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)

Інтелектуальна система для ідентифікації людини по голосу
(тема)

Виконав: студент 2 курсу, групи СКСм-18-2

_____Андрєєв М.А._____
(прізвище, ініціали)

Спеціальність *123 _ Комп'ютерна інженерія*

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

Освітня програма_____

*Спеціалізовані комп'ютерні системи*_____
(повна назва освітньої прграми)

Керівник роботи _____доц. Шкіль О.С.___
(посада, прізвище, ініціали)

Допускається до захисту

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

2019 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_____

(шифр і назва)

Тип програми _____Освітньо-професійна_____

(освітньо-професійна або освітньо-наукова)

Освітня програма _____Спеціалізовані комп'ютерні системи_____

(повна назва

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
_____ (підпис)
«_____»_____ 20 ___р.

ЗАВДАННЯ

НА АТЕСТАЦІЙНУ РОБОТУ

студентові_____Андрєєву Михайлу Андрійовичу_____

(прізвище, ім'я, по батькові)

1. Тема роботи (проекту)_____

Intelligent system for identification human by voice_____

затверджена наказом по університету від "_04_"__11__2019 р. № _1624 Ст_____.___

2. Термін подання студентом роботи до екзаменаційної комісії _____20.12.2019_____

3. Вихідні дані до роботи (проекту)_____

Мова програмування Matlab_____

IDE Matlab_____

Wav файли_____

4. Перелік питань, що потрібно опрацювати у роботі _____

Сфери використання систем ідентифікації людини_____

Методи ідентифікації людини по голосу_____

Методи виділення характерних ознак звукового сигналу_____

Методи і засоби вирішення задачі класифікації_____

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів) _20 слайдів_ _____
_____
_____
_____
_____
_____

6. Консультанти розділів роботи (проекту)

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

7. Дата видачі завдання 03.09.2019_____

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи (проекту)	Термін виконання етапів проекту (роботи)	Примітка
1	Видача теми проекту, узгодження і затвердження	03.09.2019 -10.09.2019	
2	Аналіз проблемної галузі, постановка задачі, вибір інструментальних засобів	10.09.2019 -30.09.2019	
3	Проектування систем логічного управління	30.09.2019 -15.10.2019	
4	Дослідження існуючих методів розв'язання задачі ідентифікації	15.10.2019 -15.11.2019	
5	Реалізація обраного алгоритму	15.11.2019 - 25.11.2019	
6	Програмна реалізація алгоритму	25.11.2019 -05.12.2019	
7	Перевірка якості роботи системи	05.12.2019 -15.12.2019	
8	Оформлення пояснювальної записки	15.12.2019 -20.12.2019	
9	Захист проекту	20.12.2019 -25.12.2019	

Студент _____
(підпис)

Керівник роботи (проекту) _____ _____доц. Шкіль О.С.___
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка містить 73 сторінки, 23 рисунки, 2 додатки, 13 джерел за переліком посилань.

НЕЙРОННІ МЕРЕЖІ, ІДЕНТИФІКАЦІЯ І ВЕРИФІКАЦІЯ ПО ГОЛОСУ, ДИСКРЕТИЗАЦІЯ, АУДІОЗАПИС, СЕГМЕНТАЦИЯ, ОБРОБКА ЗВУКОВОГО СИГНАЛА, ВІКОННА ФУНКЦІЯ, ПЕРЕТВОРЕННЯ ФУР'Е.

Метою роботи є розробка системи ідентифікації людини по голосу. В роботі було досліджено існуючі методи розв'язання задачі ідентифікації особистості по голосу, способи оцінки їх якості, а також існуючі проблеми та обмеження. Був реалізований один з кращих алгоритмів ідентифікації особистості по голосу. В IDE Matlab було створено програмний продукт, який має можливість працювати wav-файлами. Проведено експериментальне дослідження розробленого алгоритму.

ABSTRACT

Bachelor’s thesis contains 73 pages, 41 figures, 2 appendixes, 13 sources according to the list of links.

NEURAL NETWORKS, IDENTIFICATION AND VERIFICATION BY VOICE, DISCRETIZATION, AUDIO RECORDING, SEGMENTATION, PROCESSING OF THE AUDIO SIGNAL, WINDOW FUNCTION, FOURIER TRANSFORM.

The purpose of the work is to develop a system of identification of a person by voice. One of the best voice identification algorithms has been implemented. IDE Matlab has created a software product that can handle wav files. An experimental study of the developed algorithm was conducted.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ПЗ – програмне забезпечення;

ЦП – цифрові пристрої;

IDE – Integrated Development Environment (інтегроване середовище розробки);

ІТ – Інформаційні технології ;

ОС – Операційна система;

API – Application programming interface, прикладний програмний код;

Мел - одиниця виміру висоти сприйманого звуку;

Спектр - розподіл деякої фізичної величини за іншою величиною;

Кепстр - спектр спектра сигнала;

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

GMM – Gauss Mixture Model модель Гауссовых суміщів;

HMM – Hidden Markov Model скриті Марковскі моделі;

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.

ВСТУП

Мова - важливий елемент людської діяльності, що дозволяє людині пізнавати навколишній світ, передавати свої знання і досвід іншим людям. Усна складова мови проявляється у вигляді висловлювань в звуковій формі, які можливі завдяки голосовому апарату людини.

Кожна людина має індивідуальні голосові характеристики, які визначаються особливостями будови його голосових органів. У процесі спілкування люди здатні на підсвідомому рівні розрізняти голоси інших людей, однак для обчислювальної техніки ця задача не є тривіальною.

Завдання розпізнавання особистості по голосу було поставлено більше 40 років тому, але дослідження в цій області тривають і в даний час. За останні роки спостерігається значне підвищення якості розпізнавання мовної інформації, однак основна проблема автоматичного розпізнавання диктора в будь-якому середовищі все ще далека від вирішення. Саме тому актуальні як дослідження вже існуючих алгоритмів, так і пошук нових рішень в даній області.

Завдання розпізнавання особистості по голосу зводиться до того, щоб виділити, класифікувати і відповідним чином відреагувати на людську мову з вхідного звукового потоку. При цьому зазвичай виділяють дві підзадачі: ідентифікація і верифікація.

Ідентифікація - процес визначення особистості за зразком голосу шляхом порівняння даного зразка з шаблонами, збереженими в базі. Результатом ідентифікації зазвичай є ім'я людини, зареєстрованого в системі, шаблон якого найбільш ймовірно відповідає вхідному зразком голосу.

Верифікація - процес, при якому за допомогою порівняння представленого зразка з збереженим в базі шаблоном перевіряється ідентичність. З визначення випливає, що при верифікації разом зі зразком голосу передається ідентифікатор користувача, зареєстрованого в системі.

Результатом є підтвердження особи або негативна відповідь системи. Крім того, системи розпізнавання можуть бути розділені на тексто-залежні та текстне-залежні. При тексто-залежному розпізнаванні можуть використовуватися як фіксовані фрази, так і фрази, згенеровані системою і запропоновані користувачеві. Тексто-незалежні системи призначені обробляти будь-які фрази. У даній роботі розглядається задача автоматичної ідентифікації диктора і реалізується алгоритм, вирішуючий завдання тексто-незалежної ідентифікації.

Існують наступні проблеми і обмеження завдання розпізнавання особистості по голосу, які слід враховувати при побудові рішення:

– емоційний стан диктора;
– складна акустична обстановка (шуми і перешкоди);
– різні канали зв'язку при навчанні і розпізнаванні;
– природні зміни голосу.

Розпізнавання особистості по голосу знаходить застосування в багатьох сферах:

– криміналістика і судова експертиза;
– безпека;
– банківські технології;
– електронна комерція;
– телематика.

Мова являє собою складний сигнал, що утворюється в результаті перетворень, що відбуваються на кількох рівнях: семантичному, мовному, артикуляційному (рівні голосового апарату людини) і акустичному (рівні фізичних властивостей звуку). Відмінності в цих перетвореннях тягнуть за собою відмінності у властивостях мовного сигналу. При вирішенні задачі розпізнавання диктора за голосом всі ці відмінності можуть бути використані для того щоб виділити індивідуальні характеристики голосу кожної людини.

1 ВИКОРИСТАННЯ СИСТЕМ ІДЕНТИФІКАЦІЇ У РІЗНИХ ГАЛУЗЯХ

Розпізнавання можна розділити на два основних напрямки ідентифікацію та верифікацію. У першому випадку система повинна самостійно встановити особу користувача по голосу; у другому випадку система повинна підтвердити або спростувати ідентифікатор, який пред'являє користувач. Визначення досліджуваного диктора складається в попарному порівнянні голосових моделей, які враховують індивідуальні особливості мови кожного диктора. Таким чином, необхідно для початку зібрати базу даних. А за результатами цього порівняння може бути сформований список фонограм, які є з певною ймовірністю мовою, цільового користувача.

1.1 Біометричні системи контролю доступу

1.1.1 Статичні методи ідентифікації людини

Статичні методи біометричної аутентифікації ґрунтуються на фізіологічній (статичній) характеристиці людини, тобто унікальній характеристиці, даної йому від народження і невід'ємною від нього.

1. По відбитку пальця. В основі цього методу лежить унікальність для кожної людини малюнка папілярних візерунків на пальцях. Відбиток, отриманий за допомогою спеціального сканера, перетворюєтья в цифровий код (згортку), і порівнюється з раніше введеним еталоном. Дана технологія є найпоширенішою в порівнянні з іншими методами біометричної аутентифікації.

2. За формою долоні. Даний метод, побудований на геометрії кисті руки. За допомогою спеціального пристрою, що складається з камери і декількох діодів (включаючись по черзі, вони дають різні проекції долоні), будується тривимірний образ руки за яким формується згортка і розпізнається людина;

3. За розташуванням вен на лицьовій стороні долоні. З допомогу інфрачервоної камери зчитується малюнок вен на лицьовій стороні долоні або кисті руки, отримана картинка обробляється і по схемі розташування вен формується цифрова згортка.

4. По сітківці ока. Точніше цей спосіб ідентифікації по малюнку кровоносних судин очного дна. Для того, щоб цей малюнок став видний людині потрібно подивитися на віддалену світлову крапку, і таким чином підсвічене очне дно сканується спеціальною камерою.

5. За райдужною оболонкою ока. Малюнок райдужної оболонки ока також є унікальною характеристикою людини, причому для її сканування досить портативної камери зі спеціалізованим програмним забезпеченням, що дозволяє захоплювати зображення частини особи, з якого виділяється зображення ока, з якого в свою чергу виділяється малюнок райдужної оболонки, за яким будується цифровий код для ідентифікації людини.

6. За формою обличчя. В даному методі ідентифікації будується тривимірний образ обличчя людини. На обличчі виділяються контури брів, очей, носа, губ, обчислюється відстань між ними і будується не просто образ, а ще безліч його варіантів на випадки повороту особи, нахилу, зміни виразу обличчя. Кількість образів варіюється в залежності від цілей використання даного.

7. За термограмою особи. В основі даного способу аутентифікації лежить унікальність розподілу на обличчі артерій, що постачають кров'ю шкіру, які виділяють тепло. Для отримання термограми, використовуються спеціальні камери інфрачервоного діапазону. На відміну від попереднього - цей метод дозволяє розрізняти близнят.

8. По ДНК. Переваги даного способу очевидні, проте використовувані в даний час методи отримання і обробки ДНК - працюють настільки довго, що такі системи використовуються тільки для спеціалізованих експертиз.

1.1.2 Динамічні методи ідентифікації людини

Динамічні методи біометричної аутентифікації ґрунтуються на поведінковій (динамічній) характеристиці людини, тобто побудовані на особливостях, характерних для підсвідомих рухів в процесі відтворення якоїсь дії:.

По рукописному почерку. Для цього виду ідентифікації людини використовується її розпис (іноді написання кодового слова). Цифровий код ідентифікації формується, в залежності від необхідного ступеня захисту і наявності обладнання (графічний планшет), двох типів:

a) по самому розпису, тобто для ідентифікації використовується просто ступінь збігу двох картинок;

b) по розпису і динамічним характеристикам написання, тобто для ідентифікації будується згортка, в яку входить інформація по безпосередньо підпису, часовим характеристикам нанесення розпису і статистичним характеристикам динаміки натиску на поверхню.

1. За клавіатурним почерком. Метод в цілому аналогічний вищеописаному, але замість розпису використовується якесь кодове слово (коли для цього використовується особистий пароль користувача, таку аутентифікацію називають двухфакторною) і не потрібно ніякого спеціального обладнання, крім стандартної клавіатури. Основною харатеристико по якій будується згортка для ідентифікації - динаміка набору кодового слова.

2. За голосом. Одна з найстаріших технологій, в даний час її розвиток прискорився, оскільки передбачається її широке використання в побудові «інтелектуальних будівель». Існує досить багато способів побудови коду ідентифікації по голосу, як правило це різні поєднання частотних і статистичних характеристик голосу.

1.2 Криміналістика і судова експертиза

Необхідність у визначенні того, чи належить голос підозрюваного записам мови в телефонних каналах виникає при аналізі телефонних дзвінків в разі помилкових повідомлень, нарко-діяльності, вимагання або сексуальних домагань. При цьому, на відміну від верифікації, предметом аналізу можуть бути лише записи мовних сигналів, що підлягають порівнянню, або знову виконані записи мови підозрюваного. В останньому випадку підозрюваний зазвичай не зацікавлений в його ідентифікації, та його мова може бути свідомо спотворена. До того ж, умови такого запису, виконаного, наприклад, в тихій кімнаті для допитів, можуть сильно відрізнятися від умов, в яких мовні сигнали, що підлягають порівнянню, були згенеровані і передані по каналу зв'язку, а записані фрази можуть бути різними. У криміналістиці підозрюваного можуть попросити прочитати текст, що відповідає транскрипції раніше записанної мови, але, як показав досвід, цей прийом не дуже ефективний.

Представники органів криміналістики зацікавлені в тому, щоб отримати однозначну відповідь від приналежності біометричних параметрів. Наприклад, дослідницька група Федерального бюро розслідувань США стверджує, що стосовно відбитків пальців допустимо тільки однозначне рішення - "збігається або не збігається", і не повинні використовуватися ніякі оцінки типу "можливо, ймовірно, може бути". Але навіть і в відношенні відбитків пальців така позиція мало обгрунтована. Вважається, що ймовірність помилкового збігу відбитків пальців порядку 10^{-6}, хоча на цей рахунок відсутні статистично достовірні дослідження. Що ж стосується автоматичного розпізнавання відбитків пальців, то ймовірність помилкового впізнання набагато вище - близько 2% для 4 пальців (Fingerprint Verification Competition, 2004). Не випадково при верифікації особистості в важливих випадках потрібні відбитки всіх десяти пальців. Рішення про ідентичність тільки по одному відбитку взагалі має високий ризик помилки.

Загальна думка полягає в тому, що ідентифікація по голосу відрізняється від відбитків пальців і генетиці, де варіації дуже малі, і немає абсолютно надійного методу для визначення того, чи належать мовні сигнали одній і тій же людині. У криміналістиці розпізнавання диктора може мати тільки імовірнісний характер, тобто із зазначенням правдоподібності того, що два мовних сигнали належать одній і тій же людині. В умовах телефонного каналу проблематично навіть розпізнавання статі або віку. В силу малої вибірки мовних сигналів довірчий інтервал оцінки правдоподібності приналежності двох записів мови одного й того ж диктора настільки великий, що однозначне рішення неможливо.

Спеціальний тест з парним порівнянням мовних сигналів тривалістю 5 с показав 53% правильного розпізнавання фонетистами, яким було дозволено користуватися будь-якими технічними засобами, і 46% - не фонетистами. В інших тестах діапазон становив 38-76%. Ці оцінки наочно показують ступінь невизначеності прийняття рішень.

