УДК 004

В. Гуровський, магістр гр. КІ-22М-1
Центральноукраїнський національний технічний університет

ДОСЛІДЖЕННЯ ТА ПРОГРАМНА РЕАЛІЗАЦІЯ СИСТЕМИ ДИСТАНЦІЙНОГО ГОЛОСОВОГО КЕРУВАННЯ РОБОТОТЕХНІЧНИМ КОМПЛЕКСОМ

У статті розроблено програмне забезпечення, яке призначено для системи дистанційного голосового керування робототехнічним комплексом. Метою розробки є дослідження та програмна реалізація системи дистанційного голосового керування робототехнічним комплексом. Об’єктом дослідження є процес дистанційного голосового керування робототехнічним комплексом. Предметом дослідження є методи дистанційного голосового керування робототехнічним комплексом. Методи дослідження базуються на методах теорії Інтернету речей, методах математичної статистики, методах розробки програмного забезпечення. Результат роботи – програмна реалізація системи дистанційного голосового керування робототехнічним комплексом. В процесі роботи над програмною моделлю виконано аналіз існуючих апаратних та програмних засобів. В повній мірі описані всі компоненти розробленого програмного забезпечення.

Постановка проблеми. У цей час у міру росту обсягів інформації комп'ютерна техніка усе більше й більше проникає в людське життя. Відбувається вдосконалювання інтерфейсу людина-комп'ютер. Винаходяться нові способи відображення інформації, модернізуються пристрої уведення, тривають пошуки такого інтерфейсу, що влаштував би всіх. На цю роль зараз претендує мовний інтерфейс. Власне кажучи, це саме те, до чого людство завжди прагнуло в спілкуванні з комп'ютером.

Роботи в цьому напрямку велися ще в той час, коли про графічний інтерфейс ніхто навіть і не думав. За порівняно короткий період був вироблений вичерпний теоретичний базис, і практичні досягнення обумовлювали тільки продуктивністю комп'ютерної техніки. В 60-70х роках були створені пристрої, здатні розпізнавати десяток мовних команд.

Сучасні розробки, як правило, ґрунтуються на біонічній моделі сприйняття мови людиною. Такі системи є ієрархічними, детермінованими, з навчанням і складаються з декількох взаємозалежних рівнів. Виділяються акустична (одержання первинних ознак мовних сигналів) і лінгвістична (робота зі словниками) складові.

Системи розпізнавання злитої мови будуються на базі імовірнісних моделей граматики мови. На словниках обсягом до 5000 слів вірогідність розпізнавання цілих фраз становить більше 95%, що вважається достатнім для забезпечення успішного мовного уведення тексту на ПК.

Для завдання голосового керування різними пристроями необхідно розпізнавання окремих мовних команд. Як правило, такий спосіб керування вимагає високої надійності (99% точності розпізнавання). Найчастіше команди вимовляються в умовах підвищеної зашумленості, наприклад на виробництві. Сучасні розробки в лабораторних умовах досягають 95% точності на словниках до 100 команд і вимагають навчальні вибірки більших обсягів (10 і більше варіантів проголошення кожного слова різними дикторами).

Таким чином, проблема побудови ефективних алгоритмів розпізнавання мовних команд є актуальною.

Аналіз останніх досліджень і публікацій. При аналізі останніх досліджень і публікацій [1-10] було виявлено певні прогалини у забезпеченні системи дистанційного голосового керування робототехнічним комплексом.

Мета й завдання дослідження. Метою роботи є дослідження та програмна реалізація системи дистанційного голосового керування робототехнічним комплексом.

Для досягнення поставленої мети визначена програма дослідження, що складається з наступних завдань:

– Огляд існуючих систем дистанційного голосового керування робототехнічним комплексом.

– Дослідження системи дистанційного голосового керування робототехнічним комплексом.

– Програмна реалізація системи дистанційного голосового керування робототехнічним комплексом.

Об'єктом дослідження є процес дистанційного голосового керування робототехнічним комплексом.

Предметом дослідження є методи дистанційного голосового керування робототехнічним комплексом.

