УДК 004.8 DOi 10.30837/ bi.2023.1(99).06

Д.С. Суворов[1], І.В. Афанасьєва[2], К.Г. Онищенко[3], О.В. Калиниченко[4]

[1]ХНУРЕ, м. Харків, Україна, daniil.suvorov@nure.ua, ORCID iD: 0009-0008-0083-1978

[2] ХНУРЕ м. Харків, Україна, iryna.afanasieva@nure.ua, ORCID iD: 0000-0003-4061-0332

[3]ХНУРЕ, м. Харків, Україна, kostiantyn.onyshchenko@nure.ua, ORCID iD: 0000-0002-7746-4570

[4]ХНУРЕ, м. Харків, Україна, olga.kalynychenko@nure.ua, ORCID iD: 0000-0003-1466-3967

ВПЛИВ РОЗМІРУ КАДРУ НА РОЗПІЗНАВАННЯ ЕМОЦІЇ ЗА МОВЛЕННЯМ

У задачі розпізнавання емоції за мовленням, як і у більшості задач машинного навчання розпізнавання за звуком, використовується так званий фреймінг. Це процес поділу вихідного аудіосигналу на кадри певного розміру, кожен з яких оброблюється окремо. У цій статті представлено порівняння впливу розміру кадрів на результат розпізнавання емоції на прикладі CNN мережі. Для експериментів використовувався набір CREMA-D із аугментаціями, використовуючи додавання шуму, розтягування у часі та зміну висоти тону. В ході досліджень вдалося досягти точності розпізнавання в 98,8% із використанням динамічного розміру кадру.

АУДІО, ЕМОЦІЇ, КАДР, МАШИННЕ НАВЧАННЯ, НЕЙРОННІ МЕРЕЖІ, РОЗПІЗНАВАННЯ, PYTHON, TENSORFLOW

D.S. Suvorov, I.V. Afanasieva, K.G. Onyshchenko, O.V. Kalynychenko. The effect of frame size on speech emotion recognition. Speech emotion recognition task, as well as most audio recognition machine learning tasks, uses the so-called framing. This is the process of dividing the original audio signal into frames of a certain size, each of which is processed separately. This article presents a comparison of the effect of frame size on the emotion recognition result using a CNN network as an example. For the experiments, the CREMA-D dataset was used with the augmentations using noise adding, time stretching, and pitch shifting. We managed to achieve a recognition accuracy of 98.8% using dynamic frame size.

AUDIO, EMOTIONS, FRAME, MACHINE LEARNING, NEURAL NETWORKS, RECOGNITION, PYTHON, TENSORFLOW

Вступ

З активним зростанням технологій штучного інтелекту, ці ж технології набувають широкого поширення на різні сфери життя людини. Однією з таких галузей є психологічний аналіз стану людини. Існують різні підходи до такого аналізу, проте найбільшого поширення наразі набули методи розпізнавання емоції за текстом, з використанням міміки та пози людини, а також за мовленням [1]. Саме розпізнавання за мовленням є темою поточного дослідження.

Подібний аналіз дає змогу за невеликим уривком запису мовлення людини визначити емоцію, з якою людина говорила. Подібний підхід може мати деякі переваги, пов'язані з мовним розмаїттям і віковою варіацією. Розроблена модель на основі однієї мови (наприклад, англійської) може бути всього лише донавчена з використанням додаткового набору даних іншої мови (наприклад, української). Однак, навіть без додаткового розширення вибірки, створена модель уже може працювати з різноманітними мовами (хоча й можуть бути певні винятки, пов'язані з культурними особливостями, ідеологією і просто специфічною говіркою).

Проте, подібного роду аналіз представляє досить перспективний підхід для різних систем, таких як розумні будинки або системи екстреного реагування. Але не варто забувати, що для повноцінного емоційного аналізу необхідно використовувати багатофакторний аналіз, який би містив кілька джерел, що давало б більш об'єктивну оцінку (візуальна інформація, інформація ЧСС і так далі).

У задачі розпізнавання емоції за мовленням важливим моментом у вилученні параметрів з аудіосигналу є поділ цього сигналу на фрагменти [2], кожен з яких окремо обробляється. Саме цей аспект попередньої обробки аудіо і буде детально розглянуто в статті, щоб отримати повне уявлення впливу розміру таких фрагментів на точність моделі.

1. Опис предметної галузі

Перш ніж перейти до опису вилучення параметрів з аудіо, для початку розглянемо звук загалом [3]. У нашому звичайному (аналоговому) світі, звук є безперервною хвилею (див. рис. 1). Однак, для обробки за допомогою ЕОМ необхідно аналоговий звук оцифрувати.

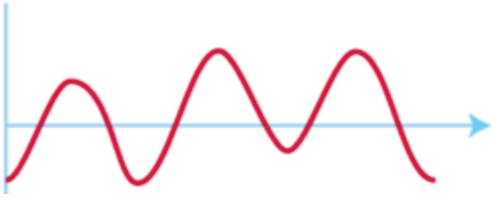

Рис. 1. Аналогова звукова хвиля

Це відбувається за допомогою АЦП — аналогово-цифрового перетворення. Тоді сигнал починає виглядати трохи інакше (див. рис. 2).