Відповідно до цією думки, у судовій практиці США, Великобританії і Франції експертний висновок про ідентичність записів мови не приймається в якості юридичного доказа. В практиці кримінального розслідування при візуальній ідентифікації особистості потрібно порівняння з деякою кількістю схожих осіб, тоді як рішення про ідентичність голосів, засноване тільки на порівнянні перехоплених записів мовного сигналу і голоса підозрюваного, без порівняння з голосами безлічі інших дикторів, містить високий ризик помилки. Цей ризик може не зупинити від прийняття рішення в деяких випадках, як це було описано в книзі А. І. Солженіцина "У колі першому", але обов'язок наукового співтовариства полягає в тому, щоб попередити про відсутність підстав для категоричних рішень.

1.3 Радіо-розвідка, контррозвідка, антитерористичний моніторинг

У цих областях ідентифікація диктора не носить юридичного характеру. Тому рішення про ступінь близькості голосів приймається на основі імовірнісних кількісних оцінок, і саме по собі не є двозначним. У такій постановці державних організацій завдання ідентифікації голосів має певну специфіку, пов'язану спотвореннями і перешкодами в каналах зв'язку. Оскільки фонетичний зміст порівнюваних мовних сигналів зазвичай відрізняється, то державні організації зацікавлені в дослідженнях розпізнавання диктора незалежно від контексту.

Звичайно, і в цьому випадку розробка методів ідентифікації голосів містить негативний аспект, пов'язаний зі зловживаннями у вигляді втручання в приватне життя або нагляд за опозицією правлячого режиму.

2 МЕТОДИ І ЗАСОБИ ІДЕНТИФІКАЦІЇ ДИКТОРА ЗА ГОЛОСОМ

В існуючих системах процес ідентифікації складається з двох основних етапів - це виявлення ознак у вихідних даних і створення на їх основі моделей, за якими визначається ступінь схожості претендента на зареєстрованого користувача. Нижче будуть розглянуті методи і засоби, які застосовуються системах голосової ідентифікації.

2.1 Етапи роботи с сигналом

У задачі голосової ідентифікації застосовують різні математичні, алгоритмічні, технічні методи, починаючи з етапу записи голосу і закінчуючи етапом класифікації. практично кожна система ідентифікації містить чотири основні етапи: отримання сигналу, попередня обробка сигналу, знаходження ознак і класифікація ознак. Розглянемо ці етапи для нашої задачі.

Етап отримання сигналу Метод отримання або записи голосового сигналу, в більшості випадків, є записом сигналу з допомогою мікрофона і вираження сигналу в цифровому вигляді за допомогою аналого-цифрового перетворювача. Як аналого-цифровий перетворювач, зазвичай використовують звукову карту персонального комп'ютера або цифровий диктофон. Цифрові дані кодуються сигнал PCM і поміщаються в формат файлу-контейнера (Waveform Audio File Format) для зберігання запису відцифрованого аудіопотока. Параметри звукового запису зазвичай такі: бітність відліку - 16 біт, частота дискретизації - 22500 Гц. Так як сучасні цифрові мобільні пристрої зазвичай мають вбудований мікрофон і продуктивні апаратні засоби, то створення система аутентифікації по голосу з залученням більш витратних за обчисленнями методів цілком вирішуване завдання для мобільних платформ. Проте, забезпечити мінімальні обчислювальні витрати при збереженні точності, завадостійкості до різних

видів перешкод і достатню надійність при поширених апаратних засобах все ж необхідно.

Етап попередньої обробки. Процес дискретизації по часу має на увазі отримання значень сигналу після перетворення, результатом якого є сигнал з певним часовим кроком, який називається кроком дискретизації.

Чим менше крок дискретизації, тим частіше беруться значення амплітуди. За визначенням, частота дискретизації - кількість замірів амплітуди в одну секунду.

Рисунок 2.1 – Процес дискретизації звукового сигналу

Квантування по амплітуді - це процес отримання наближених значень з сигналу із заданою точністю. Залежно від вибраної кількості і розташування рівнів квантування залежить точність округлення: чим ближче рівні квантування один до одного і чим їх більше, тим на меншу величину доводиться округляти отримані значення амплітуди, тим самим отримуючи меншу похибку.

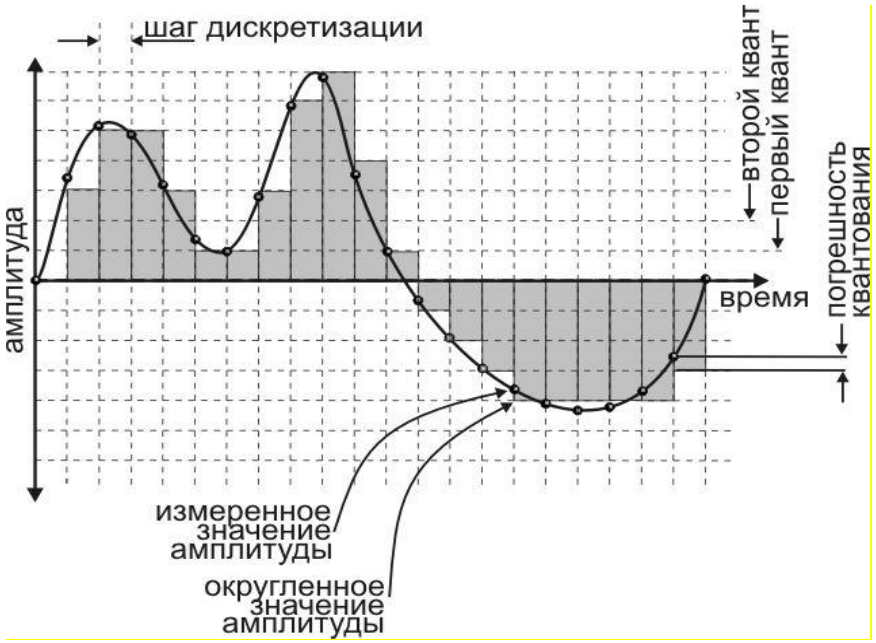

Рисунок 2.2 – Процес квантування звукового сигналу

Дослідним шляхом виявлено, що досить точну відповідність цифрового сигналу аналоговому отримується, якщо частота дискретизації буде вдвічі вище максимальної звукової частоти, тобто складе не менше 40 кГц. (Теорема Найквіста, або також як теорема Котельникова, яка стверджує, що на кожен цикл найвищої частоти має припадати як мінімум два виміру амплітуди). На практиці, найчастіше, значення частоти дискретизації становить 44,1 кГц або 48 кГц. Щоб отримати більш якісний звук, потрібно збільшувати частоту дискретизації. Це означає, що щоб зберегти оригінальну якість аудіо матеріалу необхідно вибирати високі значення параметрів відцифровки.

На даному етапі роботи, найбільш підходящий аудіоформат для перерахованих вище умов і аналізу звукового сигналу є WAVE. Даний аудіоформат зберігає не стислі звукові дані, на відміну від повсюдно поширеного MP3, при стисненні яких втрачається частина цих же звукової інформації.

Етап отримання ознак. отримання ознак зазвичай відбувається за допомогою Фур'є-перетворення, вейвлет-перетворень, лінійного

передбачення та інших. Коефіцієнти перетворень виступають в якості ознак.

В даний час точно не визначені голосові ознаки, за якими можна однозначно ідентифікувати особу людини. Вибір ознак впливає також на надійність ідентифікації. Існують, методи які описують інтегральні характеристики людського голосу і служать для вилучення тонів, динаміки мови. Такими методами є перетворення Фур'є (амплітудно-частотний розподіл), кепстральних перетвореннень (Амплітудно-часовий розподіл), перетворення лінійного передбачення (амплітудно-частотний розподіл). існують також і методи виділення фонем. Використання вейвлет-коеффіцієнтов не забезпечує значної переваги, до того ж облік частотної і тимчасової складових тягне додаткові обчислювальні витрати. Для ефективного використання частотної і часової складових голосового сигналу в завданні ідентифікації особистості по голосу необхідно проводити додаткові дослідження. Переваги тих чи інших голосових ознак виявляються в конкретних випадках при певних умовах і на певній мовній базі даних. Для різних методів класифікації ефективніше можуть виявитися набори коефіцієнтів різних перетворень.

З перерахованих вище голосових ознак, в різних комбінаціях, формується вектор ознак у вигляді послідовності чисел. Формування векторів ознак для одного або декількох дикторів входить в етап калібрування або навчання системи ідентифікації диктора. На виході етапу калібрування будується загальна модель одного або декількох дикторів, яка співвідноситься з вихідними мовними даними. На етапі тестування (перевірки) вектор ознак витягується з звукової хвилі і порівнюється з побудованою моделлю. Кожен вектор ознак повинен формуватися оптимально, враховуючи обчислювальні витрати, з одного боку, і інформативність ознак з іншого. Так, при високій інформативності вектора ознак, а саме, великій кількості ознак, обчислювальні витрати зростають, і навпаки, при малій кількості ознак обчислювальні витрати зменшуються. Однак при збільшенні кількості ознак у векторі інформативність не завжди

збільшується. деякі оцінки інформативності ознак можна отримати на етапі тестування з допомогою помилки класифікації (ідентифікації), тобто мінімум помилки буде відповідати більшої інформативності. Тому часто проводять окремі дослідження щодо формування оптимального вектора ознак для кожного диктора, тобто досліджують залежність поєднань різних наборів ознак з вихідних даних (коефіцієнтів будь-яких перетворень) від помилки класифікації. Сполучення різних наборів ознак може досліджуватися як простим перебором, так і з використанням методів оптимізації (наприклад, генетичних алгоритмів). З іншого боку, самі ознаки змінюються в деякому діапазоні, що викликане або впливом зовнішніх шуму і перешкод, або впливом внутрішніх спотворень голосового апарату людини. Тому відмінність тестових і калібрувальних умов класифікації може привести до формування різних векторів ознак для одного і того ж диктора, а отже – до зниження вірогідності методів класифікації.

Етап класифікації ознак. В цей етап входить застосування математичних методів класифікації, за допомогою яких здійснюється прийняття рішення, а також розрахунок помилок класифікації.

2.2 Поняття мелу

Висота звуку сприймається людським слухом не пов'язана лінійно з його частотою, навпаки, її величина пов'язана ще з рівнем гучності і тембром. Тому для її аналізу була створена кількісна оцінка звуку - Мел, одиниця виміру висоти сприйманого звуку заснована на психофізичних параметрах сприйняття. Так як на сьогоднішній день поки не представляється можливим виміряти АЧХ людського слуху у вигляді мозкової активності безпосередньо, при визначенні залежності Мілове від частоти була застосована статистична обробка великої кількості даних про суб'єктивне сприйняття висоти звуку.

Мел зручно застосовувати в цілях аналізу мови людини, так як його використання «наближає» алгоритми обробки даних до людських параметрів

сприйняття, що добре позначається на якості розпізнавання. На рисунку 2.3 зображений графік залежності мел-шкали від частоти коливань звукового сигналу.

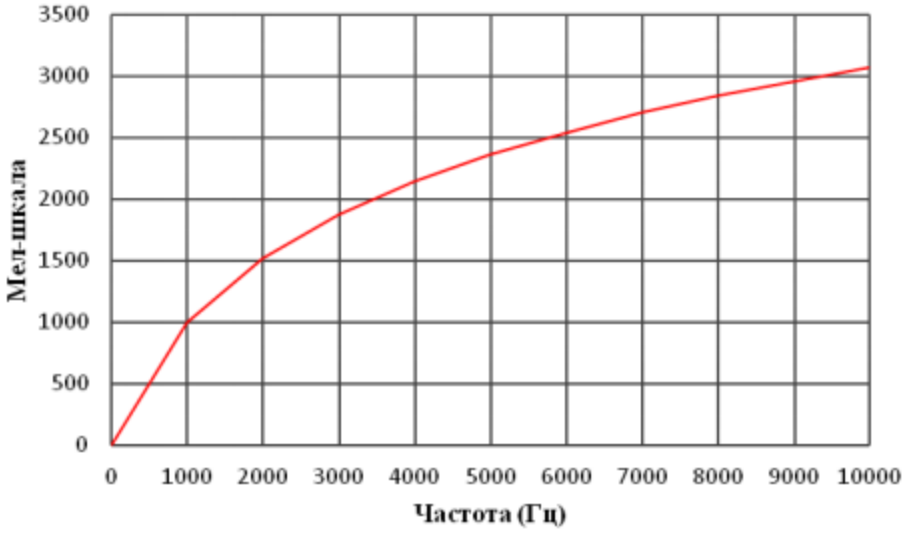

Рисунок 2.3 – Графік залежності висоти звуку в Гамелах від частоти коливань

Залежність висоти сприйманого звукового сигналу від його частоти описується формулою:

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) = 1127 \log_{e}\left(1 + \frac{f}{700}\right), \quad (2.1)$$

де m – висота звука в мелах; f – частота в герцах.

2.3 Спектр

У загальному сенсі спектром називають розподіл деякої фізичної величини за іншою величиною. Найбільш часто під спектром мається на увазі спектр електромагнітного випромінювання у вигляді розподілу енергії випромінювання за частотами. Але спектральний аналіз може застосовуватися до багатьох сигналів, в тому числі і мовних. Для обчислення спектра сигналу до нього застосовують перетворення Фур'є формула якого

наведена нижче:

$$f(w) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-ixw} dx \tag{2.2}$$

При обчисленні перетворення Фур'є на комп'ютерах виникають складності, так як при обчисленні звичайним способом потрібно підсумовувати нескінченний ряд чисел. Тому в реальних обчисленнях ЕОМ використовується дискретне і швидке, яке є оптимізованою версією дискретного. Його обчислюють за формулою:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} kn}, \; k = 0, \ldots, N-1 \tag{2.3}$$

де N – це розмірність дискретного відрізку сигналу; x_n – амплітуда n-го сигналу; X_k – N амплітуд синусоїдальних сигналів, які складають основний сигнал.

Спектральне перетворення сигналу полегшує розуміння і аналіз природи звукових сигналів.