Методи дослідження базуються на методах теорії Інтернету речей, методах математичної статистики, методах розробки програмного забезпечення.

Виклад основного матеріалу. Метод розрахунку ознак мовного сигналу – ЛСК. Мовний сигнал описується в термінах лінійних дискретних систем зі змінними параметрами й передатною функцією в частотній області виду**:**

$$H(z) = \frac{S(z)}{U(z)} = G \cdot \frac{1 + \sum_{l=1}^{q} b_l \cdot z^{-l}}{1 + \sum_{k=1}^{p} a_k \cdot z^{-k}}. \quad (1)$$

Найбільше широко для опису мовного сигнал (МС) застосовується полюсна модель лінійного пророкування, що представляється у вигляді:

$$H(z) = \frac{1}{A(z)} = G \frac{1}{1 + \sum_{i=1}^{N} a_i z^{-i}}, \quad (2)$$

де N – порядок моделі.

Параметрами такої моделі є коефіцієнти лінійного пророкування $\{a\}$, що обчислюються на кожному кадрі мовного сигналу, або еквівалентні їм параметри – ЛСК, запропоновані Ітакурой.

Корінь у загальному випадку можуть бути отримані в результаті рішення двох рівнянь:

$$\mathrm{Re}\{z^R A_N(z)\}_{z=e^{j\hat{\omega}}} = 0,\ \mathrm{Im}\{z^R A_N(z)\}_{z=e^{j\hat{\omega}}} = 0 \text{ при } R \geq (N/2), \quad (3)$$

де. $A_N(z) = 1 + \sum_{i=1}^{N} a_i z^{-i}$

При цьому на підставі нової теорії ЛСК, запропонованої А.А. Ланне, коріння можуть розраховуватися по-різному залежно від параметра R. У рамках цієї теорії виділено кілька приватних випадків розрахунку ЛСК.

Модель розрахунку ЛСК для R = N

У даній роботі розглядається випадок, коли $R = N = 10$. Досить вирішити тільки одне рівняння порядку N, щоб по його корінням знайти всі коефіцієнти вихідного багаточлена.

Задається порядок моделі (ступінь апроксимуючого полінома) *ORD*. На вхід надходить відрізок сигналу (кадр) тривалості *FRM*:

$$SG = \{sg_0, sg_1, ..., sg_{FRM}\}. \quad (4)$$

Для усунення граничних ефектів виробляється згладжування ваговою функцією Хеммінга:

$$sh_i = sg_i \cdot (0{,}54 + 0{,}46 \cdot \cos(\frac{2\pi i}{FRM - i})), \quad (5)$$

де $i = 0...FRM$.

Виконується розрахунок коефіцієнтів передатної функції за допомогою методу найменших квадратів і алгоритму Левинсона-Дарбіна.

Первинна ініціалізація:

$$E_0 = \sum_{i=1}^{FRM} sg_i{}^2. \quad (6)$$

У циклі від 1 до *ORD* виробляються наступні обчислення:

- Обчислення коефіцієнта автокореляції:

$$R_i = \sum_{l=1}^{FRM} sg_l sg_{l-i}. \quad (7)$$

- Обчислення коефіцієнта відбиття:

$$r_i = \frac{R_i - \sum_{k=1}^{i} a_k{}^{(i-1)} R_{i-k}}{E_{i-1}}. \quad (8)$$

- Завдання первісного наближення:

$$a_i{}^{(i)} = r_i. \quad (9)$$

- Уточнення значень коефіцієнтів:

$$a_k{}^{(i)} = a_k{}^{(i-1)} - r_i a_{i-k}^{(i-1)}, \quad (10)$$

де $1 \le k \le i - 1$.

- Обчислення поточної помилки пророкування:

$$E_i = (1 - r_i^2) E_{i-1}. \quad (11)$$

- На останньому кроці циклу виходить остаточне рішення:

$$a_k = a_k^{(k)}, \text{ де } 1 \le k \le i - 1. \quad (12)$$

Далі розрахунок коефіцієнтів відбиття по формулах кратних дуг:

$$G = \{g_0, g_1, ..., g_{ORD}\}. \quad (13)$$

Пошук корінь полінома методом Ньютона:

$$G(x) = \sum_{i=0}^{ORD} g_i x^i. \quad (14)$$

Розрахунок набору ЛСК:

$$w_i = \arccos(x_i), \qquad (15)$$

де $i = 0...ORD - 1$.

