Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерної інженерії та управління
(повна назва)

Кафедра Автоматизації проектування обчислювальної техніки
(повна назва)

АТЕСТАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)
(рівень вищої освіти)

Система розпізнавання голосу в режимі реального часу
(тема)

Виконав: студент 2 курсу, групи СКСм-19-1

Михайліченко І.В.
(прізвище, ініціали)

Спеціальність 123 Комп'ютерна інженерія
(код і повна назва спеціальності)

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

Освітня програма

Спеціалізовані комп'ютерні системи
(повна назва освітньої програми)

Керівник роботи доц. Рахліс Д.Ю.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ Чумаченко С.В.
(підпис) (прізвище, ініціали)

2020 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерної інженерії та управління
Кафедра Автоматизації проектування обчислювальної техніки
Рівень вищої освіти другий (магістерський)
Спеціальність 123 – Комп'ютерна інженерія
Тип програми Освітньо-професійна
Освітня програма Спеціалізовані комп'ютерні системи
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 20 ___ р.

ЗАВДАННЯ
НА АТЕСТАЦІЙНУ РОБОТУ

студентові Михайліченко Ігорю Володимировичу
(прізвище, ім'я, по батькові)

1. Тема роботи Система розпізнавання голосу в режимі реального часу

затверджена наказом по університету від 30 10 2020 р. № 1489 Ст
2. Термін подання студентом роботи до екзаменаційної комісії 20 12 2020 р.
3. Вихідні дані до роботи
Мова програмування Python
IDE PyCharm
Wav файли

4. Перелік питань, що потрібно опрацювати в роботі
1) Сфери використання систем розпізнавання голосу
2) Методи розпізнавання мови людини
3) Методи виділення характерних ознак звукового сигналу
4) Методи і засоби вирішення задачі класифікації

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів)

14 слайдів в форматі *.pptx

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

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка конс	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Видача теми проекту, узгодження і затвердження теми	03.09.2020 – 10.09.2020	
2	Аналіз предметної галузі, постановка задачі, вибір інструментальних засобів	10.09.2020 – 30.09.2020	
3	Проектування систем логічного управління	30.09.2020 – 15.10.2020	
4	Дослідження існуючих методів розв'язання задачі ідентифікації	15.10.2020 – 15.11.2020	
5	Реалізація обраного алгоритму	15.11.2020 – 25.11.2020	
6	Програмна реалізація алгоритму	25.11.2020 – 05.12.2020	
7	Перевірка якості роботи системи	05.12.2020 – 15.12.2020	
8	Оформлення пояснювальної записки	15.12.2020 – 18.12.2020	
9	Перевірка виконаного проекту керівником	18.12.2020 – 20.12.2020	
10	Захист проекту	23.12.2020	

Дата видачі завдання 03.09.2020

Студент _____ _____
(підпис)

Керівник роботи _____ доц. каф. АПОТ Рахліс Д.Ю.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка містить 79 сторінок, 35 рисунків, 12 джерел посилання.

НЕЙРОННІ МЕРЕЖІ, РОЗПІЗНАВАННЯ КОМАНД ПО ГОЛОСУ, ДИСКРЕТИЗАЦІЯ, АУДІОЗАПИС, СЕГМЕНТАЦИЯ, ОБРОБКА ЗВУКОВОГО СИГНАЛА, КЛІЄНТ-СЕРВЕР, ПЕРЕТВОРЕННЯ ФУР'Е.

Метою атестаційної роботи є розробка системи розпізнавання команд людини по голосу в режимі реального часу. В роботі було досліджено існуючі методи розв'язання поставленої задачі, способи оцінки їх якості, а також існуючі проблеми та обмеження. Програмна реалізація алгоритму здійснена на мові програмування Python. В IDE PyCharm було створено програмний продукт, який має можливість працювати з wav-файлами. Проведено експериментальне дослідження розробленого алгоритму.

Моделювання розробленої системи проводилося у вигляді макетування спрощеної схемної реалізації системи та реалізацією бездротового зв’язку між сервером та апаратною платформою.

ABSTRACT

Master’s thesis contains 79 pages, 35 figures, 12 sources according to the list of links.

NEURAL NETWORKS, VOICE COMMAND RECOGNITION, DISCRETIZATION, AUDIO RECORDING, SEGMENTATION, SOUND PROCESSING, CLIENT-CUSTOMER.

The purpose of the work is to develop a system of human voice recognition in real time. The paper examines the existing methods of solving the problem, ways to assess their quality, as well as existing problems and limitations. The software implementation of the algorithm is carried out in the Python programming language. With PyCharm IDE was created a software product that can work with wav files. An experimental testing of the developed algorithm was performed.

The modeling of the developed system was carried out in the form of a layout of a simplified circuit implementation of the system and the implementation of wireless communication between the server and the hardware platform.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

API – Application programming interface (прикладний програмний код);

GMM – Gauss Mixture Model (модель Гауссовых суміщів);

HMM – Hidden Markov Model (скриті Марковскі моделі);

IDE – Integrated Development Environment (інтегроване середовище розробки);

LPCC – Linear Predictive Cepstral Coefficients (кодування з лінійним предиктором);

MFCC – Mel-Frequency Cepstrum Coeffcents (мел-частотні кепстральні коефіцієнти);

SVM – Support vector machine (метод опорних векторів);

Wav – Waveform Audio File Format (формат файлу-контейнера для зберігання записи оцифрованого аудіопотоку);

ІТ – інформаційні технології ;

Кепстр – спектр спектра сигнала;

Мел – одиниця виміру висоти сприйманого звуку;

НМ – нейронна мережа;

ОС – операційна система;

ПЗ – програмне забезпечення;

Спектр – розподіл деякої фізичної величини за іншою величиною;

ЦП – цифрові пристрої;

ШІ – штучний інтелект.

ВСТУП

У нашому світі технології розвиваються з такою швидкістю, що прогресивні, на перший погляд, підходи та реалізації сьогодні, вважаються застарілими вже на наступний день. Інформаційні технології та комп’ютерне забезпечення набуло таких обертів, що, майже, нема чим дивувати. Здавалось би, більше нікуди рости. Але відповідь на такий вислів є – системи зі штучним інтелектом та голосове управління. Наприклад, сучасний автомобіль – має змогу побудувати маршрут, прокласти шлях, довезти без участі водія. Водієві майже нічого не потрібно робити, лише правильно натискати кнопки. Наступний етап – позбутися навіть їх. Розробити такий програмний додаток, який буде сприймати ваш голос та відштовхуючись від команд, буде виконувати їх. Інший приклад – це «розумний» помічник, який має змогу управляти вашим розумним будинком, спілкуватися з вами та бути особистим асистентом.

Звучить досить прогресивно та багатообіцяюче, але така задача не є достатньо простою. Одним з важливих завдань, для якого знаходиться рішення під час розробки подібних керуючих систем – завдання поліпшення рівня точності розпізнавання команд за голосом. Підвищення якості ведеться у напрямку покращення її надійності, зниження впливу фонового шуму на якість розпізнавання.

Тому пропозиція на такі системи зараз обмежена, хоча попит зростає кожен день, через складність побудови і проблем пов'язаних з темою голосової безпеки.

Було прийнято рішення спроектувати програмне забезпечення, яке володіє достатніми параметрами стійкості до шуму та високою точністю обробки мови. Головною метою є знаходження шляхів покращення якості розпізнавання, фільтрація шумів і деяких збудників, які можуть впливати на якість розпізнавання мови.

1 ГОЛОСОВЕ УПРАВЛІННЯ У НАШІ ДНІ

1.1 Інтелектуальне управління та штучний інтелект

Інтелектуальне управління забезпечує автоматизацію за допомогою емуляції біологічного інтелекту. Вона або прагне замінити людину, яка виконує контрольне завдання (наприклад, оператор хімічного процесу), або вона запозичує ідеї від того, як біологічні системи вирішують проблеми, і застосовує їх до вирішення проблем управління. Зазвичай, у такому підході не можливо обійтися без нейронних мереж, які є частиною штучного інтелекту. Нейронні мережі схожі з подібним завданням, яке полягає у застосуванні комп'ютерів для кращого зрозуміння людського інтелекту, хоча не зовсім завжди у якості основи можуть обрати біологічно схожі методи.

Штучний інтелект (ШІ) можна відокремити у моделювання людського мислення у машинах, які і створювались для того, щоб міркувати як люди та мати змогу імітувати їх дії. Таке визначення можна використати для будь-якої системи, яка має змогу відокремлювати риси, які відображають людський розум, наприклад: навчання та рішення проблем.

Найбільш точною характеристикою ШІ, мабуть, є здатність раціоналізувати та виконувати дії, які просто тут і зараз можуть найбільш ефективно та за короткий період часу досягти кінечної мети. Якщо людина чує термін, в розмові, штучний інтелект, скоріш за все, те, про що вона одразу думає, – це роботи. Так виходить через те, що великі фільми та романи поєднують історії про машини, які мають з людиною спільні риси, та які одразу ж руйнують землю.

Штучний інтелект полягає у підході, згідно до якого людський розум можна представити так, щоб система могла з легкістю його копіювати та виконувати завдання, як найпростіші, так і найскладніші. До цілей ШІ завжди відносять – навчання, міркування та сприйняття. Під час еволюції технологій,

нові показники витісняли попередні та робили їх застарілими. Розглянемо приклад, коли системи, які вираховують основні функції або зчитують текст опираючись на оптимальне розпізнавання символів, вже не можуть вважатися продуктом зі штучним інтелектом під капотом, так як ця здатність тепер вважається належним, вбудованою функцією комп’ютера.

Термін штучний інтелект був вперше введений Джоном Маккарті в 1956 році, коли він провів першу академічну конференцію з цього питання. Але пошуки відповіді на питання «Чи можуть машини справді мислити?», розпочались набагато раніше. В основі роботи Ванневара Буша «Як ми можемо думати» він запропонував систему, яка збільшує знання та розуміння самих людей. П'ять років потому, Алан Тьюрінг написав статтю про машини, які здатні імітувати людей та здатність робити розумні речі, наприклад, грати в шахи. Ніхто не може спростувати здатність комп’ютера обробляти логіку, але для багатьох невідомо, чи може машина мислити. Існувала навіть певна суперечка щодо того, чи це можливо взагалі. Наприклад, існує так званий аргумент «китайська кімната». Уявіть, що група людей замикається в кімнаті, куди їм передають записки китайською мовою. При використання цілої бібліотеки правил та таблиць пошуку вони змогли б дати правильні відповіді китайською мовою, але чи справді вони «розуміли» б цю мову? Аргумент полягає в тому, що, оскільки комп'ютери завжди застосовують пошук за фактом, вони ніколи не можуть "зрозуміти" ні контекст ні тему предмета. Цей аргумент багато разів спростовувався дослідниками, але він підриває віру людей у машини та так звані «експертні системи» у життєво важливих додатках.

Штучний інтелект постійно розвивається на користь багатьох різних галузей. Однією з таких галузей є розпізнавання людської мови.

Якщо проаналізувати історію штучного інтелекту, можна виділити один з важливих напрямків під назвою «моделювання міркувань». Впродовж останніх років розвиток цієї науки рухався саме цим шляхом, тому зараз це один з найрозвиненіших напрямків у сучасному штучному інтелекті.

Міркування на основі моделі – це теорія, яка намагається описати психологічні процеси, які використовуються при логічному висновку з заданого набору передумов. Психологічні моделі – це схематичні зображення можливих результатів, які узгоджуються з передумовами, використовуючи внутрішні лексеми для представлення класів подій або об'єктів. Семантика логічних сполучників (якщо, або тощо) визначає спосіб структури представлення. Це дає обмежений набір можливих результатів, з яких можна прочитати потенційні висновки. Моделювання міркувань займається розробкою та втіленням символьних систем – такі системи, на вході яких подається певна задача, а на виході отримується рішення. Зазвичай, завдання вже повністю чітко описане, це означає, що воно переведене в математичну форму, але рішення поки-що немає, або він дуже трудомісткий, складний. До даного підходу належить: прийняття рішень, доведення теорем, планування і диспетчеризація, прогнозування.

Міркування про психологічні моделі спираються на якісні відносини, а не на кількісні відносини. Люди можуть добре міркувати про те, що одна величина менше іншої, не посилаючись на точні значення величин. Цей принцип лежить в основі якісної теорії процесів.

Психологічні моделі часто включають у себе розумове моделювання: відчуття можливості запустити ментальну модель внутрішньо, щоб можна було спостерігати, як вона буде поводитись і яким буде результат процесу. Процеси, що лежать в основі розумового моделювання, все ще вивчаються. Однак є вагомі докази того, що люди здатні в певних межах міркування імітувати поведінку пристрою, навіть якщо їм просто показують статичний дисплей. Існує очевидний компроміс між онлайн-моделюванням та пошуком збережених результатів. По мірі того, як люди знайомляться з системою, вони більше не виконують повного моделювання поведінки у всіх випадках, а натомість просто отримують доступ до своїх збережених знань про результат.

Інший висновок дослідження психологічних моделей полягає в тому, що люди здатні тримати дві або більше несумісні моделі в одному домені,

зразок, який називають пастишними моделями або знаннями. Наприклад, Коллінз і Гентнер (1987) виявили, що багато суб'єктів-початківців мали "пастишні" моделі міркувань. Новачок, який навчається, може дати одне пояснення того, через що рушник сохне на сонці, і зовсім інше пояснення того, із-за чого калюжа води випаровується, не складаючи жодного зв’язку між цими явищами. Початківці часто використовують узгоджені на місцевому рівні, але суперечливі у всьому світі висновки, часто досить тісно пов'язані з деталями конкретного прикладу. Ця закономірність підкреслює тенденцію початківців навчатись консервативно, адже знання кешовані в дуже конкретних категоріях, що мають контекст. До тих пір, поки кожна модель має вузький доступ в контексті, притаманному їй, невідповідності можуть ніколи не потрапляти в поле зору учня.

Одним з найважливіших напрямків є обробка мови, в рамках якого проводиться аналіз можливості обробки, розуміння, і генерації текстів на основі людської мови. Головною задачею цього напрямку ставиться мета досягнення такої обробки природної мови, яка була б в змозі отримати знання самостійно, шляхом аналізу існуючого тексту, доступного через Інтернет. Існують такі застосування обробки природної мови, що базуються на інформаційному пошуку і машинному перекладі.

Розглянемо деякі з найвідоміших ШІ-систем.

1. Siri та Alexa – голосові асистенти. Проблема з голосовими асистентами, які відіграють дедалі зростаючу роль у бізнесі, полягає в тому, що їм потрібно насправді розуміти людську мову. Ще важче: їм потрібно насправді розуміти людей. Ось тут з’являється ШІ. Хоча системні інженери можуть створювати цих голосових помічників, вони не можуть вбудувати в них величезну кількість людської ідіосинкразії під час запуску. Отже, ці системи широко використовують машинне навчання, щоб дати їм змогу покращуватись та виконувати надзвичайно складні завдання інтерфейсу людина-машина. Озброївшись штучним інтелектом, голосові асистенти стануть дедалі здатнішими перебирати Інтернет, допомагаючи нам робити

покупки та надавати підказки. Існує сподівання, що ця голосова технологія забезпечить домашніх помічників для допомоги у догляді за літніми людьми.

2. Amazon та Інтернет-торгівля.

Концепція системи, яка реагує на вказівки споживача, сама по собі не є прикладом штучного інтелекту. Наприклад, ті оголошення для сорочок, які стежать за вами в Інтернеті після того, як ви випадково перевірили сорочки, не обов'язково є вдосконаленим додатком ШІ. Але у випадку з рекомендаційною системою Amazon це потужність транзакційної платформи ШІ. Ви, напевно, спостерігали його здатність вчитися – і продовжувати вчитися – сезон за сезоном. Загалом, велика армія покупців «навчають» систему штучного інтелекту Amazon краще показувати ймовірні товари для продажу. Тобто, збіг одного предмета з іншим, що був показаний у минулому, спричинить продаж. Відповідно до цього, цій сучасній системі ШІ потрібна величезна обчислювальна платформа для обробки всіх цих даних.

1.2 Поняття голосового управління

Термін «голосове управління» з'явився у науці досить недавно. Він має на увазі під собою перетворення людської мови в цифрову інформацію. Таких перший пристрій з'явився в 1952 році, що мав змогу розпізнавати вимовлені людиною цифри.

Насправді, перша в історії зафіксована спроба технології розпізнавання мови датується 1000 роком н. е. Завдяки розробці інструменту, який нібито міг відповісти «так» чи «ні» на прямі запитання.

Хоча цей експеримент технічно не передбачав обробки голосу в будь-якій формі, ідея, що лежить в його основі, залишається частиною фундаменту технології розпізнавання мовлення: використання природної мови як введення для ініціювання дії. Через століття лабораторії Белла працювали над розробкою системи "Одрі", системи, здатної розпізнавати цифри 1-9, промовлені одним голосом. Пізніше IBM розробила пристрій, який міг

розпізнавати та розрізняти 16 вимовлених слів. Ці успіхи призвели до більшого поширення технологічних компаній, що зосереджуються на мовленнєвих технологіях. Дійсно, навіть Міністерство оборони хотіло взяти участь у цій акції. Повільно, але впевнено розробники рухались до мети, щоб машини могли розуміти та реагувати на все більше і більше наших вербалізованих команд.

Оскільки простота можливості спілкуватися з цифровими асистентами вводить в оману, розпізнавання мови насправді неймовірно складна річ, навіть зараз. Подумайте, як дитина вивчає мову.

З першого дня вони чують, як навколо них вживають слова. Батьки розмовляють зі своєю дитиною, і, хоча дитина не реагує, вони поглинають всілякі словесні сигнали: інтонація, флексія та вимова, їх мозок формує схеми та зв'язки на основі того, як батьки використовують мову.

Хоча може здатися, що людей важко слухати і розуміти, ми насправді все життя тренуємось розвивати цю так звану природну здатність.

Технологія розпізнавання мови працює по суті однаково. Хоча люди вдосконалили цей процес, ми все ще знаходимо найкращі практики роботи з комп'ютерами. Ми повинні навчати їх так, як навчали нас наші батьки та вчителі. І це навчання передбачає багато інноваційного мислення, робочої сили та досліджень.

Удосконалення таких систем розпізнавання мовлення займе набагато більше часу і набагато більше польових даних. Зрештою, існують тисячі мов, наголосів та діалектів.

Це не означає, що ми не прогресуємо – станом на травень 2017 року алгоритми машинного навчання Google досягли 95% рівня точності слів для англійської мови. Цей поточний показник також є порогом для людської точності, майте на увазі.

У центрі голосового управління знаходиться досить зрозуміла схема функціонування – передавач, приймач, вихід. Передавачем голосової інформації може бути окремий модуль, який має змогу передавати голосові

команди, такий як: смартфон з відповідним програмним додатком, персональний комп'ютер, за умови, що є мікрофон та схожа програма [1].

