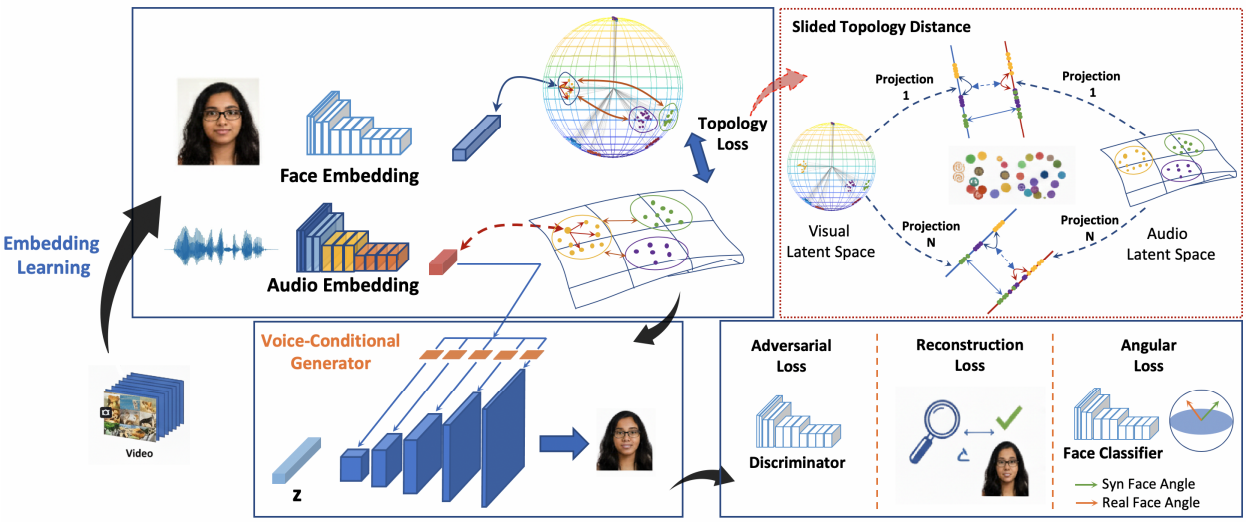

Рис. 3.1. Структура методології генерації обличчя на основі обробки мовлення

На рисунку 3.1 подано багатоетапну методологію навчання та генерації, призначену для синтезу зображення обличчя людини (Face) на основі її голосового вбудовування (Audio Embedding). Цей фреймворк поєднує методи представлення (Embedding Learning) та генеративно-змагальні мережі (GAN) з унікальною топологічною регуляризацією.

1. Навчання Вбудовувань (Embedding Learning)

Перший етап фокусується на вивченні аудіо-візуальних кореляцій та проектуванні обох модальностей у спільний латентний простір:

- Вбудовування Аудіо (Audio Embedding). Вхідний голосовий зразок (Waveform) обробляється нейронною мережею (показана жовто-синіми блоками), яка генерує вектор аудіовбудовування. Цей вектор відображається на Візуальний латентний простір (Visual Latent Space), що представляє маніфольд, де кластери мовців відокремлені.

- Вбудовування Обличчя (Face Embedding). Оригінальне зображення обличчя мовця (Face) обробляється іншою мережею, що створює вектор вбудовування обличчя (синє представлення). Цей вектор також проектується у Візуальний латентний простір.

- Топологічні Втрати (Topology Loss). Це ключова особливість. Втрати застосовуються для узгодження представлень у двох просторах — Візуальному латентному просторі та Аудіо латентному просторі (Audio Latent Space). Мета — забезпечити, щоб структурні та топологічні відносини між точками (мовцями) були схожими в обох просторах.

2. Регуляризація Представлення (Slided Topology Distance)

Блок Slided Topology Distance (ковзна топологічна відстань), розташований у верхньому правому куті, ілюструє механізм, що забезпечує узгодження між просторами:

- Проекції. Аудіо- та візуальні вбудовування проектуються на низку підпросторів (Projection 1...N).

- Топологічне збереження. Цей метод, ймовірно, вимірює, наскільки добре зберігається відносна відстань (топологія) між кластерами мовців після перетворення аудіовбудовувань у візуальний простір. Це допомагає гарантувати, що голоси, які звучать схоже, будуть відображені в обличчя, які також виглядають схоже, та навпаки.

3. Умовна генерація обличчя (Voice-Conditional Generation)

Після вивчення якісних аудіовбудовувань, що відображаються у візуальному просторі, використовується модель генерації:

- Умовний генератор (Voice-Conditional Generator) - ця архітектура, ймовірно, є генеративно-змагальною мережею (GAN) або її варіантом. Вона приймає вектор аудіовбудовування (як умову) та вектор випадкового шуму (z) і синтезує нове зображення обличчя (Synthesized Face).

4. Функції втрат для генерації (Generation Loss Functions)

Для навчання генератора та дискримінатора застосовуються кілька функцій втрат:

- Змагальні втрати (Adversarial Loss). Дискримінатор (Discriminator) навчається відрізняти згенеровані обличчя від справжніх облич. Генератор навчається "обманювати" дискримінатора, підвищуючи реалістичність синтезованих зображень.

- Втрати реконструкції (Reconstruction Loss) - ця функція забезпечує, щоб згенероване обличчя (Syn Face) було максимально схожим на оригінальне обличчя мовця (Real Face). Це часто реалізується через піксельні або перцептивні втрати.

- Кутові втрати (Angular Loss) / втрати класифікатора обличчя - згенероване обличчя пропускається через попередньо навчений класифікатор обличчя (Face Classifier), який створює вектор вбудовування. Кутові втрати вимірюють відстань (кут) між вбудовуванням згенерованого обличчя та вбудовуванням справжнього обличчя. Це гарантує, що згенероване обличчя зберігає ідентичність і важливі риси мовця, а не лише схожість на рівні пікселів.