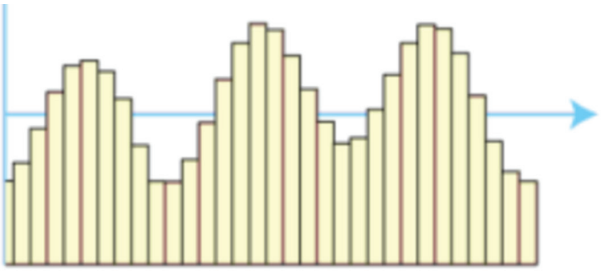

Рис. 2. Цифрова звукова хвиля

У чому ж полягає процес АЦП і чому сигнал стає ступінчастим (дискретним)? Завдяки записуючим пристроям та аналогово-цифровим перетворювачам звукова хвиля зчитується з певною частотою, вимірюваною в Герцах (Гц). Наприклад, 10 Гц означає, що перетворювач 10 разів на секунду зчитує звукову хвилю. Але для оцифрування звуку 10 Гц — це дуже мало. Мало тому, що дуже велика частина інформації буде загублена. Тому поширеними частотами є 22,050 Гц або 44,100 Гц. Така частота дає змогу перетворювати аналоговий сигнал у досить якісну цифрову версію. Частота 22,050 Гц часто використовується в машинному навчанні, оскільки дає змогу захопити достатньо деталей у звуці, водночас отримуючи файли відносно невеликого розміру.

Отже, уявімо, що ми зчитали сигнал із частотою 22,050 Гц. Тепер файл із такою частотою (її називають sample rate) лежить у сховищі комп'ютера. Тепер із ним можна працювати. По ходу опису методу обробки, розглянемо специфічну для звуку термінологію.

Ми вже говорили про таке поняття, як фреймінг — розбиття аудіосигналу на фрагменти. Ці фрагменти називаються кадрами. Або, по-іншому, оскільки в цифровому вигляді сигнал — це послідовність семплів, то кадр — це підпослідовність цих семплів. Насамперед необхідно зрозуміти, для чого застосовується цей фреймінг.

Звук — це непостійний сигнал. Однак багато методів аналізу сигналу (зокрема, за допомогою перетворення Фур'є) призначені для інтерпретації тільки постійних сигналів. Тому, щоб застосувати методи до звуку, ми працюємо з кадрами. Тривалість кадру вибирається залежно від уявлення про те, як швидко змінюється зміст сигналу. Передбачається, що в кадрі сигнал постійний, і тому ми можемо застосувати до нього подібного роду аналіз. Таким чином, фреймінг є обов'язковим інструментом при аналізі природних звукових сигналів.

Отже, перший крок в обробці сигналу ми розібрали — розбиття на кадри. Але й у цього процесу є особливість. А саме перекриття (див. рис. 3).

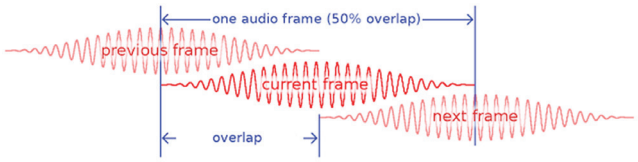

Рис. 3. Перекриття кадрів

Вигода перекриття проявляється на наступних етапах обробки. Але в загальних рисах перекриття дає змогу:

– зберігати залежність сигналу при переході від кадру до кадру

– зберігати дані сигналу після застосування windowing function

Якщо узагальнювати, то перекриття завжди використовується під час фреймінгу. Для розміру кадру і перекриття використовуються такі терміни як frame size і hop size (або frame length та hop length) (див. рис. 4).

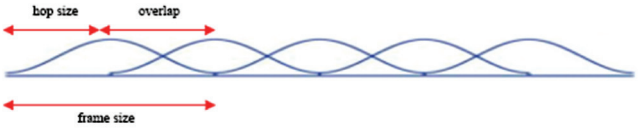

Рис. 4. Frame size, hop size та overlap

Якщо уявити, що ми здійснюємо фреймінг за допомогою ковзного вікна, то hop size — це крок зсуву ковзного вікна. Досить часто hop size роблять удвічі меншим за розмір кадру, що дає змогу зберегти достатньо інформації в обох кадрах.

Розглянемо згадану вище windowing function [4]. Зараз мимохідь згадаємо, що в цій роботі будуть використані виключно Мел-частотні кепстральні коефіцієнти. Це означає, що необхідно перевести сигнал у часово-частотну область, для чого застосовується перетворення Фур'є (а саме STFT). У цьому процесі може відбуватися так звана spectral leakage, для усунення якої і використовується windowing function.

Найпоширенішими такими функціями є Hamming і Hann (див. рис. 5 і 6).