Приймач – це такий пристрій, який потрібен для голосового управління. Він має змогу та задачу не тільки приймати інформацію, але також перетворювати її у цифровий вигляд сигналу або дані для коректного виконання команди або показу тексту. На рис. 1.1 показаний приклад роботи такої системи для голосового управління.

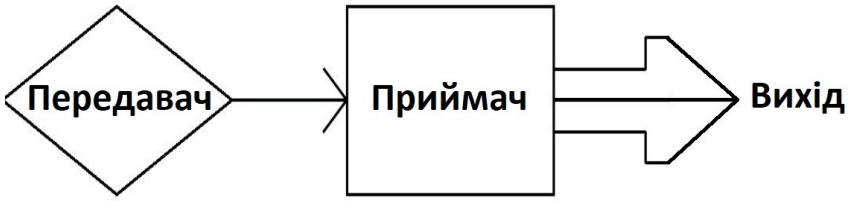

Рисунок 1.1 – Робоча схема голосового управління

1.3 Призначення пристроїв розпізнавання мови

Технологія розпізнавання мовлення дозволяє комп’ютерам приймати звукове аудіо, інтерпретувати його та генерувати з нього текст. Але як комп’ютери розуміють людську мову? Коротка відповідь – чудо обробки сигналів. Мова – це просто серія звукових хвиль, створюваних нашими голосовими акордами, коли вони викликають вібрування повітря навколо них. Ці звукові хвилі записуються мікрофоном, а потім перетворюються в електричний сигнал. Потім сигнал обробляється за допомогою передових технологій обробки сигналів, ізолюючи склади та слова. З часом комп’ютер може навчитися розуміти мову з досвіду завдяки неймовірним останнім досягненням у галузі штучного інтелекту та машинного навчання. Але обробка сигналів – це те, що робить це можливим.

Отже, які переваги технології розпізнавання мови? Чому, власне, нам потрібні комп’ютери, щоб розуміти нашу мову, коли введення тексту швидше? Мова є природним інтерфейсом для багатьох програм, які не

запускаються на комп’ютерах, що стають все більш поширеними. Ось декілька важливих способів, за допомогою яких технологія розпізнавання мовлення відіграє життєво важливу роль у житті людей. Ціллю таких пристроїв для розпізнавання команд по голосу та просто мовлення людини полягає у тому, щоб облегшити доступ, наприклад, до різноманітної техніки у будинку, виробництві, машині, офісній роботі і іншого.

Розмова з роботами: ви можете не думати, що спілкування з роботами є загальним заняттям. Але роботів все частіше використовують у ролях, колись виконуваних людьми, включаючи розмову та інтерфейс. Наприклад, фірми вже досліджують використання роботів та програмного забезпечення для проведення первинних співбесід. Оскільки співбесіди мають бути розмовними, дуже важливо, щоб робот міг інтерпретувати те, що говорить співбесідник. Для цього потрібна технологія розпізнавання мови.

Контроль цифрових пристроїв: цифрові особисті помічники, такі як Alexa та Google Home, очевидно, потребують усного спілкування між людьми та комп’ютерами. Вони також є чудовими прикладами того, як комп’ютери використовують машинне навчання, щоб з часом краще зрозуміти вашу мову через досвід. Але для цього ключовою є технологія розпізнавання мови, що можлива за допомогою обробки сигналів.

Допомога людям із вадами зору та слуху: є багато людей із вадами зору, які покладаються на засоби зчитування з екрану та системи диктування тексту в мову. А перетворення звуку в текст може стати важливим інструментом спілкування для людей із вадами слуху.

Увімкнення технології Hands Free: коли ваші очі та руки зайняті, наприклад, коли ви за кермом, мова надзвичайно корисна. Можливість спілкуватися із Apple Siri або Google Maps, щоб доставити вас туди, куди потрібно, зменшує ваші шанси загубитися та позбавляє потреби натягуватись та керувати телефоном чи читати карту.

На даний момент, голосове управління все більше і більше отримує розповсюдження у медицині, бізнесі, роботі в офісі. Прикладом розвитку

управління голосом, у сфері бізнесу, є звичайно, телефонія: автоматизований аналіз вхідних та вихідних дзвінків за допомогою впровадження мовленнєвих систем для самообслуговування, типу: інформаційне консультування, онлайн купівля / продаж, зміни конфігурації сучасних послуг, проведення голосування, збору інформації, анкетування, інформування та інші.

1.4 Види голосового управління

Існує декілька типів голосового управління. Одні різняться за типом з'єднання, а інші функціонально. Однак, усі вони є пристроями розпізнавання мови.

За типом з'єднання класифікація наступна: з'єднання по кабелю, Bluetooth, Wi-Fi з'єднання. У одному із прикладів такої системи, для передачі голосу та мовлення може використовуватися окремий модуль для прийому команд через звук, у свою чергу, він поєднаний з модулем обробки і перетворення голосових команд у сигнал за допомогою кабелю, або може знаходитись у його складі у якості єдиного модуля.

Звичайно, що у наші часи, найбільш поширеною схемою використання такого програмного забезпечення є використання на смартфонах, комп'ютерах, тому передача голосу для обробки здійснюється за допомогою Bluetooth або Internet по Wi-Fi. З'єднання по Bluetooth не потребує підключення до мережі Internet та є простішим у використанні, аніж Wi-Fi, хоча мінусом такого підходу є відносно невелика дальність передачі сигналу.

Передача даних через Internet по Wi-Fi зі смартфона чи комп'ютера потребує наявності підключення, але суттєвою перевагою такого методу є те, що управління системою можливе з будь-якої точки земної кулі.

Зазвичай, при голосовому управлінні використовують тільки окремі слова та вислови, адже при надиктовці тексту потрібно витримувати певну паузу між словами, щоб надрукувати цей текст було можливо швидше. Хоча люди з обмеженими можливостями або травмами постійно використовують

голосовий набір тексту. Сучасні системи для голосового управління підтримують широкий вибір операційних систем на комп'ютері, смартфонах, планшетах. Але обов'язковою умовою задля отримання голосового управління є наявність мікрофона, за допомогою якого планується у майбутньому передача голосової інформації.

1.5 Схема пристроїв розпізнавання голосу

Системи або пристрої, які займаються перетворенням голосу в відцифрований сигнал або інформацію мають унікальну архітектуру всередині. В єдиному модулі частіше за все, можуть бути поєднані як елемент очищення від шуму, так і елемент виділення корисного сигналу. Кожному окремому звукові відповідає новозбудована статистична модель, яка описує як саме звучить він у людській мові. Бібліотека, яка містить у собі набори голосу, збирається основуючись на окремій мові, тому її складність буде прямо пропорційна такій мові. Як можна зрозуміти, створити голосову бібліотеку для англійської мови значно простіше, ніж для українській. Бібліотека для голосу та роботи з ним є невід’ємною при наборі речень, текстів, словосполучень чи слів. Декодер – є елементом, призначеним для розпізнавання голосу, який може поєднувати отримані голосові дані і на основі такої моделі та бібліотеки може визначати межі для кожного слова, тим самим розпізнавати мову, яка звучить [2].

Системи розпізнавання голосу працюють, аналізуючи звуки та перетворюючи їх у текст. Програмне забезпечення спирається на широкий словниковий запас та знання того, як розмовляють англійською, щоб визначити, що, швидше за все, сказав спікер. У деяких програмах є словниковий запас спеціалістів або часто вживані слова, такі як імена, можна розширити запас, надавши йому документи, списки слів або використовуючи сторонні плагіни.

Програма розпізнавання голосу фіксує та перетворює мову через

мікрофон. Деякі комп’ютери мають вбудовані мікрофони, але більшість спеціалізованих програм розпізнавання голосу також мають мікрофонну гарнітуру. Її можна підключити до комп'ютера або через гніздо звукової карти, або через USB (або подібне) з'єднання.

Також можна використовувати відповідний ручний цифровий магнітофон для диктування записів – те, що може бути особливо корисним для мобільної роботи. Деякі програми розпізнавання голосу можуть транскрибувати записи з багатьох форматів (включаючи wav, mp3 та wma).

Голос і фраза кожного звучать дещо по-різному, тому найефективніша програма використовує простий, одноразовий процес, який називається «реєстрація». Це займає лише хвилину і просто передбачає читання короткого тексту з декількох рядків. Однак, не усі програми для розпізнавання мови використовують реєстрацію.

1.6 Ознаки в системах розпізнавання голосу

Основні ознаки, за якими можна характеризувати людське мовлення, відносяться до форми, розмірів, динаміки зміни голосу і за допомогою яких можна ідентифікувати емоційний стан людини. Таким чином їх поділяють на чотири групи ознак, за якими можна розрізнити голосові зразки: спектрально-часові, кепстральні, амплітудно-частотні, ознаки нелінійної динаміки.

Спектральні ознаки:

- середнє значення спектра голосового сигналу, що аналізується;
- нормалізовані середні значення спектра;
- відносний час знаходження сигналу у моментах спектра;
- нормалізований час знаходження сигналу у моментах спектра;
- медіанне значення спектра голосу в проміжках;
- відносна потужність спектра мови в проміжках;
- варіація огинаючих спектрів мовлення;

– нормалізовані величини варіації огинаючих спектрів мовлення;

– коефіцієнти кросскорреляції спектральних огинаючих між проміжками спектра.

Часові ознаки діляться на: тривалість сегмента, фонеми, висоту сегмента, коефіцієнт форми сегмента. Спектрально-часові ознаки можуть характеризувати голосовий сигнал в його фізико-математичної суті, опираючись на наявність таких компонентів як: періодичних (тональних) проміжків звукової хвилі, неперіодичних проміжків хвилі звуку, проміжків, що не мають голосових пауз.

Спектрально-часові признаки дають можливість побачити своєрідність форми для спектру і тимчасового ряду імпульсів голосу у різних користувачів та особливості функцій, які базуються на фільтрації їх голосових каналів. Можна охарактеризувати деякі особливості промовляння у голосі, які відносяться до перебудови динаміки мовленнєвих органів голосу людини, а також являються інтегральними характеристиками потоку голосу, що мають змогу показувати особливість взаємозв'язку або злагодженість руху мовленнєвих органів людини.

Розглянемо кепстральні ознаки:

– мел-частотні кепстральні коефіцієнти;

– головні лінійні коефіцієнти, але з поправкою на те, що людське вухо не рівномірно чутливе;

– коефіцієнти для позначення потужності частоти фіксувань;

– коефіцієнти для спектра лінійного прогнозування;

– коефіцієнти кепстра для лінійного прогнозування.

Велика кількість сучасних автоматизованих систем для розпізнавання людської мови зосереджуються на отриманні частотної характеристики голосу людини, відкидаючи при цьому характеристики сигналу збудження. Це можна пояснити тим, що коефіцієнти моделі забезпечують найкращу роздільність між звуками. Задля того, щоб відділити сигнал збудження від сигналу голосового тракту вдаються до кепстрального аналізу.

Амплітудно-частотні ознаки:

– інтенсивність, амплітуда;

– енергія;

– частота основного тону;

– фомантні частоти;

– джіттер – тремтіння частотної модуляції основного тону;

– шіммер – модуляція за допомогою амплітуди.

Амплітудно-частотні признаки дають можливість отримувати дані, інформація і ключі з яких мають змогу змінюватися в залежності від параметрів дискретного перетворення Фур'є, а також при незначних зрушеннях проміжків по вибірці. Голосовий сигнал акустично поширюється за допомогою навколишнього повітряного середовища, непрості за своєю природою коливання звуку у якому аналізуються відповідно їх частоти (відповідає за кількість коливань у секунду), тривалості та інтенсивності (тобто амплітуди збуджень). Частотно-амплітудні признаки вміщують у себе достатні та необхідні дані для людини по голосовому сигналу за умов мінімального часу на розпізнавання [3]. Попри все, використання таких ознак не може повністю дозволити користуватися ними, як інструментом для розпізнавання емоційно-забарвленого голосу.

Характеристики для нелінійної динаміки:

– відображення за Пуанкаре;

– графік рекурентності;

– показник максимальності Ляпунова – відповідає за емоційний стан диктора;

– портрет за фазою (аттрактор);

– Каплана-Йорк розширення – міра, яка відповідає за кількість значень емоційного стану диктора, починаючи від «спокою» до «гніву».

Щоб можливо було розглянути ознаки нелінійної динаміки, мовний сигнал розглядається як скалярна величина, яка знаходиться у центрі

голосового апарату диктора. Саме тому процес мовостворення вважається нелінійним, а аналізують, зазвичай, його за допомогою методів нелінійної динаміки. Метою нелінійної динаміки є знаходження і дослідження базових математичних моделей і реальних систем, які виходять з найбільш типових пропозицій про властивості окремих елементів, що складають систему, і закони взаємодії між ними. На даний момент підходи для нелінійної динаміки основуються на фундаментальному математичному апараті, яка у свою чергу базується на теоремі Такенса, яка підкладує складний математичний базис під підходи нелінійної авторегресії, доводячи можливість відновлення аттрактора на місцевому ряді або на одній з його координат (у даному випадку під аттрактором розуміють велику кількість точок або піделементи у просторі фази, до якої наближається траєкторія фази одразу після того, як перехідні процеси загасли). Аналіз показників сигналу з відновлених мовних траєкторій можуть бути використані у побудові нелінійних розрізнених у часі, фазово-просторових схем тимчасового ряду, який спостерігається. Після виявлення різностей у формі портретів за фазою, використовують для діагностичних правил і ознак, що дозволяють нетільки розпізнати, але і правильно ідентифікувати різні емоції в емоційно голосовому сигналі.

Мова – це акустична хвиля, яка надходить з системи органів: легень, бронхів і трахеї, а потім перетворюється в голосовому тракті. Якщо зробити припущення, що джерела збудження і форма голосового тракту незалежні, тоді голосовий апарат людини необхідно показати у вигляді декількох генераторів для сигналу тональності, фільтрів, шумів. Якщо показати на схемі, то отримаємо результат на рисунку 1.2, де:

- генератор імпульсної послідовності (тонів);
- генератор випадкових чисел (шумів);
- коефіцієнти цифрового фільтра (параметри голосового тракту);
- нестаціонарний цифровий фільтр.

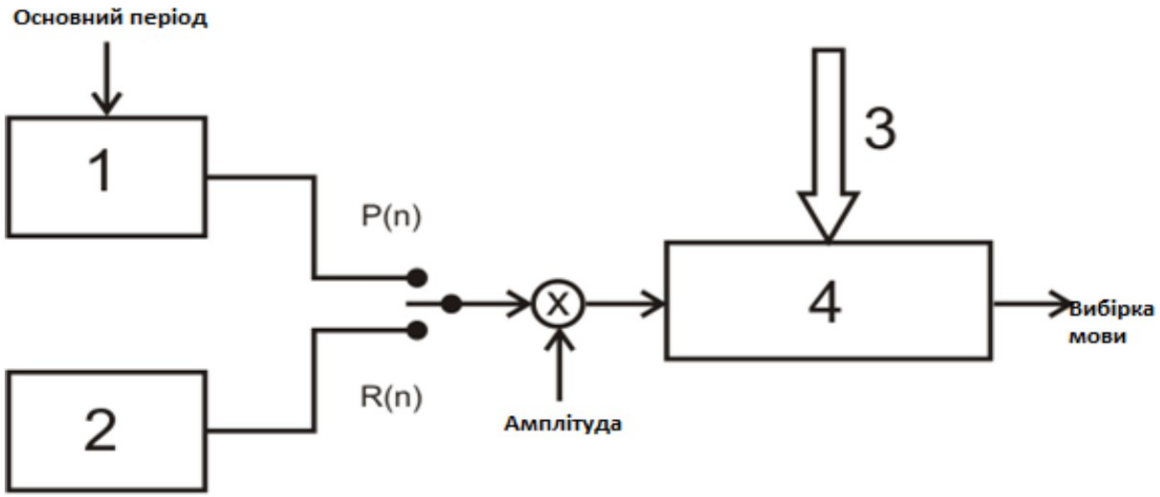

Рисунок 1.2 – Схема голосового апарату людини

1.7 Якість мови та її синтез

Параметри якості мови:

– складова розбірливість мови;

– фразова розбірливість мови;

– якість мови в порівнянні з еталоном;

– якість мови в умовах роботи реального часу.

Основні поняття включають у себе наступні пункти.

1) Чіткість голосу – кількість точно сприйнятих фонем голосу (звуків, слів, фраз), виражається у відсотках від усієї кількості елементів, що передаються.

2) Якість мови – відображає суб'єктивну оцінку того, як звучить мова в середовищі випробуваної системи для передачі голосу.

3) Нормальний темп мови – висловлювання мови з такою швидкістю, при якій середня тривалість вимовлюваної фрази дорівнює 2,4/с.

4) Пришвидшений темп мовлення – вимовлення фрази з такою швидкістю, при якій середня тривалість контрольного елементу дорівнює 1,5-1,6 с.

5) Розпізнання голосу конкретної людини – дає змогу слухачеві

порівняти як звучить голос, з конкретною людиною, яку слухач знав до експерименту.

6) Смислова розбірливість – відображає ступінь коректного відтворення змісту інформації у промові.

7) Інтегральна якість – вказує на загальне враження слухача від почутої мови.

Якість мови та чіткість промовлювання є ключовим фактором в діалозі між людиною та голосовою системою, так як, точно зрозуміти яка фраза або команда буде оголошена – найважливіше завдання, вирішення якого допоможе уникнути помилок та розпізнавання невірної команди або розпізнання неправильного тексту.

У подібних областях ідентифікація диктора не носить критичного характеру. Саме через це – рішення про ступінь подібності голосів приймається на основі імовірнісних кількісних оцінок. У такій постановці завдання про ідентифікацію різних голосів має певну специфіку, пов'язану зі спотвореннями і шумами в каналах зв'язку [4]. Так як фонетичний зміст порівнюваних голосових сигналів, як правило, відрізняється, то такі системи зацікавлені в дослідженнях розпізнавання унікальності диктора, не дивлячись на контекст.

Синтез мови – це відновлення за допомогою параметрів голосового сигналу, формування голосу напряму з друкованого тексту. Застосування для синтезу мови можна знайти всюди там, де одержувачем є людина. Якість такого синтезатора мови напряму залежить від його схожості з реальним людським голосом та зрозумілою мовою. Такий підхід до якості дозволяє людям зі сліпотою слухати улюблені письмові твори на комп'ютері чи смартфоні.

Синтез мови знайшов своє застосування у наступних випадках:

- інформаційно-довідкові системи;
- видача інформації про поточні технологічні процеси;
- створення музики.

На даний момент існує декілька способів синтезу мови:

- параметричний синтез;
- компіляційний синтез;
- синтез за правилами;
- предметно-орієнтований синтез.

Параметричний синтез мови є однією з кінцевих операцій в системах кодування, де голосовий сигнал відображається за допомогою набору невеликого числа безперервно параметрів, які змінюються. Такий тип синтезу правильно застосовувати в тих випадках, коли набір голосових елементів кінечний і змінюється рідко. Головною перевагою такого способу є можливість записати голос на будь-якій мові і від будь-якого диктора. Якість параметричного синтезу може досягати високого рівня (залежіть від ступеня стиснення інформації в параметричному представленні). Однак параметричний синтез не може застосовуватися для довільних, не заданих попередньо сигналів.