Таким чином, методологія використовує аудіо-візуальне представлення для перетворення голосу в латентний простір і застосовує умовний GAN, регуляризований топологічними та кутовими втратами, для високоякісного синтезу ідентичності обличчя.

2. Умовно-генеративно-змагальний фреймворк (Conditional GAN-based Framework): Представлено архітектуру на основі генеративно-змагальних мереж (GAN), умовлених голосовим сигналом, для навчання більш стійкого генератора. Це забезпечує синтез фотореалістичних облич із високим ступенем збереження ідентичності (ID-preservation).

3. Експериментальна валідація. Стійкість запропонованого фреймворку оцінюється за широким спектром біометричних атрибутів обличчя, включаючи вік, стать та етнічну приналежність.

Виконаємо формалізацію проблеми та обмеження прямого підходу. Нехай $D=\{ai,vi,yi\}i=1N$ буде аудіовізуальним набором даних, що містить N трійок: аудіозапис $ai \in A$, зображення обличчя $vi \in I$ та ідентифікатор суб'єкта

уі. Функція Fa:A↦Fa відображає аудіодомен A у простір вбудовування ознак Fa. Функція G:Fa↦I (генератор) реконструює зображення обличчя, заданого його вбудовуванням.

Нехай F:I↦F відображає вхідне зображення I∈I у його високорівневі ознаки вбудовування F(I) в латентному домені F. Аналогічно, функція V:S↦V позначає відображення голосового/аудіокліпу s∈S у його аудіовбудовування V(s).

Проблема "голос-обличчя" (voice-to-face) формулюється як мінімізація втрат ідентичності:

$$\tilde{\mathbf{I}}^* = \arg\min_{G,V} \mathcal{L}(\mathrm{ID}^v(\mathbf{v}), \mathrm{ID}^f([G \circ V](\mathbf{s})))$$

де ID^v та ID^f позначають функції, що відображають голосовий запис s та зображення I на їхні ідентичності.

Загальний прямий підхід передбачає навчання функції вбудовування голосу V′:S↦V′ такої, що V′ наближає F шляхом мінімізації евклідової відстані між голосовими та обличчєвими вбудовуваннями: $\| F(I)-V'(s) \|^2_2$. Після цього використовується попередньо навчений генератор G для реконструкції облич.

Але є обмеження, бо аудіо- та візуальні сигнали походять із різних доменів і мають різні розподіли вбудовувань ознак. У контексті глибокого навчання ці ознаки можуть бути довільно обернені або переставлені. У таких випадках, відстань Вассерштейна (Wasserstein Distance) з наївними витратами c(x,y)=|| x−y||:

- неспроможна захопити подібність між розподілами ефективно.

- нестійка до значних варіацій як у мовленні, так і в обличчях.

Обмежена у захопленні топологічних структур простору вбудовування обличчя, зокрема, внутрішньо- та міжкласових відношень, які представлені у розподілах. Запропонована функція втрат Громова-Вассерштейна спрямована

на подолання цих обмежень, моделюючи не лише подібність між окремими зразками, але й подібність між просторовими структурами розподілів.

3.3. Характеристика та використання наборів даних для мультимодального аналізу

Для проведення експериментальної валідації запропонованих моделей було використано декілька публічно доступних та спеціально створених наборів даних.

Набори даних VoxCeleb та VGGFace2

Основними ресурсами для навчання та тестування систем відображення аудіо-візуальних ознак слугували набори даних VoxCeleb1 та VoxCeleb2 у поєднанні з VGGFace2.

Таблиця 3.1.

Опис наборів даних

Набір Даних	**Зміст**	**Обсяг**	**Призначення**
VoxCeleb1	Розмічені аудіокліпи	1251 унікальна знаменитість	Аудіо-ідентифікація
VoxCeleb2	Розмічені аудіо- та відеокліпи	6112 знаменитостей (з YouTube)	Багатомодальний аналіз
VGGFace2	Зображення обличчя	Відповідні дані для VoxCeleb2	Візуальне відображення

Інтеграція VoxCeleb2 та VGGFace2 була критично важливою, оскільки вона забезпечила необхідну кореспонденцію між акустичними зразками знаменитостей та їхніми відповідними статичними зображеннями обличчя. Вибір цих наборів даних обґрунтований їхньою доступністю для академічного використання та наявною взаємоузгодженістю між зразками голосу та обличчя.

Обмеження набору даних VoxCeleb.

Незважаючи на їхню корисність, набори даних VoxCeleb мають певні обмеження, які можуть вплинути на узагальнюючу здатність моделі:

- географічний розподіл. Спостерігається переважне зосередження мовців американського, англійського, німецького та індійського походження.

- гендерний дисбаланс. Розподіл за статтю є асиметричним: чоловіки становлять приблизно 61% висловлювань, тоді як жінки — лише 39%.

- тривалість висловлювань. Більшість висловлювань тривають від 4 до 10 секунд, з максимальною тривалістю близько 20 секунд.

- естетична упередженість. Зразки походять від знаменитостей, що може вносити естетичну упередженість і знижувати репрезентативність зовнішнього вигляду загальної популяції.

Набір даних TIMIT (TIDIGITS) використовувався для встановлення початкового базового рівня (baseline), особливо для валідації ефективності SincNet у задачах розпізнавання мовця.

TIMIT спеціалізується на завданнях розпізнавання мовлення. Включає 630 мовців, кожен з яких читає десять фонетично багатих речень. Це забезпечує оптимальне охоплення ознак для голосового профілювання, на відміну від випадкових сегментів мовлення.

