УДК \_\_\_\_\_

М. Ю. Гапей, студент

М. А. Фесенко, кандидат технічних наук

Державний університет інформаційно-комунікаційних технологій, Київ

**Проведено аналіз ефективності методики мультиголосового поділу на основі нейронних мереж. Дослідження зосереджується на прояві переваг та потенціалу методики мультиголосового поділу в різних сферах. Було розглянуто сучасні методи, алгоритмни машинного навчання та результати експериментів, щодо покращення мотодик звукового поділу. Висвітлюються перспективи подальшого застосування методики в складному штучному інтелекті, машинному навчанні та робототехніці.**

**Ключові слова:** AI - artificial intelligence; NN - neural network; ML - machine learning; RoPE - rotary position embedding.

**Аналіз ефективності методики мультиголосового поділу на основі нейронних мереж**

**Вступ:**

Розвиток сучасних технологій поступово наближає нас до життя, де люди будуть менш зайняті рутиною, виконання важкої праці буде покладатися на міцному фундаменті робототехніки. Ще до недавна ми могли тільки мріяти, на що здатний AI, а зараз він пише картини, музику, короткі книги і не тільки. Однак такий AI вважається слабким, він здатний працювати лише в конкретному середовищі. Також існує термін - сильний AI, простими словами його можна охарактеризувати як з'єднаних між собою слабких AI. Наприклад сильним AI можна вважати монолітний модуль людино-подібного робота. Деякі складові можуть існувати незалежно, такі як тіло, зір, слух, голос та інші. За допомогою навчання з підкріпленням, проблема ходьби робота, доволі посильне завдання, достатньо створити середовище, з урахуванням фізики та інших факторів, які можна зчитати з середовища та параметрів робота. Хоча ходьба доволі важлива характеристика людино-подібного робота, такими якостями наділені не лише люди. Існують NN здатні моделювати та розпізнавати людську мову, що є наділяє робота властивостями більш притаманними людям. Але давайте створимо умови, коли одночасно будуть розмовляти декілька людей, або на фому буде грати музика, або ми будемо їхати в машині з відкритими вікнами, чи зможе NN адекватно моделювати та прогнозувати? Нажаль ні, ось приклад більш поширенішої ситуації: коли користувач вмикає VPN, то деякі веб-сайти пропонують пройти перевірку на робота. Людині доволі просто пройти перевірку, а от вірусні программи, які мають у своєму арсеналі NN та інші алгоритми ML, наврядчи зможуть потрапити у веб простір, оскільки картинки, які використовуються при перевірці, містять шум та інші поверхнево-псуючі елементи. Додаючи шум в інформацію, ми перешкоджаємо роботі NN, що може бути корисним, та поганим з іншого боку. Технології розпізнавання голосу та інших звукових елементів, з самого їх початку, подолали доволі довгий і складний шлях. Зараз ми можемо розділити вокал від музики, і навіть розділити музичну композицію на інструментальні складові (піаніно, барабан, гітара та інші), також зовсім донедавна стала можливість мультиголосового поділу через мітки. Однак даний підхід не позбавлений недоліків. Щоб вилучити потрібний голос, потрібна наявність цього голосу в окремому джерелі звуку, що не завжди є можливим. Все ускладнюється тим, що AI не відомо які голоса поєднуються, тому простий метод "грубої сили", може бути доволі тривалим та ресурсоємним процесом, що може негативно вплинути на автономність роботів.

Дослідження показало, що впровадження мультиголосового поділу, стане ще одним кроком до створення сильного AI. Можливість аналізу та синтезу людської мови дозволить AI ординарно та швидко обробляти інформацію в середовищі із різною кількості людських голосів. Методика MusicSourceSeparation дозволяє отримувати зі звуку складові компоненти музичних інструментів, в порівнянні із запропонованою методикою мультиголосового поділу ці технології доволі схожі за змістом, але виконують протилежні один одному завдання [1]. Наближена методика AudioSep частково покриває сферу застосування мультиголосового поділу, але не вирішує проблему нерозмічених данних [2]. Використання методики VoiceFilter в парі з мультиголосовим поділом дозволить отримувати більш точні та очікувані результати, оскільки попередня обробка виконується доволі швидко [3].