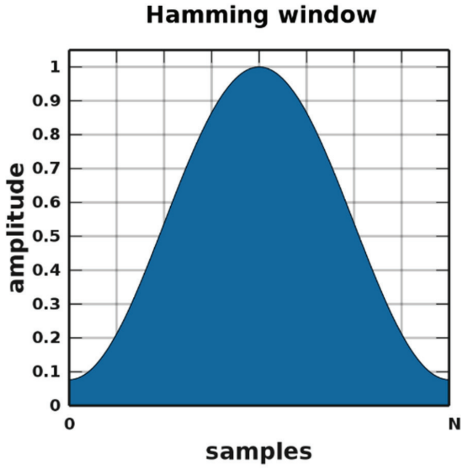

Рис. 5. Hamming windowing function

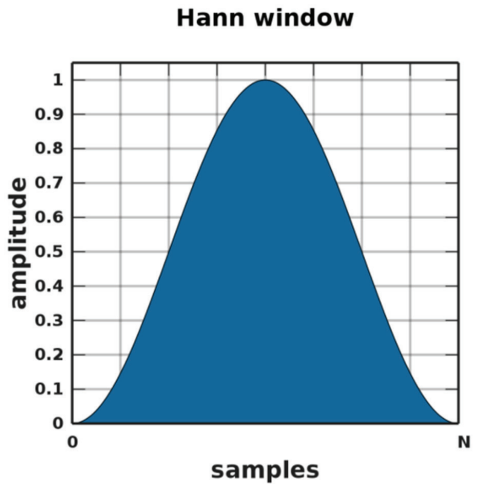

Рис. 6. Hann windowing function

Отже, чому ж без windowing function не вийде адекватного перетворення?

Більшість реальних аудіосигналів неперіодичні, тобто реальні аудіосигнали, як правило, не повторюються в точності протягом будь-якого заданого проміжку часу. Однак математика перетворення Фур'є припускає, що перетворюваний сигнал є періодичним.

Ця невідповідність між припущенням Фур'є про періодичність і реальним фактом, що аудіосигнали, як правило, неперіодичні, призводить до помилок у перетворенні, які й називаються spectral leakage та проявляються у вигляді неправильного розподілу енергії за спектром потужності сигналу.

Щоб дещо пом'якшити такі помилки в перетворенні можна попередньо помножити сигнал на windowing function, розроблену спеціально для цієї мети.

Після застосування, сигнал загасає на краях (див. рис. 7). Тут проявляється друга перевага фреймінгу. Через те, що після множення на windowing function дані сигналу на краях значно загублені, за рахунок перекриття під час аналізу всі дані вихідного сигналу будуть враховані.

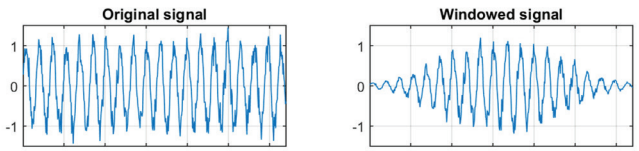

Рис. 7. Сигнал після застосування windowing function

На цьому етапі ми маємо для кожного аудіосигналу набір кадрів із перекриттям і застосованою windowing function. Саме час розглянути параметри аудіосигналу, оскільки наступним кроком буде їх вилучення.

Під час обробки аудіо виокремлюють кілька типів параметрів [5]. Кожен із типів отримують зі свого подання сигналу (часового, частотного або часово частотного).

Базовим поданням (тим, в якому спочатку представлений звуковий сигнал) є часове подання. Це те, як ми звикли бачити сигнал (див. рис. 8). Вісь x відповідає за час, а вісь y — за амплітуду сигналу. З такого подання можна отримати часові параметри, до яких належать amplitude envelope (максимальна амплітуда кадру), root mean square (середньоквадратичне значення амплітуди) тощо. І хоча подібні параметри дають деяке уявлення про характер сигналу, їх абсолютно недостатньо для повного аналізу аудіо. У всякому разі без додаткових параметрів.

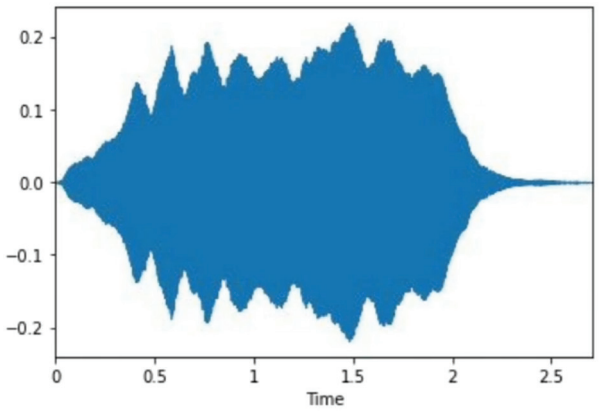

Рис. 8. Сигнал у часовому поданні

Наступне подання називається частотним. Аналогічно до часового подання, проте вісь x відповідає за частоту, а вісь y — за енергію частоти («кількість» частоти в сигналі) (див. рис. 9).

Тобто завдяки такому поданню можна проаналізувати частотний склад усього сигналу та отримати різні спектральні параметри: amplitude spectrum, spectral centroid, spectral bandwidth тощо. Ці параметри вже більш інформативні та можуть показувати набагато більше даних про сигнал, але є ще більш описове подання.

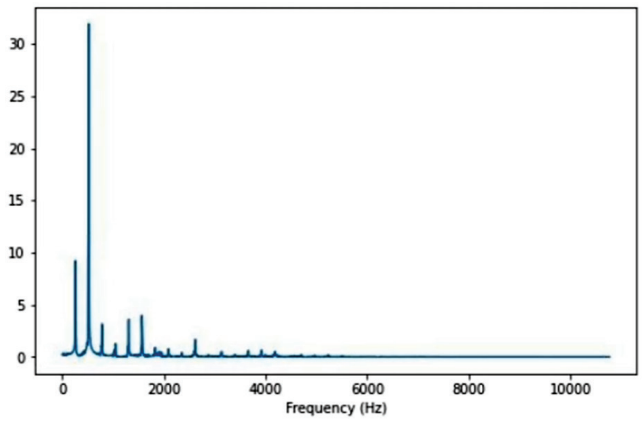

Рис. 9. Сигнал у частотному поданні

Часово-частотне подання має вигляд спектрограми (див. рис. 10). Це складніша структура, яка, однак, дає змогу отримати інформацію про частоту в конкретний проміжок часу. Іншими словами, це те саме частотне представлення, але для дуже маленьких

фрагментів аудіосигналу, а не всього сигналу загалом.