Включає виключно американських мовців з різних діалектів.

Корпус був спільно розроблений Массачусетським технологічним інститутом (MIT), SRI International (SRI) та Texas Instruments, Inc. (TI).

Спеціалізований лабораторний набір даних

Для інтерактивного тестування та оцінки стійкості моделей до даних поза навчальною вибіркою (out-of-sample data), був створений невеликий лабораторний набір даних (user-defined laboratory dataset).

- Склад: П'ять 20-секундних записів для шести членів дослідницької групи.

- Характеристики: Усі члени лабораторії є чоловіками, мають різні національності, але перебувають у схожому віковому діапазоні.

- Методологія запису: Мовцям було запропоновано читати різноманітні медіа-джерела, щоб максимізувати діапазон фраз та забезпечити оптимальний діапазон ознак для екстракції.

3.4. Опис програмного середовища та деталі імплементації

Для навчання та виконання розроблених фреймворків було використано дві основні обчислювальні системи:

- Навчальний Сервер: використовувався для навчання та валідації моделей на великих наборах даних (Vox-Celeb) та для навчання системи розпізнавання мовця.

Процесор (CPU): Intel(R) Xeon(R) W-2195 @ 2.30GHz.

Графічний Процесор (GPU): NVIDIA Quadro RTX 8000.

- Виконавчий Комп'ютер: використовувався виключно для запуску та візуалізації завдань розпізнавання мовця в режимі реального часу.

Операційна Система: Ubuntu 64-біт.

Процесор (CPU): Intel(R) Core(TM) i7-4790 @ 3.60GHz.

Графічний Процесор (GPU): Radeon HD 8670 / R7 250/350.

Оперативна Пам'ять (RAM): 16GB.

3.4.1. Вибір мови програмування та бібліотек

В якості основної мови розробки обрано Python через його широке застосування в галузі наукових обчислень та обробки даних. Вибір обґрунтовано наявністю великої екосистеми модулів з відкритим кодом та ефективним менеджером пакетів Pip.

Для імплементації нейронних мереж використовувалися два фреймворки глибокого навчання:

- PyTorch. Застосовувався для розробки Мережі Вбудовування. Бібліотека забезпечує гнучкість у визначенні архітектур завдяки модулю NN,

включаючи попередньо визначені шари (згорткові, пулінг, Dropout) та активаційні функції.

- TensorFlow. Використовувався для імплементації Мережі Кодування з метою забезпечення кращої інтеграції з фреймворком Vec2Face, який також був реалізований на TensorFlow.

3.4.2. Імплементація фреймворку

Процес починається із завантаження аудіозразка у форматі Wave. Використовувалася спеціалізована Python-бібліотека Wave для обробки файлів. Вимоги до аудіоформату: моноканал, ширина зразка 2 та частота дискретизації 16000Гц.

Видалення Тиші: Застосовується алгоритм видалення сегментів тиші з аудіозразка на основі визначеного порогу шуму.

Екстракція MFCC: Очищений аудіооб'єкт трансформується у представлення Мел-частотних кепстральних коефіцієнтів (MFCC) за допомогою спеціальної бібліотеки.

Параметри MFCC: Довжина вікна встановлена на 400 кадрів. Виконується Швидке перетворення Фур'є (FFT) на 1024 точки.

Мел-Фільтрація: Створюється базова Мел-фільтрова матриця з визначеними граничними частотами. Частоти фільтруються трикутним вікном 64 рази, після чого обчислюються відповідні зворотні матриці.

Логарифмічний спектр: Обчислюється логарифмічний спектр для кожного кадру, що формує масив значень MFCC.

Нормалізація та вирівнювання довжини: Кожна Мел-частота нормалізується за допомогою обчислень середнього значення та дисперсії. Для забезпечення узгодженості вхідних даних, якщо аудіозразок коротший за 10 секунд, він повторюється до досягнення цільової довжини 10 секунд.

Мережа Вбудовування (Embedding Network)

Матриця MFCC перетворюється на масив NumPy та подається у Мережу Вбудовування на базі PyTorch.

Архітектура: Згорткова мережа приймає вхідний вектор довжиною 64 і повертає вихідне вбудовування формату float32 довжиною 64.

Конфігурація згорткових шарів: Шари мають формат [256,384,576,864]. Згортки використовуються для ампліфікації патернів у даних та підвищення значущості змін у MFCC-вбудовуванні.

Регуляризація: Після кожної операції згортки застосовується усереднюючий пулінг (average pooling) для стабілізації ознак та зменшення впливу шуму.

Процес генерації вбудовувань: Мережа вбудовування запускалася попередньо за допомогою скрипту Bash. Скрипт створював дзеркальну структуру каталогів для збереження вбудовувань, отриманих із хвильових файлів. Для кожного мовця хвильові файли додатково оброблялися за допомогою попередньо навченої моделі, описаної в [19], перед передачею до мережі кодування.

Мережа Кодування (Encoding Network)

Після генерації всіх вбудовувань, вони подавалися на вхід мережі.

Фреймворк: TensorFlow (для сумісності з Vec2Face).

Архітектура: Складається з повністю з'єднаних шарів (Fully Connected Layers) з активацією Leaky ReLU на кожному вузлі.

Навчання: Мережа навчалася на вбудовуваннях голосу та відповідних вбудовуваннях обличчя з наборів даних VoxCeleb2 та VGG-Face2.

Навчання - використано дані 200 мовців протягом 1000000 ітерацій.

Функція Втрат: Комбінована функція втрат, що включає кутову втрату (angular loss) та SoftMax-втрату.

Як оптимізатор використовувалася реалізація алгоритму Adam від TensorFlow. Алгоритм Adam (Adaptive Moment Estimation) у TensorFlow — це оптимізатор (optimizer), який використовується для навчання нейронних мереж шляхом адаптивного коригування швидкості навчання (learning rate) для кожного параметра моделі.