Синтез на базі компіляції подягає у сумуванні повідомлення з вже існуючого та розробленого словника, який містить вихідні елементи для синтезу. При цьому, розмір для елементів, що синтезуються повинен сягати не менше слова. Звичайно, зміст повідомлень, що були синтезовані дорівнює обсягу словника. Зазвичай, кількість елементів словника не переходить за межу у декілька сотень. При такому синтезі можна побачити одну з основних проблем, яка пов’язана з обсягом пам'яті, яка використовується для зберігання самого словника.

Відповідно до цього, навчилися користуватися різними методами кодування/стиснення голосового сигналу. Компілятивний синтез, не дивлячись на перший погляд, має широке застосування на практиці. Наприклад, західні країни оснащують системами голосового управління велику кількість пристроїв (починаючи побутовими пристроями, закінчуючи військовими машинами). У нас, на Україні, системи голосового управління також стають все більше розповсюдженими у звичайному житті, ось як – в

службах довідки, операторах місцевого зв'язку, коли необхідно отримати інформацію за станом рахунку користувача.

Вичерпуючий синтез голосу за правилами (також може бути синтез за введеним текстом) дає можливість управляти усіма параметрами голосового сигналу, таким чином, є можливість генерувати мову відповідно до невідомого тексту, який був прийнятий. У такому разі налаштування, які були отримані у ході аналізу голосового сигналу, повинні зберігатися у пам'яті таким же чином, як і налаштування задля з'єднання звуків в окремі слова та фрази.

Подальший синтез голосу реалізується за рахунок моделювання голосового тракту, та застосування цифрової або аналогової техніки. При такому підході, всередині процесу синтезування, значення налаштувань та правила по'єднання фонем додають один за одним шляхом повного часового інтервалу, який може бути 5-10 мс. Метод синтезу голосу через друкований текст основується на запрограмованому понятті лінгвістичних та акустичних обмежень, при цьому не використовуючи елементи людського мовлення. Тому у схемах, що базуються на такому способі синтезу, можна виділити два підходи. Перший з них націлений на побудову схеми голосу похідного від людської системи, це називається артикуляторним синтезом.

Інший підхід – це форматний синтез за допомогою правил. Точність та правильність подібних синтезаторів можна довести до величин, які співпадають з характеристиками справжньої мови. Синтез голосу за допомогою правил з використанням збережених попередніх відрізків реального голосу називається різновидом синтезу голосу за правилами, що набула широкого резонансу у зв'язку з появою варіантів маніпулювання мовленевим сигналом, яка відображається в оцифрованій формі. Існують наступні види синтезу, які залежать від розміру вихідних елементів:

– мікросегментний (мікрохвильовий);
– аллофонічний;
– дифонний;

– напівскладовий;
– складовий;
– синтез з одиниць довільного розміру.

У більш загальних ситуаціях, в якості подібних елементів можуть використовуватися напівскладові сегменти – це такі сегменти, що містять лише частину приголосного та частину доданого до нього голосного. Під час такого підходу є можливість синтезувати голос та мову відповідно до спочатку не заданого тексту, хоча недоліком є складне керування характеристиками інтонації.

Можна зауважити, що якість подібного синтезу не може відповідати якості справжньої мови, так як на кордонах перекрещення дифонів зазвичай можуть виникати спотворення. Мовленнєва компіляція із заздалегідь занесених словоформ також не може точно вирішити подібну проблему високоякісного синтезу ризновманітних повідомлень, так як акустична складова, тривалість та інтонація та характеристики слів також можуть змінюватися через різні типи фраз та різні слова у фразі [5]. Таке розташування не може змінюватися навіть якщо використовувати масивні обсяги пам'яті, які зберігають слова.

Синтез на основі предметно-орієнтованого підходу зберігає слова, що були записані до цього, а також фрази для відтворення цілісних мовленнєвих повідомлень. Такий тип синтезу, зазвичай, використовується у додатках, в яких різноманіття слів комплексу може бути обмежене спеціальною областю/темою, а саме повідомлення про розклад відбуття літаків та прогнози погоди. Подібна технологія не складна при використанні та є часто використовуваною в комерційних задачах: вона часто використовується під час виготовлення електроніки, наприклад, годинників, які мають змогу видавати інформацію.

Точність звучання таких моделей може бути високою, у потенціалі, адже різноманіття типів пропозицій невелика та схожа з потенційною інтонацією вихідних звуків. Беручи на увагу, що такі системи обмежені через

кількість слів та фраз, що зберігаються у базі даних, тому вони у майбутньому не будуть мати змогу широкого розповсюдження у сферах впливу людини, лише через те, що вони мають змогу видавати комбінації фраз та слів, які були запрограмовані для них.

1.8 Оцифрування сигналу

Оцифрування звуку займає ключову роль в системах розпізнавання голосу та мови. Цифровий звук – це аналоговий звуковий сигнал, але представлений у вигляді значень, які мають математичну базу – звукової амплітуди. Оцифрування голосу містить у собі декілька процесів:

- дискретизація;
- амплітудне квантування.

Дискретизація сигналу у часі – це процес, при якому отримуються значення сигналу, який перетворюється, відповідно до певного часового кроку дискретизації. Частота дискретизації або частота вибірки – це ні що інше, як заміри величини сигналу, які проходять впродовж однієї секунди. Звичайно, при збільшенні кількості кроків, зростає і їх кількість і тим більш точне уявлення можна отримати щодо сигналу, який буде на виході. Таке зауваження можна підтвердити за допомогою теореми Котельникова. Відповідно до теореми, сигнал у вигляді аналогового, та який має обмежений спектр, може бути точно описаний за допомогою дискретної послідовності точок його амплітуди, при умові, що ці точки беруться відповідно до частоти, що у два рази перевищує найбільшу частоту сигнального спектра. На рисунку 1.3. можна побачити схему дії такого аналого-цифрового перетворювача.

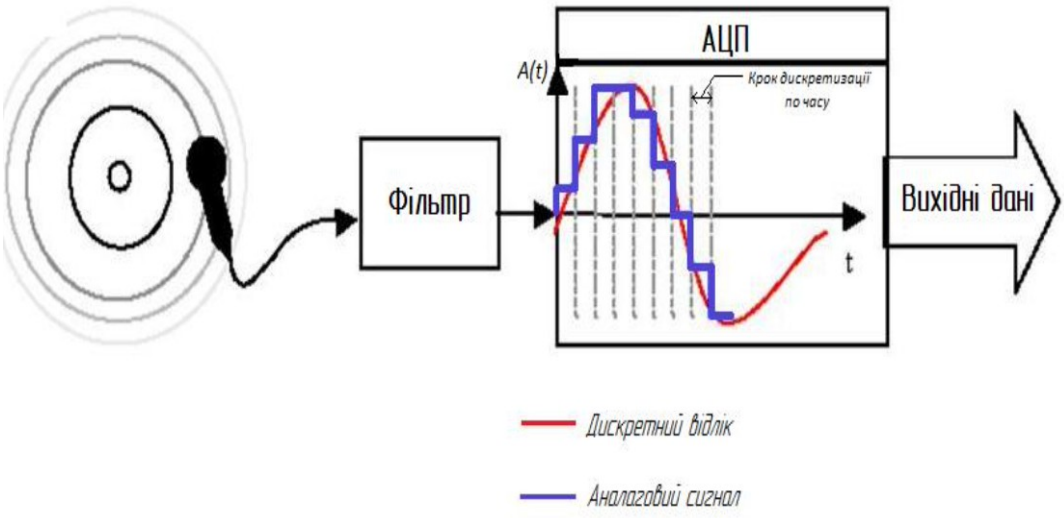

Рисунок 1.3 – Аналого-цифровий перетворювач

Таким чином, сигнал у аналоговому вигляді, який має частоту спектра у вигляді прямої, може бути достатньо точно представлений послідовністю значень амплітуди у дискретні проміжки часу. Вдаючись у. реальні дослідження можна зауважити, що для того, щоб сигнал, що оцифрували, мав дані по всьому спектрі частот, які можна почути, вихідного аналогового сигналу (0 – 20 кГц) потрібно, щоб обране число частотності дискретизації складало не менше 40 кГц. Величину амплітудних замірів в виділену секунду можна назвати частотою дискретизації. Загалом, основні труднощі в оцифруванні сигналу, містяться в неможливості збереження отриманих значеннь голосу з точністю, близькою до ідеальної. Схему оцифрування звукового сигналу можна розглянути на рис. 1.4.

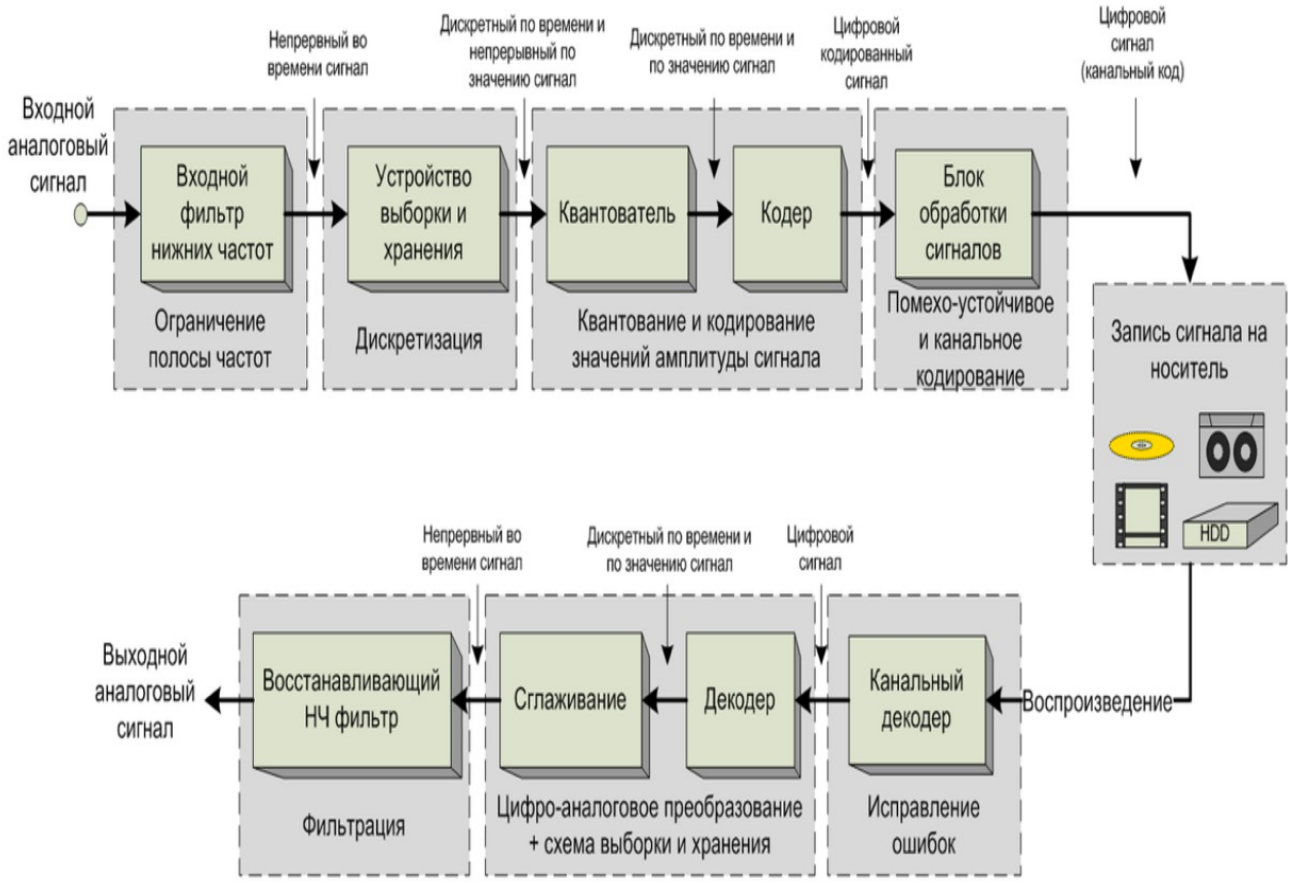

Рисунок 1.4 – Оцифрування голосу

У цифровому звуці можна виділити два основних джерела шумів.

Джитер – це таке явище, при якому відбувається випадкове відхилення звукового сигналу, зазвичай, такі явища виникають при нестабільній частоті генератору тактів або коли поширення окремих частотних налаштувань одного сигналу відбувається за різної швидкості. Така проблема нерідко виникає на етапі стадії оцифрування. Таке явище можна побачити на різних відстанях між лініями по вертикалі (рис.1.5). Тому при звукозаписі на цифровий носій, необхідно користуватися якісними кварцовими генераторами, які мають джерелами живлення з невеликою пульсацією та шумами [6]. При використанні доцільно використовувати також повністю цифрові студії, що допомагає звести вплив тремтіння до мінімуму. Прикладом такої студією може виступати персональний комп'ютер, на борту якого є звукова плата, що містить у собі гарний АЦП, який у разі зберігання,

відтворення або редагування звуку буде отриманий тільки у вигляді цифрової схеми.

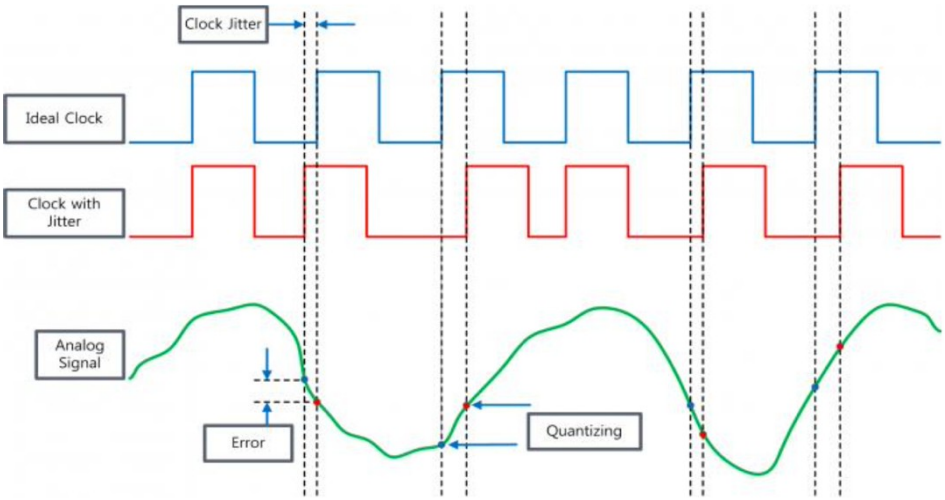

Рисунок 1.5 – Джитер

Алиасінг – таке поняття, яке вміщує у себе проблему, за якої при оцифруванні, у цифровому записі можливе появлення частотних складових, яких до сих пір не було в вихідному сигналі. Таке явище отримало назву Aliasing. Дана проблема, звичайно, пов'язана з частотою дискретизації, або, як її прийнято називати – частота Найквіста.

1.9 Аналіз сучасних систем голосового управління

На сьогоднішній день, на збільшення показників ринку розпізнавання мови та управління а допомогою голосових команд можуть впливати деякі фактори, тому однозначно не можна сказати за його розвиток. Так як системи з голосовим керуванням з’явилися досить недавно, важливим фактором у таких моделях є безпека, адже у деяких областях діяльності важливим питанням є кодування голосу користувача, це робиться задля того, щоб захиститися від зломів та надання доступу до голосових команд та управління сторонньому користувачеві. На сучасний стан, безпека голосового управління не достатньо розвинена, тому найновіші дослідження в даній

області базуються як раз на покращенні цієї ситуації.

Важливим мінусом в області розробок голосового управління є дуже висока вартість, саме через це істотно сповільнюються процеси розвинення схожих моделей. На сьогоднішній момент потрібно зазначити неможливість повної фільтрації зовнішніх шумів, що гальмують процес якісного та точного розпізнавання мовлення та не дають подібним схемам твердо закріпитися на світовому ринку, а потім істотно вплинути на неї, що є результатом того, що подібні системи займають дуже не великий відсоток на ринку технологій.

Одну з важливих ролей, як на технічному ринку, так і у розвитку систем розпізнавання голосу і голосового управління, відіграє мовленнєва біометрія, що впроваджується на секретних станціях, наприклад – бази військових або лабораторії науковців. Спільною проблемою в області схожих моделей є досить низькі показники якості, тому зараз як раз вони активно покращуються шляхом впровадження голосової біометрії та подібних систем для розпізнавання голосу.

Слово «біометрія» раніше вживалося тільки в теорії медицини. Лише через деякий час, почали зростати потреби щодо безпеки, які включали у себе біометричні технології, на сам перед у підприємствах та державних установах. Включення технологій, пов’язаних з біометрією – це одна з головних чинників на технічному світовому ринку розпізнавання голосу.

Розпізнавання мови може примінитися для перевірки автентичності користувача, адже голос у будь-якої людини індивідуальний. Такий підхід може забезпечити якісний рівень безпеки та точності. Розпізнавання мовлення користується попитом у інститутах фінансування, наприклад, банк, також на підприємствах, що базуються на охороні здоров'я. На даний момент, частина розпізнавання голосу дорівнює 3,5% від частини біометричних технологій на технічному світовому ринку, хоча, зауважимо, що вона постійно зростає. В добавок до цього, невелика вартість пристроїв біометрії веде до зростання попиту з точки зору середнього та малого бізнесу.

У більшості країн світу, військові відомства зазвичай використовують

дуже обмежені зони для того, щоб стала неможлива ситуація проникнення зловмисників. Також, задля того, щоб забезпечити секретність та безпеку в даній зоні, вони користуються системами розпізнавання мови. Подібні методи допомагають військовим інститутам одразу виявляти ситуації несанкціонованих проникнень у межу захищеної зони. Усередині системи лежить база даних голосів кожного службовця та державного чиновника, адже вони мають доступ до території, що знаходиться під захистом. Такий тип людей розпізнається системою розпізнавання мовлення, що призводить до запобігання допуску користувачів, голосів яких не має в базі даних.

Можна додати, що військовослужбовці користуються голосовими командами задля керування літаком. Також, військові інститути впроваджують розпізнавання голосу та систему Voice-to-text щоб мати можливість комунікувати з громадянами, які знаходяться в інших країнах. Розглянемо приклад, американські військовослужбовці, вже зараз користуються системами розпізнавання голосу в завданнях в Афганістані та Іраку. Аналізуючи ці дані, можемо зауважити, що є високий попит на розпізнавання голосу та мовлення для військових цілей.

Наслідки від проблем, які з’являються перед ринком, як очікується, зведе під нуль наявність різноманітних тенденцій, що починають з'являтися на ринку. Прикладом однієї з таких тенденцій може бути зростання попиту на розпізнавання голосу на мобільних пристроях.