Саме таке подання є найбільш описовим і дає змогу отримувати найрепрезентативніші параметри аудіо, такі як спектрограма, Мел-спектрограми, MFCCs (Мел-частотні кепстральні коефіцієнти). Саме останній параметр найчастіше використовується в подібного роду обробці аудіо, зокрема в завданні розпізнавання емоції за мовленням. Результати з використанням цих коефіцієнтів є найвищими серед подібних досліджень.

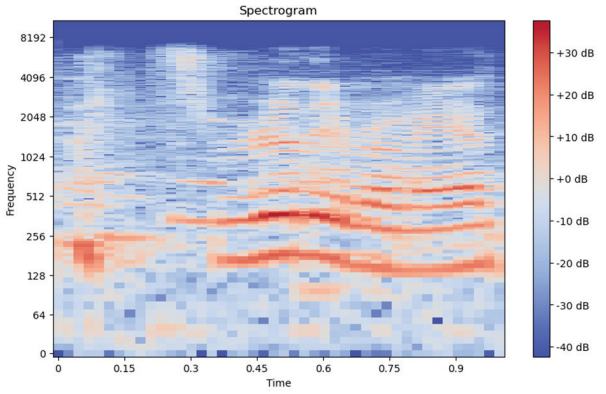

Рис. 10. Сигнал у часово-частотному поданні

Повернемося до того, що ми на даний момент зробили з сигналом. У нас є часове подання сигналу, розділене на кадри з перекриттям і застосованою windowing function. Щоб отримати MFCCs, необхідно кожен кадр перевести в часово-частотне подання. Для цього використовується вже згаданий раніше механізм під назвою перетворення Фур'є.

Перетворення Фур'є (FT) — це метод, який дає змогу розкласти складний сигнал на його базові частоти. В основі цього методу полягає ідея, що будь-який складний сигнал можна уявити як суму кількох простіших сигналів із різними частотами.

Простіше кажучи, можна уявити, що у нас є складний звук, наприклад, музична мелодія або голос. Перетворення Фур'є розбиває цей звук на його складові частини — від найнижчих до найвищих звукових частот, які присутні в цьому сигналі.

Є кілька варіацій цього перетворення: FT, FFT, DFT, STFT. Але всі вони необхідні для однієї цілі — розбити сигнал на частотні складові.

У нашому випадку, необхідний саме STFT (Short-Time Fourier Transform). Цей різновид перетворення дає змогу розбивати сигнал на частоти саме для невеликих фрагментів — кадрів.

Більш детально розглянемо перетворення Фур'є [6]. Можна виділити 3 етапи:

1. Для кожної частоти перетворення Фур'є обчислює комплексні експоненти, які є основними функціями синуса і косинуса

2. Комплексні експоненти множаться на значення вихідного сигналу. Це відбувається для кожної з розглянутих частот

3. Результати множення комплексних експонент на вихідний сигнал підсумовуються для кожної частоти. Це створює спектральні компоненти, що представляють амплітуди і фази кожної частоти у вихідному сигналі.

Таким чином на виході ми отримуємо в графічному поданні спектрограму (у разі застосування STFT) (див. рис. 10).

Перед тим, як отримати MFCCs є ще кілька кроків. Але перед цим розглянемо, що таке Мел і для чого ця шкала використовується в аудіообробці.

Шкала Мел використовується в аудіообробці для представлення частот в більш інтуїтивно зрозумілому вигляді. Вона ґрунтується на сприйнятті звукових частот людиною, відображаючи нелінійний спосіб сприйняття звуку. Тобто не всі частоти рівномірно розподілені. Людський слух більш чутливий до частот нижче 1000 Гц, ніж до частот вище. У шкалі Мел частотні діапазони, що відповідають низьким частотам, розтягнуті, а ті, що відповідають високим частотам, стиснуті. Це дає змогу краще враховувати особливості сприйняття звуку людиною.

Шкала Мел [7] була розроблена на основі досліджень психоакустики [8], яка вивчає сприйняття звуку людиною. Таким чином, вона краще відповідає реальному сприйняттю частоти, ніж лінійні шкали. І одними з найпопулярніших галузей, де використовується ця шкала, є аналіз аудіоданих.

Отже, щоб отримати необхідні MFCCs, потрібно, для початку, перетворити отриману після STFT спектрограму в Мел-спектрограму.

Перетворення частоти f у герцах на частоту m у Мелах представлено у формулі 1.

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

де f – частота в Герцах; m – частота в Мелах.

Частотний діапазон від мінімальної частоти $f_{\min}$ до максимальної частоти $f_{\max}$ ділиться на n Мел фільтрів. Ці фільтри розташовуються на шкалі Мел рівномірно.