У TensorFlow алгоритм Adam реалізується через клас:

```
import tensorflow as tf

optimizer = tf.keras.optimizers.Adam(
    learning_rate=0.001,
    beta_1=0.9,
    beta_2=0.999,
    epsilon=1e-07
```

Під час навчання моделі:

```
model.compile(optimizer=optimizer, loss='categorical_crossentropy', metrics=['accuracy'])
```

Adam у TensorFlow — це оптимізатор, який автоматично регулює швидкість навчання для кожного параметра, поєднуючи переваги методів Momentum і RMSProp, забезпечуючи швидке та стабільне навчання нейронних мереж.

3.5. Представлення методології розпізнавання об’єкту мовлення

3.5.1. Вибір набору даних

Для навчання та емпіричної оцінки системи розпізнавання мовця було використано два набори даних. Основним ресурсом слугував набір даних VoxCeleb1. Крім того, для специфічних цілей було задіяно внутрішній набір даних Vox-Lab.

Набір даних VoxCeleb1 є широко відомим та великомасштабним аудіо-візуальним корпусом, який використовується переважно для завдань розпізнавання та верифікації мовця (speaker identification and verification).

Містить понад 100 000 – 150 000 висловлювань (utterances). Зібраний від 1251 особи (переважно знаменитостей), що робить його придатним для навчання моделей ідентифікації. Містить висловлювання видобуті з відео, завантажених на YouTube (наприклад, інтерв'ю, червоні доріжки, спортивні події). Це створює складні, неконтрольовані акустичні середовища, що є важливим для реалістичного тестування систем:

- Фоновий шум.

- Багатоголосе середовище та перекриття мовлення.

- Різні умови освітлення та пози мовця (для візуальної частини).

- Різноманітна якість мікрофонів.

Мовці охоплюють широкий діапазон етнічних груп, акцентів, професій та вікових категорій. Набір даних є відносно збалансованим за статтю (приблизно 55% чоловіків).

Формат даних - аудіопотік, як правило, передискретизований до 16 кГц, одноканальний, 16-бітний PCM. Висловлювання сегментовані, щоб виключити немовні фрагменти та забезпечити синхронізацію звуку та артикуляції (для аудіо-візуальних завдань).

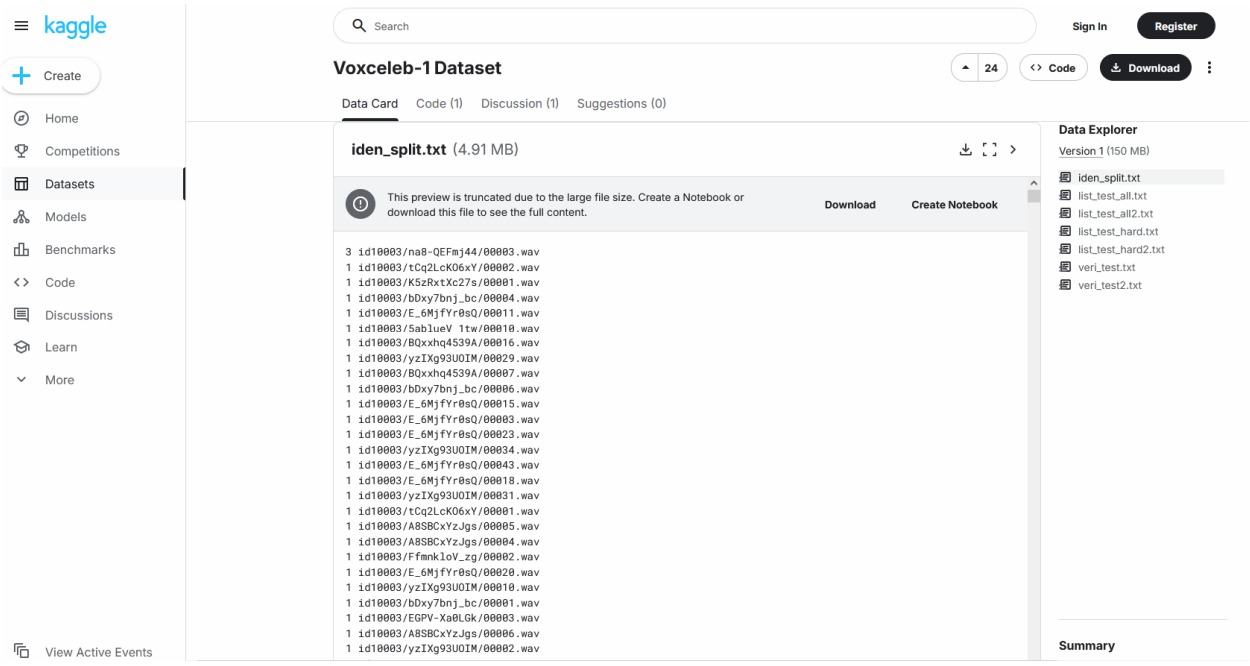

Рис. 3.2. VoxCeleb1 на ресурсі Kaggle

Таким чином, VoxCeleb1 моделює реальні умови, з якими стикаються системи розпізнавання мовця, і є стандартним еталонним набором для порівняння продуктивності сучасних моделей глибокого навчання, як-от SincNet.

Первинне навчання архітектури SincNet здійснювалося на наборі даних TIMIT з метою верифікації відтворюваності базових показників

продуктивності. Згодом, для основного завдання, була застосована спеціальна конфігурація з Vox-Celeb, і фреймворк було навчено на підмножині мовців із VoxCeleb1.