Використання ЛСК як інформативні ознаки МС

При розрахунку ЛСК на тривалому МС (рис. 1), виробляється його розбивка на кадри з перекриттям. У результаті розрахунків виходить набір значень ЛСК (рис. 4).

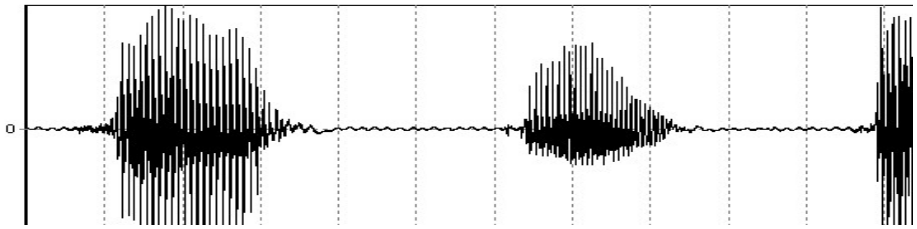

Рисунок 1 – Часова діаграма голосних фонем «а», «і», «о»

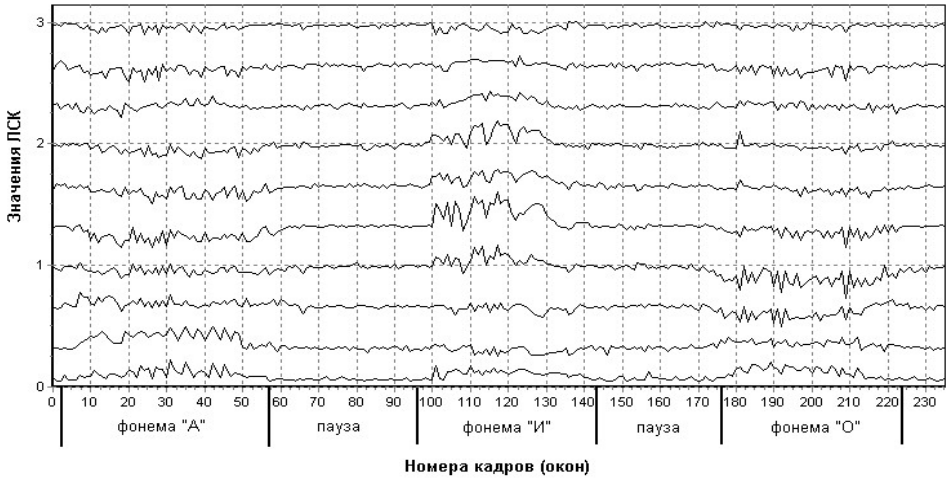

Рисунок 2 – Набір ЛСК для трьох фонем (порядок моделі – 10 корінь)

На рис 2 спостерігається порушення певних корінь при проголошенні фонем. Це обумовлено тим, що ЛСК несуть у собі спектральну інформацію про МС. Порушення корінь відбувається в області формантних частот голосних звуків.

Значення кожного ЛСК використовується як координата в N-мірному просторі ознак. На рис 3 і 4 показані образи трьох фонем у двомірному й тривимірному підпросторах. Сполучні лінії між точками відображають послідовність кадрів МС. Для деяких комбінацій ЛСК спостерігається впевнений поділ фонем – точки групуються в межах однієї області. Ця властивість дозволяє використовувати ЛСК як інформативні ознаки в СРМ.

Далі розглянемо моделі побудови словників еталонів, методики пошуку по них, проводиться критерій для оцінки вірогідності розпізнавання мовної команди.