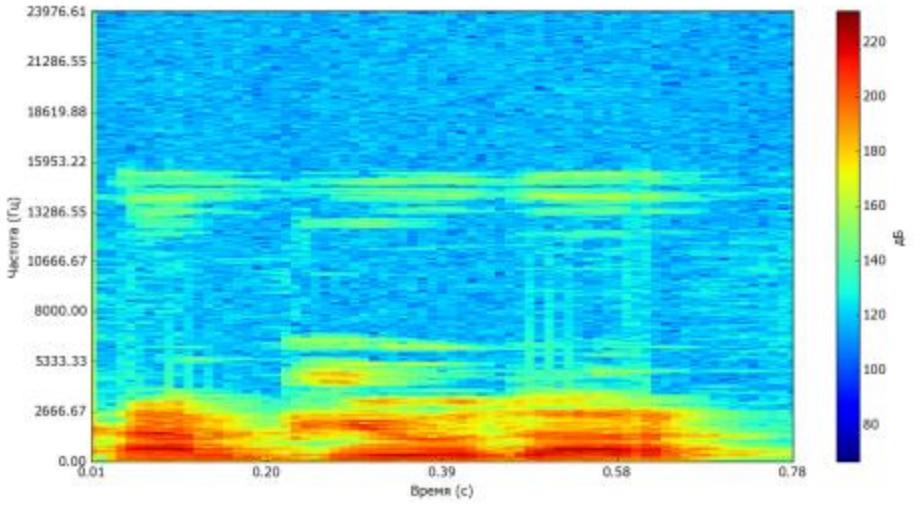

Рисунок 2.4 – Спектрограма слова «комп’ютер»

2.4 Кепстр

Найчастіше при аналізі отриманих даних не достатньо зробити висновок про їх інформативність тільки по спектру сигналу. Тоді в таких випадках застосовується кепстр, або іншими словами спектр спектру вихідного сигналу. Він використовується коли на спектрограмі не помітні потайливі, але справді існуючі періодичності в сигналі. Суть обчислення кепстра в тому, щоб представити вже наявний спектр не як розподіл деякої величини вихідних даних, а я самостійний сигнал. Завдяки цьому значна спектральна інформація може надаватися більш компактно, що полегшує її аналіз. У загальному вигляді кепстр обчислюється за формулою:

$$C_s(q) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \ln|S(w)|^2 e^{iwq} dw \tag{2.4}$$

2.5 Виділення характерних ознак

При постановці завдання ідентифікації особистості по її біометричним даними, одним з головних етапів є виділення цінної інформації з вхідних сигналів. Цей процес називається виділенням значущих ознак. Найчастіше для подібних задач біометричної ідентифікації потрібні різні методи. наприклад, при розпізнаванні мови потрібно якомога краще позбутися ознак, які характеризують окрему особистість, щоб полегшити роботу алгоритмам вилучення фонем. У розпізнаванні диктора навпаки - важливо підкреслити ознаки, що відповідають за індивідуальність висловлювань. Для розпізнання особистості по голосу використовуються два основних алгоритми: це MFCC і LPCC. Розглянемо нижче дані алгоритми.

2.5.1 Мел-частотні кепстральні коефіцієнти

При виконанні завдання розпізнавання мови або диктора після первинної фільтрації сигналу від шумів потрібно виділити з нього акустичні параметри і ознаки. Для цього застосовуються різні методи. Одним з таких методів є обчислення мел-частотних кепстральних коефіцієнтів (від англ. Mel-Frequency Cepstrum Coeffcents, MFCC). Його суть полягає в тому, щоб використовуючи шкалу переведення частоти сигналу в його висоту в Гаммелах обчислити багатовимірні вектори ознак, з якими надалі будуть працювати алгоритми класифікації. Нижче наведено порядок обчислення мел-частотних кепстральних коефіцієнтів.

Спочатку вхідний сигнал розбивається на кадри (фрейми) таким чином, щоб вони перекривали наступні і попередні за ними. Довжина фреймів безпосередньо впливає на роботу алгоритму: при збільшенні довжини відрізків підвищується точність, але падає швидкість роботи алгоритму.

В основному приймаються значення в діапазоні від 20 до 40 мілісекунд. Потім для кожного фрейма обчислюється його спектр за допомогою дискретного перетворення Фур'є.

Отримані спектральні коефіцієнти фреймів накладаються на мел-частотні вікна. Дані вікна зосереджуються ближче до низьких частот, тому що це найближче до механізму сприйняття висоти звуку: чим нижче частота, тим менше відрізняються сусідні частоти. Це видно на рисунку 2.5.

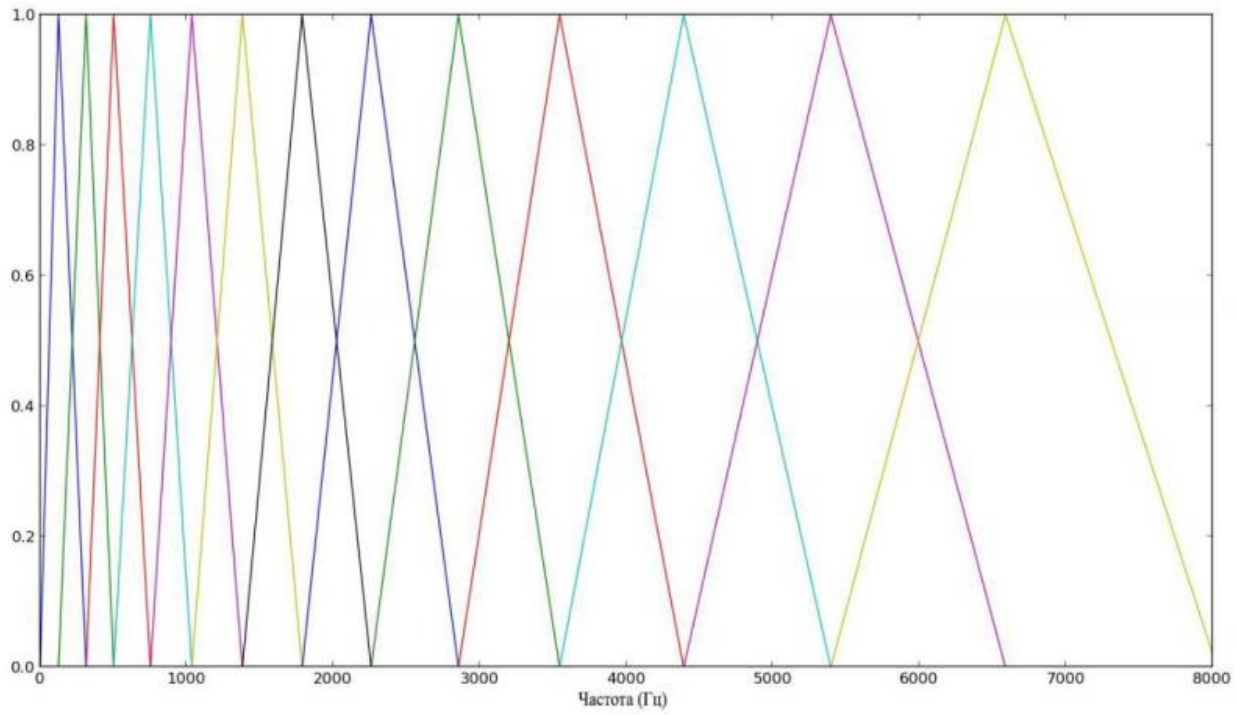

Рисунок 2.5 – Приклад накладання вікон на мел-шкалу

Проводиться обчислення енергії кожного кадру за формулою:

$$S(m) = \ln(\sum_{k=0}^{N-1} |X_a[k]|^2 H_m[k]),\ \ 0 \le m \le M \tag{2.5}$$

В кінці застосовується дискретне косинусне перетворення, яке дає на виході багатовимірний вектор ознак сигналу. Вони і є мелчастотними кепстральними коефіцієнтами.

2.5.2 Коефіцієнти лінійного передбачення

Алгоритм кепстральних коефіцієнтів лінійного передбачення (від англ. Linear Predictive Cepstral Coefficients, LPCC) починається з обчислення коефіцієнтів $\{a_q\}_{q=1}^{p}$ авторегресивної моделі для кожного фрейму. Після знаходження всіх параметрів моделі обчислюються кепстральні LPCC-коефіцієнти по рекурсивній функції, яка виглядає наступним чином:

$$c_n = \begin{cases} \ln(G), & q = 0 \\ a_q + \sum_{k=1}^{q-1} \frac{k-q}{q} a_k c_{q-k}, & 1 \le q \le p \\ \sum_{k=1}^{p} \frac{k-q}{q} a_k c_{q-k}, & p < q < Q \end{cases}, \quad (2.6)$$

де Q - це кількість коефіцієнтів LPCC; c_n - багатовимірний вектор коефіцієнтів LPCC.

Метод LPCC схожий на MFCC багато в чому, але головна відмінність в тому, що він використовує лінійну шкалу переведення частоти звуку в його висоту сприйняту мозком. Цей спосіб добре працює в області низьких частот, так як в цій зоні залежність висоти звуку від його частоти практично лінійна. Дана особливість дозволяє досягти схожих результатів при добуванні ознак в області низьких частот, однак приблизно після кордону в 1000 Гц висота звуку сприймається людиною не лінійно, і в зв'язку з цим погіршується якість роботи вищеописаного алгоритму.

2.6 Методи і засоби вирішення задачі классифікації

В даному розділі розглядаються основні існуючі рішення задачі автоматичної ідентифікації диктора за голосом. Незважаючи на те, що методи багато в чому відрізняються, в цілому можна виділити наступні основні етапи, характерні для кожного з розглянутих методів:

– отримання ознак із вхідного мовного сигналу;

– побудова моделі (шаблону) диктора на основі отриманих на попередньому етапі векторів ознак.

Процес визначення диктора, зареєстрованого в системі, за вхідним мовним сигналом у всіх розглянутих методах полягає в пошуку найбільш відповідної моделі на основі будь-яких критеріїв.

2.6.1 Dynamic Time Warping

Dynamic Time Warping (DTW) – метод динамічного програмування, що дозволяє знайти близькість між двома послідовностями вимірювань за деякий проміжок часу. В загальному випадку ці послідовності можуть бути різної довжини, і вимірювання можуть проводитися з різною швидкістю.

В якості моделі, що зберігається в даному методі виступає послідовність векторів ознак вхідного мовного сигналу з навчальної вибірки $Q = \{q_1, \dots, q_n\}$. Нехай $C = \{c_1, \dots, c_m\}$ – послідовність векторів ознак вхідного мовного сигналу з тестової вибірки. Також вводяться поняття матриці вирівнювання двох послідовностей $M_{m \times n}$, в позиції (i, j) якої міститься значення вирівнювання між елементами c_i та q_j послідовностей C та Q відповідно, та набору індексів суміжних елементів цієї матриці $W = \{w_1, \dots, w_T\}$, що визначає відповідність між елементами послідовностей, що порівнюються. При цьому елементи набору W повинні відповідати наступним вимогам:

1. $w_1 = (1, 1)$, $w_T = (m, n)$.

2. Якщо $w_{t-1} = (a', b')$, то $w_t = (a, b)$, де $a - a' \leq 1$, $b - b' \leq 1$.

Метою алгоритму DTW є знаходження такого набору W, що задовольняє умовам 1 та 2, при якому сумарне спотворення послідовності щодо послідовності було б мінімальним, тобто:

$$DTW(Q,C) = \min\left\{\frac{1}{T}\sqrt{\sum_{t=1}^{T} M(w_t)}\right\} \quad (2.7)$$

Значення цього виразу і визначатиме міру близькості послідовностей Q та C. Для знаходження значення $DTW(Q, C)$ застосовується метод динамічного програмування, де на кожному етапі обчислюється значення за формулою:

$$M(i,j) = d(i,j) + \min\{M(i-1,j-1), M(i-1,j), M(i,j-1)\} \quad (2.8)$$

При цьому $M(0, 0) = 0$, а всі інші елементи стовбця та строки з індексом 0 ініціалізуються значенням ∞. $d(i, j)$ визначає евклідову відстань між елементами c_i та q_j. Результатом роботи алгоритму є значення $DTW(Q, C) = M(m, n)$.

Для визначення диктора обчислюється значення $DTW(Q_i, C)$, для кожного збереженого шаблону. Значення i, при якому досягається мінімум, визначає номер диктора, зразок голосу якого найбільш близький до зразка вхідного мовного сигналу. Основною перевагою алгоритму DTW є простота реалізації. Проте, даний алгоритм непридатний для вирішення завдання тексто-незалежної ідентифікації диктора.

2.6.2 Hidden Markov Model

Hidden Markov Model (HMM) - статистична модель, яка може використовуватися для вирішення задачі класифікації прихованих параметрів на основі спостережуваних. HMM являє собою кінцевий автомат, в якому переходи між станами здійснюються з певною ймовірністю, і заданий стартовий стан, з якого починається процес. Через дискретні моменти часу може здійснюватися перехід в нові стани. При цьому кожному з прихованих

станів із заданою вірогідністю відповідає спостережуваний стан. Крім того, поточний стан автомата залежить тільки від кінцевого числа попередніх, а закон зміни станів не змінюється в часі. У даній роботі розглядається випадок, коли поточний стан залежить тільки від попереднього (модель першого порядку).

Формально HMM визначається наступними параметрами:

- множина прихованих станів Q = {q0, … , qN}, де q0 – початковий стан, qend – кінцевий стан;
- множина спостережень O = {o1, … , oM};
- вихідне розподілення станів π = { πi }, 1≤i ≤ N, який визначає ймовірність почати роботу у стані i;
- матриця ймовірностей переходів між прихованими станами AN×N:
- A(i, j) = aij = P(qi, qj), 1 ≤ i, j ≤ N;
- матриця ймовірностей спостережень BN×M: B(i, j) = bij = P(oj|qi),
- 1 ≤ i ≤ N, 1 ≤ j ≤ M

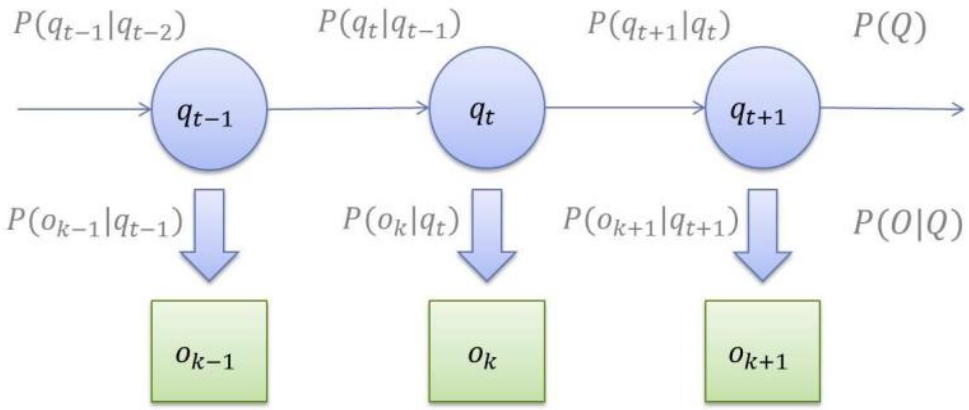

Рисунок 2.6 –. Hidden Markov Model

Виділяються три основні задачі, що вирішуються за допомогою HMM:

1. Обчислення ймовірності послідовності спостережень. Необхідно визначити ймовірність появи заданої послідовності спостережень O = {o_1, … ,o_R}. Для цього використовується наступний алгоритм:

$$\alpha_0(i) = \alpha_{0i}, 1 \le i \le N, \tag{2.9}$$

$$\alpha_j(r) = \sum_{i=1}^{N} \alpha_i(r-1) a_{ij} b_{jt}, 1 \le j \le N, 1 \le r \le R,$$, (2.10)

$\alpha_j(r)$ – ймовірність того, що на r-ому кроці модель виявиться у стані j. Тоді шукана ймовірність визначається за формулою:

$$P(O|A,B) = \sum_{i=1}^{N} \alpha_i(R)$$ (2.11)

2. Знаходження найбільш правдоподібної послідовності прихованих станів для спостережуваної послідовності**.** Потрібно знайти найбільш правдоподібну послідовність прихованих станів $Q = \{q_1, \dots, q_R\}$ для заданої послідовності спостережень $O = \{o_1, \dots, o_R\}$, при якій досягається max P (O|Q). Ця задача вирішується за допомогою алгоритму, подібному алгоритму з попереднього пункту з тієї лише різницею, що на кожному кроці запам'ятовується стан i, в якому $\alpha_i(r)$ приймає найбільше значення. В результаті обирається послідовність станів, для якої $\alpha_i(R)$ приймає найбільше значення. Цей алгоритм називається алгоритмом Вітербі.

3. Навчання параметрів моделі по заданій послідовності спостережень та множини прихованих станів. Необхідно обчислити для заданої моделі матриці A і B. Для вирішення даної задачі застосовується алгоритм Баума-Велша.

Для задачі розпізнавання диктора прихованими станами є вектори ознак мовного сигналу з навчальної вибірки, в якості спостережень - вектори ознак мовного сигналу з тестової вибірки. В якості збережуваної моделі тут виступають матриці A і B.