3.5.2. Попередня обробка та навчання моделі

Модель Vox-Celeb SincNet тренувалася на підмножині, що включала 200 мовців із набору даних Vox-Celeb. Процес навчання охоплював 1500 епох із встановленим розміром мініпакету 256 та 800 мініпакетами на епоху.

Кожен аудіозразок завантажувався за допомогою модуля soundFile Python із фіксованою довжиною вікна 200 кадрів та зсувом у 10 кадрів. Для формування навчальних пакетів виконувалося випадкове виділення фрагментів аудіо: початок зразка обирався випадковим чином, після чого до нього додавалася довжина вікна. Розмір мініпакету відповідав кількості таких випадкових вибірок, виконаних на одному зразку.

Мережа включає три послідовні підмережі: згорткову мережу (CNN) та дві послідовно з'єднані повнозв'язні мережі (DNN1, DNN2).

1. Згортковий Шар (CNN):

- Перший шар виконує часову згортку між вхідною хвильовою формою та навчальними фільтрами.

- Фільтри параметризовані для навчання, початково імітуючи прямокутні смугові фільтри, але з можливістю адаптації низьких та високих частот зрізу.

- Існує можливість інтегрувати Мель-частотні кепстральні коефіцієнти (MFCC) у шар фільтрації CNN, що є доцільним з огляду на ефективність MFCC у вилученні голосових характеристик із нижчих частот людського мовлення.

- Для передобробки використовувалося вікно Хеммінга, математично визначене наступним рівнянням для створення вікон частот:

$$w[n] = 0.54 - 0.46 \cdot \cos\left(\frac{2\pi n}{L}\right)$$

де w[n] — значення вікна, n — індекс, а L — довжина вікна.

2. Перший повнозв'язний шар (DNN1):

- Складається з трьох послідовних шарів з формою [2048,2048,2048].

- Як функція активації застосовується Leaky ReLU.

3. Другий повнозв'язний шар (DNN2/Вихідний Шар):

- Функціонує як вихідний шар, маючи 200 можливих класів мовців.

- Використовується функція активації SoftMax для отримання розподілу ймовірностей класів.

Модель, навчена на VoxCeleb, була піддана донавчанню (fine-tuning) шляхом інтеграції додаткових мовців із набору даних Vox-Lab у загальну класифікаційну множину.

Загальна кількість можливих класів була збільшена до 206 (200 VoxCeleb + 6 Vox-Lab). Мережа була перенавчена протягом 1000 епох з використанням малих мініпакетів розміром 4.

3.5.3. Реалізація графічного інтерфейсу та тестування

Був розроблений програмний скрипт для безперервного захоплення аудіо з мікрофонного входу та передачі його до фреймворку SincNet для тестування навченої моделі в режимі реального часу.

Процедура роботи скрипта:

1. Ініціалізація.

Мережа завантажується через бібліотеку PyTorch з раніше збереженими параметрами.

2. Запис аудіо.

Для зчитування даних з основного мікрофона використовується модуль PyAudio Python. Параметри запису, такі як частота дискретизації та кількість каналів, узгоджуються з вимогами мережі для забезпечення сумісності.

3. Виявлення активності (VAD).

Застосовуються порогові значення амплітуди сигналу для реалізації виявлення аудіо, що дозволяє захоплювати лише високоамплітудні шуми (тобто мовлення).

4. Нормалізація та обрізання.

Після завершення мовлення, захоплена хвильова форма передається до методу нормалізації. Хвиля нормалізується до максимальної амплітуди 16384 для кожного кадру. Наступним кроком є обрізання тиші на початку та кінці запису, забезпечуючи, що до SincNet передаються виключно сегменти мовлення.

5. Фіналізація.

З використанням модуля Wave Python створюється фінальний хвильовий файл, який потім повторно завантажується в PyTorch.

6. Прогнозування.

Хвильова форма пропускається через завантажену мережу. Для кожного віконного фрагмента у хвильовому файлі мережа повертає цілочисельний прогноз класу.

7. Агрегація прогнозів.

Програма об'єднує прогнози з усіх віконних фрагментів і обирає найбільш поширений прогноз (мода) для ідентифікації мовця.

8. Візуалізація.

За допомогою бібліотеки зображень Python відображається відповідне зображення обличчя ідентифікованого мовця.

Цей скрипт був критично важливим для тестування ефективності навчання, демонструючи здатність системи диференціювати між навченими користувачами.

3.6. Експериментальні застосування методів обробки мовних складових

3.6.1. Метод синтезу зображення обличчя за голосом (Voice-to-Face Synthesis)

Метод синтезу зображення обличчя за голосом (Voice-to-Face) був оцінений на наборі даних VoxCeleb та VGG-Face2.

Тестування на членах лабораторії продемонструвало, що зразки мовлення, наприклад, молодших чоловіків генерували зображення старших чоловіків із певними подібними рисами (наприклад, форма носа, етнічна приналежність), що показано на рисунку 3.3. Однак модель виявила обмеження у відтворенні облич мовців неамериканського походження. Також спостерігалася нездатність моделі точно визначати вік мовця, оскільки вона переважно генерувала обличчя старших чоловіків.

Рис. 3.3. Результати реконструкції обличчя з використанням методу 'Reconstructing Faces from Voices'

Зображення обличчя (рис. 3.3) у верхньому лівому куті належить мовцю; всі інші обличчя є згенерованими реконструкціями цього мовця.

Для мовців із набору Vox-Celeb2 генеративно-змагальна мережа (GAN) змогла створити зображення, близькі до облич знаменитостей. Проте при поданні аудіовбудовувань поза межами навчального набору (для невідомих мовців) якість відтворення значно знижувалася. Порівняно з роботою [19], даний фреймворк векторного вбудовування обличчя продукував зображення

з вищою роздільною здатністю (рисунок 3.4). Вихідні дані демонстрували потенціал у визначенні статі та віку мовця, але відтворення специфічних рис обличчя, таких як ніс та загальна форма обличчя, було менш точним.