Так як мобільні пристрої мають величезний потенціал, виробники на технологічному світовому ринку щодо розпізнавання мови розвивають найновіші додатки, спеціалізовані на роботі для мобільних пристроїв. Можна точно сказати, що це один з майбутніх чинників, які матимуть рушійну силу. Попит на мовленнєву автентифікацію для банкінгу на мобільних пристроях є однією з позитивних тенденцій на світовому ринку для розпізнавання мови. Нижче наведені деякі з головних тенденцій на технологічному ринку щодо розпізнавання мовлення.

1) Зростання попиту щодо програм для розпізнавання голосу на

мобільних пристроях.

2) Розвиток попиту для послуг мовленнєвої автентифікації на ринку мобільного банкінгу.

3) Впровадження голосової ідентифікації та розпізнавання голосу.

4) Зріст поглинань та злиттів.

Велика частина попиту з систем розпізнавання голосу та голосового керування приходиться на мобільні додатки, у головному чині на смартфони, сучасні музикальні пристрої, на системи розумного будинку, та автомобілебудування, що сильно полегшують доступ до обраної техніки в секторах для життя. Активно, подібні методи розглядаються для сфер діяльності військових інститутів, медицини та виробництва для промисловості.

Так як число правил для дорожнього руху, які забороняють користуватися мобільними пристроями під час водіння зростає, це привело до збільшення попиту на додатки, що займаються розпізнаванням голосу. Приведемо декілька країн, у яких вже існують суворі обмеження: Філіппіни, Австралія, Великобританія, США, Індія та Чилі. Наприклад, у Сполучених Штатах, майже в 13 з них, не дивлячись на діючі правила щодо використання мобільних гаджетів, можна користуватися гучним зв'язком під час поїздки.

Так чином, можна побачити, що користувачі все частіше обирають мобільні девайси, які мають на борту додатки для розпізнавання голосу, адже вони мають змогу допомогти їм мати доступ до телефону без необхідності відволікатися на нього. Для того, щоб задовільнити зростаючий попит на системи розпізнавання голосу, в мобільних гаджетах, виробники почали збільшувати кількість дослідно-конструкторських та науково-дослідних досліджень, щоб розвинути голосові опції для команд, що розташовуються на мобільних пристроях. Як наслідок, зростаюча кількість додатків для розпізнавання голосу були додані до мобільних пристроїв, ось деякі з них – управління музичним списком, зчитування інформації абонента, зчитування даних, голосові повідомлення. Якщо порівняти компанії на технологічному

ринку, які займаються поширенням систем розпізнавання голосу та голосовим управлінням, найбільш конкурентноздатними стануть такі, які матимуть системи з найбільшою точністю, або комбіновані системи (якісне розпізнавання голосу буде підтверджуватися та супроводжуватися розпізнаванням відео). Таким чином, якщо не звертати увагу на неоднозначні тенденцій ринку розпізнавання голосу та мовленнєвого управління, він плавно розвивається. На період з 2011 по 2015 ріки, частка на ринку подібних методів збільшилась в 1,5 – 2 рази, що залежить від того, на що конкретно націлена конкретна система. На рис. 1.7 можна побачити основні тенденції розвитку голосового управління на ринку.

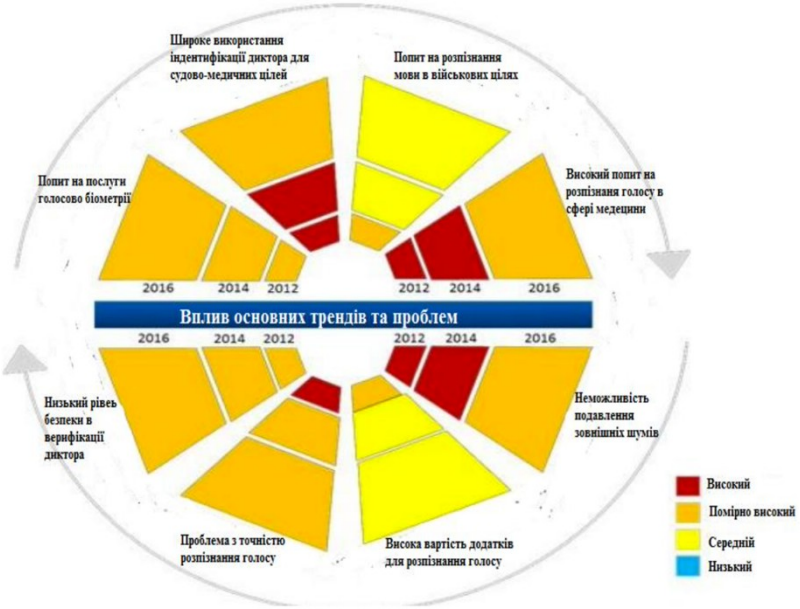

Рисунок 1.6 – Актуальність пристроїв для розпізнавання голосу

Перші методи розпізнавання голосу були основані на математичному апараті, який називається «приховані марківські моделі». Відомий математик Андрій Марков, який дав їм назву, під час дослідження проблем обробки літературних текстів, приблизно, у початку двадцятого століття, оцінював імовірність знаходження усіх літер у тексті, що залежить від контексту, в якій вона знаходиться. Задля того, щоб спростити обчислення, ним було допущено, що подібні ймовірності залежать тільки від конкретної літери –

попередньої, потім це стало Марківською властивістю. З’ясувалося, що оцінювання імовірностей переходу між двома сусідніми буквами на різних фрагментах одного і того ж письма, майже однакові. Потім виявилась унікальність параметрів такої схеми для кожного користувача, що дало змогу приміняти їх в проблемах щодо визначення автора письма.

Для подібних систем, тексти уявляються послідовністю букв та символів, а також станів марковського ланцюга. Так само, для усного голосу окреме слово може бути описаним за допомогою транскрипції фонетики, тобто послідовності фонем. Хоча, якщо під час обробки письма їх символи не відомі, тоді в звуковому голосі можна побачити не самі фонеми (стани ланцюга), а лише їх реалізації, або, як ще називають, голосові сигнали, які подаються у вигляді залежності від часу звукового тиску. Саме через це, стани фонеми тепер приховані, тобто невідомо, яка саме фонема в реальності була проголошена, адже ми можемо знати лише її реалізацію. Також, у зв'язку з різноманітністю голосу із кожної фонеми витікає безліч реалізацій. Отже, через це, можна сказати, що до звичайної для марковських ланцюгів проблеми щодо оцінки імовірностей переходу між різними фонемами можна додати точне моделювання залежності сигналу, який спостерігається, від тієї чи іншої фонеми.

У практичних системах для розпізнавання голосу використовуються не звичайні фонеми, а більш комплексні мінімальні звукові частки, о мають назву трифони – це така реалізація фонеми, де кожна з них може бути описана за допомогою користувацької прихованої марковської системи. Тому проблема побудови акустичної системи, яка залежить від акустичних характеристик реалізації голосових сигналів та від типу звукової частини, являє собою одну з найбільш складних задач при автоматичному розпізнаванні голосу.

Майже до 2011 року за основу брали модель гауссових сумішей задля рішення питання розподілу сигналу, який аналізується, та залежить від фонеми. Щоб досягти такого ефекту, звуковий сигнал необхідно поділити на

малі ділянки розміром 10-50мс. У якості застосування основної обробки звукових сигналів у області частоти для усіх ділянок всередині сигналу виконується швидке перетворення Фур'є. Після цього етапу було примінено логарифмування отриманого спектра, відносно відомого логарифмічного сприйняття людським каналом та масштабу сигналу. В кінці кінців, користуючись допомогою дискретного косинусового перетворення логарифма для спектру були отримані майже незалежні ознаки, такі як, кепстральні коефіцієнти, для яких був отриманий розподіл, та були записані у вигляді декількох гауссових випадкових векторів, що мають діагональні коваріційні матриці.

Після цього етапу, згідно до революції у глибинному навчанні, був замінений традиційний підхід, який використовується для отримання характерних ознак та їх опису, на модель гауссових сумішей, з ціллю побудови акустичної моделі мови, а також почали користуватися глибинними нейронними мережами. Для вирішення задачі розпізнання голосу використовувалися традиційні мережі прямого поширення, але з доданням великого числа шарів, та були навчені, у режимі без вчителя, послідовно один за одним шар мережі. Після отриманих результатів були отримані висновки, що застосування такого методу схоже з підходом прихованих марковських моделей, які мають имовірнисний перехід від однієї фонеми до наступної, на десятки відсотків покращують точність розпізнавання реального голосу. Такий підхід у наші дні реалізований майже на всіх новітніх програмних бібліотеках розпізнавання голосу.

Приблизно з появою новітньої акустичної схеми голосу другим інноваційним моментом стали нові лінгвістичні, голосові моделі. В їх основі лежить наступний принцип – потрібно передбачити наступне фразу чи букву знаючи попередні дані, таке завдання є типовим для обробки письма. Раніше, для основних систем застосовувались схеми типу n-грам, принципи яких засновувалися на великій кількості письма, велась оцінка розподілу імовірності появи даних (слова чи фрази), залежно від кількості попередніх

слів. Задля того, щоб отримати якісні оцінки щодо розподілів, параметр n має набувати малих значень: декілька слів, такі як схеми біграм, уніграм або триграм.

Поява методів глибинного навчання та розвиток рекурентних НМ для обробки письма дали змогу істотно покращити точність лінгвістичної моделі за рахунок обчислення контексту та відсутності обмежень на використання лише декількох попередніх слів. Як наслідок, вийшло ще більше покращити якість остаточного розпізнавання голосу, тобто на звучання можуть бути розпізнані не всі слова, а допущені елементи необхідно прогнозувати по контексту, так само, як це робить користувач. Мовленнєві моделі, які базуються на рекурентних НМ, що мають змогу точно реалізувати подібну поведінку, на теперішній день повсюдно використовують в світовій індустрії.

Якщо проаналізувати ринок подібних систем, можна зробити висновок, що неодноразово можна зустріти рішення, що використовуються в розумних будинках, але є проблема – такі системи мали не високу якість розпізнавання вимовлених команд. Достатньо розповсюдженою реалізацією таких систем є реалізація на платформах Arduino, які використовують датчики, що розпізнають та зчитують звук. У представленій системі реалізується зчитування звуку з декількох джерел, таких як: модуль розпізнавання мови та сприйняття звукового сигналу через мобільний пристрій, який дозволяє керувати віддалено.

У питаннях підвищення якості розпізнавання мови у системах є рішення використання глибинних НМ, адже саме вони за останні роки не один раз демонстрували значні результати у процесах класифікації, прогнозування, розпізнавання образів, рукописного письма та голосу. Через це, використання глибинних нейронних мереж та їх різновидів у задачах розпізнавання голосу – є актуальною задачею на сьогоднішній день.

У попередні 35 років лідерами рішень у задачах розпізнавання мовлення вважались системи, в основі яких лежать приховані ланцюги Маркова та моделі Гауса (рис. 1.7).

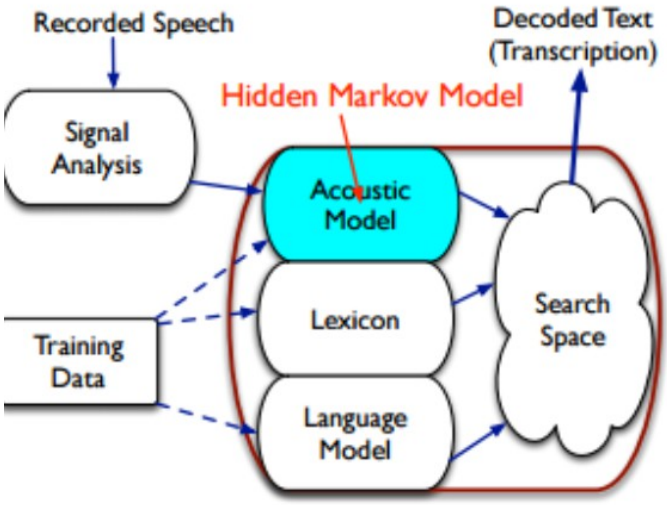

Рисунок 1.7 – Системи на основі ланцюгів Маркова у розпізнаванні мови

Звук, який записується, необхідно розділити на короткі, приблизно, 10 мс фрагменти, після чого проаналізувати на вміст у них частот. Після чого, отримується вектор характеристик, який необхідно пропустити через акустичну систему, що може видавати набір імовірнісних розподілень серед усіх існуючих фонем. HMM дає змогу показати послідовні ланцюги у даному наборі розподілів імовірностей (рис. 1.8).

Головний підхід тут повинен аналізувати слова в імовірнісній схемі, у якій фонеми можуть сприяти слову та показують стан ланцюгів Маркова, в той час як імовірності переходу були б еквівалентні ймовірністі наступної фонеми, яка проголошується, у яких моделі для фраз, що складають частину словника, з’являються у фазі навчання.

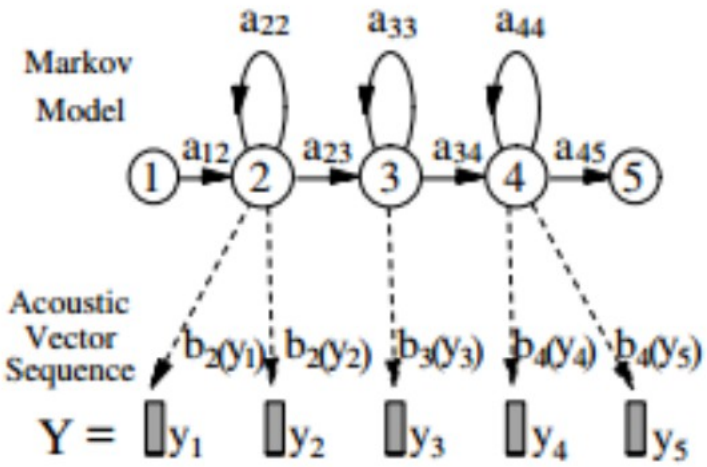

Рисунок 1.8 – Послідовні структури

Перевагами подібних систем (на основі HMM) можна вважати:

– задача розпізнавання повністю вирішена;

– повне розпізнавання слів, які складаються із різних букв, що не мають чіткого смислового значення;

– простота в плані реалізації або навчанні.

Приведемо мінуси систем на основі прихованих маркових ланцюгів:

– невелика точність;

– погана стійкість до шуму.

1.10 Постановка цілей та задач дослідження

Одразу після того, як був проведений аналіз сучасних методів розпізнавання голосу користувача, були розглянуті традиційні типи проблем, що не можуть бути життєздатними без схем розпізнавання. За допомогою теоретичного підходу були обґрунтовані методи та схеми аналізу та розпізнавання голосових сигналів декількох змінних. Були представлені алгоритми, методи та обчислювальні підходи для аналізу звукових сигналів, що базуються на параметричних функціях систем, що створюють сигнал.

Наразі відомо, що існує немало підходів щодо вирішення подібних

задач, хоча ні один підхід не є бездоганним, тому їх точність не переходить за межу 85%.

Ціллю та метою даного дослідження є побудова системи розпізнавання голосу, що базується на нейронних мережах, та яка має достатні характеристики щодо стійкості до шумів та відрізняється високою точністю обробки голосу, знаходження методів покращення якості, позбавлення від шумів та подібних факторів, що можуть впливати на хід розпізнавання мовлення. Створити та провести тестові експерименти, враховуючи різні умови роботи системи.

Після аналізу були поставлені такі задачі:

- проаналізувати типи нейронних мереж;
- розробити модель системи для розпізнавання голосу;
- реалізувати програмну частину;
- протестувати програмну частину.

2 МОДЕЛЬ РОЗПІЗНАВАННЯ ГОЛОСУ В РЕЖИМІ РЕАЛЬНОГО ЧАСУ

2.1 Оцифрування звукових сигналів

Після проведення аналізу сучасних методів, за допомогою яких проводиться розпізнавання голосу, і які включають у себе нейронні мережі, було встановлено, що найбільш відповідний метод для навчання – це підхід через завантаження файлів у форматі аудіо, які будуть аналізуватися (рис. 2.1).

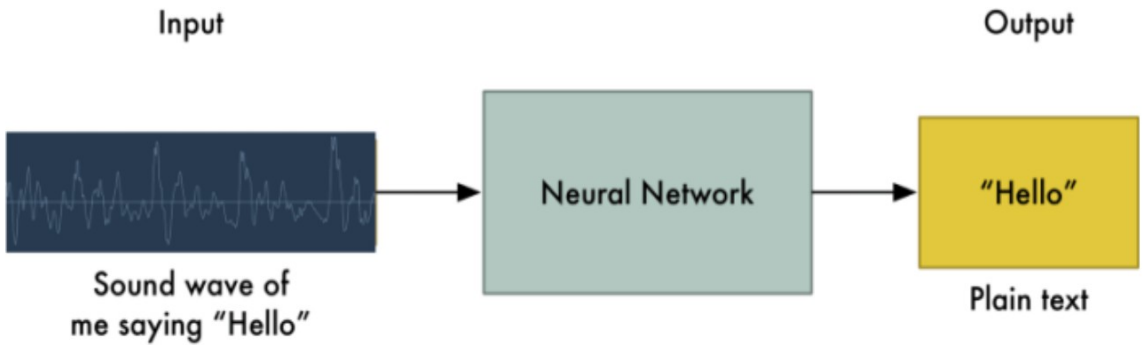

Рисунок 2.1 – Схема розпізнавання голосу нейронною мережею

Якщо використовувати такий підхід, то рано чи пізно, спливе проблема, яка полягає у швидкості розмови людини, яка проводить тестування. Наприклад, розглянемо ситуацію – перший користувач промовляє слово «Привіт» з великою швидкістю, а другий «пррриииивввііііт» з малою швидкістю, за таких умов, створюється більший файл аудіо формату, який містить у собі багато даних [7]. Через це випливають труднощі щодо розпізнавання, так як обидва записи повинні ідентифікуватися, як слово «Привіт». Одним зі складних завдань є зведення цих двох різних файлів до однакової довжини фіксованого розміру. Для рішення цієї проблеми, необхідно впровадити дещо спеціалізовані підходи та залучитися глибинною нейронною мережею задля підвищення точності розпізнавання.

На першому етапі в розпізнаванні голосу стоїть коректна передача аудіо на персональний комп'ютер. Опираючись на джерела про нейронну мережу можна зазначити, що для того, щоб проаналізувати дані, спочатку необхідно звести їх у такий формат, який був би сприйнятий коректно комп’ютером, наприклад, як це показано при розпізнаванні цифри 8 із зображення (рис.2.2).