Кожен Мел-фільтр являє собою трикутний фільтр, який охоплює певний діапазон частот. Центр фільтра відповідає центральній частоті на шкалі Мел, а його межі сходяться до нуля на сусідніх центральних частотах.

Припустимо, у нас є спектрограма $S(f, t)$, де f — частота, а t — час. Для кожного фільтра $H_i(f)$ і кожної часової точки t розраховується енергія у фільтрі шляхом підсумовування значень спектрограми, помножених на значення фільтра. Тоді за формулою 2 можна отримати матрицю i-го фільтру у певний час.

$$M(i,t) = \sum_f S(f,t) * H_i(f), \quad (2)$$

де M – матриця; i – номер фільтру; t – час; $S(f, t)$ – спектрограма; $H_i(f)$ – Мел-фільтр.

Для поліпшення сприйняття й аналізу до отриманих значень енергії застосовується логарифмічна шкала.

Таким чином, ми отримуємо Мел-спектрограму (див. рис. 11).

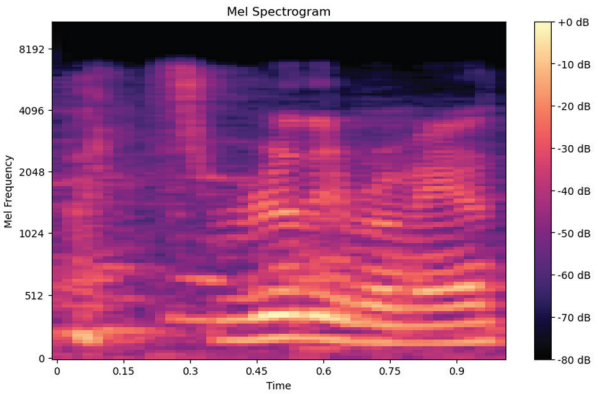

Рис. 11. Мел-спектрограма

Подальші кроки є фінальними для отримання Мел-частотних кепстральних коефіцієнтів. Ці коефіцієнти — це набір параметрів, які являють собою компактний та інформативний опис аудіосигналу, що відображає особливості людського сприйняття звуку. У графічній репрезентації MFCCs можна зобразити у вигляді спектрограми (див. рис. 12).

Отже, для отримання MFCCs необхідно до Мел-спектрограми застосувати дискретне косинусне перетворення (DCT). Це математичне перетворення, яке використовується для перетворення послідовності чисел у набір коефіцієнтів. Воно схоже на перетворення Фур'є, але замість використання комплексних експонентів використовує косинуси. У чому ж перевага такого підходу?

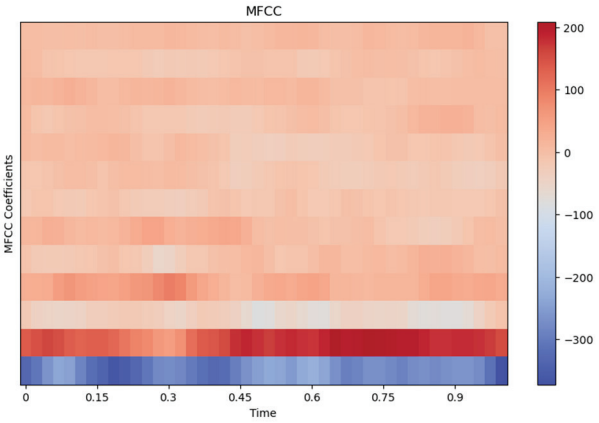

Рис. 12. Представлення MFCCs у вигляді спектрограми

DCT зменшує кореляцію між ознаками, що допомагає поліпшити продуктивність алгоритмів машинного навчання, які використовують ці ознаки.

Також, оскільки DCT концентрує інформацію в декількох коефіцієнтах, можна використовувати тільки перші кілька коефіцієнтів, відкидаючи інші без значної втрати інформації.

MFCCs, отримані з використанням DCT, менш схильні до шуму та викривлень, що робить їх надійнішими для завдань розпізнавання мови та звуків.

Щодо кількості коефіцієнтів MFCC, можна сказати таке. Невелика кількість коефіцієнтів дає змогу зменшити розмір даних і підвищити швидкість обробки таких даних. Однак, значна частина даних за невеликої кількості коефіцієнтів може бути загублена.

Велика ж кількість збільшує розмір даних, зменшує швидкість опрацювання, проте може містити набагато більше потрібної інформації. Таким чином варто досліджувати оптимальну кількість коефіцієнтів для конкретного завдання для отримання найкращих результатів.

Підбиваючи підсумки, ми описали всі основні етапи роботи з аудіо. Визначили необхідну термінологію, розглянули кроки отримання параметрів з аудіосигналу для подальшого їхнього опрацювання в нейронній мережі. Ми визначили роль фреймінгу та кадрів у процесі обробки і подальші експерименти спрямовані на вивчення впливу розміру кадрів, а також розміру зсуву/перекриття на якість моделей, які можна отримати.

2. Інструменти розробки

Для проведення експериментів з метою аналізу впливу розміру кадру на точність моделі нейронної мережі насамперед використовували бібліотеку librosa, що має великий функціонал для роботи з аудіо, зокрема:

– зчитування аудіо
– вилучення параметрів
– аугментація

Використовувалася мова програмування Python і середовище розробки JupyterLab. А конфігурація системи, на якій проводилося навчання моделей, така:

– NVIDIA RTX 3060 12GB
– Ryzen 5 3600X
– RAM 32GB 3200MHz

3. Експериментальні дослідження

Проведення експериментів почалося з вилучення параметрів аудіо, про які йшлося раніше, а саме MFCCs.