HMM досить прості в розумінні, мають досить високу точність розпізнавання, але, як і DTW, застосовуються в основному для завдань тексто-залежної ідентифікації диктора.

2.6.3 Vector Quantization

Задача векторного квантування з кодовими векторами $W = \{ w_1, w_2, \dots , w_n\}$ для послідовності вхідних векторів $C = \{c_1, c_2, \dots , c_m\}$ ставиться як задача мінімізації спотворення при заміщенні кожного вектора з відповідним кодовим вектором. Моделлю диктора в даному методі є множина кодових векторів, що отримується з вхідної послідовності векторів ознак мовного сигналу. Для побудови цієї множини вихідна послідовність векторів ознак розбивається на L кластерів, і в якості кодових векторів беруться їх центри.

Процес визначення диктора по вхідному мовному сигналу відбувається наступним чином. Для кожного тестового вектора c_i визначаються k найближчих кодових векторів. Нехай k_{ij} - кількість векторів, що належать диктору, серед знайдених найближчих. Тоді ймовірність того, що вектор c_i належить дикторові S_j, визначається формулою:

$$P(S_j|c_i) = \frac{k_{ij}}{k} , \qquad (2.12)$$

Таким чином, послідовність тестових векторів може бути класифікована за правилом:

$$S = \underset{1 \le j \le N}{\operatorname{argmax}} \prod_{i=1}^{L} P(S_j|c_i) , \qquad (2.13)$$

Для згладжування похибки вимірювання, зв'язаної з близькими до нуля ймовірностями, враховуючи постійності числа k, частіше використовують правило:

$$S = \underset{1 \le j \le N}{\operatorname{argmax}} \sum_{i=1}^{L} P(S_j|c_i) = \underset{1 \le j \le N}{\operatorname{argmax}} \sum_{i=1}^{L} k_{ij} , \qquad (2.14)$$

Метод векторного квантування простий в реалізації, може застосовуватися до задачі тексто-незалежної ідентифікації диктора, проте не завжди дає високу точність розпізнавання.

2.6.4 Support Vector Machine

Support vector machine (метод опорних векторів) – бінарний класифікатор, який будує в просторі ознак роздільну функцію, що задає гіперплоскість, виду:

$$f(x) = w \cdot x + b, \tag{2.15}$$

Нехай задана послідовність точок простору ознак $X = \{x_1, x_2, \dots, x_n\}$ з мітками $Y = \{y_1, y_2, \dots y_n\}$, $y_i \in \{-1, 1\}$, $1 \le i \le n$, що відповідають двом класам. У разі лінійної розділимості даних, умови для знаходження функції записуються у вигляді:

$$\begin{cases} w \cdot x_i + b \ge 1, & y_i = 1 \\ w \cdot x_i + b \le -1, & y_i = -1 \end{cases} \Leftrightarrow y_i(w \cdot x_i + b) - 1 \ge 0, 1 \le i \le n, \tag{2.16}$$

Для надійного розподілу класів необхідно щоб відстань між розділяючими гіперплощинами була якомога більшою. Відстань обчислюється як $\frac{2}{\|w\|}$, отже, задачу пошуку розділяючої гіперплощини можна звести до мінімізації $\|w\|^2$ при зазначених умовах. Ця задача може бути вирішена за допомогою методу множників Лагранжа.

У разі лінійно-нероздільних множин вводиться функція ядра. Основна ідея полягає в тому, щоб відобразити початковий простір у просторі більш високої розмірності, в якому множини вже можуть бути розділені лінійно. При цьому в силу того, що всюди в алгоритмі ознаки використовуються не окремо, а у вигляді скалярних похідних, немає необхідності будувати дане

перетворення в явному вигляді. Достатньо задати функцію ядра, що визначає скалярну похідну в новому просторі:

$$K(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j), \tag{2.17}$$

В якості моделі диктора, що зберігається, в методі опорних векторів виступають параметри розділяючої функції $f(x)$, а також параметри функції ядра. Параметри ядра зазвичай визначають шляхом перебору деякого множини значень та оцінкою методом крос-валідації.

Після того, як вирішальна функція $f(x)$ обчислена, приналежність вектора x' відповідному класу визначається знаком виразу $f(x)'$.

Для застосування методу опорних векторів до задачі багатокласового розпізнавання використовується стратегія «один проти інших». Для цього будуються q класифікаторів, кожен з яких навчається відрізняти один конкретний клас від всіх інших. При розпізнаванні об'єкт приписується до того класу, чий класифікатор видав найбільше значення розділяє функції $f(x)$. Метод опорних векторів дає високу точність класифікації, має теоретичне обгрунтування, дозволяє застосовувати різні підходи до класифікації згідно з вибором функції ядра. Серед недоліків слід відзначити проблему вибору ядра, а також повільне навчання у випадку задачі багатокласового розпізнавання.

2.6.5 Gaussian Mixture Model

Модель гауссових сумішей являє собою взважену суму M компонент та може бути описана виразом:

$$P(\bar{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\bar{x}), \tag{2.18}$$

де $\bar{x}$ – D-мерний вектор випадкових величин, p_i, $1 \le i \le M$ – ваги компонентів моделі, $b_i(\bar{x})$, $1 \le i \le M$ – функції щільності розподілення складових моделі:

$$b_i(\bar{x}) = \frac{1}{(2\pi)^{D/2} |\sum_i|^{1/2}} exp\left\{-\frac{1}{2}(\bar{x} - \bar{\mu}_i)^T \sum_i^{-1} (\bar{x} - \bar{\mu}_i)\right\}, \quad (2.19)$$

де $\bar{\mu}_l$ – вектор математичного очікування та $\sum_i$ – коваріаційна матриця. При цьому ваги суміші задовольняють умові:

$$\sum_{i=1}^{M} p_i = 1, \quad (2.20)$$

Повністю модель гауссової суміші визначається векторами математичного очікування, коваріаційними матрицями та вагами сумішей для кожного компонента моделі:

$$\lambda = \left\{p_i, \bar{\mu}_i, \sum_i\right\}, \; i = 1, \dots M, \quad (2.21)$$

При використанні даного методу кожен диктор представляється моделлю гауссових сумішей λ.

Для побудови моделі диктора необхідно оцінити її параметри, які найкращим чином відповідають розподілу векторів ознак навчального висловлювання. Найбільш популярним і широко використовуваним методом вирішення цього завдання є метод оцінки максимальної правдоподібності. Метою оцінки максимальної правдоподібності є знаходження параметрів моделі, які максимізують правдоподібність цієї моделі, при заданих навчальних даних.

Для послідовності навчальних векторів $X = \{\overline{x_1}, \dots, \overline{x_T}\}$ правдоподібність моделі гауссових сумішей може бути записано у вигляді:

$$P(X|\lambda) = \prod_{t=1}^{T} P(\overline{x_t}|\lambda), \quad (2.22)$$

Цей вираз уявляє собою нелінійну функцію від параметрів λ, і її безпосереднє обчислення неможливо, тому зазвичай для оцінки параметрів застосовується ЕМ-алгоритм.

Нехай – $S = \{S_1, S_2, \ldots, S_N\}$ - група дикторів, які представлені набором моделей гаусових сумішей. При ідентифікації диктора потрібно знайти модель, яка має найбільше значення апостеріорної ймовірності для заданого висловлювання:

$$S = \underset{1 \le k \le N}{\operatorname{argmax}} P(\lambda_k | X) = \underset{1 \le k \le N}{\operatorname{argmax}} \frac{P(X|\lambda_k)P(\lambda_k)}{P(X)} = \underset{1 \le k \le N}{\operatorname{argmax}} P(X|\lambda_k) \quad ,(2.23)$$

Використовуючи логарифм і незалежність між спостереженнями, система ідентифікації диктора в результаті обчислює:

$$S = \underset{1 \le k \le N}{\operatorname{argmax}} \sum_{t=1}^{T} \log P(\bar{x_t} | \lambda_k) \quad , \qquad (2.24)$$

Моделі гаусових сумішей є ефективним алгоритмом з високою точністю розпізнавання. Разом з тим виникає ряд проблем, пов'язаних з вибором числа компонентів моделі та ініціалізацією її початкових параметрів.

2.6.6 Нейромережі

Найбільш відомою і використовуваною моделлю нейронної мережі є багатошаровий перспетрон. Схема багатошарового перспетрона представлена на рисунку 2.5.

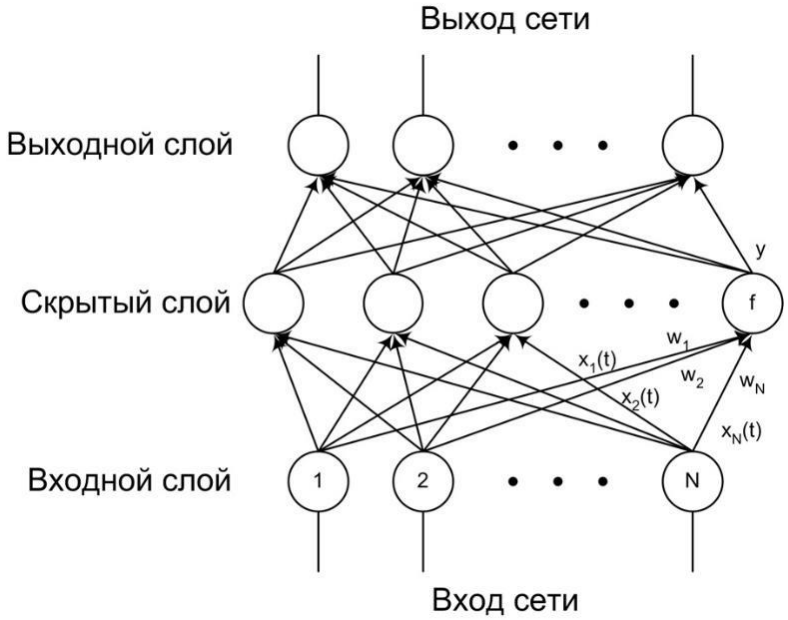

Рисунок 2.7– Трьохшаровий персептрон

Елементи багатошарового персептрона розділені на кілька шарів, всередині шару елементи можна вважати лінійно впорядкованими і не взаємодіючими між собою. Кожен нейрон мережі (крім нейронів вхідного шару - рецепторів) отримує вхідний сигнал від кожного нейрона попереднього шару і вихідний сигнал нейрона (крім останнього шару) надходить на вхід нейронів наступного шару. Таким чином, МП є моделлю зі зв'язками забезпечують поширення сигналу тільки вперед (без зворотних зв'язків) - від входу до виходу мережі. Елементи проміжних шарів називаються прихованими елементами, а шари - прихованими шарами. Самі нейрони найчастіше функціонують відповідно до моделі Маккаллок-Пітса (рисунок 2.8).

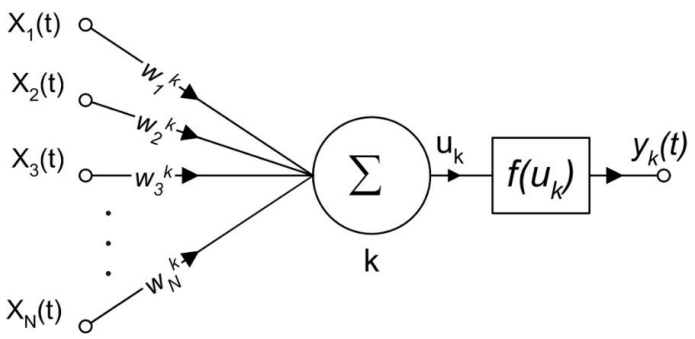

Рисунок 2.8 – Модель нейрона

Ця модель функціонує в такий спосіб: на вхід нейрона подається вхідний вектор $x(t) = \{x_1(t), x_2(t), \dots, x_N(t)\}$ який скалярно множиться на ваговий вектор $w_k(t) = \{w_{1k}, w_{2k}, \dots, x_{Nk}\}$ або, іншими словами, компоненти вектора $x_i(t)$ зважуються ваговими коефіцієнтами w_{ik} відповідно до формули

$$u_k(t) = \sum_{i=1}^{D} w_{ik} x_i(t), \tag{2.25}$$

Вихідний сигнал нейрона визначається як

$$y_k(t) = f(u_k(t)), \tag{2.26}$$

де $f(u_k(t))$ є функцією активації нейрона

На даний момент існує невелика кількість методів, що дозволяють вирішувати задачу тексто-незалежної ідентифікації диктора за голосом, причому кожен з наведених методів має свої переваги і недоліки.

В результаті аналізу методів, описаних в першому розділі, для подальшої роботи було обрано метод мел-частотних кепстральних коефіцієнтів для аналізу мовного сигналу і для розв'язання задачі ідентифікації диктора - метод нейронних мереж.

Головною перевагою методу кепстральних коефіцієнтів є простота реалізації при хорошій якості розпізнавання, яке не поступається іншим популярним алгоритмам. Також даний метод передбачає менший аналіз даних, що значно скорочує час навчання нейронної мережі.

Незаперечною перевагою ідентифікації диктора за допомогою нейронних мереж є здатність до навчання системи, що підвищує точність розпізнавання при більшій кількості циклів навчання.

3 МЕТОД КЕПСТРАЛЬНИХ КОЕФІЦІЄНТІВ РОЗПОДІЛЕННИХ ПО МЕЛ-ШКАЛІ

Основна ідея даного методу - отримати ознаки мовного сигналу, відкидаючи індивідуальні особливості вимови слів диктором.

Перевага кепстральних коефіцієнтів перед згаданими раніше методами - простота реалізації при досить високій якості розпізнавання мови.

3.1 Ділення на фрейми

Безперервний сигнал мови не задовольняє ні умові існування, ні умові визначеності на нескінченному часовому інтервалі, тому необхідно «розбити» цей сигнал на сегменти («фрейми») деякої довжини, спектри яких залишаються відносно незмінними протягом обраного періоду часу. Зазвичай за такий період приймається тривалість 5-100 мс. «Розбиття» вихідного сигналу на фрейми заданої тривалості робиться з половинною перекриттям для боротьби з спотвореннями, які можуть бути викликані розташованими поруч кадрами, або без перекриття, що значно економить обчислювальні ресурси. В межах отриманого кадру вхідний сигнал можна вважати стаціонарним. З точки зору динаміки мови, найшвидші зміни можуть відбуватися лише за кілька мілісекунд, тоді як деякі голосні звуки залишаються відносно стабільними протягом 100-200 мс. Найбільш часто використовуються фрагменти довжиною 10-25 мс.

3.2 Застосування перетворення Фур'є

Після розбиття сигналу на фрейми, до кожного відрізку застосовується вагова функція, а потім перетворення Фур'є. Прикладом ваговій функції може служити вікно Хеммінга:

$$w_n = 0.54 - 0.46 \cdot \cos\left(2\pi \cdot \frac{n}{N-1}\right), n = 0, \dots, N-1, \quad (3.1)$$

де N - довжина вікна.

Вагова функція необхідна для зменшення спотворень при перетворенні Фур'є (3.2), викликаних кінцівкою вибірки.

$$X_k = \sum_0^{N-1} \mathrm{x[n]} \cdot exp(-\frac{2\pi i}{N} kn). \quad (3.2)$$

Якщо підставити формулу 3.2 в 3.3, то отримаєм:

$$X_k = \sum_0^{N-1} w_n \mathrm{x[n]} \cdot exp(-\frac{2\pi i}{N} kn). \quad (3.3)$$

Значення індексів k відповідають частотам:

$$f_k \frac{fs}{N} k, \quad (3.4)$$

де F_s- частота дискретизації сигнала.