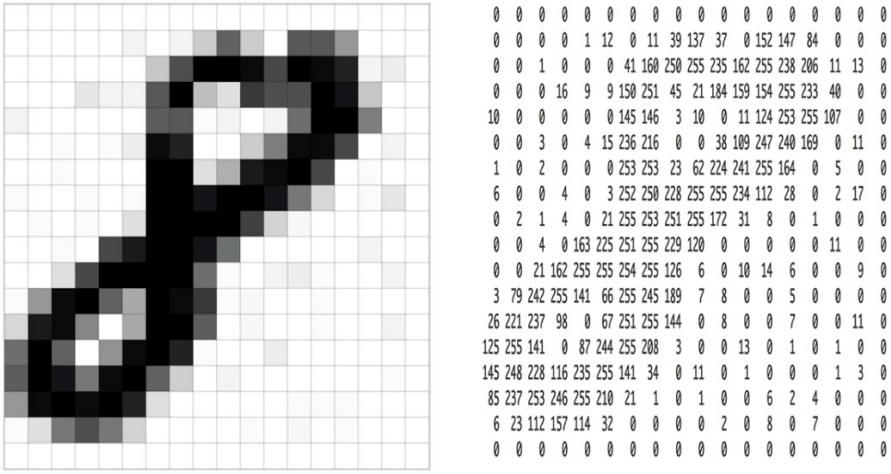

Рисунок 2.2 – Форма передачі даних для аналізу

Звукові дані дещо відрізняються від зображень, через це їх складніше оцифровувати. Для прикладу виконаємо промовляння та запис слова «привіт» (рис. 2.3).

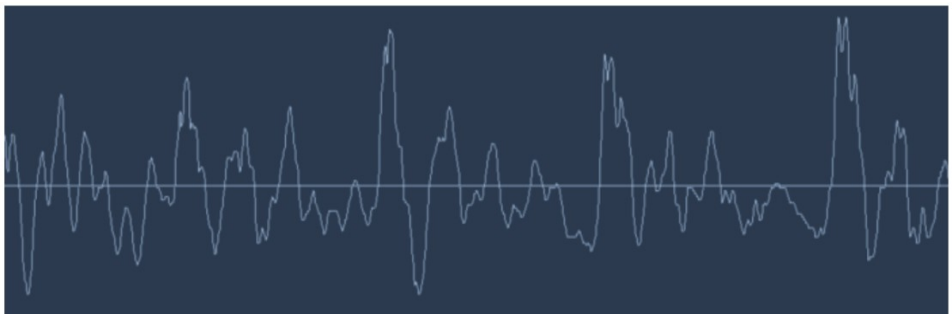

Рисунок 2.3 – Звуковий сигнал зі словом «привіт»

На рисунку видно, що звукова хвиля є лише одномірною. Можна побачити, що в однакові проміжки часу, вони мають лише одне значення, що залежить від амплітуди цієї хвилі. Якщо збільшити, наприклад, першу

частину звукової хвилі та придивитися, можна зробити висновок, що для її конвертації у формат чисел, необхідно записати показники амплітуди у точках, які лежать на однаковій відстані одна від одної (рис. 2.4).

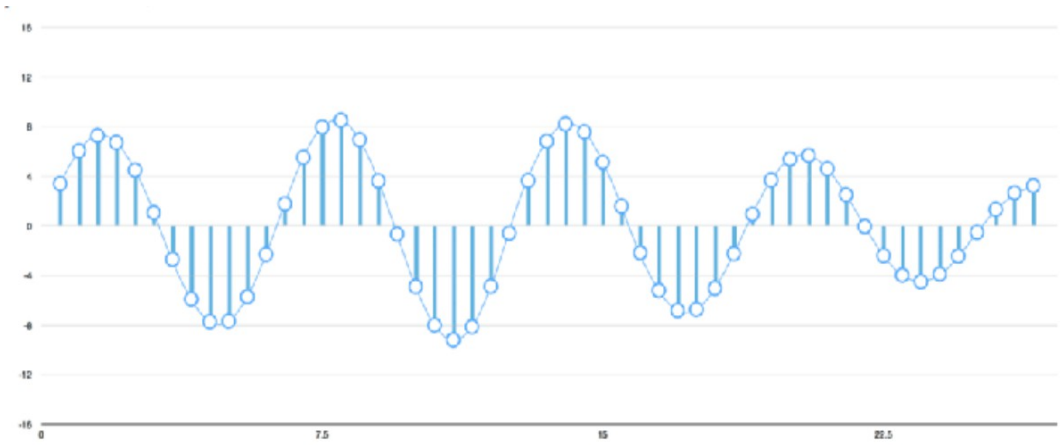

Рисунок 2.4 – Запис амплітуди хвилі у кожній точці

Подібний підхід відомий як дискретизація. Обчислення точок проходить багато разів на секунду, паралельно при цьому відбувається зчитування чисел, які відповідають амплітуді у цих точках хвилі. Після такої обробки, на виході отримуємо аудіофайли у форматі wav. Аудіо, яке заноситься на носій інформації, отримує дискретизацію з частотою приблизно 44,1кГц (44 100 замірів на секунду). Це не маленьке число, але для задачі розпізнавання голосу буде достатньо 16000 замірів на секунду, що відповідає 16 кГц, адже людський голос можна почути не в такому великому діапазоні. Опираючись на даний факт, зробимо заміри та оцифрування на аудіозаписі, на якому слово «Привіт», з частотою 16000 Гц/с. Після даної процедури на виході буде отриманий масив з точками.

Можна припустити, що такий підхід відтворює саме приблизний аналог вхідної хвилі звуку, через те, що відбувається зчитування на різних проміжках, що не залежать один від одного, також у проміжках між тактами дані можуть втрачатися. Цю проблему вирішує теорема Котельникова, яка свідчить, що для максимально наближеного відтворення вхідної звукової хвилі достатньо просто задати дискретизації таку частоту, яка усього в два

рази відрізняється від найвищої частоти звуку, який був записаний. Хоча, існує багато новачків, які помилково стверджують, що для підвищення якості звуку необхідно завжди брати більш високі частоти для дискретизації.

2.2 Аналіз отриманих даних

Наразі маємо масив із чисел, які відповідають за амплітуду звуковій хвилі у проміжки часу рівних 1/16000 секунди.

Звичайно, тепер можна виконати навчання нейронної мережі на даному масиві чисел, хоча за таких умов, розпізнавання певних моделей мови є. задачею не із легких. Для того, щоб облегшити задачу, необхідно провести деяку обробку над вхідними звуковими данними.

Першим кроком необхідно згрупувати такти у формат фрагментів у часовому проміжку 20ти мілісекунд. Один із прикладів такого фрагменту: [-1264, -1352, -1060, -976, -682, -624, -276, 124, -56, -42, -168, -457, -440, -531, -751, -1017, -1331, -1037, -942, -635, -479, -438,-387, -222, 183, 124, -16, -112, 127, 251, 188, 380, 451, 762, 958, 1351, 1874, 2524, 3693, 4978, 5539, 6157, 6781, 7202, 7540, 7213, 6129, 5361, 4520, 4453, 3511, 2540, 2104, 1339, 1168, 1985, 921, 312, -252, -489, -467, -727,- 1346, -1877, -2257, -2335, -2645, -2555, -2635, -2510, -2238, -1538, -960, -374, 12, 250, 484, 7876, 1121, 937, 746, 512, 322, 314, 315, 330, 113, -112, 63, 174, 92, -250, -423, -612, -978, -1239, -1237, -1424, -1115, -1625, -1241, -951, -979, -763,-251, 52, 220, 132, 172, -82, -335, -429, -539, -720, -917, -897, -663, -413, -190, -13, -11, 30, 90, -46...]. Після того, як ми відобразимо ці числа на графіку (рис.2.5), можна приблизно побачити зображення вхідної звукової хвилі у проміжку 20ти мілісекунд:

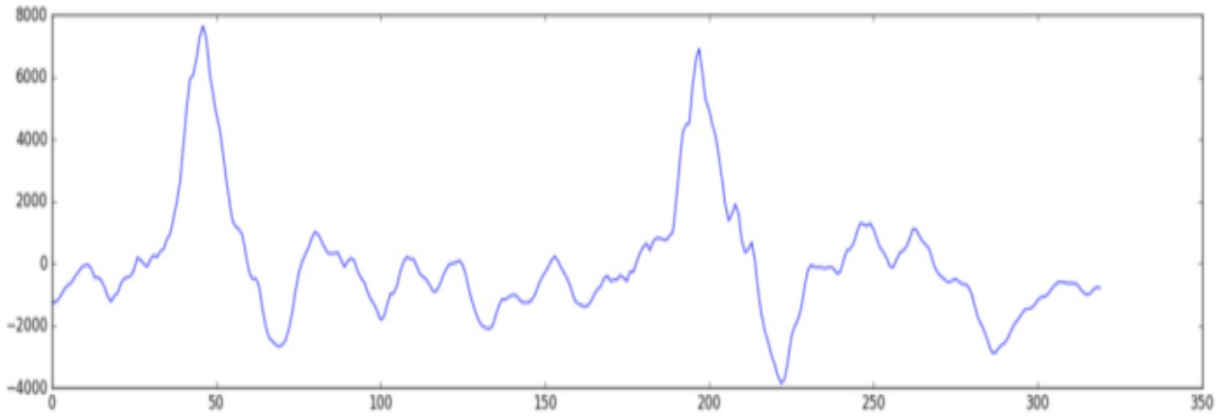

Рисунок 2.5 – Графік вихідного оцифрованого голосового сигналу

Якщо подумати, то такий аудіо запис триває всього навсього 1/50 від секунди, але на ньому відображена велику кількість різних фрагментів, що відрізняються частотою, та складають складну суміш. На ньому є: пару звуків з низькою частотою, звуки з середньою частотою а також декілька високих звуків. Не зважаючи на кількість, усі вони міксуються разом – саме так можна отримати звук людського вимовляння. Задля спрощення обробки подібної інформації, для передачі у НМ, необхідно розділити таку звукову хвилю на частини, які складають її, спочатку беручи найнижчі частоти. Наступним кроком, після обчислення потужності звуку на кожному періоді частоти, необхідно створити картину частот голосу. Наприклад, візьмемо запис деякого акорду – До-мінор на фортепіано. Даний звук вміщує у себе суміш з декількох музичних нот – До, -Ре і -Мі – саме які, при змішуванні, видають єдиний складний звук. Наступним кроком є розділення такого складного звуку на відокремленні ноти, задля виявлення вихідних нот. Таке перетворення можливе при математичний операції, яка відома як «перетворення Фур'є». При перетворенні проходить роз'єднання комплексної хвилі зі звуком на більш простіші звукові хвилі, які є її складовою. Коли ми маємо різні хвилі звуку, далі комбінуються потужності звуку, які знаходяться у кожній з них. В кінці кінців, виходом є оцінювання ролі кожного із частотного діапазону, починаючи від низьких до високих частот. Дані, які представлені далі, показують потужність звуку на кожному проміжку по 50

Гц на виході фрагменту рис.2.6:

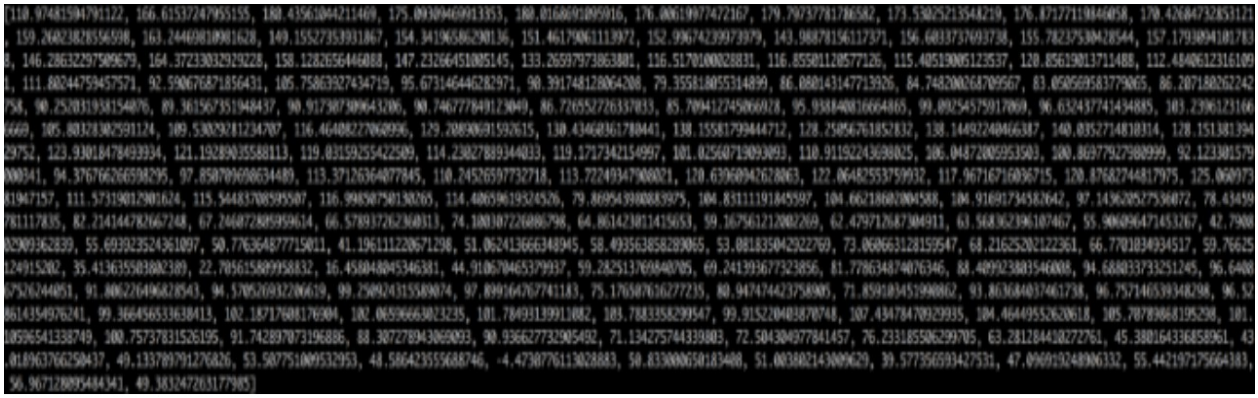

Рисунок 2.6 – Масив даних з значеннями потужності звуку в смузі по 50 Гц

При умовах, що такий процес буде мати повторення на кожному 20-ти мілісекундному фрагменті з вхідного аудіо, то ми матимемо змогу отримати спектрограму (стовпець з ліва на право відповідає 20-мілісекундному фрагментові) (рис. 2.7).

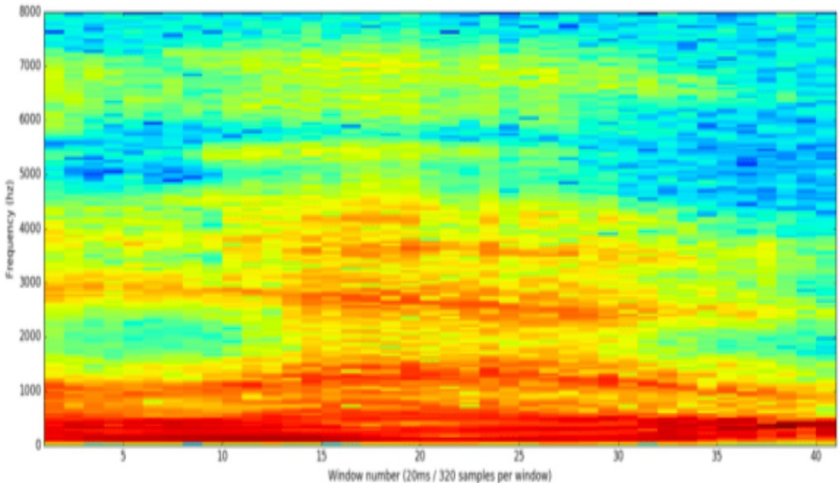

Рисунок 2.7 – Спектрограма запису

2.3 Алгоритм розпізнавання

Після отримання звуку в такому форматі, з ним можна надалі з легкістю працювати, а саме маємо змогу навчати на такій інформації глибинну нейронну мережу. Вона матиме змогу на вході отримувати аудіо-данні у

довжину 20мс. Далі НМ для кожного з цих фрагментів зробить спробу вказати, яка саме буква була вимовлена (рис. 2.8).

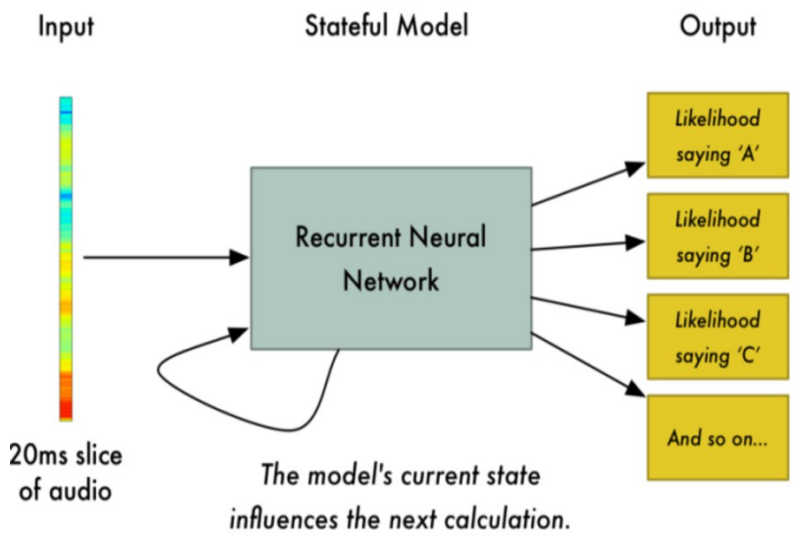

Рисунок 2.8 – Схема обробки мови

Найбільш доцільним варіантом буде використання рекурентної НМ, саме такої, яка на наступному кроці алгоритму буде враховувати дані, отримані з попередніх етапів. До переваг такої системи можна віднести здатність мати вплив на розпізнану наступну букву, адже попередня була вже визначена НМ. Якщо користувач, наприклад, сказав скорочене «прив», то у такому випадку, більш за все, наступним буде ідти «іт», для того, щоб завершити вислів «привіт». Найменша ймовірність того, що буде промовлені букви, які зовсім не відносяться до слова, наприклад, «РНК». Тобто, НМ має можливість при умові, що попередні результати будуть запам'ятовані, відтворити найбільш коректні прогнози у наступному кроці.

Вже після обробки нейронною мережею усього аудіо файлу, ми зможемо отримати роз'єднання абсолютно кожного фрагменту звуку на букви, які, скоріш за все, були вимовлені у цьому фрагменті. Як відображається «привіт» показано на рисунку 2.9.

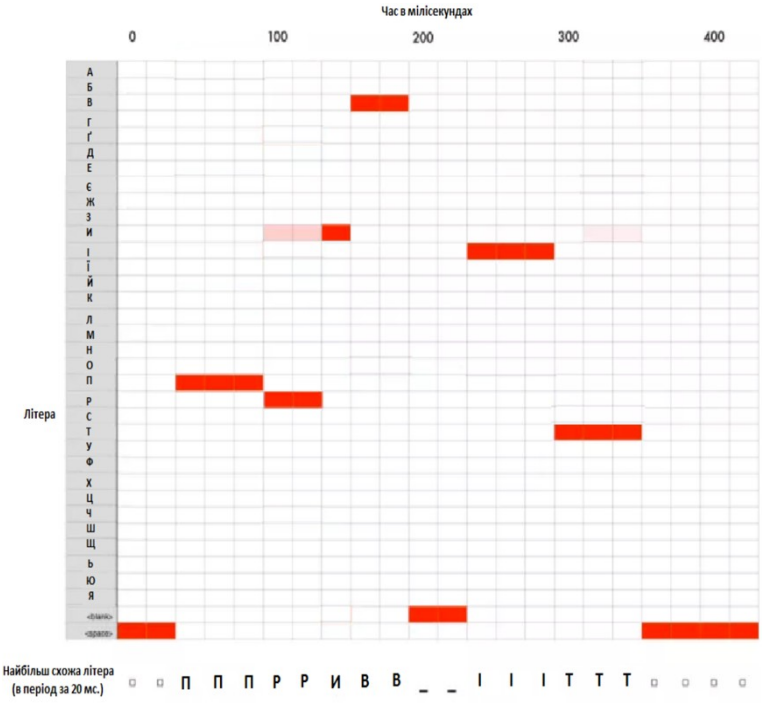

Рисунок 2.9 – Розкладання слова на літери

НМ далі проводить прогнозування на предмет того, що було озвучено саме «ПППРРРРРИВВ__ІІІТТТТ». Хоча не можна відкидати той факт, що нейронна мережа подумає, що звук відповідає «ПППРРРРИИВВВ__В_ІІІДДД» або також «ПРІІИУУУУ_УУІІІИТ». Вихідні дані далі перетворюються у декілька етапів. Перш за все, заміна будь-яких символів, що повторюються на один:

– ПППРРРРРИВВ__ІІІТТТТ стане ПРИВ_ІТ;
– ПРРРРРИИВВВВ__ВВВ_ІІІД стане ПРИ_В_ІД ;
– ПРРРІУУУУУ_УУУІІІІТ стане ПРІ_У_ІТ.