Результати дослідження показали, що впровадження мультиголосового поділу є комплементарною складовою технологій розпізнавання інформації усіх видів, якы стрімко розвиваються, що відкриває нові можливості для інженерів, які можуть використовувати їх у власних системах. Дослідження також порівняло мультиголосовий поділ із методиками MusicSourceSeparation та AudioSep. Було встановлено, що, хоча ці методики мають спільні риси, вони вирішують різні завдання. Нещодавні досягнення у розпізнаванні звукової інформації вже дозволяють машинам легко розрізняти голоси людей. Однак, коли додаються фонові шуми, такі як звуки дощу або крик немовляти, завдання стає складнішим. Інтеграція мультиголосового поділу та VoiceFilter може ефективно вирішити цю проблему, забезпечуючи більш надійну роботу AI у реальних умовах. Розвиток та інтеграція технологій мультиголосового поділу та інших сучасних методів аудіообробки відкриває нові можливості для створення більш потужних та універсальних систем AI, здатних ефективно взаємодіяти з людьми в різних умовах. Це не тільки підвищує функціональність існуючих технологій, але й сприяє розвитку нових інноваційних рішень у сфері робототехніки та автоматизації.

**Метою дослідження є** -аналіз ефективності методики мультиголосового поділу на основі нейронних мереж.

Основні завдання даного дослідження включають:

* Оцінка поточних методик звукового поділу з використанням їх слабких та сильних сторон;
* Дослідження інноваційних підходів до використання звукового поділу в сферах штучного інтелекту, автоматизації та робототехніки;
* Визначення оптимальної методики на основі нейронних мереж для досягнення найкращих результатів мультиголосового поділу;
* Розробити методику мультиголосового поділу на основі нейронних мереж.

**Основна частина:**

**Оцінка поточних методик звукового поділу.**

AudioSep. Сильні сторони: Універсальність - підходить для різноманітних завдань, таких як виділення вокалу, інструментів або мовлення. Сліпе розділення джерел, яке виконується на основі текстового опису. Висока точність завдяки використанню спектральних і часових характеристик аудіосигналу. Слабкі сторони: Може вимагати значних обчислювальних ресурсів. Залежність від якості спектрального аналізу. Може не працювати ефективно в реальному часі.

VoiceFilter. Сильні сторони: Ефективно ізолює людське мовлення від різного шуму. Використання глибокого навчання - моделі навчені на великих наборах даних змішаних аудіозаписів і чистих мовних сигналів. Слабкі сторони: Вузька спеціалізація - менш ефективний для інших типів аудіоджерел. Високі вимоги до даних для навчання моделі.

MusicSourceSeparation. Сильні сторони: Оптимізовано для музичних джерел - відокремлює вокал, ударні, бас та інші інструменти. Використовує різноманітні методики, включаючи спектральне моделювання та машинне навчання. Підходить для створення музики, реміксування, автоматичної транскрипції музики. Слабкі сторони: Може мати труднощі з відокремленням подібних за характеристиками інструментів. Ефективність може залежати від складності музичного твору.

**Інноваційні підходи до звукового поділу.**

* Використання комбінованих методів, таких як поєднання глибокого навчання з традиційними методами (наприклад, спектральне моделювання), може покращити результати.
* Збільшення обсягів даних для навчання моделей глибокого навчання може суттєво покращити точність та адаптивність моделей.
* Розробка оптимізованих моделей, здатних працювати в реальному часі, може розширити застосування у сферах AI, автоматизації та робототехніки.
* Використання систем, здатних адаптуватися до нових або незнайомих звукових середовищ, може підвищити ефективність у різних умовах.

**Оптимальна методика на основі нейронних мереж.**

Для досягнення найкращих результатів у мультиголосовому поділі, можна використовувати комбінацію згорткових та рекурентних NN, оскільки згордкові NN добре обробляють спектральні характеристики аудіо данних, а рекуретні NN ефективно працюють з часовими характеристиками.

**Розробка методики мультиголосового поділу.**

1. Збір даних - зібрати великий набір даних змішаних аудіозаписів та відповідних чистих компонентів.
2. Попередня обробка даних - виконати спектральний аналіз аудіозаписів, перетворивши їх у спектрограми.
3. Умовна архітектура моделі:
   * Convolutional NN шар: Використовується для виділення спектральних характеристик.
   * Recurrent NN шар: Використовується для аналізу часових залежностей у даних.
   * Fully Connected NN шар: Для об’єднання ознак та остаточного прогнозування.
4. Навчання моделі: Використати великий набір даних для навчання моделі з використанням оптимізатора (наприклад Adam) та функції втрат (наприклад MSE - mean squared error).
5. Валідація та тестування: Розділити дані на тренувальний, валідаційний та тестовий набори. Валідувати модель на валідаційному наборі та оцінити точність на тестовому наборі.
6. Оптимізація: Впровадити механізми регуляризації та техніки зменшення перенавчання, такі як Dropout.