3.3 Складання банку трикутних фільтрів

Наступний крок при обчисленні кепстральних коефіцієнтів - накладення банку трикутних перекривающихся мел-фільтрів (рис. 3.1). Мел-фільтри дають можливість виділяти найбільш інформативні частоти для людського слуху. Фактично, на даному етапі обчислюється сума енергій в певному частотному діапазоні.

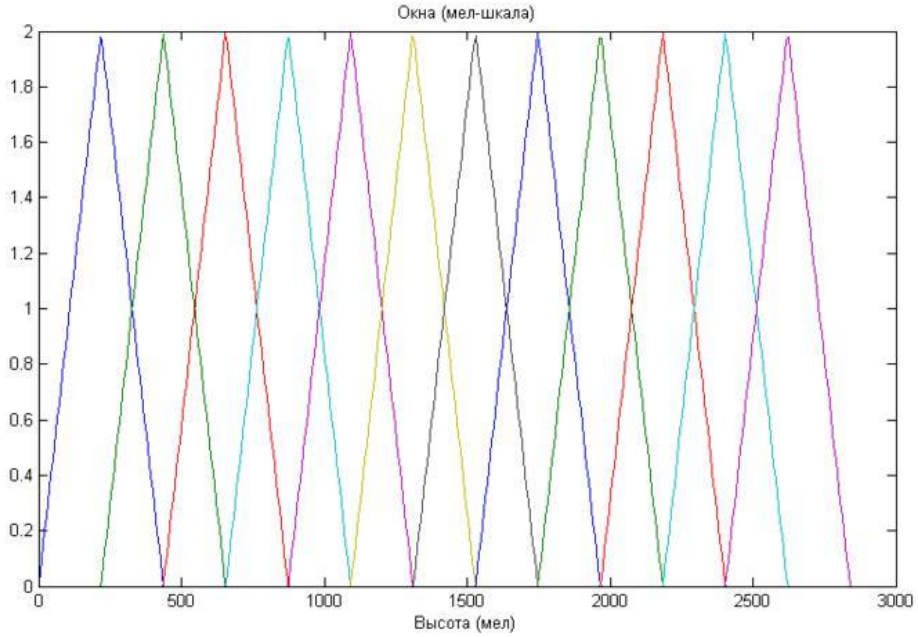

Рисунок 3.1 – Банк фільтрів на мел-шкалі

Для переходу від частоти Гц до висоти мел-звуку, використовуються формула:

$$m = 1125\ln\left(1 + \frac{fs}{700}\right). \quad (3.4)$$

При переході від мел шкали до частотної шкалою, фільтри будуть збиратися в області низьких частот, тим самим забезпечуючи більш високу роздільну здатність там, де воно необхідне для розпізнавання (рис. 3.2)

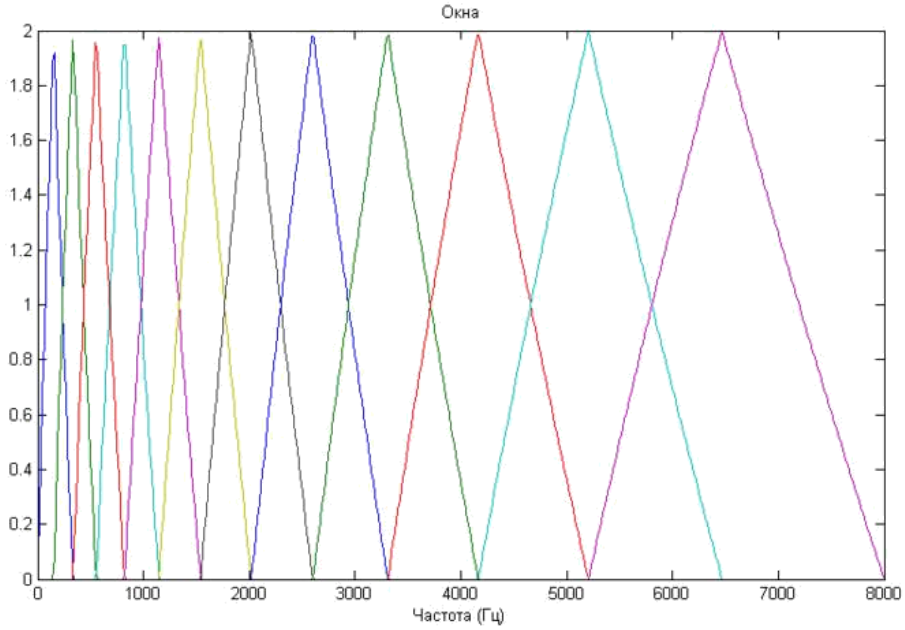

Рисунок 3.2 – Банк фільтрів в частотній шкалі

Розташування фільтрів визначається наступною формулою:

$$F(k) = f\left(M_{min} - \frac{M_{max} - M_{min}}{K+1}\right), k = 0,1,2,\dots,K-1, \quad (3.5)$$

де M_{max} - мел-вираження максимальної частоти M_{min} мел-вираження мінімальної частоти, K - кількість фільтрів

3.4 Обчислення вагових коефіцієнтів

Для обчислення вагових коефіцієнтів можна використовувати отримані частоти, або ж перейти до індексів семплів:

$$Fn(k) = \frac{N}{fs}F(k), \quad (3.6),$$

де fs - частота дискретизації, N – кількість циклів

Перехід до індексів семплів спрощує подальші обчислення, оскільки позбавляє від необхідності визначати частоту кожного семпли. Отримані індекси підставляємо в формулу обчислення вагових коефіцієнтів:

$$W(k,n) = \begin{cases} 0, n < Fn(k-1), \\ \frac{n-Fn(k-1)}{Fn(k)-Fn(k-1)}, Fn(k-1) \le n \le Fn(k), \\ \frac{n-Fn(k+1)}{Fn(k)-Fn(k+1)}, Fn(k) \le n \le Fn(k+1), \\ 0, Fn(k+1) \le n; \end{cases} \quad (3.7)$$

де $k = 0,1,2 \dots, K-1$ – індекс фільтра, $n = 0,1,2 \dots, N-1$ – індекс відліка.

3.5 Отримання кепстральних коефіцієнтів

Останнім кроком при отриманні мел-частотних кепстральних коефіцієнтів є логарифмування енергії частотної області, помноженої на ваговий коефіцієнт і дискретне косинусне перетворення Фур'є. Обчислюємо енергію для кожного вікна. Це зроблено тому, що необхідно застосувати мел-фільтри не до значень спектра, а до його енергії. Так як фільтри застосовуються до квадратах модулів коефіцієнтів перетворення Фур'є, отримані результати логарифмуються:

$$L(k) = (ln \sum_{n=0}^{N-1} |X_k|^2 W(k,n)), k = 0,1,2, \dots, K-1. \quad (3.8)$$

Далі застосовується дискретне косинусне перетворення Фур'є, в результаті якого отримують шукані мел-частотні кепстральних коефіцієнти:

$$C_i(n) = \sum_{k=0}^{K-1} L(k) \cos\left(\frac{\pi n}{K}(k+\frac{1}{2})\right), n = 0,1,2, \dots, K-1, \quad (3.9)$$

Коефіцієнт C_0 не використовується, тому що являє енергію сигналу. Кількість коефіцієнтів K зазвичай, вибирають від 12 до 30. На рисунку 3.3 представлений графік мел-частотних кепстральних коефіцієнтів, отриманих для вихідного сигналу.

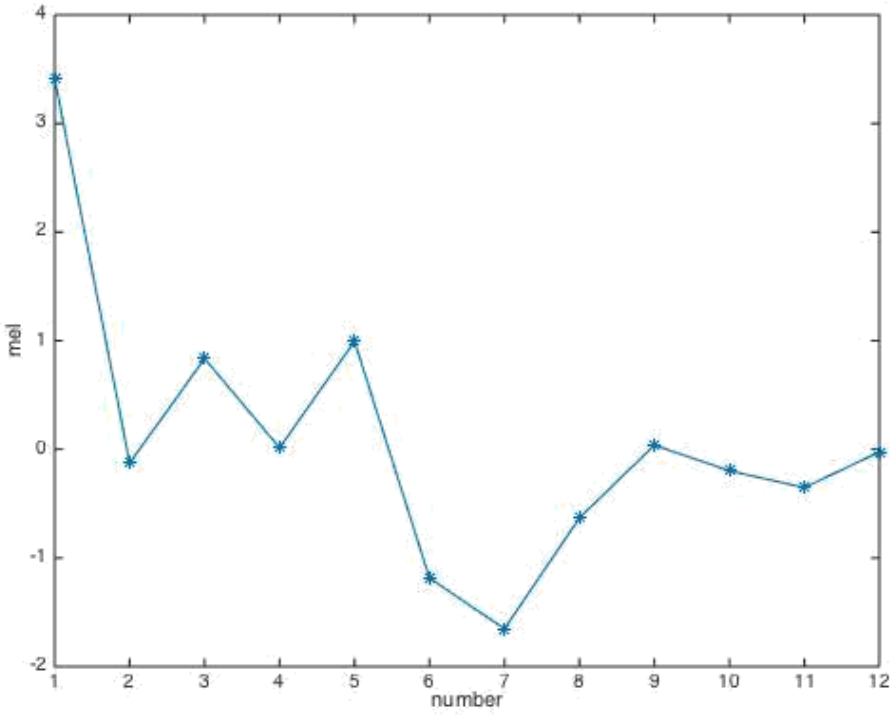

Рисунок 3.3 – Отриманні коефіцієнти

4. РОЗРОБКА АЛГОРИТМУ І ПРОГРАМИ ДЛЯ ІДЕНТИФІКАЦІЇ ДИКТОРІВ В СЕРЕДОВИЩІ РОЗРОБКИ MATLAB

4.1 Формулювання вимог

Для виконання поставлених завдань програма повинна мати такі функції:

– зчитування звукових файлів WAV формата;

– обчислення MFCC;

– ідентифікація дикторів.

Для написання програми обрано середовище розробки Matlab. Розробка графічного інтерфейсу недоцільна, так як розробляється програма не призначена для широкого використання, а необхідна для про ведення дослідницької роботи самим розробником.

4.2 Структура Wave файлу

Розглянемо файл аудіозаписи формату WAVE. Даний аудіоформат є підвидом RIFF (Resource Interchange File Format - формат файлу обміну даними). Він являє собою дві області, чітко розмежовані між собою. Одна з них - являє собою коротку інформацію про аудіофайл, інакше кажучи заголовок файлу, інша - область даних. Зазвичай, в заголовку файлу зберігається інформація:

1) розмір файлу;

2) кількість каналів;

3) частота дискретизації;

4) глибина звучання (або кількість біт в кожному семпли).

Однак, в заголовку файлу може зберігатися додаткова інформація про аудіофайл, наприклад: кількість байт в області даних, формат стиснення і так далі.

Звук складається з коливань, які при оцифрування набувають ступінчастий вигляд. Цей вид обумовлений тим, що комп'ютер може відтворювати в будь-який короткий проміжок часу звук певної амплітуди (гучності) і цей короткий момент не нескінченно короткий. Тривалість цього проміжку і визначає частота дискретизації. Сукупність амплітуди і короткого проміжку часу носить назву семпл. Амплітуда виражається числом, яке може займати в файлі 8, 16, 24, 32 біт (або більше). Таким чином, чим більше зарезервовано місце в пам'яті під числову характеристику амплітуди, тим більший діапазон значень можна записати. При розробці програмної реалізації важливо враховувати момент, що в одноканальному аудіофайл значення амплітуди розташовані послідовно. Найчастіше, в стерео спочатку йде значення всіх амплітуд для лівого каналу, потім для правого каналу. Але в залежності від параметрів форматування, можливо подання до стерео форматі наступним чином: один семпл для лівого каналу, наступний семпл для другого каналу і так далі.

На рисунку 4.1 представлена структура файлу формату WAVE

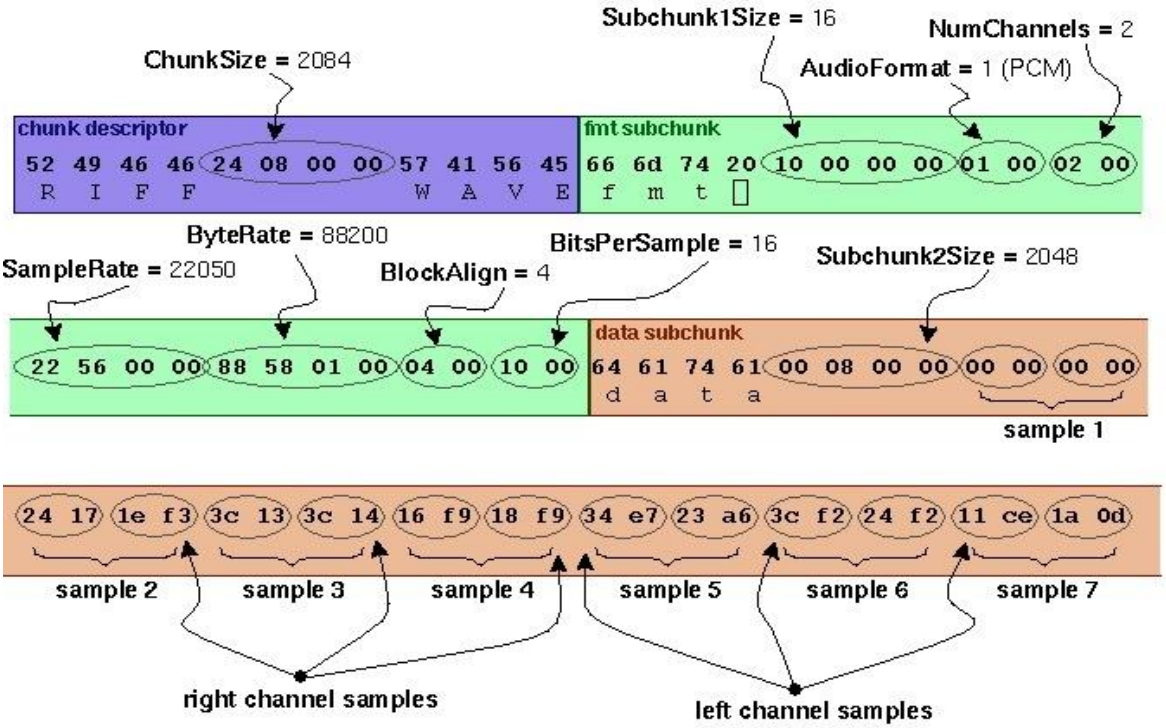

Рисунок 4.1 – Структура файла форматуWAVE

Перші чотири байти в файлі (синя зона) - назва основного шматка «RIFF». Всі сигнатури записані в кодуванні ANSII. Наступні 4 байта (іменовані найчастіше як «chunkSize») являє лишившийся розмір ланцюжка, починаючи з цієї позиції. Чотири наступних байта задають формат RIFF-файлу («format»). Так як ми розглядаємо WAVE файл, то відповідно в цьому полі записано «WAVE» в кодуванні ANSII.

Чотири байта (початок зеленої зони) («subchunk1Id») - назва нового шматка «fmt», розмір якого задають наступні за ними чотири байти (subchunk1Size). Наступні два байта («audioFormat») задають ступінь стиснення. Якщо там число, відмінне від одиниці, значить присутнє стиснення файлу. Наступні два байта («numChannels») - кількість каналів (1 - моно звук, 2 - стерео звук і так далі). Наступні два байта («sampleRate») - частота дискретизації. Наступні два байта («byteRate») - кількість байт, переданих за секунду відтворення. Наступні два байта («blockAlign») - кількість байт для одного семпли, включаючи всі канали. І два байта (кінець зеленої зони «bitsPerSample») - кількість біт в семпли, «глибина» звучання.