По-друге, відкидання усіх пропущених частин:

– ПРИВ_ІТ зіль'ється у ПРИВІТ;
– ПРИ_В_ІД зіль'ється у ПРИВІД;
– ПРІ_У_ІТ зіль'ється у ПРІУІТ.

Таким чином, можна побачити, що на виході є три можливих слова – «Привіт», «Привід» і «Пріуіт». При спробі їхнього вимовлювання, усі вони будуть сприйматися на звук, як «Привіт». Так як усі букви розпізнаються по

черзі, нейронна мережа також має змогу написати будь-які можливі та невимовні слова. Ось, якщо взяти висловлювання «Він не піде», НМ може показати «Уін небі те».

В основі принципу розпізнавання голосових команд за допомогою нейронних мереж, лежить принцип, уточнювання наступних передбачень, шляхом порівняння їх з досить не малою базою даних цифрового тексту (журнали, новинні статті і т.п.). Букви з найменшою вірогідністю відхиляються та використовуються ті, які відображають найбільш реалістичну картину слова.

Розглянувши різні варіації «Привіт», «Привід» і «Пріуіт», звичайно, «Привіт» буде вказана у базі даних текстових слів найбільш часто (а також у різноманітних аудіо файлах), саме тому, більш за все, це слово буде вірним варіантом. На такій підставі, можна обрати «Привіт» як кінцевий запис слова з аудіо фрагменту.

Надалі, опираючись на такі результати, виникає цікава ситуація – а що, якщо користувач промовить «Привід»? Таке слово можливе, тому «Привіт» не буде обране у якості остаточного результату, що не є правильниим варіантом. Однак, диктор і справді може вимовити «Привід», а не «Привіт». У такому разі, що система навчається на мові літератури, для розпізнавання голосу, не зможе обрати «Привід» у якості вірного слова. Також, якщо користувач мовить «Привід», але все ще хоче сказати «Привіт», навіть якщо він акцентує увагу на літері «Д». Можете і самі протестувати це – якщо ваш телефон підтримує функції для розпізнавання голосу, потренуйтесь вимовити слово «Привід». Звичайно, телефон відмовиться сприйняти вас, тому буде відображати це як «Привіт».

2.4 Опис алгоритму для аналізу голосу

Проаналізуємо алгоритм, за допомогою якого можна вирішити одне з практичних завдань, пов’язаних з управлінням девайсів та відображення

даних при участі голосових команд. Звернемо увагу, що вибір команд, а саме окремих слів, побудований таким чином, що необхідно, щоб слова були схожі за вимовою. Це робиться для оцінювання придатності до моментів, при яких готового вирішення розпізнання команди не має. Для того, щоб команда була отримана, використовується звичайний мікрофон, що підключається за допомогою аудіо адаптеру до робочої машини, яка працює за допомогою операційної системи. Вирішення даної проблеми можна розбити на етапи:

1) Відокремлення з загальної звукової осцилограми, що вже оціфрована та триває 2 секунди, даних про конкретне слово – команду.

2) Розділення осцилограми на різні частки у довжину приблизно 15-23 мілісекунд (довжина команди, тобто, слова приблизно 0.65-0.90 секунди).

3) Вираховування дискретного перетворення Фур'є для кожної ділянки слова (вивести спектр сигналу на кожній ділянці).

4) Відокремлення на кожній з ділянок n – точок, які відповідають за локальні максимуми амплітуд, та за їх показами частот (виділення форм голосового сигналу).

5) Пошук, опираючись на попередньо отримані результати, такого числа n, при якому відразу після відновлення голосу за допомогою синусового перетворення Фур'є виконається точне суб'єктивне розпізнання слова виходячи з набору отриманих команд.

6) Устаткування масиву інформації для кожного фрагменту з ділянок слова, які будуть відображати відповідне слово.

7) При ситуації, що одне і те ж слово, промовлене навіть одним користувачем, має різні осцилограми, буде створений набір з декількох сотень масивів, відповідних для одного і того ж звуку.

8) Після отримання окремих наборів масивів, що характерні для відповідних слів, застосовується математичний апарат НМ для того, щоб розпізнати конкретне введене за допомогою мікрофона слово.

Виходячи з цього, можна навести алгоритм розпізнавання команди з вхідного голосу людини (рис. 2.10).

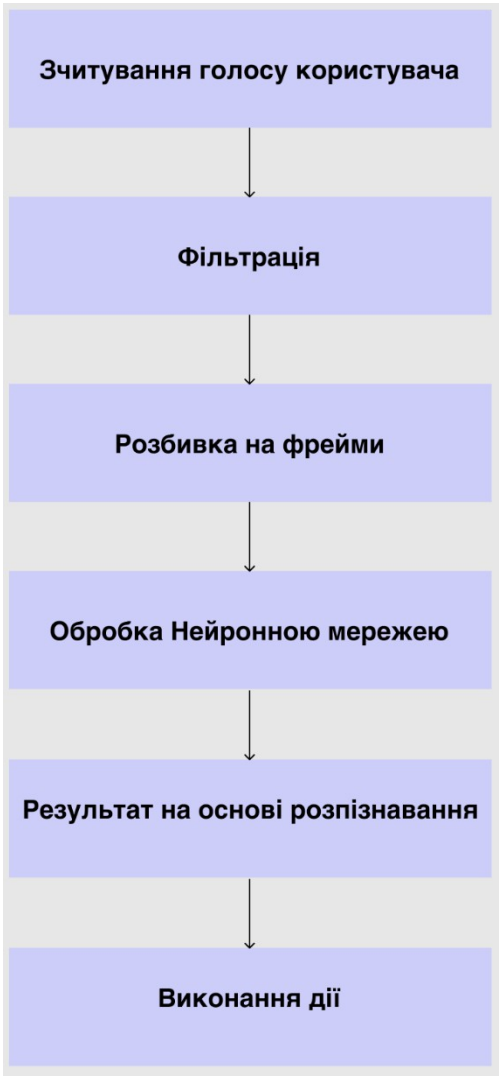

Рисунок 2.10 – Блок-схема роботи системи

3 РЕАЛІЗАЦІЯ СИСТЕМИ РОЗПІЗНАВАННЯ ГОЛОСУ

3.1 Технічне завдання

Задля того, щоб можливо було виконати поставлені задачі, програмний комплекс повинен включати у себе такі функції:

- сприйняття звукових сигналів у форматі WAV;
- розбивка голосового сигналу на фрейми;
- ідентифікація тексту.

Для написання програми обрано середовище розробки PyCharm. Графічний інтерфейс та його проектування не є першочерговою задачею, адже програма, яка розробляється не буде використана задля масового використання, вона необхідна для проведення аналітичної роботи в умовах розробки.

3.2 Структура Wave файлу

Розглянемо файл аудіозаписи формату WAV. Подібний формат аудіо запису є підвидом RIFF (Resource Interchange File Format – формат запису, який призначений для обміну даними). В його основі лежить декілька областей, які є розмежованими одна між одною. Перша область – це область, що містить у собі певну невелику кількість інформації про сам аудіо запис, тобто іншими словами, заголовок файлу, друга – область даних. Традиційно, заголовок запису містить у собі наступну інформацію:

- розмір файлу;
- кількість каналів;
- частота дискретизації;
- глибина звучання (або кількість біт в кожному семпли).

Хоча, заголовок може містити у собі не тільки такі дані, але і додаткову

інформацію щодо аудіофайлу, таку як: кількість байт, яка характеризує область даних, формат для стиснення, тощо.

В основі звуку лежать коливання, що при оцифруванні перетворюються у вигляд ступінчастої інформації. Подібний вигляд може бути отриманий завдяки тому, що комп'ютер має змогу показувати звуковий запис певної амплітуди у будь-який не тривалий проміжок часу, при цьому такий малий момент має скінченну тривалість. Частота дискретизації визначає тривалість такого проміжку. Семпл – це поєднання короткого проміжку часу та амплітуди сигналу. Амплітуда може бути представлена у вигляді числа, яке може бути в файлі величиною в 8, 16, 24, 32 біт та більше. Дивлячись на це, можна стверджувати, що чим більше відведено місця в пам'яті для відображення числової характеристики амплітуди, то тим більш великий діапазон уявлень числа може бути збережений. Під час розробки програмної реалізації такої системи необхідно враховувати той факт, що в аудіофайлі з одним каналом уявлення амплітуди розміщуються послідовно [8]. Зазвичай, у стерео форматі, першим йде показник усього списку амплітуд для лівого каналу, а вже потім для правого. Хоча якщо опиратися на параметри форматування, можна представити стерео формат у вигляді, у якому один семпл використовується для лівого каналу, а інший семпл для правого каналу. На рисунку 3.1 представлена структура файлу формату WAV.

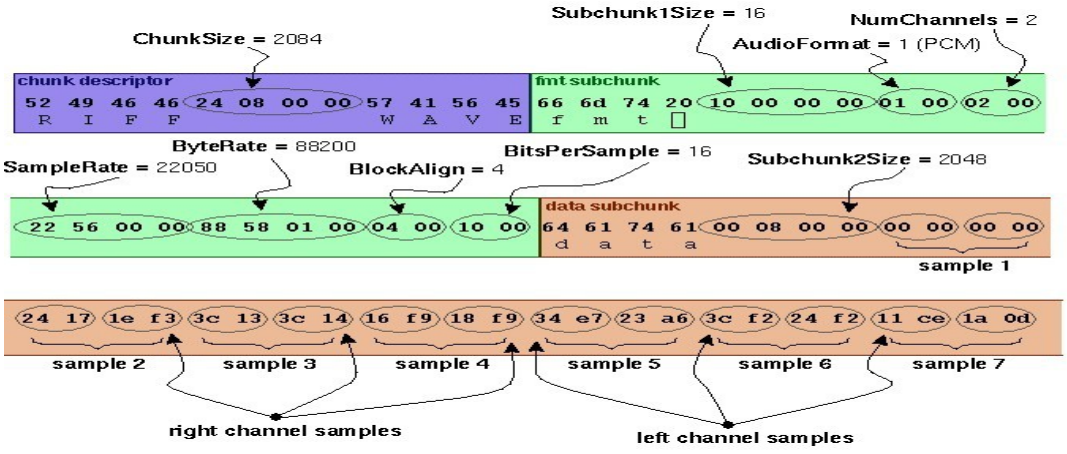

Рисунок 3.1 – Структура файла формату WAV

Синя зона на рисунку, яка включає у себе перші 4 байти – це назва

головної частини «RIFF». Усі представлені дані зазвичай записуються в режимі кодування ANSII. Чотири байти, які розташовуються після, представляють собою розмір ланцюга, який залишився, починаючи від установленої позиції. Наступні 4 байти означають формат RIFF-файлу. Опираючись на те, що наразі розглядається файл формату WAV, то звичайно, в цій комірці є надпис «WAV» в кодуванні ANSII.

Чотири байти на початку зеленої зони, що мають ідентифікатор «subchunk1Id» є іменем нової частини «fmt», розмір якої відображається у наступних за ним 4х байтах. Два байти, що йдуть далі показують ступінь стиснення. У разі, наявності у них числа, яке відрізняється від одиниці, це є ознакою того, що має місце стиснення файлу. Два байти, що йдуть далі та мають назву «numChannels» показують кількість каналів, наприклад: 1 – це монозвук, 2 – відображає стерео звук і тому подібне. Два байти, що йдуть далі та мають назву «sampleRate» показують частоту дискретизації. Два байти, що йдуть далі та мають назву «byteRate» – кількість байт, що були передані за одну секунду відтворення. Два байти, що йдуть далі та мають назву «blockAlign» відображають кількість байт для лише одного семпла, що включає усі доступні канали. У кінці зеленої зони 2 байта «bitsPerSample» показують кількість біт у семплі або глибину звучання.

На початку коричневої зони чотири байти показують назву наступного блоку під назвою «data», а розмір цього блоку відображають наступні 4 байти «subchunk2Size». За ними йдуть байти, що відображають дані звукової хвилі.

Але можна побачити неточність – якщо підрахувати кількість байт, що йдуть до звукових даних, а потім просумувати їх з кількістю байтів, що були відведені для даних щодо звуку, то можна побачити, що іноді може проявитися недолік приблизно у 260-310 байтів у розмірі файлу. Таке явище можливе через те, що аудіо файл додатково зберігає різну інформацію, яка не суттєва для відтворення – дату створення файлу, автора, користувача, який створив файл, жанр і подібні дані [9]. Головна проблема полягає в тому, що зовсім не однакові програми для форматування можуть розміщувати такі дані

у досить різних місцях. Розглянемо на прикладі Audacity, що може зберігати їх одразу до чи після розділу «data», а ось інша програма Freemake Audio Converter кладе їх до розділу з даними. Такі різні підходи можна побачити на рисунках 3.2 і 3.3 відповідно.

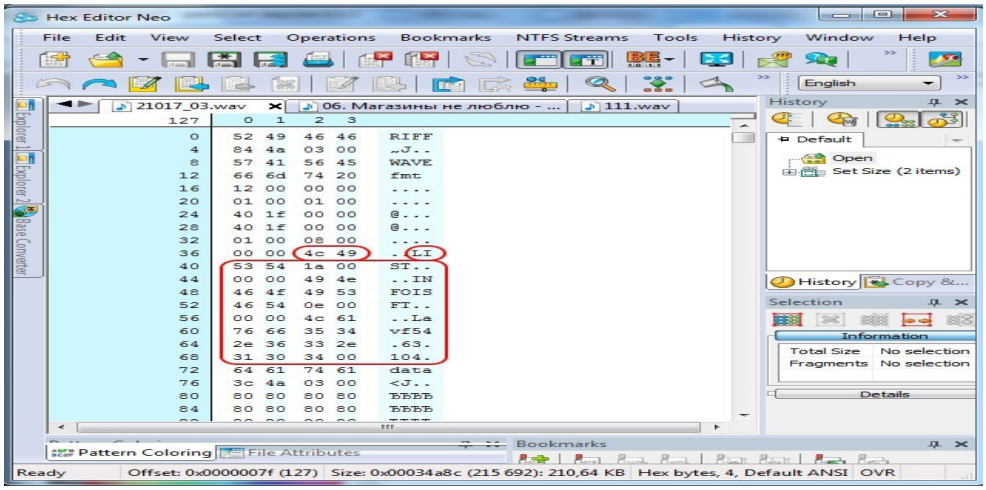

Рисунок 3.2 – Допоміжні данні на початку файлу

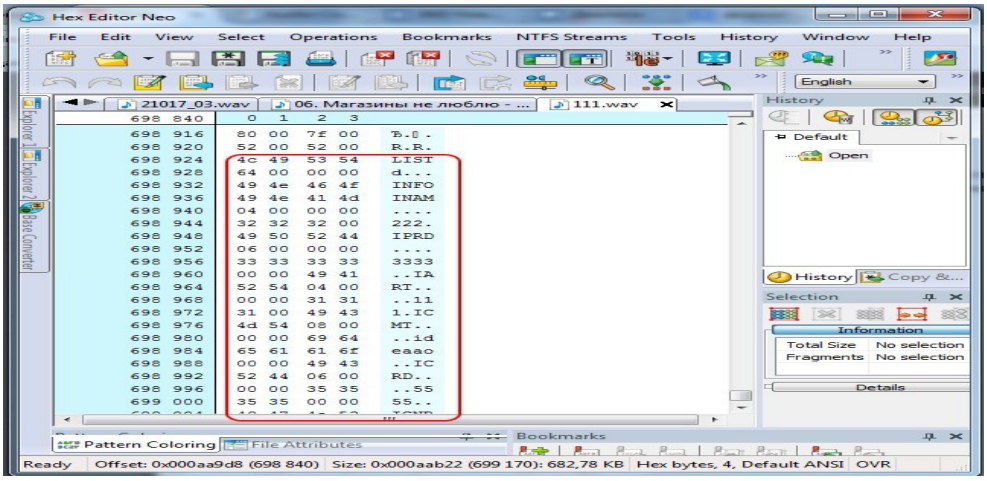

Рисунок 3.3 – Додаткові дані в кінці файлу

Опираючись на ці дані, необхідно звертати увагу на цю особливість під час зчитування основної послідовності звукових даних.

3.3 Робота з wav файлами у Python

Для забезпечення мінімізації обчислювального навантаження, необхідно користуватися принципом зчитування сигналів у вигляді звуку на пряму з файлу. Щоб мати змогу зчитувати звукові дані з файлів у середовищі розробки для Python – PyCharm, тут заздалегідь розроблена функція wavread. Дана функція завантажує файл зі звуком та повертає такі дані як: fs – частота дискретизації, sig – сигнал у вигляді звуку о приймає значення від -1 до 1, b – кількість біт.

Можна додати, що є можливість прослухати завантажений у програму файл зі звуком, для такої ситуації можна користуватися функцією sound (sig, fs), у якій fs – частота дискретизації, sig – звуковий сигнал.

3.4 Вибір нейронної мережі

Нейронні мережі прямого поширення (feed forward neural networks, FF або FFNN) і перцептрони (perceptrons, P) дуже прямолінійні, вони передають інформацію від входу до виходу. Нейронні мережі часто описуються у вигляді листкового торта, де кожен шар складається з вхідних, прихованих або вихідних клітин. Клітини одного шару не пов'язані між собою, а сусідні шари зазвичай повністю пов'язані. Найпростіша нейронна мережа має дві вхідних клітини і одну вихідну, і може використовуватися в якості моделі логічних вентилів. FFNN зазвичай навчається за методом зворотнього поширення помилки, в якому мережа отримує безлічі вхідних і вихідних даних. Цей процес називається навчанням з учителем, і він відрізняється від навчання без учителя тим, що в другому випадку безліч вихідних даних мережу становить самостійно. Вищезазначена помилка є різницею між введенням і висновком. За ситуації, у якій у НМ міститься достатня кількість прихованих нейронів, така нейронна мережа має змогу змоделювати відносини між даними, що отримуються на вході та виході. Хоча на практиці, подібні НМ використовуються не часто, але їх часто поєднують разом з іншими нейронними мережами задля отримання нових.

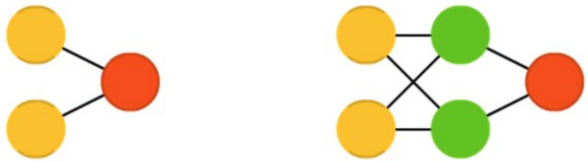

Рисунок 3.4 – Нейронні мережі прямого поширення

Нейронна мережа Хопфілда (Hopfield network, HN) – це НМ, яка має повнозв’язну структуру разом з зеркальною матрицею зв'язків. У момент реєстрації вхідної інформації кожен вузол являється вхідним, потім, під час навчання він поступово стає прихованим, після чого – виходом. За таких уявлень, НМ навчається наступним чином: спершу значення нейронів реєструються відповідно до наближених у шаблоні, далі обчислюються ваги, адже у майбутньому вони не змінюються. Наступним етапом, після того, як нейронна структура навчилася на позначених шаблонах, вона і надалі буде звертатися до одного з них, хоча і не завжди до правильного. Мережа нормалізується в залежності від загальної «температури» і «енергії» мережі. Для усіх нейронів існує індивідуальний поріг активації, який змінюється в залежності від температури, під час проходження якого нейрон займає одне з двух значень (при нормальних умовах -1 або 1, рідше 0 або 1). Подібні мережі зазвичай вважаються системами з асоціативної пам'яттю: як людина, бачачи половину таблиці, може представити другу половину таблиці, так і ця мережа, отримуючи таблицю, наполовину зашумленну, відновлює її до повної.