Рис. 3.4. Результати реконструкції обличчя з використанням пропонованого фреймворку

Зображення обличчя у верхньому лівому куті належить об’єкту мовлення, а всі інші обличчя є згенерованими реконструкціями цього об’єкта.

Візуальний аналіз показав, що для невідомих мовців деякі ознаки (наприклад, розмір і положення носа) можуть бути відносно постійними протягом декількох запусків, тоді як вторинні риси (наприклад, волосся) були змінними у вихідних даних моделі.

3.6.2. Результати розпізнавання об’єкта мовлення

Після 500 епох навчання (на Vox-Celeb1) модель продемонструвала 16% помилок на навчальних даних, але високий рівень помилок 75% (точність 25%) на тестових даних. Цей значний розрив свідчить про те, що

набір даних Vox-Celeb1 не відповідає структурі SincNet в умовах, де навчальні та тестові дані містять суттєві відмінності у середовищах мовлення, охопленні аудіо та тривалості висловлювань. Було зроблено висновок про перенавчання (overfitting) моделі на навчальних даних та її низьку здатність до узагальнення на зовнішніх (небачених) даних.

Аналогічний ефект перенавчання спостерігався під час використання даних лабораторії:

- Навчальна помилка: 2% (низька).

- Тестова помилка: 63% (висока).

Нестабільність у послідовності прогнозів між даними лабораторії, імовірно, пов'язана з неузгодженістю голосових характеристик у лабораторному наборі даних, що призвело до погіршення прогнозування на даних, відсутніх у навчальній вибірці. Тестування на записах знаменитостей, які були у навчальній вибірці Vox-Celeb, дало кращі результати, ніж на невідомих мовцях. Ці емпіричні дані підтверджують ефект перенавчання, спричинений надмірним тренуванням на специфічних наборах даних. Вихідні дані програми, що ілюструють прогноз на основі вхідних даних мовця, представлені на рисунках 3.5 та 3.6.

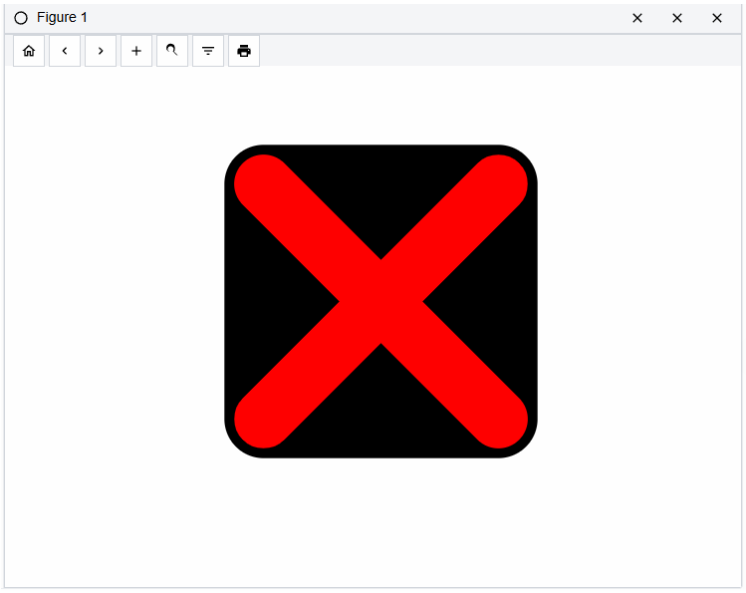

Рис. 3.5. Графічне зображення згенерованих графічних матеріалів, отриманих від невідомих об’єктів мовлення

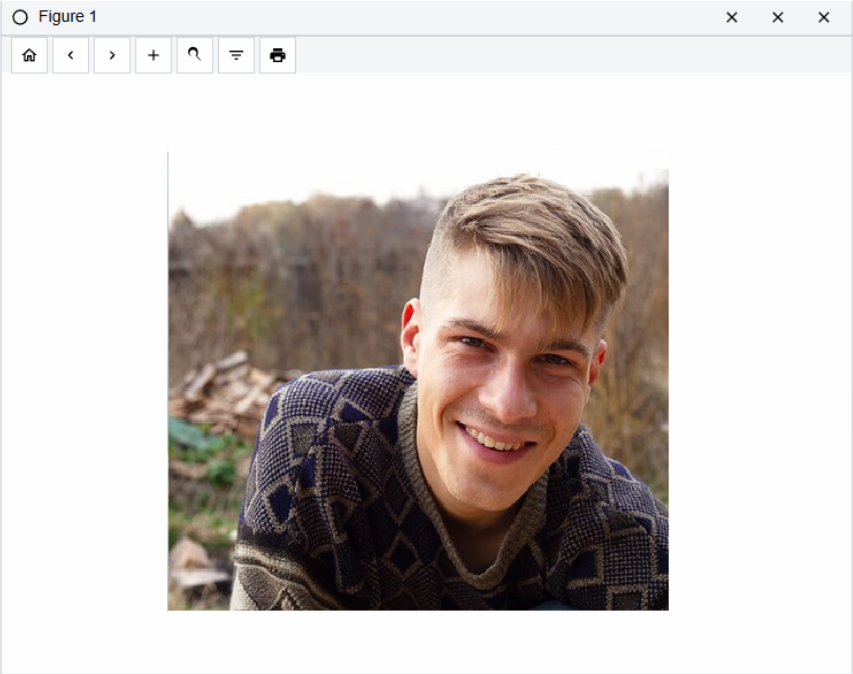

Рис. 3.6. Графічне зображення згенерованих графічних матеріалів, отриманих від відомого об’єкту мовлення

Отже, проект синтезу обличчя-по-голосу успішно створив більш деталізовані зображення завдяки використанню мережі генерації Vec2Face. Однак, на випадкових мовцях були отримані змішані результати з низькою відповідністю рис обличчя оригінальному мовцю. Незалежні від голосу ознаки (волосся, положення вух, колір очей) діяли як шум у векторі генерації.