Як набір даних для задачі було обрано CREMA-D — великий набір аудіо- та візуальних даних для задачі розпізнавання емоції за аудіо [9]. Трохи деталей про цей набір:

– містить аудіозаписи понад 90 професійних акторів, кожен з яких зачитує 12 фраз на 4 рівнях емоційності

– містить записи 6 емоцій (злість, щастя, відраза, страх, смуток і нейтральний стан)

– має рівномірний розподіл емоцій у наборі та містить загалом 7,442 аудіозаписи

Цей набір є одним із найбільш репрезентативних серед усіх проаналізованих наборів знайдених у відкритому доступі для задачі розпізнавання емоції за мовленням. Цей набір балансує між достатньою кількістю даних і якістю цих даних. Для дослідження було взято 4 емоції (щастя, злість, смуток і нейтральний стан), чого буде достатньо для поставленої задачі.

Отже, перед безпосередньо проведенням експериментів було проведено підготовку, до якої входить кілька кроків:

1) вилучення MFCCs із записів
2) аугментація даних
3) створення архітектури моделі згорткової нейронної мережі

Вилучення звукових параметрів було проведено за схемою, описаною раніше. Однак, уже тут для кожного експерименту була своя особливість. Оскільки метою дослідження є вивчення впливу розміру кадру на точність моделі, а розмір кадру вказується вже на цьому етапі, було підготовлено такі набори параметрів:

– кадр 2048, зсув 1024 (перекриття 50%)
– кадр 2048, зсув 512 (перекриття 75%)
– кадр 1024, зсув 512 (перекриття 50%)
– кадр 1024, зсув 256 (перекриття 75%)
– кадр 512, зсув 256 (перекриття 50%)
– кадр 512, зсув 128 (перекриття 75%)
– кадр динамічний, кадр в 2 рази менший за розмір кадру (перекриття 50%)

Таким чином, було отримано 7 наборів параметрів (MFCCs). Далі будемо позначати кожен із цих наборів як набір із розміром кадру X/динамічним і перекриттям N%.

Одночасно з вилученням параметрів було проведено аугментацію для кожного набору параметрів.

Аугментація — це особливий спосіб розширення набору даних, який використовує вже наявні дані для створення нових шляхом застосування до цих даних спеціальних операцій.

Одними з найпоширеніших таких операцій (методів аугментації) для аудіоданих є:

– додавання шуму
– розтягнення і стиснення в часі
– зміна висоти тону

Метою даних експериментів не є вивчення впливу різних методів аугментації на якість одержуваних моделей, тому в експериментах цієї роботи використовувалися всі три методи. Тобто для кожного набору параметрів вибірку CREMA-D було розширено шляхом додавання шуму до кожного запису, прискорення та сповільнення запису, підвищення та зниження висоти тону.

Важливо зазначити, що через фіксований розмір кадру не кожен запис може цілком потрапити у фінальний набір. Наприклад, якщо наш аудіозапис складається з 1300 семплів, a frame size і hop size дорівнюють 256 і 128 відповідно, то тільки 10 кадрів можна вилучити з такого запису (128 * 10 = 1280 семплів). 20 семплів запису просто не потраплять у навчання. Подібне можна вирішити, «дорощуючи» необхідну кількість семплів для цілого кадру, наприклад, нулями. Тоді весь запис буде враховано в навчанні, однак також буде враховано і додані нулі, які не є частиною запису, що, теоретично, може негативно впливати на якість моделі.

Проте в цій роботі не було використано додавання нулів. Однак, був придуманий альтернативний спосіб врахування повного аудіозапису для навчання.

Оскільки моделі, які будуть використані, вимагають фіксований розмір даних на вхід, необхідно задовольнити цю умову. Якщо в першому підході обмеження виходять від розміру кадру і фіксованої тривалості аудіозапису, який можна отримати, то наступний підхід не вимагає встановлювати обмеження на тривалість запису, хоча і передбачає приблизно однакову тривалість. Таким чином було придумано підхід динамічного фреймінгу. У такому разі фіксованою стає кількість кадрів, яку необхідно вилучити з аудіозапису. У цьому дослідженні кількістю кадрів було встановлено 128. Це дало змогу отримувати кадри розміром від 500 семплів до 2000-3000 семплів, що є доволі адекватним розміром, співмірним із тим, що було використано у наборах з фіксованим розміром кадру.

Таким чином, набори параметрів із фіксованим розміром кадрів враховують записи не повністю, тоді як набори з динамічним розміром кадрів — повністю. Варто нагадати, що розмір кадру виходить з тієї думки, що впродовж усього кадру сигнал статичний і не змінюється. Це і буде перевірено.

Отже, після отримання наборів параметрів з урахуванням аугментації, було розроблено архітектуру моделі нейронної мережі на основі згорткових шарів [10] (див. рис. 13).