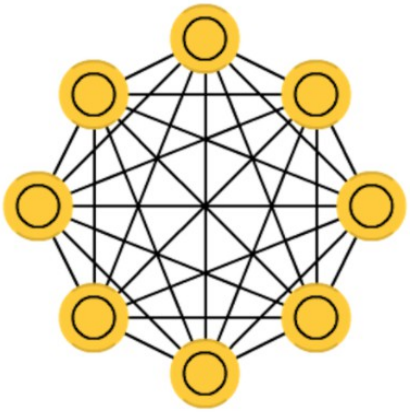

Рисунок 3.5 – Нейронна мережа Хопфілда

Ланцюги Маркова (Markov chains або discrete time Markov Chains, DTMC) – це попередні підходи до машин Больцмана і мереж Хопфілда (HN). Їхній зміст можна пояснити так: які мої шанси потрапити в один з наступних вузлів, якщо я перебуваю в даному? Кожне наступне стан залежить тільки від попереднього.

Машина Больцмана майже ідентична з мережею Хопфілда, головна відмінність у тому, що деякі нейрони у ній маркуються як вхідні, а інші – як приховані. Вхідні нейрони у майбутньому стають вихідними. Машина Больцмана – це стохастична мережа. Навчання відбувається по методології зворотного поширення помилки або за допомогою алгоритму порівняльної розбіжності. В цілому процес навчання дуже схожий на такий, як у мережі Хопфілда.

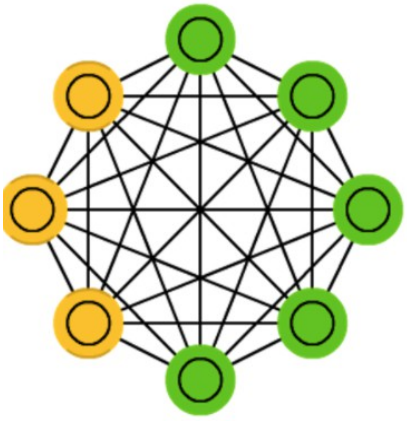

Рисунок 3.6 – Машина Больцмана

Обмежена машина Больцмана (restricted Boltzmann machine, RBM) дуже схожа на машину Больцмана і, отже, на мережу Хопфілда. Єдиною різницею є її обмеженість. У ній нейрони одного типу не пов'язані між собою. Обмежену машину Больцмана можна навчати, але з одним нюансом: замість прямої передачі даних і зворотного поширення помилки потрібно передавати дані спершу в прямому напрямку, потім в зворотному. Наступним кроком є етап проходження навчання за методом прямого та зворотного поширення

помилки [10].

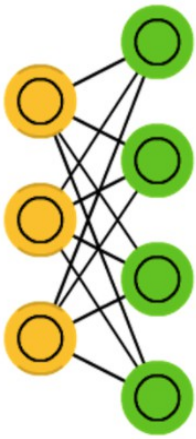

Рисунок 3.7 – Обмежена машинаБольцмана

Розріджений автокодувальник (sparse autoencoder, SAE) – в якомусь сенсі протилежність звичайного. Замість того, щоб навчати мережу відображати інформацію в меншому «обсязі» вузлів, ми збільшуємо їх кількість. Замість того, щоб звужуватися до центру, мережа там роздувається. Мережі такого типу корисні для роботи з великою кількістю дрібних властивостей набору даних. Якщо навчати мережу як звичайний автокодувальник, нічого корисного не вийде. Тому крім вхідних даних подається ще і спеціальний фільтр розрідженості, який пропускає тільки певні помилки.

Варіаційні автокодувальник (variational autoencoder, VAE) мають схожу архітектуру, але навчають їх іншому: наближенню імовірнісного розподілу вхідних зразків. У цьому вони беруть початок від машин Больцмана. Проте, вони спираються на Байєсову математику, коли мова йде про імовірнісні висновки і незалежність, які інтуїтивно зрозумілі, але складні в реалізації. Якщо узагальнити, то можна сказати що ця мережа приймає до уваги вплив нейронів. Якщо щось одне відбувається в одному місці, а щось інше – в іншому, то ці події не обов'язково пов'язані, і це повинно враховуватися.

Мережа типу «deep belief» – це мережа, у якій представлена архітектура, основу якої складає зв’язка з декількох з'єднаних НМ. Подібні

структури навчаються у блочному режимі, саме тому кожному блоку необхідно тільки мати змогу закодувати попередній. Подібний підхід відомий як жадібне навчання, який базується на виборі локальних оптимальних рішень, які не можуть забезпечити точно найкращій кінцевий результат. У додатку, НМ можна навчити (за допомогою методу зворотного поширення помилки) показувати інформацію у вигляді ймовірнісної схеми. При умові, що буде використане навчання без вчителя, стабільна мережа може бути використана для генерації нової інформації.

Згорткові НМ (convolutional neural networks, CNN) та глибинні згорткові нейронні мережі (deep convolutional neural networks, DCNN) дуже різняться з іншими типами мереж. Зазвичай, їх використовують задля обробки малюнків, рідше для аудіо чи відео. Традиційним способом застосування згорткових нейронних мереж є класифікація зображень, наприклад. якщо зображення містить у собі кішку, НМ покаже «кішка», у разі собаки – «собака». Подібні системи містять у собі так званий «сканер». Припустимо, що ми маємо зображення 200 на 200 пікселів, звичайно, ми не будемо одразу обчислювати всі 40 тисяч точок. Натомість, така НМ отримує квадрат розміром у 20 на 20 пікселів, опорною точкою для якого є лівий верхній кут, далі пряумє на 1 піксель та зберігає новий квадрат, такий ланцюг продовжується до кінця. Отримана вхідна інформація надалі передається через згорткові нейронні утворення (шари), у яких не усі вузли поєднані один з одним. Данні шари мають особлівисть стискуватися з глибиною, при цьому найчастіше використовується ступень двійки: 32, 16, 8, 4, 2, 1. На практиці, для подальшої обробки даних до кінця CNN прикріплюють FFNN. Зазвичай ці мережі маю назву - глибинні (DCNN).

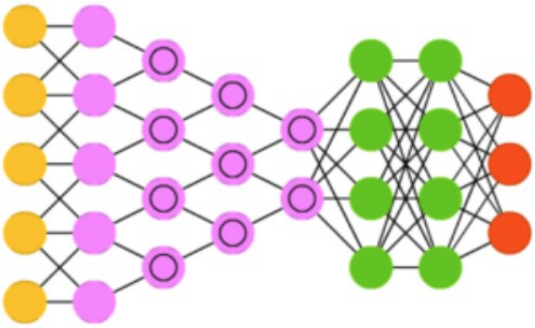

Рисунок 3.8 – Згорткові нейронні мережі

Деконволюційні нейронні мережі (deconvolutional networks, DN), які також називають зворотними графічними мережами, є зворотним до згорткових нейронних мереж. Уявіть, що ви передаєте мережі слово «кішка», а вона генерує картинки з кішками, які подібні на реальні зображення котів. Слід зазначити, що є можливість об'єднувати DNN з FFNN. В переважній кількості випадків мережі бінарний вектор, передається не рядок: наприклад, <0, 1> – це кішка, <1, 0> – собака, а <1, 1> – і кішка, і собака.

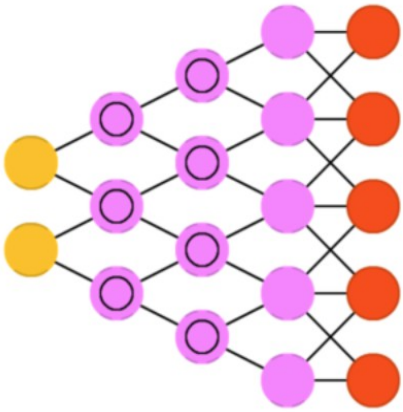

Рисунок 3.9 – Деконволюційні нейронні мережі

В якості методу розпізнавання голосу була обрана однонаправлена мережа прямого розповсюдження, бо саме такі мережі найчастіше застосовують для розпізнавання образів, прогнозування та апроксимації нелінійних функцій.

3.5 Вибір типу навчання нейронної мережі

Навчання з учителем означає, що система має наявність повного словнику проіндексованих даних задля того, щоб натренувати моделі на кожному з етапів її побудови. Існування повністю розмічених словників може означати, що для кожного запиту в наборі для навчання залежить відповідь, яку алгоритм має отримати у майбутньому. Тому, такий розмічений набір даних, що містить фотографії квітів, може навчити НМ, де зображені ромашки, рози або кульбаби [11]. На момент коли нейронна мережа отримає нове зображення, вона поетапно порівняє його з існуючими даними з навчальної бази, у цілях передбачення відповіді.

У більшості разів навчання з учителем використовується для рішення декількох видів задач: регресії та класифікації. Для задач, що стосуються класифікації, алгоритм може передбачати дискретні значення, які дорівнюють порядковим номерам класів, до яких можна віднести об'єкти. Навчальна база даних із зображеннями тварин містить містки для кожого зображення, такі як, «кролик», «собака» або «криса». Точність алгоритму визначається тим, наскільки якісно він може коректно розбити на класи нові зображення з собаками і кроликами.

Хоча завдання регресії тісно пов'язані з безперервною інформацією. Наприклад, лінійна регресія має змогу обчислювати очікуване значення змінної y, при цьому враховувати істині значення x.

Найбільш утилітарні проблеми машинного навчання включають у себе не малу кількість змінних. Наприклад, НМ може передбачати ціну дому в Києві на основі її площі та місця розташування або доступності міського транспорту. Програма з подібним алгоритмом може виконувати роботу експерта, що повинен обчислювати ціну квартири опираючись на ті ж дані.

Можна зробити висновок, що навчання з учителем більше всього підходить для проблем, коли є великий набір точних даних для навчання програми. Хоча це не постійне явище.

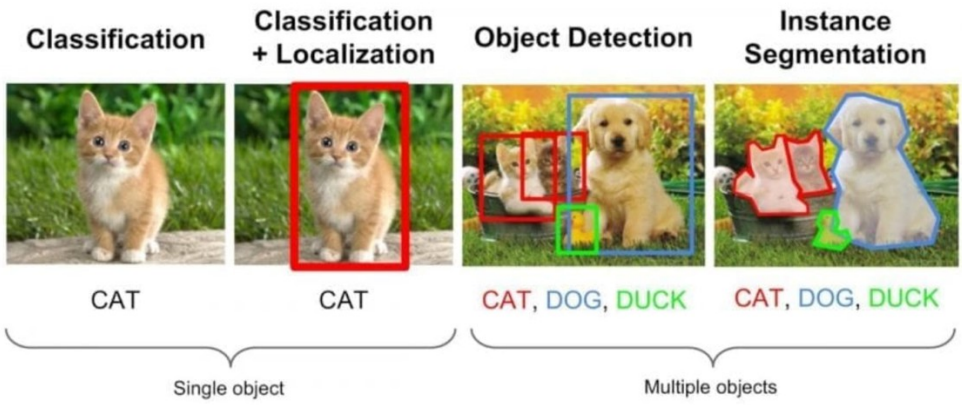

Рисунок 3.10 – Навчання з учителем

Бездоганно розмічені та якісні відфільтровані дані отримаи нелегко. Саме через це інколи перед алгоритмом стоїть проблема знаходження результатів на заздалегідь не відомі дані. Це головна сфера, де точно потрібне навчання без учителя.

В основі навчання без учителя (unsupervised learning) у моделі лежить набір даних, при цьому немає ніяких правил, які б описували, що з ним робити. НМ намагається своїми силами відновити кореляцію у даних, вирізняючи корисні ознаки та намагаючись аналізувати їх. Опираючись від завдання система аналізує дані по-різному.

Кластеризація. Звичайно, зрозуміло, що навіть без спеціальних знань експерта-орнітолога людина може глянути на колекцію зображень та класифікувати їх на групи за типами птахів, опираючись на колір опиріння, форму або розмір дзьоба. У цьому і полягає принцип кластеризації, що є найбільш поширеним завданням для навчання без учителя. Алгоритм обирає подібні дані, та знаходить спільні ознаки, а потім групує їх один з одним.

Асоціації. Наприклад, ви берете в Інтернеті товар для малюків, малинове пюре та дитячу кухоль, після чого сайт порекомендує вам внести радіоняню або нагрудник до корзини замовлень. На такому принципі і засновуються асоціації: деякі характеристики об'єкта зпівпадають з іншими характеристиками. Під час аналізу декількох головних ознак об'єкту, система може передбачити наступні, з якими існує певний зв’язок.

Виявлення аномалій. Банки мають змогу виявляти шахрайські методи, при цьому виявляти незвичайні дії в купівельних алгоритмах дій клієнтів. Звичайно, буде підозрілим, якщо одна кредитна карта спочатку використовується у Києві, а потім у Парижі в один і той же день. Приблизно таким чином навчання без участі вчителя може використовуватися для знаходження аномалій в обраній інформації.

Автоенкодери. Автоенкодери мають змогу приймати вхідну інформацію, кодувати її, а потім стараються показати початкові дані з вхідного коду. Існує не так багато дійсних ситуацій, коли може бути використаний звичайний автоенкодер. При цьому необхідно додати шари і можливості розширюватися, при цьому використовуючи зашумлені вхідні версії даних для навчання, автоенкодери мають змогу вирізати шум з відео контенту, цифрових зображень або медичних знімків, для підвищення якісті даних.

При методі у якому використовується навчання без учителя дуже не просто обчислити якість алгоритму, адже у них не має коректних відповідей або міток. Хоча при цьому розмічена інформація зазвичай ненадійна або її дуже складно отримати. При таких ситуаціях, якщо надавати моделі свободу дій для пошуку схожостей, можна отримати точні дані.

Існує навчання з частковим залученням вчителя, яке характеризується своєю назвою: навчальний набір даних включає у себе як розмічені, так і розподілені дані. Подібний метод дуже корисний, за умови, коли важко отримати з інформації головні ознаки або розмітити усі набори даних – це є склданим завданням.

Подібний метод машинного навчання дуже часто використовується для аналізу медичних сканів, таких як комп'ютерна томографія. Якісний рентгенолог має змогу розмітити малу підмножину зображень, на яких виявлені недоліки та захворювання. Звичайно, розмічати вручну усі зображення – то є занадто трудомісткою та складною задачею. Не дивлячись на це, НМ має змогу отримати дані з малої частини поміченої інформації та

повисити якість прогнозів у порівнянні з системою, яка може навчатися тільки на нерозмічених даних [12].

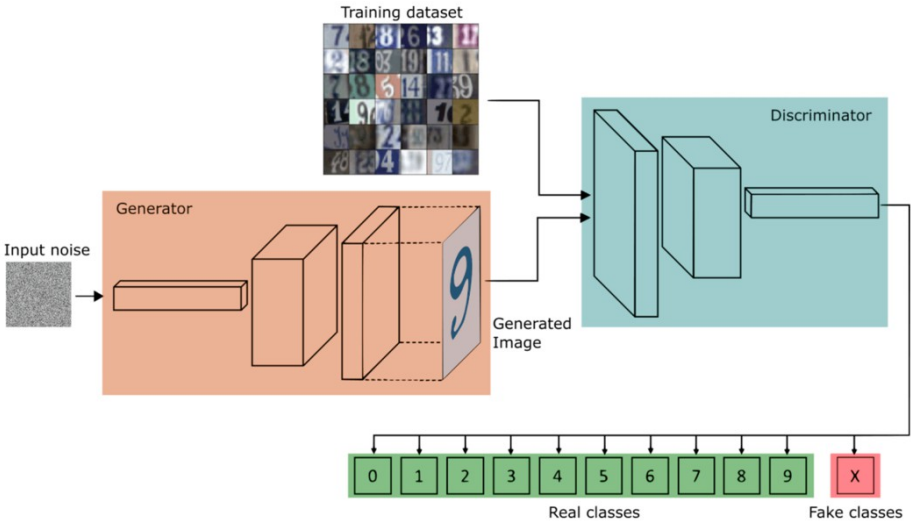

Рисунок 3.11 – Навчання з частковим залученням вчителя

Один із популярних методів навчання, для якого необхідний малий набір розміченої інформації, базується на використанні генеративно-змагальної мережі.

Розглянемо змагання двох нейронних систем, у якому кожна з них намагається виграти іншу за рахунок хитрощів. Перша з нейронних мереж, що є генератором, намагається відтворити нові частини даних, які можуть імітувати навчальну вибірку. Друга нейронна мережа, що називається дискримінатором, у свою чергу, оцінює, чи є така згенерована інформація реальною або сгенерованою. Нейронні мережі взаємодіють одна з одною та постійно удосконалюються, так як дискрімінатор намагається більш точно розділяти оригінали від підробок, а генератор, у свою чергу, намагається генерувати переконливі дані.

Відеоігри засновані на системі стимулів. Завершіть рівень і отримаєте нагороду. Переможете всіх монстрів і заробите бонус. Потрапили в пастку – кінець гри, не потрапляйте. Ці стимули допомагають гравцям зрозуміти, як краще діяти в наступному раунді гри. Без зворотного зв'язку люди б просто

брали випадкові рішення і сподівалися перейти на наступний ігровий рівень.

Навчання з підкріпленням (reinforcement learning) діє за тим же принципом. Відеоігри – популярне тестове середовище для досліджень. Агенти нейронних мереж намагаються знайти найкращій спосіб досягнення цілі або покращення продуктивності для кожного середовища. У той час, коли агент проводить дії, які можуть сприяти досягненню цілей, він отримує винагороду. Найвищою ціллю є передбачення таких кроків, з метою досягти максимальну винагороду у кінці кінців.

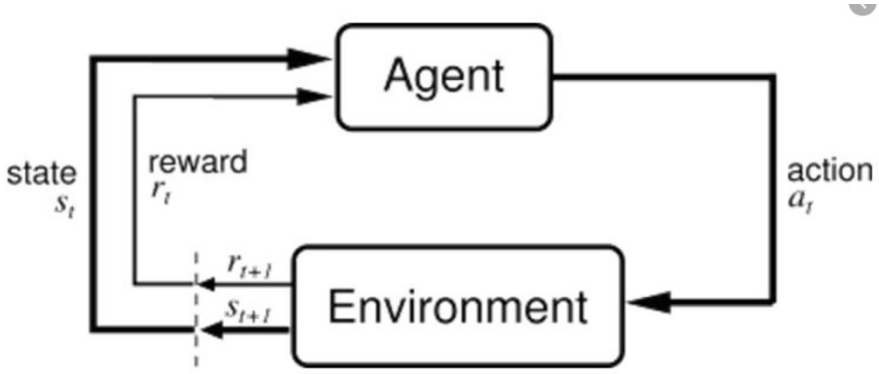

Рисунок 3.12 – Навчання з підкріпленням

Під час прийняття рішення, агент починає вивчати зворотний зв'язок, а також прогресивні тактики рішення яких, здатне привести до кращого результату. Такий підхід користується довгостроковою стратегією, наприклад, як у шахах: попередній найкращий хід не завжди може допомогти у виграші у кінцевому рахунку. Саме через це агент старається максимізувати нагороду, яка буде у сумі.