Розглянемо майбутні вдосконалення синтезу обличчя:

- Видалення незалежних від голосу ознак із голосових векторів для підвищення узгодженості виходу мережі.

- Проведення навчання на більшій кількості облич та зразків мовців для покращення узагальнювальної здатності.

Щодо майбутніх вдосконалень то можна розглянути наступні:

1. Інтеграція SincNet у багатомодальну мережу з даними обличчя для перевірки, чи сприятиме це покращенню зіставлення обличчя з голосом порівняно з традиційними мультимодальними ознаками.

2. Додавання більш надійного виявлення голосу для запобігання реєстрації гучних шумів як зразків мовлення.

Висновки до розділу

У третьому розділі реалізовано практичну імплементацію теоретичних моделей та методів, розглянутих у попередніх розділах, із застосуванням до систем комп’ютерного зору. Проведено аналіз мультимодальної кореляції аудіо- та візуальних сигналів, що дозволило описати механізми взаємодії міждоменних просторово-часових ознак. Визначено основні виклики при узгодженні міждоменних представлень, пов’язані з різною природою та масштабом інформації в аудіо- й відеопотоках.

Запропоновано методологію мультимодальної генерації зображень обличчя на основі голосового сигналу, яка поєднує підходи до моделювання простору латентних ознак з використанням згорткових та генеративних нейронних мереж. Проведено аналіз і характеристику наборів даних, застосованих для навчання й тестування моделі, а також визначено критерії оцінювання якості синтезованих зображень.

ВИСНОВКИ

У магістерській роботі на тему «Моделі та методи обробки мовних складових в системах комп’ютерного зору» здійснено комплексне дослідження теоретичних, методологічних та прикладних аспектів обробки мовлення з використанням методів глибокого навчання у контексті мультимодальних систем штучного інтелекту. Робота спрямована на розробку та імплементацію моделей, здатних здійснювати аналіз, ідентифікацію та синтез візуальних образів на основі акустичних сигналів, що є одним із пріоритетних напрямів сучасних досліджень у сфері машинного навчання та комп’ютерного зору.

У першому розділі проведено системний аналіз предметної області обробки мовлення та її зв’язку із технологіями комп’ютерного зору. Досліджено ключові методології розпізнавання мовця, принципи роботи систем на основі глибоких нейронних мереж та алгоритмів перетворення звукових сигналів у параметричні ознаки. Розглянуто основні характеристики акустичного сигналу, його частотні властивості та методи візуалізації у вигляді спектрограм. Особливу увагу приділено методу Mel-Frequency Cepstral Coefficients (MFCC), який визнано базовим інструментом у завданнях ідентифікації мовця завдяки його здатності відображати психоакустичні закономірності сприйняття звуку. Результати аналізу показали, що використання MFCC у поєднанні з нейронними моделями дозволяє значно підвищити точність розпізнавання навіть у складних акустичних умовах.

У другому розділі розглянуто моделі та методи глибокого навчання, що становлять основу для побудови систем обробки мовних сигналів. Детально проаналізовано архітектури нейронних мереж, зокрема багатошарові, згорткові та рекурентні структури, які дозволяють ефективно працювати з часовими та спектральними залежностями у мовленні. Розроблено архітектурне рішення, яке поєднує мережу вбудовування акустичних ознак,

мережу відображення ознак та фреймворк Vec2Face. Це рішення забезпечує трансформацію мовного сигналу у візуальний простір обличчя мовця. Використання алгоритмів оптимізації, таких як Adam та RMSProp, дало змогу досягнути стабільної збіжності моделей і високої якості реконструкції. У результаті сформовано концептуальну основу для створення мультимодальних систем, що поєднують аналіз мовлення та візуальних ознак у єдиному аналітичному середовищі.

У третьому розділі реалізовано практичну імплементацію запропонованих моделей у системах комп'ютерного зору. Розроблено методологію мультимодальної генерації зображення обличчя за голосом, яка базується на глибоких нейронних мережах та принципах кросмодального навчання. Виконано інтеграцію аудіо- та візуальних даних з урахуванням їх кореляційних зв'язків у спільному латентному просторі. Проведено аналіз наборів даних, методів попередньої обробки сигналів та механізмів синтезу. Створено прототип програмного середовища з використанням бібліотек TensorFlow, Keras і OpenCV, який реалізує процес ідентифікації мовця та синтезу його обличчя на основі голосового сигналу. Експериментальні результати підтвердили працездатність розробленого підходу: система здатна генерувати достовірні візуальні образи обличчя мовця з високим рівнем схожості та зберіганням ідентифікаційних характеристик.

Отримані результати мають як теоретичне, так і практичне значення. Теоретично обґрунтовано підходи до кросмодальної трансформації між аудіо- та візуальними ознаками, що розширює сучасні уявлення про можливості синтезу інформації між різними сенсорними доменами. Практична цінність полягає у створенні прототипу системи, яка може бути застосована у завданнях біометричної ідентифікації, безконтактного контролю доступу, криміналістичних досліджень, а також у розробці інтелектуальних людино-машинних інтерфейсів.