Модель містить усі основні компоненти базової згорткової мережі. Згорткові блоки (повторюваний набір шарів, де відбувається операція згортання) складаються з:

– згортковий шар (розмір фільтрів 5 на 5)
– шар нормалізації (дає змогу прискорити навчання)
– шар об'єднання (який виділяє найбільш значущі патерни характеристик у даних)
– шар відсіву (відключає деякі нейрони під час навчання, що перешкоджає перенавчанню нейронної мережі та сприяє підвищенню якості моделі)

Модель містить 6 таких згорткових блоків, після яких йде шар Flatten, для перетворення параметрів після згорткових шарів у вектор. Після якого

йде повнозв'язний шар мережі з активацією ReLU. Останнім шаром мережі є шар із 4 нейронів, кожен з яких відповідає одній з емоцій. Цей останній шар із функцією активації Softmax відповідає за безпосередньо визначення ймовірності класу аудіозапису [11].

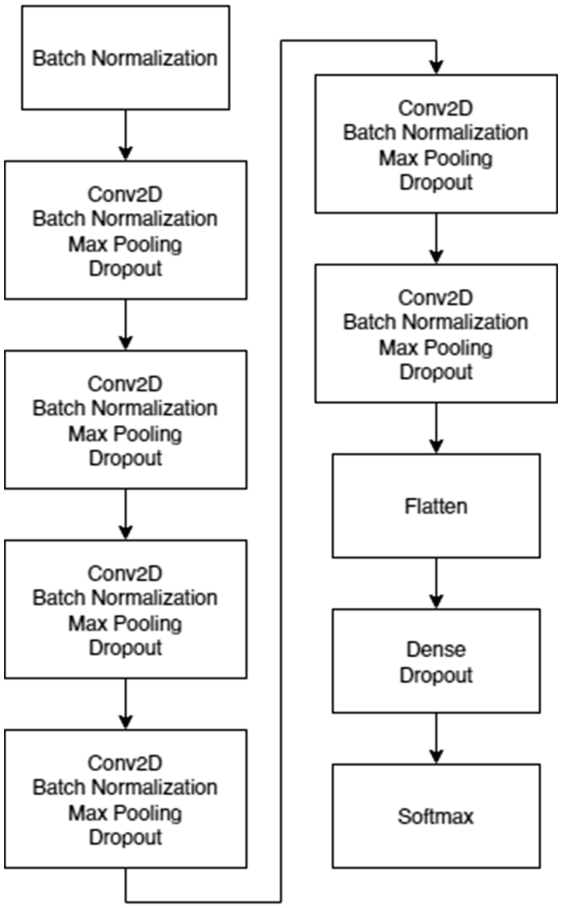

Рис. 13. Графічне зображення архітектури моделі нейронної мережі

Варто описати методику оцінювання моделей.

По-перше, для зняття метрик, які будуть описані нижче, використовувався підхід K-Fold Cross Validation. Цей метод оцінювання моделі користується широкою популярністю в машинному навчанні й дає змогу найбільш об'єктивно оцінити якість математичної моделі. Цей метод також менш чутливий до нерівномірного розподілу класів у наборі, якщо такий є.

Основна суть цього підходу оцінки полягає в тому, що набір даних розбивається на K частин. Далі протягом K ітерацій 1 з K частин виступає як тестова вибірка для моделі, а на решті K-1 частин виконується навчання. Таким чином усі дані використовуються для тестування, при цьому зберігається їхня незалежність.

На кожній із цих K ітерацій знімають метрики, які потім усереднюють і описують оцінку якості моделі на наборі даних.

У цьому дослідженні використовували популярні стандартні метрики під час оцінювання моделей нейронних мереж:

– accuracy
– precision
– recall
– f1-score

F1-score (або F-міра) є однією з найефективніших метрик, оскільки враховує і precision, і recall.

Отже, усі підготовчі етапи описано. Далі було проведено безпосередньо експерименти, які складаються з:

– навчання нейронної мережі з використанням методу K-Fold Cross Validation
– зняття та усереднення метрик

Таким чином, було проведено 7 експериментів для кожного набору параметрів. Для більш компактного відображення результатів зобразимо в таблиці тільки f1-score для всіх 7 наборів.

Таблиця 1 містить значення F-міри для чотирьох емоцій на яких проводилися експерименти.

Також до комірок таблиці застосоване умовне форматування, щоб наочно було видно кольором, як саме змінюється якість моделі в залежності від розміру кадру.

Варто також сказати, що, не дивлячись на рівномірність даних у наборі, після вилучення параметрів для тих утворених наборів параметрів із фіксованим розміром кадру розподіл екземплярів класів (емоцій) вже не такий рівномірний, оскільки не усі аудіозаписи змогли задовольнити умови розміру кадру та тривалості запису (аудіозапис повинен складати не менше 2 секунд).

Таблиця 1

Порівняння якості моделей розроблених при різному розмірі кадру

	frame 2048		**frame 1024**		**frame 512**		**dynamic frame**
overlap	**50%**	**75%**	**50%**	**75%**	**50%**	**75%**	**50%**
anger	0,958	0,956	0,954	0,962	0,953	0,960	0,995
happiness	0,922	0,918	0,913	0,930	0,908	0,922	0,989
sadness	0,934	0,943	0,933	0,945	0,925	0,940	0,986
neutral	0,911	0,921	0,906	0,924	0,899	0,916	0,981
avg	0,931	0,935	0,927	0,940	0,921	0,934	0,988

Тож, для фіксованих кадрів найбільша кількість екземплярів для тестування була для емоцій злості та суму, нейтральний стан йшов на останньому місці.