Такий процес є ітеративним. Чим більше рівнів зі зворотного зв'язку, тим більш краще розвивається стратегія даного агенту. Подібний підхід завжди корисний для навчання роботизованих систем, що мають змогу керувати автономними транспортними установами або інвентарем на складі.

Як можна бачити, різні алгоритми навчають НМ по-різному, для нашого випадку був обраний тип навчання без учителя, так як у нас не має ідеально розмічених та чистих звукових даних.

3.6 Розробка нейронної мережі

Для того, щоб створити НМ використовувалася функція newff (PR, [S1 S2 ... SNl], {TF1 TF2 ... TFNl}, BTF, BLF, PF). У якості вхідних даних функція використовує:

– – Rx2 матриця максимальних та мінімальних значень строк матриці входу с розмірністю RxQ;

– – кількість нейронів в і – тому шарі, – кількість шарів;

– – функції активації і-го шару, за стандартом = 'tansig';

– – навчальна функція зі зворотним поширенням, за замовчуванням ='trainlm';

– – алгоритм підстроювання ваг і зміщень (навчальний алгоритм), за замовчуванням = 'learngdm';

– PF – функція оцінки того, як працює мережа, за замовчуванням = 'mse'.

А потім повертає односпрямовану нейронну мережу, що складається з декількох шарів.

У даній атестаційній роботі була використана НМ, яка складається з трьох розподілених шарів. Окремий шар нейронної мережі складається з нейронів. Нейрон є головним елементом для обчислення у нейронній мережі. В основі нейрона лежать суматори, помножувачі та нелінійний перетворювач даних. Синапси у свою чергу, дають зв'язок між цими нейронами та множать вхідні дані з сигналу на коефіцієнт, який вказує на вагу зв'язку – силу синапсів.

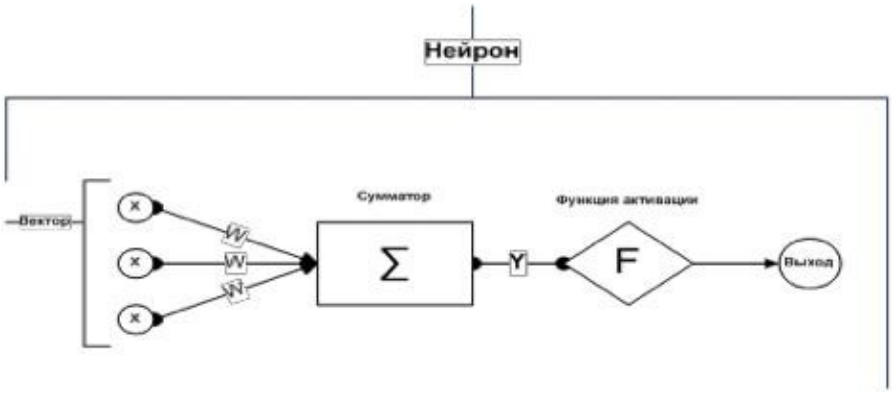

Рисунок 3.13 – Структура нейрона

Суматор у свою чергу, робить сумування даних із сигналів, що приходять від інших нейронів або зовнішніх сигналів. Нелінійний перетворювач являю реалізацію нелінійної функції одного аргументу, яка є виходом суматора. Подібна функція має назву «Функція активації» окремого нейрона.

Математичну модель для нейрона можна описати як:

(3.1)

де w_i – вага синапса (), – компонент вхідного вектора ()

Головним завданням в процесі розробки НМ є навчання подібної системи, або як можна сказати, коригування ваг для мережі у цілях зменшення помилки на виході НМ.

У якості функції активації була використана функція логарифмічної активації

(3.2)

Для функції навчання була обрана функція traingda. Це такий метод, що має в основі адаптивну швидкість навчання. Сенс методу заключається у налаштуванні ваг та зміщень відносно до методу градієнтного спуску.

Тренування НМ виконується відносно до характеристик функції

навчання traingda. Дані характеристики та їх значення при першому налаштуванні, наведені в таблиці нижче

Таблиця 3.1 – Параметри навчальної функції

Функція	Значення	Опис
net.trainParam.epochs	1000	Максимальна кількість епох навчання
net.trainParam.goal	0	Умови зупинки щодо відхиленню від еталону
net.trainParam.lr	0.01	Швидкість навчання
net.trainParam.max_fail	6	Максимальна кількість помилок на контрольному масиві
net.trainParam.min_grad	1e-5	Мінімальний градієнт
net.trainParam.show	25	Кількість епох поміж графіками
net.trainParam.showCommandLine	False	Показати командну строку
net.trainParam.showWindow	True	Показати графік тренування
net.trainParam.time	Inf	Максимальний час тренування в сек.

Виходячи з інформації вище, головним завданням у навчанні НМ є зведення до мінімуму квадратичної помилки. Така помилка вектора входу визначається за сумою подібних помилок у кожному нейроні виходу:

(3.3)

Задля мінімізації квадратичної помилки був використаний алгоритм градієнтного спуску. При такому виборі, тренування нейронної системи продовжується до тих пір, поки не буде пройдена одна з умов зупинки з таблиці 3.1.

Для ознайомлення, алгоритм навчання нейронної системи можна побачити на рисунку 3.14.

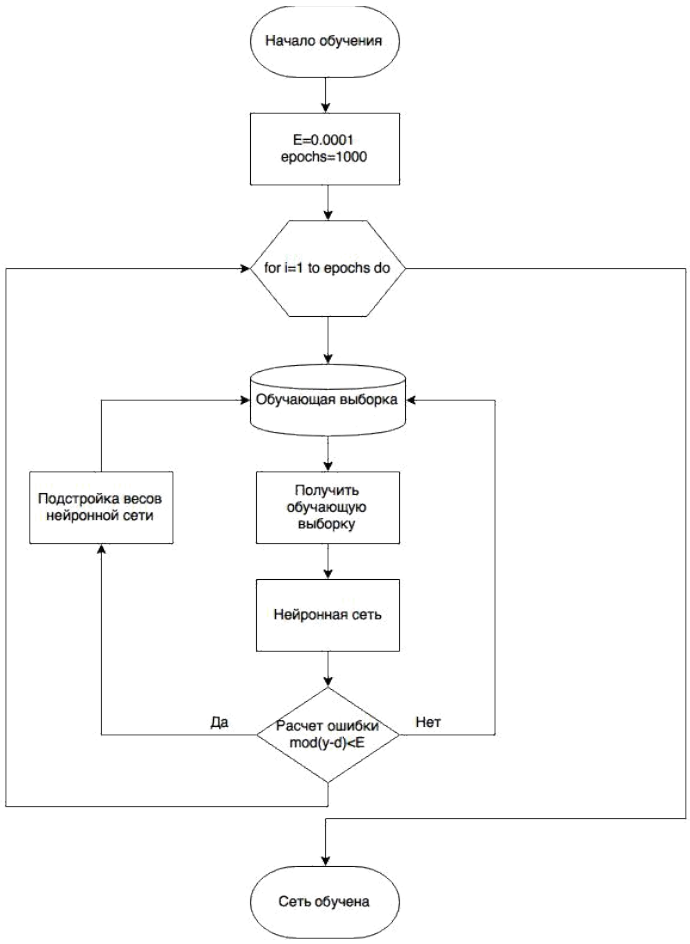

Рисунок 3.14 – Схема алгоритму навчання нейронної мережі

Як можна побачити на схемі, нейронна мережа може бути складена з декількох шарів. Перший шар приймає ваги, які приходять зі входу. Наступні шари отримують ваги від шарів, що йдуть попереду. Усі шари з нейронами мають зсуви. Третій шар являється виходом з мережі.

4 ТЕСТУВАННЯ СИСТЕМИ РОЗПІЗНАВАННЯ

4.1 Характеристики тестового стенду

Експеримент проводився на комп’ютері MacBook Pro.

Технічні характеристики:

- процесор: 2.2 GHz Intel Core i7;
- оперативная пам’ять: 16Gb 1600 MHz DDR3;
- графічний адаптер: Intel Iris Pro1536 Mb;
- операційна система: macOS Big Sur.

4.2 Результати експерименту

В процесі навчання на 500 епохах, найкращий вихід розпізнавання був отриманий при останній з епох для навчання, частка помилкового розпізнавання становить 30%. Час, витрачений на процес навчання склав 30 секунд. На наступному графіку показана зміна середньоквадратичної помилки рис. 4.1.

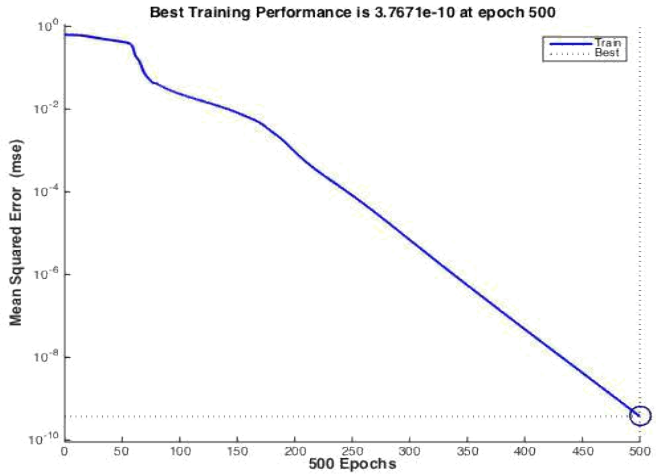

Рисунок 4.1 – Графік зміни середньоквадратичної помилки

При збільшенні епох навчання до 1000, частка помилкового

розпізнавання стала 22%. Даний результат на 8% кращий за попередній. Час, витрачений на процес навчання склав 35 секунд. Графік зміни середньоквадратичної помилки показаний на рисунку 4.2.

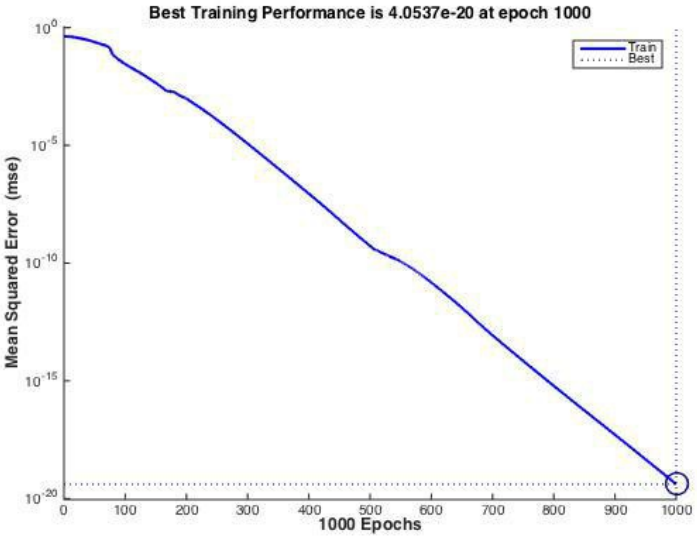

Рисунок 4.2 – Графік зміни середньоквадратичної помилки

При збільшенні епох навчання до 5000, частка помилкового розпізнавання становить 15%. Даний результат на 7% кращий, від навчання при 1000 епохах. Найкращий вихід розпізнавання був отриманий при 16000 з епох для навчання. Час, витрачений на процес навчання, склав 1 хвилину. Графік зміни середньоквадратичної помилки:

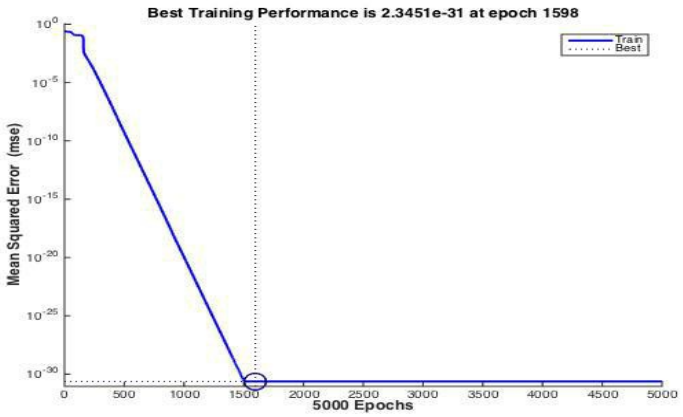

Рисунок 4.3 – Графік зміни середньоквадратичної помилки

За результатами проведених експериментів наочно видно, як

змінюється точність розпізнавання голосових команд системою при зміні параметрів: кількості епох навчання. Виходячи з цього треба зробити висновок, що при збільшенні кількості епох навчання та збільшенні тестового набору для навчання якість розпізнавання зростає.

Якість розпізнавання залежить від мікрофона, диктора, або навіть розташування апарату вводу від диктора. Для того, щоб вирішити ці задачі потрібно завести велику базу даних текстів зі словами, які будуть промовлені різними користувачами та використовуючи різні системи вводу, за різних положень мікрофонів.

Щоб досить точно (майже 100% вірного проголошення слова) провести розпізнавання слова зі звукового сигналу, необхідна відсутність сторонніх звуків коли вимовляється слово. Загалом, сторонні шуми маєть вплив на розпізнавання слова з 2-х секундного файлу формату wav. Тому можна ввести припущення, що якщо розпізнавання слова, за присутності шумів, буде вірним, то НМ для різних аналогів слів з шумами матиме змогу коректно виконати розпізнавання.

За наявності великої кількості вхідних даних, потрібно зіставити досить довгий масив інформації для навчання нейронної мережі, за умови використання бібліотеки. Можна припустити, що це є основним обмеженням використання бібліотеки методом розпізнавання за словами.

4.3 Інструкція користувача

Головним файлом програми є main.py. Для запуску системи необхідно запустити даний файл, попередньо відкривши його в PyCharm.

Так само для коректної роботи програми в PyCharm повинні знаходитися файли функцій neural.py, адже такі функції використовуються в системі та не є стандартними для середовища PyCharm.

Для того, щоб почати розпізнавання, необхідно запустити файл, включити мікрофон, почати говорити.

Для навчання нейронної мережі в папку voices необхідно додати нові данні команд у звуковому форматі. Далі потрібно змінити розмір транспонуючої матриці T та змінити величину сигналів в функції, яка відповідає за створення нейромережі.

Якщо необхідно протестувати мережу на нових тестових наборах або просто їх змінити, необхідно додати певну кількість записаних звуків та команд різних типів голосів та типів мовлення. Після чого запустити навчання, а далі спробувати розпізнати команди.

ВИСНОВКИ

В результаті виконання даної роботи був проведений аналіз методів розпізнавання мови та голосу. Детально досліджені існуючі методи вирішення задач розпізнавання команд. Проведений аналіз існуючих математичних апаратів, що вирішують дану проблему.

Найбільш перспективним вбачається варіант побудови пристрою на основі розпізнання голосу за допомогою нейронних мереж. Такий підхід дасть змогу досягти високої точності та стійкості в моменти розпізнавання команд із голосу.

Побудова пристрою розпізнавання мови на основі нейронних мереж дає можливість збільшення кількості голосових команд за допомогою звичайного додавання таких даних у базу нейронної мережі, що є дуже істотним, так як під’єднання великої кількості пристроїв є вагомою перевагою.

Було спроектоване програмне забезпечення, яке володіє достатніми параметрами стійкості до шуму та високою точністю обробки мови. Важливим предметом досліджень стало пошук варіантів підвищення точності, видалення шумів та різноманітних факторів, які мають змогу впливати на процес розпізнавання мови.

Для перевірки працездатності системи, було також проведене тестування за різних умов, у яких працювала програм. Після отримання результатів проведених тестувань, можна зробити висновок, що при збільшенні кількості епох навчання точність зростає.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Фролов, А.В. Синтез и распознавание речи. Современные решения [Электронный ресурс] / А.В. Фролов, Г.В. Фролов. – 2003. – Режим доступа: http://www.frolov-lib.ru/books/hi/index.html / 20.11.2020 г. – Загл. с экрана.
2. Квитко, М.В. Распознавание речи с помощью глубоких рекуррентных нейронных сетей [Электронный ресурс] / М.В. Квитко // IASA – 2016. – 223 с. – Режим доступа: http://sait.kpi.ua/media/filer_public/73/32/7332a68e-e93b-4c57-a3c8-66f11ee074cd/sait2016ebook.pdf . – Дата доступу: 15.11.2020 г. – Загл. с экрана.
3. Mohri, M. Speech recognition with weighted finite-state transducers. In Springer handbook of speech processing / M. Mohri, M. Pereira, F. Riley // Springer Berlin Heidelberg. – 2008. – P. 559-584.
4. Hinton, G. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups / G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, et al. // IEEE Signal Processing Magazine. – 2012. – V. 29, № 6, P. 82-97.
5. Холоденко, А.Б. О построении статистических языковых моделей для систем распознавания русской речи / А. Б. Холоденко // Интеллектуальные системы, 2002. – Т.6., вып. 1-4. – С. 381-394.
6. Сравнительный анализ систем распознавания речи с открытым кодом [Электронный ресурс] / Международный научно-исследовательский журнал. – Режим доступа: https://www.research-journal.org/technical/. Дата доступа: 16.03.2018 г. – Загл. с экрана.
7. Mermelstein, P. Distance measures for speech recognition, psychological and instrumental / P. Mermelstein // Pattern recognition and artificial intelligence. 1976. – v. 116. – P. 374-388.
8. Funahashi, K. On the approximate realization of continuous mappings by neural networks / K. Funahashi // Neural Networks, 1989. – v. 2, № 3. – P. 183-

191.

9. Карандашев, Я. М. Обобщённая модель Хопфилда и статфизический подход: общий случай / Я. М. Карандашев, Б. В. Крыжановский, Л.Б. Литинский // Нейроинформатика-2011. XIII Всероссийская научно-техническая конференция. Сборник научных трудов, ч.3, М., НИЯУ МИФИ, 2010. – С.181-190.

10. Hornick, K. Multilayer feedforward networks are universal approximators / K. Hornick, M. Stinchcombe, H. White // Neural Networks, 1989. – v. 2, № 5. – P. 359-366.

11. Cybenko, G. Approximation by Superpositions of a Sigmoidal Function / G. Cybenko // Mathematics of Control, Signals and Systems, 1989. – V. 2, № 4. – P. 303-314.

12. Funahashi, K. On the Approximate Realization of Continuous Mappings by Neural Networks / K. Funahashi // Neural Networks, 1989. – V. 2, № 3. – P. 183-191.