Чотири байта (початок коричневої зони) - назва наступного блоку «data», розмір якого задають наступні чотири байти («subchunk2Size»). Наступні байти - звукові дані.

Якщо порахувати кількість байт до звукових даних і скласти їх з кількістю байтів, відведених під звукові дані, то можна помітити, що в ряді випадків може виявитися недолік в 250-300 байтів до розміру файлу. Це відбувається через те, що в файлі також зберігаються додаткові дані, які не важливі при відтворенні. Наприклад, автора, дату створення файлу, хто створив файл, жанр і так далі. Важливий момент полягає в тому, що різні програми форматування розташовують ці дані в різних місцях. Наприклад, Audacity їх зберігає після розділу «data», Freemake Audio Converter їх зберігає до розділу «data». Наочна демонстрація даної особливості показана на рисунках 4.2 і 4.3.

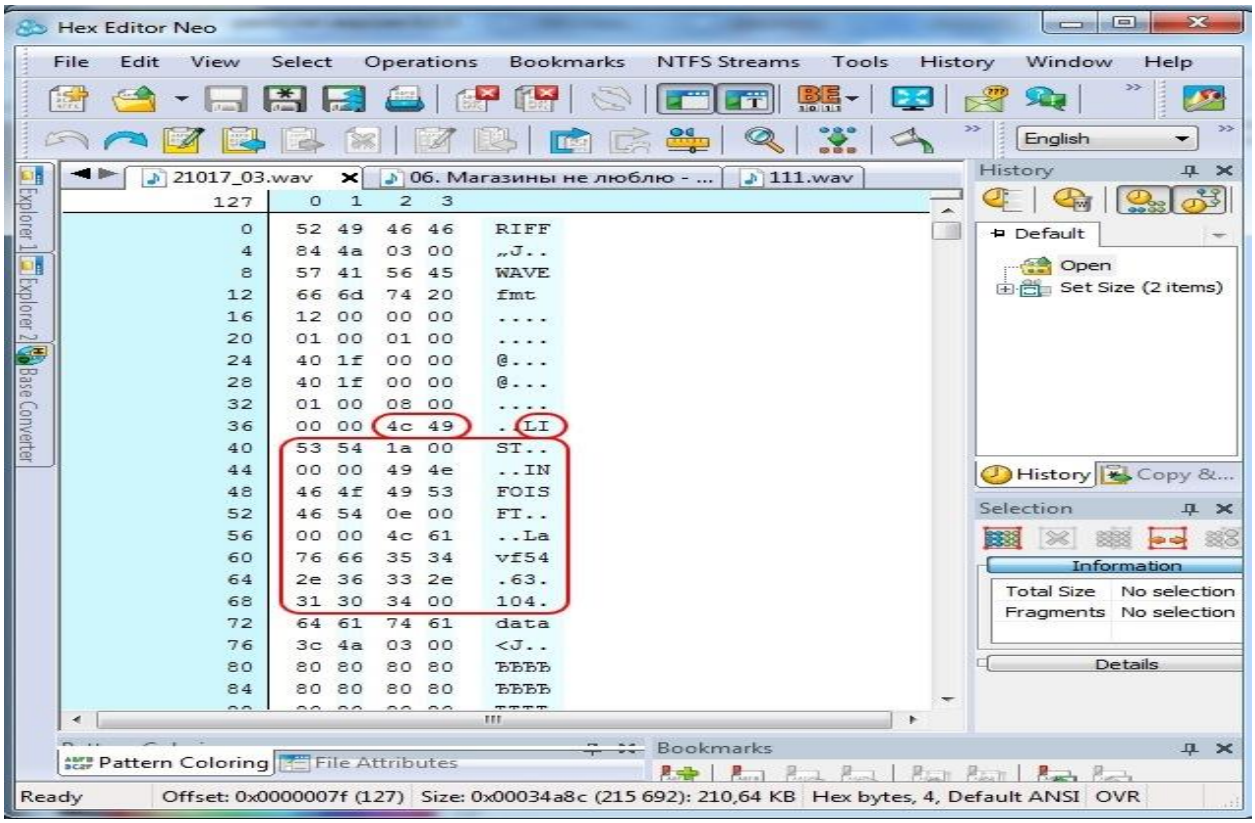

Рисунок 4.2 – Допоміжні данні на початку файлу

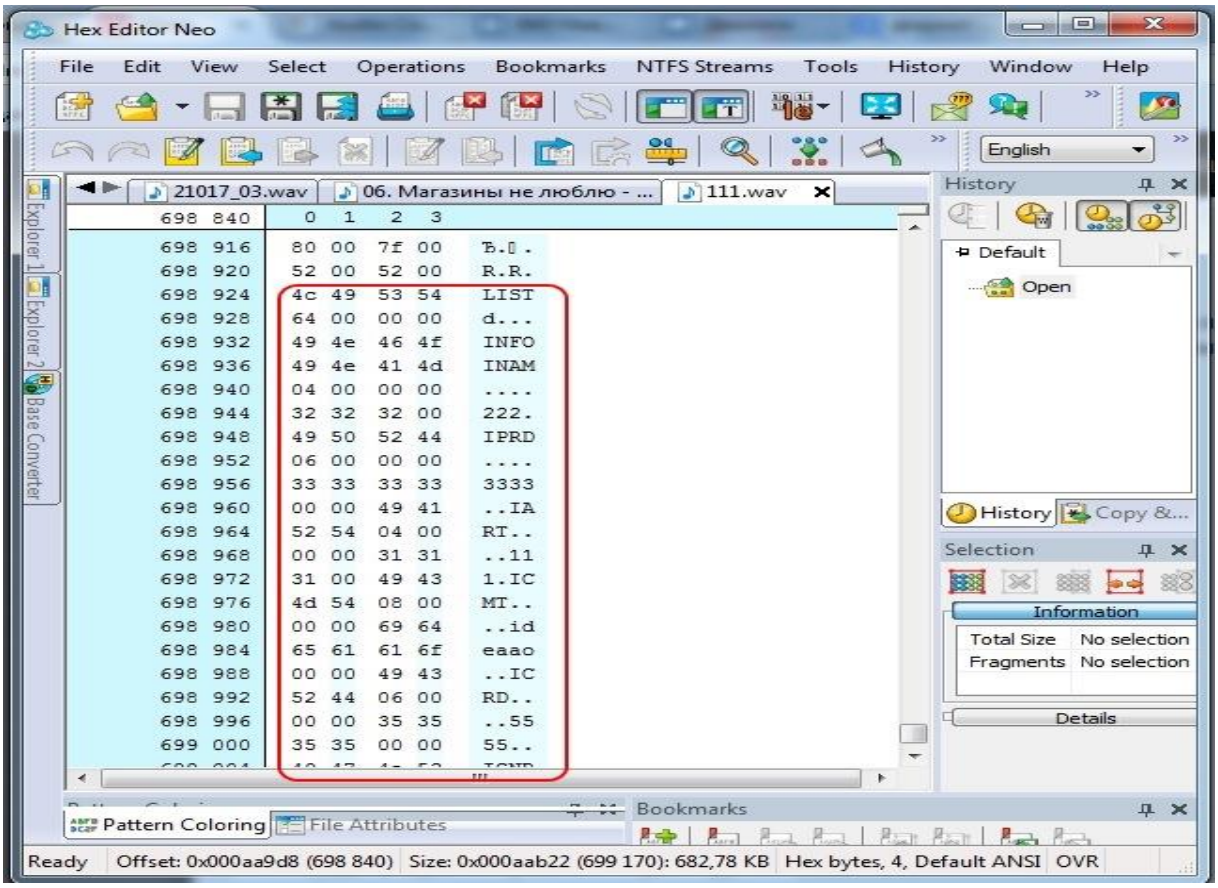

Рисунок 4.3 – Додаткові дані в кінці файлу

У зв'язку з цим, потрібно враховувати цю особливість при зчитуванні самого набору звукової інформації.

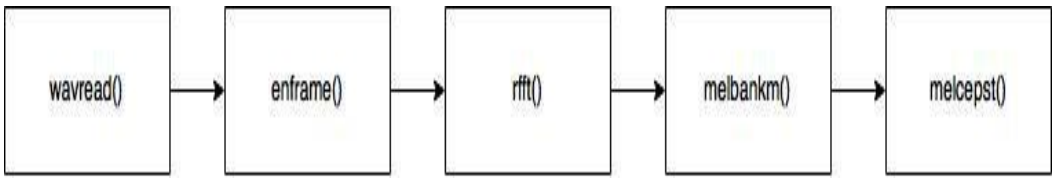

Рисунок 4.4 – Алгоритм програми

На рисунку 4.4 представлений алгоритм програми, який необхідно реалізувати. Показання в цьому алгоритмі функції не є стандартними для середовища розробки Matlab. Опис функцій і вхідні параметри будуть розглянуті нижче.

4.3 Розробка алгоритму

4.3.1 Зчитування і відтворення wav-файлів

Для того, щоб зменшити обчислювальне навантаження, використовується читання звукових сигналів безпосередньо з файлу. Для читання звукових файлів в середовищі розробки Matlab передбачена функція [sig, fs, b] = wavread('filename.wav'). Функція wavread завантажує звуковий файл і повертає вибрані дані: sig - звуковий сигнал зі значеннями від -1 до 1, fs - частота дискретизації, b - кількість біт на одну вибірку

Так само існує можливість прослухати завантажений нами звуковий файл, для цього використовується функція sound (sig, fs), де sig - звуковий сигнал, fs - частота дискретизації

4.3.2 Розрахунок мел-частотних кепстральних коефіцієнтів

В даному розділі буде розглянуто розрахунок мел-частотних коефіцієнтів, розподілених за мел-шкалою. Алгоритм виглядає наступним чином:

1. Ділимо мовний сигнал на фрейми тривалістю 20 мс. Причому фрейми повинні йти не один за одним, а «внахлест», тобто кінець одного повинен накладатися на початок іншого. У нашому випадку зміщення становить *N*/ 2, де - кількість відліків з вибірки. Перед тим як розділити сигнал на фрейми, з нього видаляються всі відрізки, що не несуть інформації про голос диктора, тобто місця, де диктор мовчить. Для цього був написаний наступний код:

```
frame_duration = 0.2;
frame_length = frame_duration *Fs;
N = length( sig );
num_frames = floor( N / f_length);
inc = floor(f_length /2);
```

```
    new_sig = zeros(length(sig), 1);
    count = 0;
    for k = 1: num_frames
        frame = sig((k-1)*frame_length+1: f_length *k);
        max_val = max(frame);
        if max_val > 0.1
            count = count + 1;
            new_sig((count-1)*f_length + 1:
frame_length*count) = frame;
        end
    end
```

Функція enframe (sig, win, inc) - використовується для розбиття вихідного сигналу на фрейми, а також для множення на віконну функцію Вхідні параметри: sig - вхідний сигнал, win - віконна функція, inc - зміщення.

2. Далі для кожного фрейму необхідно розрахувати спектр сигналу за допомогою перетворення Фур'є . Перетворення Фур'є в дійсній області реалізується в середовищі Matlab за допомогою функції rfft (sig), де sig помножений на віконну функцію сигнал.

3. Для складання банку трикутних фільтрів використовується функція melbank (p, n, fs, fl, fh, w). Вона створює набір трикутних мел-фільтрів на основі вхідних параметрів: p - кількість фільтрів, n - довжина перетворення Фур'є, fs - частота дискретизації, fl/fh- нижня / верхня межа фільтрів, w - тип фільтрів.

4. Функція melcepst (sig, fs, w, nc, p, n, inc, fl, fh) - повертає мел кепстральні коефіцієнти. Вхідні параметри: sig - сигнал, fs - частота дискретизації, w - віконна функція, nc - кількість коефіцієнтів, p - кількість фільтрів, inc - зміщення, fl/fh - нижня / верхня межа фільтрів.

Для наочності покажемо схему алгоритму обчислення мелчастотних кепстральних коефіцієнтів:

Рисунок 4.5 – Схема алгоритму обчислень мел-кепстральних коефіцієнтів

4.4 Розробка нейронної мережі

В якості методу ідентифікації диктора була обрана нейронна мережа. В даному випадку обрана однонаправлена мережу прямого поширення, так як такі мережі найчастіше використовуються для прогнозування, розпізнавання образів і апроксимації нелінійних функцій.

Для створення нейронної мережі була використана функція newff (PR, [S1 S2 ... SNl], {TF1 TF2 ... TFNl}, BTF, BLF, PF). В якості вхідних параметрів вона використовує:

– PR - Rx2 матриця мінімальних і максимальних значень строк вхідної матриці с розмірністю RxQ;
– S_i – кількість нейронів в і – тому шарі, $N1$ – кількість шарів;
– TF_i - функциі активації і - го шару, за замовчуванням = 'tansig';
– BTF – навчальна функція зворотного поширення, за замовчуванням ='trainlm';
– BLF – алгоритм підстроювання ваг і зміщень (навчальний алгоритм), за замовчуванням = 'learngdm';
– PF функція оцінки функціонування мережі, за замовчуванням = 'mse'.

І повертає односпрямовану мережу, що складається з шарів.

Для даної роботи використовується мережа, що складається з трьох шарів (рис 4.6). Кожний шар мережі складається з нейронів. Нейрон - це головний обчислювальний елемент нейронної мережі. До складу нейрона входять помножувачі, суматори і нелінійний перетворювач. Синапси здійснюють зв'язок між нейронами і множать вхідний сигнал на число, яке вказує на силу зв'язку - ваги синапсів.

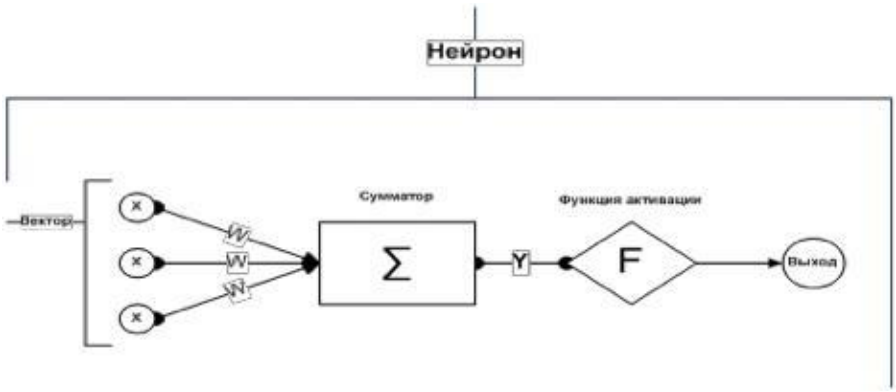

Рисунок 4.6 – Структура нейрона

Суматор виконує додавання сигналів, що надходять з інших нейронів або зовнішніх сигналів. Нелінійний перетворювач реалізує нелінійну функцію одного аргументу - вихід суматора. Ця функція називається «Функція активації» нейрона.

Математична модель нейрона описується як:

$$Out_i = F(\sum_{i=1}^{n} w_i x_i), \quad (4.1)$$

де w_i - вага синапса ($i = 1, \dots, n$), x_i - компонент вхідного вектора ($i = 1, \dots, n$)

Основне завдання в процесі розробки нейронної мережі - навчання цієї мережі, тобто коригування ваг мережі для того, щоб мінімізувати помилки на виході нейронної мережі.

Як функція активації була використана логарифмічна функція активації logsig

$$logsin(n) = \frac{1}{1+\exp(-n)}. \quad (4.2)$$

В якості функції для навчання була обрана функція traingda. Це метод з адаптивною швидкістю навчання. Сенс методу полягає в модифікації ваг і зміщень відповідно до методу градієнтного спуску.