Отже, які висновки можна зробити із наданої таблиці?

По-перше, видно, що значення f-міри для фіксованого розміру кадру майже не відрізняються для усіх варіацій кадрів. Тобто для усіх них точність моделей тримається на рівні 92-94%.

По-друге, можна побачити, що збільшення перекриття (або зменшення hop size) дійсно позитивно впливає на якість моделей. Для усіх проведених експериментів якість моделей при перекритті на 75% вище за моделі із перекриттям 50%.

По-третє, якщо розглядати лише експерименти із фіксованим розміром кадру, то розмір кадру в 1024 семпли з перекриттям 75% демонструє найвищу якість моделі.

Розглянемо порівняльний графік отриманих результатів (див. рис. 14).

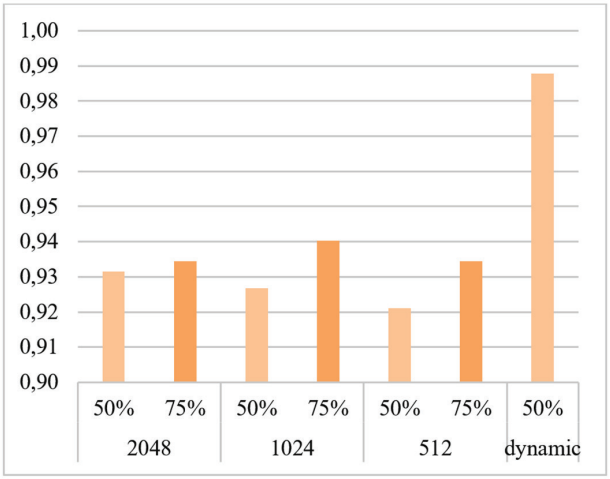

Рис. 14. Значення f1-score при різному розмірі кадру та різному перекритті кадрів

З діаграми чітко видна перевага більшого перекриття. Також видно, що динамічний розмір кадру відіграє значну роль і дозволяє підвищити якість отриманої моделі на 5%. Хоча це значення може здатися невеликим, із динамічним розміром ми отримали модель, яка пропонує точність розпізнавання майже 99%. Звісно, на чотирьох емоціях.

Висновки

У статті було проаналізовано різні механізми, які використовуються під час обробки звуку. Були розглянуті особливості та підходи попередньої обробки аудіосигналу для подальшого їхнього аналізу з використанням нейронної мережі. Як нейронну мережу використовували згорткову мережу, яку самостійно було спроектовано. Результати показали, що фіксований розмір кадру демонструє показники точності значно нижчі, ніж динамічний. При цьому різниця в точності між усіма протестованими (фіксованими) розмірами кадрів дуже незначна. Це може свідчити про те, що для завдання розпізнавання емоції за мовленням розмір кадру відіграє меншу роль для точності, на відміну від повного охоплення аудіосигналу для навчання.

Як подальші дослідження можуть бути проаналізовані інші типи звукових характеристик та їхній вплив на точність, а також вплив типів аугментації (їхній внесок) на загальну точність навченої моделі.

Список літератури

[1] What is speech emotion recognition? – klu. Design, Deploy, and Optimize LLM Apps with Klu – Klu.ai. URL: https://klu.ai/glossary/speech-emotion-recognition (дата звернення: 13.04.2024).

[2] Bevor Sie zu YouTube weitergehen. URL: https://www.youtube.com/@ValerioVelardoTheSoundofAI (дата звернення: 06.03.2024).

[3] Valerio Velardo — The Sound of AI. Understanding audio signals for machine learning, 2020. YouTube. URL: https://www.youtube.com/watch?v=daB9naGBVv4 (дата звернення: 21.03.2024).

[4] Windowing signals – telecommunication engineering. Telecommunication Engineering – My WordPress Blog. URL: https://telecommunicationengineering.softecks.in/535/ (дата звернення: 20.05.2024).

[5] Valerio Velardo — The Sound of AI. Types of audio features for machine learning, 2020. YouTube. URL: https://www.youtube.com/watch?v=ZZ9u1vUtcIA (дата звернення: 03.04.2024).

[6] Valerio Velardo — The Sound of AI. Short-Time fourier transform explained easily, 2020. YouTube. URL: https://www.youtube.com/watch?v=-Yxj3yfvY-4 (дата звернення: 20.05.2024).

[7] Mel. Simon Fraser University. URL: https://www.sfu.ca/sonic-studio-webdav/handbook/Mel.html (дата звернення: 27.04.2024).

[8] Minard A. Psychoacoustics: understanding the listening experience. Ansys Blog. URL: https://www.ansys.com/blog/understanding-psychoacoustics/ (дата звернення: 11.03.2024).

[9] GitHub — cheyneycomputerscience/crema-d: crowd sourced emotional multimodal actors dataset (CREMA-D). GitHub. URL: https://github.com/CheyneyComputerScience/CREMA-D (дата звернення: 17.05.2024).

[10] Basic CNN architecture: explaining 5 layers of convolutional neural network | upgrad blog. upGrad blog. URL: https://www.upgrad.com/blog/basic-cnn-architecture/ (дата звернення: 09.02.2024).

[11] Emotional speech recognition using deep neural networks. PubMed Central (PMC). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8877219/ (дата звернення: 26.05.2024).

Надійшла до редколегії 28.08.2023