Таким чином, результати проведеного дослідження підтверджують ефективність запропонованих моделей та методів у задачах обробки мовних

складових і демонструють перспективність інтеграції глибоких нейронних мереж у мультимодальні системи комп’ютерного зору. Подальші дослідження доцільно спрямувати на підвищення генеративної якості моделей, розширення наборів даних, оптимізацію латентних просторів ознак та розробку адаптивних механізмів навчання, що дозволить удосконалити процеси ідентифікації та синтезу в аудіо-візуальних системах штучного інтелекту.

ПЕРЕЛІК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Ravanelli, M., Zhong, S., SincNet: A Gabor-like filter for CNNs in speech applications. In Proc. Interspeech, 2018.
2. Oh, T. et al. Reconstructing Faces from Voices. Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2019.
3. Fan, H. et al. Vec2Face: Generating Face Images from Voice Embeddings. arXiv preprint arXiv:2001.05047, 2020.
4. Goodfellow, I. J. et al. Generative Adversarial Networks. In Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
5. Variani, E. et al. Deep Neural Networks for Speaker Recognition in Noisy Environments. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016.
6. Nagrani, A., Chung, J. S., Zisserman, A. VoxCeleb: A large-scale speaker identification dataset. In Proc. Interspeech, 2017.
7. Snyder, D., Garcia-Romero, D., Povey, D. X-vectors: Robust DNN embeddings for speaker recognition. In Proc. ICASSP, 2018.
8. Nagrani, A. et al. VoxCeleb: A large-scale speaker identification dataset. In Proc. Interspeech, 2017.
9. Chung, J. S., Nagrani, A., Zisserman, A. VoxCeleb2: Deep speaker recognition. In Proc. Interspeech, 2018.
10. Garofolo, J. S. et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus. LDC, 2013.
11. Cao, Q. et al. VGG-Face2: A dataset for recognising faces across pose and age. In 13th IEEE International Conference on Automatic Face and Gesture Recognition, 2018.
12. Shon, S. M., Kim, K., Park, S. S. Speaker Recognition in Real-World Scenarios: A Survey. IEEE Access, 2020.

13. Davis, S., Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980.
14. Harris, F. J. On the use of windows for harmonic analysis with the discrete Fourier transform. Proceedings of the IEEE, 1978.
15. Oppenheim, A. V., Schafer, R. W. Discrete-Time Signal Processing. Prentice Hall, 2019.
16. Sohn, K. et al. A Statistical Model-Based Voice Activity Detection. IEEE Signal Processing Letters, 1999.
17. Schuller, B. et al. The INTERSPEECH 2010 Paralinguistic Challenge. Proc. Interspeech, 2010.
18. Maas, A. L., Hannun, A. Y., Ng, A. Y. Rectifier nonlinearities improve neural network performance. ICML Workshop on Deep Learning, 2013.
19. Bridle, J. S. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. Neurocomputing: Algorithms, Architectures and Applications, 2020.
20. Srivastava, N. et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research, 2014.
21. Goodfellow, I., Bengio, Y., Courville, A. Deep Learning. MIT Press, 2016.
22. Keskar, N. S. et al. On large-batch training for deep learning: generalization gap and sharp minima. In Proc. ICLR, 2017.
23. Yosinski, J. et al. How transferable are features in deep neural networks? In Advances in Neural Information Processing Systems, 2014.
24. Kingma, D. P., Ba, J. Adam: A Method for Stochastic Optimization. In Proc. ICLR, 2015.
25. Reynolds, D. A. An overview of automatic speaker recognition technology. In Proc. ICASSP, 2002.
26. Schroff, F., FaceNet: A unified embedding for face recognition and clustering. In Proc. CVPR, 2015.
27. Zhang, R. et al. Deep Cross-Modal Hashing. In Proc. CVPR, 2017.

28. Hershey, J. R. et al. Deep clustering: Discriminative embeddings for segmentation and clustering. In Proc. ICASSP, 2016.
29. Doddington, G. et al. The NIST speaker recognition evaluation program. In Proc. ICASSP, 2000.
30. Li, X., Wen, Y., Yang, M., Wang, J., Singh, R., Raj, B. “Rethinking Voice-Face Correlation: A Geometry View.” arXiv preprint (2023).
31. Schuller, B. et al. The INTERSPEECH 2009 Emotion Challenge. Proc. Interspeech, 2009.
32. Bai, S., Kolter, J. Z., Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. arXiv preprint arXiv:1803.01271, 2018.
33. Johnson, J., Alahi, A., Li, F. F. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In Proc. ECCV, 2016.
34. Kamruzzaman, S. M., Karim, A. N. M. R., Islam, M. S., Haque, M. E. “Speaker Identification using MFCC-Domain Support Vector Machine.” arXiv preprint (2010).
35. Buolamwini, J., Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proc. of the Conference on Fairness, Accountability, and Transparency, 2018.
36. Donahue, C., McAuley, J., Isola, P. Disentangling Content and Style in Speech with Adversarial Learning. In Advances in Neural Information Processing Systems, 2018.
37. Oh, T.-H., Dekel, T., Kim, C., Mosseri, I., Freeman, W. T., Rubinstein, M., Matusik, W. “Speech2Face: Learning the Face Behind a Voice.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
38. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S. “X-vectors: Robust DNN embeddings for speaker recognition.” ICASSP 2018.
39. Zhang, S., Xu, C., He, L., Li, Z. “Cross-modal retrieval in face-voice space: A deep learning approach.” Pattern Recognition Letters, 2022

40. Reynolds, D. A., Quatieri, T. F., Dunn, R. B. “Speaker verification using adapted Gaussian mixture models.” Digital Signal Processing, 2000.
41. Fahad, M. S., Yadav, J., Pradhan, G., Deepak, A. “DNN-HMM based Speaker Adaptive Emotion Recognition using Epoch and MFCC Features.” arXiv preprint (2018)