Тренування нейронної мережі виконується відповідно до параметрів функції навчання traingda. Ці параметри і їх значення за замовчуванням наведені в таблиці нижче

Таблиця 4.1 – Параметри навчальної функції traingda

Функція	Значення	Опис
net.trainParam.epochs	1000	Максимальна кількість епох навчання
net.trainParam.goal	0	Умови зупинки по 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 між графіками
net.trainParam.showCommandLine	False	Відкрити командну строку
net.trainParam.showWindow	True	Показати графічне уявлення тренування
net.trainParam.time	Inf	Максимальний час тренування в сек.

Як було написано вище, головне завдання в навчанні нейронної мережі - мінімізувати квадратичну помилку. Квадратична помилка вхідного вектора визначається за сумою квадратичних помилок в кожному вузлі виходу:

$$E = 1/2 \sum_{i=1}^{l} (y_k - d_k)^2 \quad (4.3)$$

Для мінімізації квадратичної помилки використовується алгоритм градієнтного спуску. Тренування нейронної мережі триває до тих пір, поки не буде досягнуто одна з умов зупинки з таблиці 4.1.

Для наочності алгоритм навчання нейронної мережі наведено на рисунку 4.7.

В коді мережа стврорюється наступним чином:

```
Net = newff(ma, [15 10 3], {'logsig' 'logsig' 'logsig'},
'traingda');
net.train_param.epochs = 4000;
net.train_param.min_grad = 1e-50;
```

```
net = init( net );
net = train( net, voices_for_train, T);
```

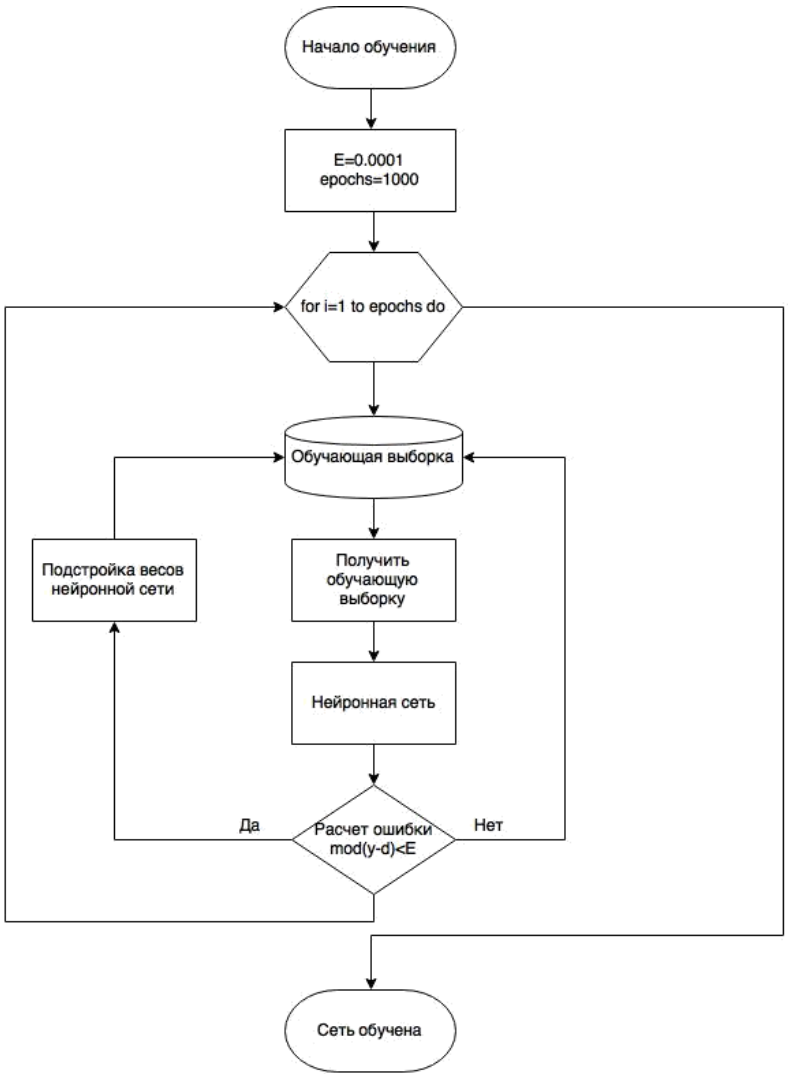

Рисунок 4.7 – Схема алгоритма навчання нейронної мережі

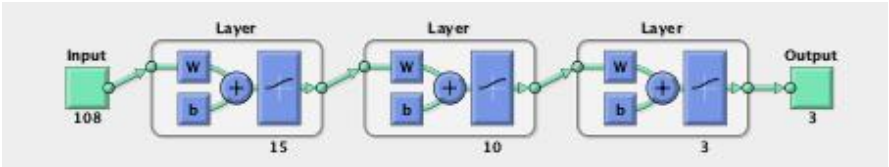

Рисунок 4.8 – Схема нейронної мережі

Як видно зі схеми, нейронна мережа складається з трьох шарів. Перший шар отримує ваги, які приходять зі входу. В даному випадку це

набір мел-частотних кепстральних коефіцієнтів. Наступні шари отримують ваги від попередніх шарів. Всі шари мають зсуви. Третій шар є виходом мережі.

В даному розділі були описані функції, використані для розробки програми з розрахунку мел -частотних кепстральних коефіцієнтів, а також для побудови нейронної мережі. Було розглянуто математичний апарат, на підставі якого ці функції працюють і складені блок-схеми алгоритмів розрахунку кепстральних коефіцієнтів і навчання нейронної мережі.

4.5 Інструкція користувача

Головним файлом програми є main.m. Для запуску системи ідентифікації диктора необхідно запустити даний файл, попередньо помістивши його в папку MATLAB. Так само для коректної роботи програми в папці MATLAB повинні знаходитися файли функцій enframe.m, melbankm.m, rfft.m, melcepst.m, так як ці функції використовуються в програмі і не є стандартними функціями середовища Matlab.

Для того, щоб додати до системи нового диктора необхідно скопіювати звукові файли з його голосом для навчання і для розпізнавання в форматі .wav в папку Matlab, потім додати їх в кінець масиву voices, а потім в циклі розрахунку мел-частотних кепстральних коефіцієнтів додати нову умову з індексом, де будуть для цих файлів розраховано коефіцієнти.

Для навчання нейронної мережі в кінець масиву voices_for_train необхідно додати змінну, що містить розраховані коефіцієнти. Також необхідно змінити розмір транспонуючої матриці T і змінити кількість сигналів в функції створення нейромережі.

Для перевірки працездатності мережі використовується функція sim (net, var), де var - обчислені коефіцієнти для звукового файлу з голосом для перевірки, а net - змінна, що містить інформацію про створену нейронної мережі.

5 ПЕРЕВІРКА РЕЗУЛЬТАТІВ РОБОТИ РОЗРОБЛЕНИХ АЛГОРИТМІВ

У даному розділі проводиться аналіз результатів роботи розробленої програми. Завдання полягало в тому, щоб простежити залежність точності розпізнавання дикторів від різних параметрів, таких як: кількість епох навчання, кількість дикторів, тексто-незалежніть розпізнавання.

Для проведення експерименту було записано 10 дикторів. Аудіозаписи, на яких будувалося навчання нейронної мережі, мають рівну тривалість - 2 секунди. Цього часу цілком достатньо для вимови контрольного слова або фрази.

Так само попередньо слід описати технічні характеристики комп'ютера, на якому проводився експеримент. Це зроблено тому, що результати дослідження на одному комп'ютері можуть сильно відрізнятися від результатів на іншому комп'ютері. Пов'язано це, в основному, з обчислювальними можливостями комп'ютерів.

5.1 Характеристики пристрою

Експеримент проводився на комп’ютері Dell inpiron 5521.
Технічні характеристики:

- процесор: 1.3 GHz Intel Core i3;
- оперативная пам’ять: 8Gb 1600 MHz DDR3L;
- графічний адаптер: Intel HD Graphics 4000 1536 Mb;
- ос: Windows 10.

5.2 Результати експерименту

В даному розділі буде представлений результат роботи програми. Експеримент розбитий на кілька етапів:

1. Тест системи з трьома дикторами;
2. Тест системи з п'ятьма дикторами;
3. Тест системи з десятьма дикторами;
4. Тест системи на тексто-незалежність с десяттю дикторами.

Умови експерименту:

- кількість дикторів – 3;
- кількість епох навчання– 500;
- кількість експериментів – 10.

Таблиця 5.1 – Результати ідентифікації для системи з трьома дикторами (500 епох навчання)

Диктор	Результат експерименту									
	1	2	3	4	5	6	7	8	9	10
1	+	+	-	+	+	+	-	+	+	-
2	-	+	+	+	-	-	-	+	+	+
3	+	+	+	-	+	+	+	+	+	+

Як можна побачити з таблиці, частка помилкового розпізнавання становить 26,6%. В процесі навчання кращий результат був досягнутий на останній епосі навчання, середньоквадратична помилка при цьому дорівнювала 3.76 * 10C% ". Час, витрачений на процес навчання склав 4 секунди. Графік зміни середньоквадратичної помилки показаний на рисунку 5.1

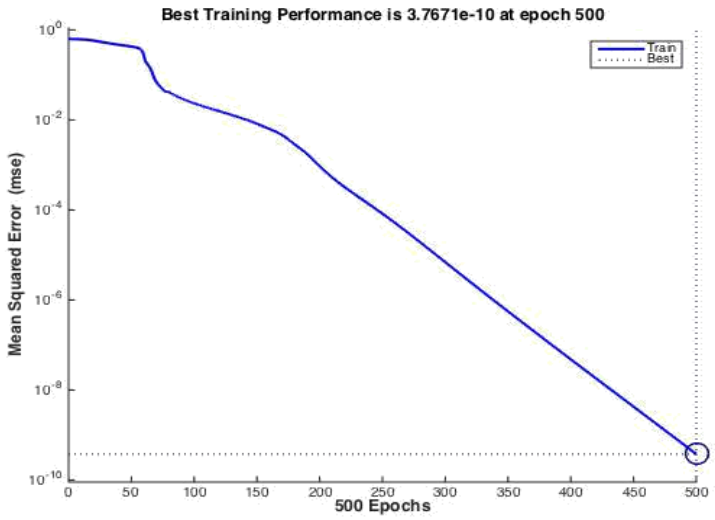

Рисунок 5.1 – Графік зміни середньоквадратичної помилки

Розглянемо випадок, коли кількість епох навчання складає 1000.

Умови експерименту:

– кількість дикторів – 3;

– кількість епох навчання– 1000;

– кількість експериментів – 10.

Таблиця 5.2 – Результати ідентифікації для системи з трьома дикторами (1000 епох навчання)

Диктор	Результат експерименту									
	1	2	3	4	5	6	7	8	9	10
1	+	+	+	+	-	+	+	+	+	+
2	-	+	-	+	-	+	+	+	-	+
3	+	+	+	+	+	+	+	+	+	+

Як можна побачити з таблиці, частка помилкового розпізнавання становить 16,6%. Даний результат на 10% кращий за попередній. В процесі навчання кращий результат був досягнутий на останній епосі навчання,

середньоквадратична помилка при цьому дорівнювала $4.05 * 10^{-20}$. Час, витрачений на процес навчання склав 8 секунд. Графік зміни середньоквадратичної помилки показаний на рисунку 5.2

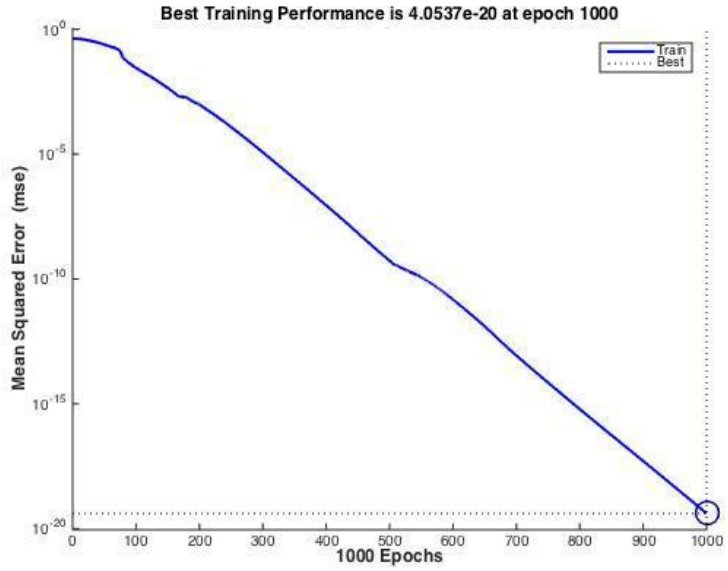

Рисунок 5.2 – Графік зміни середньоквадратичної помилки

Тепер розглянемо випадок, коли кількість епох навчання становило 5000.

Умови експерименту:

- кількість дикторів – 3;
- кількість епох навчання – 5000;
- кількість експериментів – 10.

Таблиця 5.3 – Результати ідентифікації для системи с трьомя дикторами (5000 епох навчання)

Диктор	Результат експерименту									
	1	2	3	4	5	6	7	8	9	10
1	+	+	+	+	+	+	+	+	+	+
2	+	+	+	+	+	+	+	+	-	+
3	+	+	+	+	+	-	+	+	+	+

Як можна побачити з таблиці, частка помилкового розпізнавання становить 6%. Даний результат на 10,6% краще, ніж при навчанні в 1000 епох. В процесі навчання кращий результат був досягнутий на 1598 епосі навчання, середньоквадратична помилка при цьому дорівнювала 2.3451 * 10Cé%. Час, витрачений на процес навчання, склало 23 секунди. Графік зміни середньоквадратичної помилки показаний на малюнку 5.3

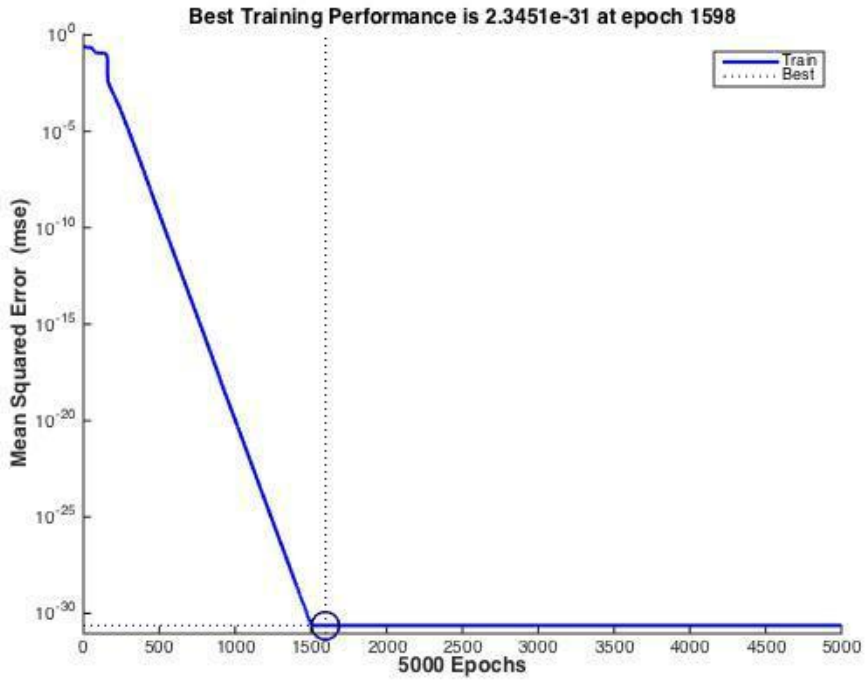

Рисунок 5.3 – Графік зміни середньоквадратичної помилки

Умови експерименту:

- Кількість дикторів – 5;
- Кількість епох навчання– 500;
- Кількість експериментів – 10.