**Реалізація в реальному часі.**

Реалізація технології мультиголосового поділу в реальному часі є складним, але досяжним завданням. Нижче наведені ключові аспекти, які необхідно враховувати для успішного виконання цього завдання.

Оптимізація моделі: Щоб забезпечити роботу в реальному часі, модель повинна бути оптимізована для швидкої обробки. Зменшення кількості шарів або параметрів у моделі, зберігаючи при цьому прийнятну точність підвищить продуктивність та зменшить час сходження алгоритму навчання. Вибір ефективних алгоритмів для обчислень, таких як оптимізовані алгоритми зворотного поширення помилки прискористь навчання та підвищить точність.

Використання апаратного прискорення: Апаратне прискорення, таке як графічні процесори або спеціалізовані чипи для машинного навчання (наприклад Tensor Processing Units), може значно прискорити обробку великих обсягів даних. Вибір правильного апаратного забезпечення може мати вирішальне значення для забезпечення роботи в реальному часі. Розподіл обчислень між кількома процесорами або ядрами може допомогти зменшити час обробки. Це особливо актуально для великих моделей, які потребують значних обчислювальних ресурсів.

Буферизація і потокова обробка: Для обробки аудіосигналів в реальному часі важливо використовувати буфер для зберігання частини вхідного аудіосигналу перед його обробкою, що дозволяє моделі працювати з потоком даних, а не чекати на завершення обробки всього сигналу.

Затримка та оптимізація часових інтервалів: Мінімізація затримки (latency) є важливою для реального часу. Це включає оптимізацію як обробки на стороні апаратного забезпечення, так і ефективне управління часовими інтервалами у програмному забезпеченні. Важливо інтегрувати модель в середовище, яке підтримує реальну часову обробку. Це може бути спеціалізоване програмне забезпечення або платформи, що дозволяють обробляти аудіосигнали в реальному часі, такі як:

* Node.js з модулями для аудіо обробки.
* Python з бібліотеками, такими як PyAudio або SoundDevice.
* C++ з бібліотеками, такими як PortAudio або RtAudio.

Приклад потоку обробки в реальному часі

Захоплення аудіо - захоплення аудіосигналу з мікрофону або іншого джерела вхідного сигналу. Буферизація - збереження захопленого аудіо у буфер для забезпечення безперервного потоку. Попередня обробка - виконання попередньої обробки, такої як нормалізація та перетворення аудіосигналу у спектрограму. Обробка моделлю - пропускання даних через модель для відокремлення компонентів аудіосигналу. Постобробка - перетворення результатів моделі назад в аудіо формат. Вивід - вивід обробленого аудіосигналу через динаміки або інші вихідні пристрої.

**Інструменти та технології для реалізації.**

* TensorFlow та TensorFlow Lite: Використовуються для створення та оптимізації моделей машинного навчання.
* ONNX Runtime: Для запуску моделей глибокого навчання з високою продуктивністю.
* PyTorch: Популярний фреймворк для розробки моделей машинного навчання.
* CUDA та cuDNN: Для прискорення обчислень на графічних процесорах.

**Висновки:**

За результатами дослідження різних методик звукового поділу було встановлено, що модель NN глибокого навчння використана в MusicSourceSeparation на базі трансформерів RoPE з роздільною смугою, та поєднання згорткових і рекурентних NN теоретично дає найкращі результати. Однак різновид людських голосів підвищує вимоги до планування більш розширенної архітектури глибокої NN. Використання методики мультиголосового поділу в різних сферах розширює можливість AI, а також доповнює арсенал обробки звукової інформації використовуючи NN.

**Список використаної літератури:**

1. Music Source Separation with Band-Split RoPE Transformer. *arXiv.org*. URL: <https://arxiv.org/abs/2309.02612>
2. Separate Anything You Describe. URL: <https://audio-agi.github.io/Separate-Anything-You-Describe/>.
3. VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking. *arXiv.org*. URL: <https://arxiv.org/abs/1810.04826>