Таблиця 5.4 – Результати ідентифікації для системи з п'ятьма дикторами (500 епох навчання)

Диктор	Результат експерименту									
	1	2	3	4	5	6	7	8	9	10
1	-	-	+	+	+	-	-	-	+	+
2	+	-	-	-	+	-	-	+	-	+
3	+	+	-	+	+	+	+	-	-	+
4	+	+	+	+	-	+	+	+	-	-
5	-	+	+	+	+	-	+	+	+	+

Як видно з таблиці, частка невірного розпізнавання становить 38%. В процесі навчання кращий результат був досягнутий на останній епосі навчання, середньоквадратична помилка при цьому дорівнювала $5.78*10^{-6}$. Час, витрачений на процес навчання склало 5 секунд.

Збільшимо кількість епох навчання до 1000 і проведемо повторний експеримент.

Умови експерименту:

– кількість дикторів – 5;

– кількість епох навчання– 1000;

– кількість експериментів – 10.

Таблиця 5.5 – Результати ідентифікації для системи з п'ятьма дикторами (1000 епох навчання)

Диктор	Результат експерименту									
	1	2	3	4	5	6	7	8	9	10
1	-	-	+	-	+	+	-	+	+	+
2	+	+	-	+	+	+	+	-	-	+
3	+	+	+	+	+	+	-	+	+	+
4	+	+	-	+	-	-	+	+	+	-
5	+	+	+	+	+	+	+	+	+	+

Як видно з таблиці, частка невірного розпізнавання становить 24%, що на 14% краще результату попереднього експерименту. В процесі навчання кращий результат був досягнутий на останній епосі навчання, середньоквадратична помилка при цьому дорівнювала $2.98 * 10^{-16}$. Час, витрачений на процес навчання склав 8 секунд.

Для наступного експерименту збільшимо кількість епох навчання до 5000.

Умови експерименту:

– кількість дикторів – 5;

– кількість епох навчання – 5000;

– кількість експериментів – 10.

Таблиця 5.6 – Результати ідентифікації для системи з п'ятьма дикторами (5000 епох навчання)

Диктор	Результат експерименту									
	1	2	3	4	5	6	7	8	9	10
1	+	+	+	+	+	+	+	-	+	+
2	-	+	-	+	+	+	+	+	-	+
3	+	+	+	+	+	+	+	+	-	+
4	+	+	+	+	-	+	+	+	+	-
5	+	+	+	+	+	+	-	+	+	+

Як видно з таблиці, частка невірного розпізнавання становить 16%, що на 8% краще результату попереднього експерименту. В процесі навчання кращий результат був досягнутий на 3787 епосі навчання, середньоквадратична помилка при цьому дорівнювала $1.01 * 10^{-32}$. Час, витрачений на процес навчання склав 27 секунд.

Збільшимо кількість дикторів в системі з п'яти до десяти і подивимося, як зміниться якість ідентифікації з різною кількістю епох навчання. Умови експеременту:

- кількість дикторів – 10;
- кількість епох навчання– 500;
- кількість експериментів – 10.

Таблиця 5.7 – Результати ідентифікації для системи з десятьма дикторами (500 епох навчання)

Диктор	Результат експерименту									
	1	2	3	4	5	6	7	8	9	10
1	+	-	+	+	+	-	+	-	+	+
2	-	-	+	-	-	-	-	-	+	-
3	+	-	-	+	+	-	-	-	+	+
4	-	+	+	-	-	-	+	-	-	+
5	+	+	+	+	-	+	+	+	+	-
6	+	-	-	-	+	-	-	-	-	+
7	-	-	-	+	+	+	+	+	+	+
8	+	+	-	-	+	-	+	-	+	-
9	-	+	+	+	-	+	+	+	-	-
10	-	-	+	-	-	+	-	-	+	+

Як видно з таблиці, частка невірного розпізнавання становить 47%. В процесі навчання кращий результат був досягнутий на останній епосі навчання, середньоквадратична помилка при цьому дорівнювала $2.79 * 10^{-6}$. Час, витрачений на процес навчання склав 8 секунд. Збільшимо кількість епох навчання до 1000.

Умови експерименту:

– кількість дикторів – 10;

– кількість епох навчання– 1000;

– кількість експериментів – 10.

Таблиця 5.8 – Результати ідентифікації для системи з десятьма дикторами (1000 епох навчання)

Диктор	Результат експерименту									
	1	2	3	4	5	6	7	8	9	10
1	+	+	+	+	+	-	+	-	+	-
2	-	+	+	-	+	+	-	-	-	+
3	+	+	+	+	+	+	-	+	-	+
4	-	+	+	-	+	+	+	+	+	-
5	+	+	+	+	-	-	+	+	+	+
6	+	-	-	-	+	+	+	+	-	-
7	+	+	+	+	+	-	+	-	+	+
8	+	+	-	-	+	+	-	-	+	-
9	+	-	-	+	+	-	+	+	-	+
10	-	+	+	-	-	+	+	+	+	-

Як видно з таблиці, частка невірного розпізнавання становить 34%. Результат експерименту на 13% краще попереднього. В процесі навчання кращий результат був досягнутий на останній епосі навчання, середньоквадратична помилка при цьому дорівнювала 2.82. Час, витрачений на процес навчання склав 12 секунд.

Тепер збільшимо кількість епох навчання до 5000 і подивимося, яким чином це відіб'ється на точності ідентифікації

Умови експерименту:

– кількість дикторів – 10;

– кількість епох навчання– 5000;

– кількість експериментів – 10.

Таблиця 5.9 – Результати ідентифікації для системи з десятьма дикторами (5000 епох навчання)

Диктор	Результат експерименту									
	1	2	3	4	5	6	7	8	9	10
1	+	-	+	+	-	+	+	-	-	+
2	-	+	+	+	-	+	+	-	+	+
3	-	+	+	+	+	+	+	-	+	+
4	-	+	+	-	+	+	+	+	+	+
5	+	+	-	-	+	+	+	+	+	+
6	+	+	+	+	+	-	+	-	+	+
7	+	+	+	+	+	+	+	+	+	-
8	+	+	+	+	+	-	-	+	+	+
9	+	+	+	+	+	+	+	+	-	-
10	+	-	+	+	-	+	+	+	+	+

Як видно з таблиці, частка невірного розпізнавання становить 22%. Результат даного експерименту на 12% краще попереднього. В процесі навчання кращий результат був досягнутий на 4897 епосі навчання, середньоквадратична помилка при цьому дорівнювала $4.28 * 10^{-32}$. Час, витрачений на процес навчання склав 36 секунд.

Для перевірки системи на тексто-незалежність було обрано три диктора. Навчання нейронної мережі відбувається по одній фразі, а ідентифікація за зовсім іншою. Кількість епох навчання дорівнює 5000

Умови експерименту:

– кількість дикторів – 3;
– кількість епох навчання– 5000;
– кількість експеременті в – 10;
– навчання та ідентифікація засновані на різних фразах.

Таблиця 5.10 – Результати ідентифікації для системи з десятьма дикторами (5000 епох навчання)

Диктор	Результат експерименту									
	1	2	3	4	5	6	7	8	9	10
1	+	+	+	+	-	+	+	+	+	+
2	+	-	+	+	+	+	-	+	+	+
3	+	+	+	+	+	+	+	+	-	+

Як можна побачити з таблиці, частка помилкового розпізнавання становить 13,3%. Даний результат на 7.3% гірше, ніж тексто-залежна ідентифікація при тих же умовах. В процесі навчання кращий результат був досягнутий на 4835 епосі навчання, середньоквадратична помилка при цьому дорівнювала $1.01 * 10^{-32}$. Час, витрачений на процес навчання склав 22 секунди. Графік зміни середньоквадратичної помилки показаний на рисунку 5.4.

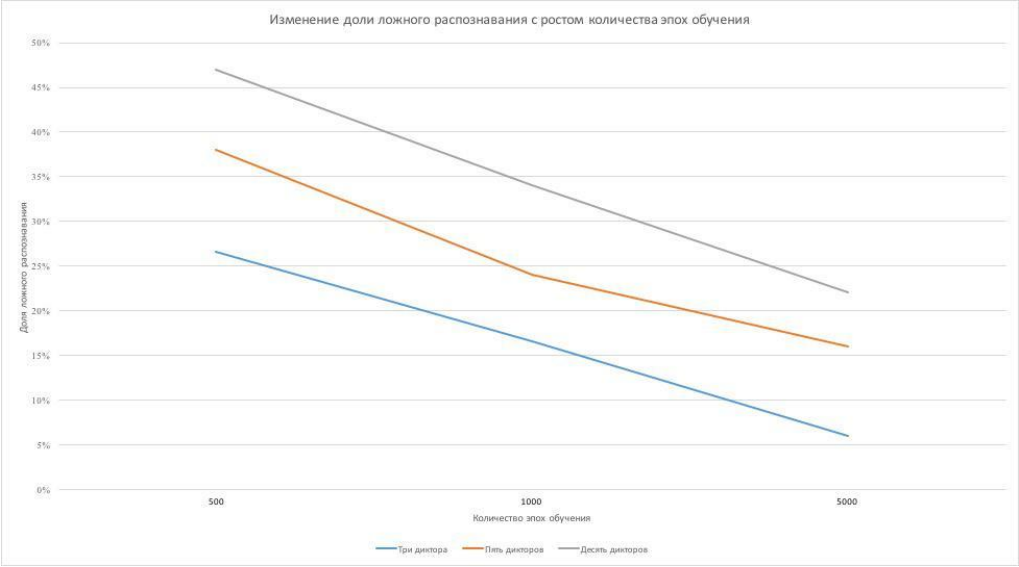

Рисунок 5.4 – Залежність точності розпізнавання від епох навчання для систем з різною кількістю дикторів

За результатами проведених експериментів був побудований графік, на якому наочно видно, як змінюється точність розпізнавання дикторів системою при зміні параметрів: кількості дикторів для розпізнавання і кількості епох навчання. На підставі цього можна зробити висновок, що при збільшенні кількості дикторів в системі точність ідентифікації зменшується, а при збільшенні кількості епох навчання зростає.

Окремо варто зауважити, що при тексто-незалежній ідентифікації дикторів точність розпізнавання зменшується. Пов'язано це з тим, що для такої ідентифікації необхідно більше епох навчання нейронної мережі, ніж для тексто-залежної, з таким же набором дикторів.

Даний експеримент доводить, що мел-частотні кепстральні коефіцієнти не прив'язуються до слів, які вимовляє диктор, а використовують при розпізнаванні «значущі» для людського вуха частоти, які характеризують людський голос.

Для збільшення точності розпізнавання дикторів нейронна мережа вимагає більш тонкої настройки, більшої кількості епох навчання і комп'ютера з великими обчислювальними можливостями.

ВИСНОВКИ

В результаті виконання даної роботи був проведений аналізу методів ідентифікації дикторів. Детально досліджені існуючі методи вирішення задач ідентифікації по голосу. Проведений аналіз існуючих математичних апаратів, що вирішують проблему ідентифікації.

Для роботи були обрані метод мел-частотних кепстральних коефіцієнтів для аналізу мовного сигналу, і для розв'язання задачі ідентифікації диктора - метод нейронних мереж.

Головною перевагою методу кепстральних коефіцієнтів є простота реалізації при хорошій якості розпізнавання, яке не поступається іншим популярним алгоритмам. Також даний метод передбачає менший аналіз даних, що значно скорочує час навчання нейронної мережі.

Незаперечною перевагою ідентифікації диктора за допомогою нейронних мереж є здатність до навчання системи, що підвищує точність розпізнавання при більшій кількості циклів навчання.

Так само були описані функції, використані для розробки програми з розрахунку мел-частотних кепстральних коефіцієнтів і для побудови нейронної мережі. Було розглянуто математичний апарат, на підставі якого дані функції працюють і складені блок-схеми алгоритмів розрахунку кепстральних коефіцієнтів і навчання нейронної мережі.

Для перевірки працездатності програми були проведені експерименти при різних умовах роботи програми. За результатами проведених експериментів можна зробити висновок, що при збільшенні кількості дикторів в системі точність ідентифікації зменшується, а при збільшенні кількості епох навчання зростає.

Окремо варто зауважити, що при тексто-незалежній ідентифікації дикторів точність розпізнавання зменшується. Пов'язано це з тим, що для такої ідентифікації необхідно більше епох навчання нейронної мережі, ніж для тексто-залежної з таким же набором дикторів.

Дана система ідентифікації дикторів може активно використовуватися для розпізнавання людей в системах безпеки і для визначення спікерів на нарадах і засіданнях.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Текстонезависимая идентификация по голосу [Электронный ресурс] / habr. – Режим доступа: www / URL: https://habr.com/ru/post/336516/ – 28.28.2017 г. – Загл. с экрана.

2. Малинин, П.В. Идентификация личности на основе комбинированных данных отпечатка пальцев и изображения лица / П.В. Малинин // Проблемы правовой и технической защиты информации: сб. ст. / под ред. В.В. Полякова, В.А. Мазурова. – Барнаул: Изд-во Алт. ун-та, 2008. – С. 163–166.

3. Где нужны технологии распознавания речи [Электронный ресурс] / postbauka. – Режим доступа: www / URL: https//postnauka.ru/faq/82575– 07.06.2018 г. – Загл. с экрана.

4. Матвеев, Ю. Н. Технологии биометрической идентификации личности по голосу и другим модальностям / Ю. Н. Матвеев – Москва: Вестн. МГТУ им. Н. Э. Баумана. Сер. Приборостроение. Специальный выпуск. Биометрические технологии. 2012. № 3(3). С. 46-61.

5. Садыхов, Р. Х. Модели гауссовых смесей для верификации диктора по произвольной речи / Р. Х. Садыхов, В. В. Ракуш // – Минск: доклады БГУИР. – 2003. – № 4. – С. 95–103.

6. Kinnunen, T. An overview of text-independent speaker recognition: From features to supervectors / Tomi Kinnunen, Haizhou Li. – University of Joensuu, Joensuu, Finland, 2010. – 39 с.

7. Сравнительный анализ систем распознавания речи с открытымм кодом [Электронный ресурс] / Международный научно-исследовательский журнал. – Режим доступа: www / URL: https:// www.research-journal.org/technical?– 16.03.2018 г. – Загл. с экрана.

8. Mermelstein, P. Distance measures for speech recognition, psychological and instrumental // Pattern recognition and artificial intelligence.

1976. Vol. 116. – С. 374–388.

9. Заковряшин, А.С. Применение распределений мел-частотных кепстральных коэффициентов для голосовой идентификации личности // П.В. Малинин, А. А. Лепендин – Барнаул: Изв. АГУ. – 5/2007. – С. 156–160.

10. Клименко Н.С. Разработка структуры текстонезависимой системы идентификации диктора // Искусственный интеллект. – Донецк,2012.. – С. 161–171.

11. Первушин Е.А. Обзор основных методов распознавания дикторов / Е.А. Первушин // Математические структуры и моделирование. − 2011− С. 41-54.

12. Сорокин В.Н. Верификация диктора по спектрально-временным параметрам речевого сигнала / В.Н. Сорокин, А.И. Цыплихин // Информационные процессы. – С. 87-104.

13. Прогресс нейронных сетей [Электронный ресурс] / postbauka. – Режим доступа: www / URL: https://postnauka.ru/talks/80077 – 15.09.2018 г. – Загл. с экрана.