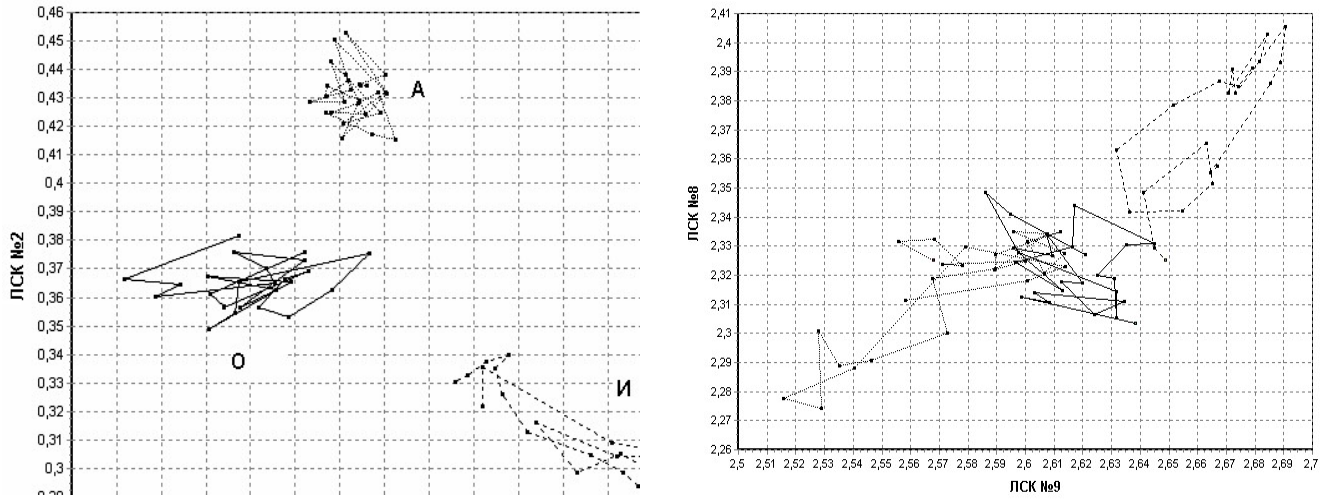

Рисунок 3 – Образи фонем у двомірному підпросторі ознак ЛСК

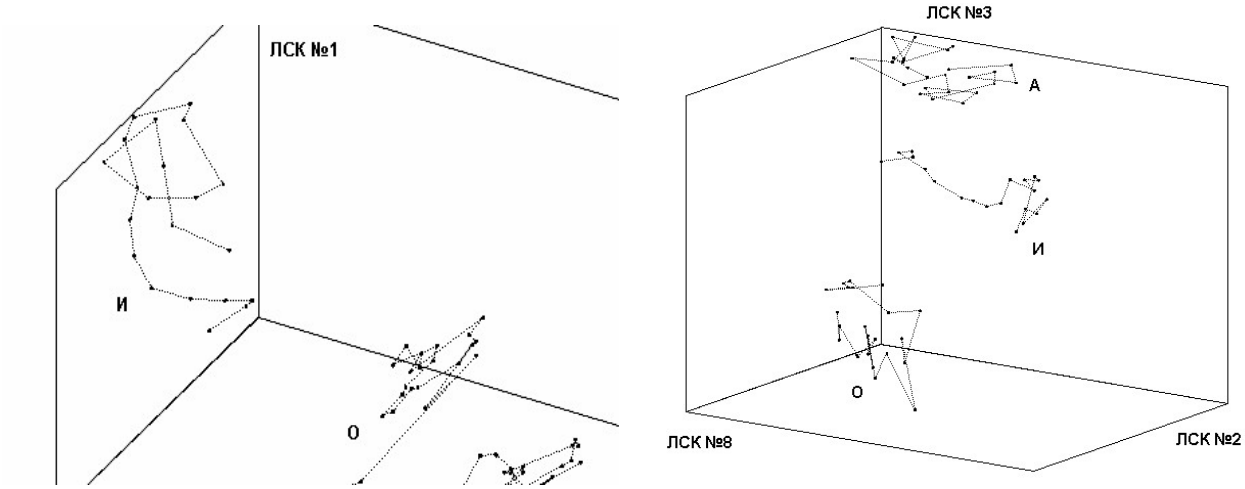

Рисунок 4 – Образи фонем у тривимірному підпросторі ознак ЛСК

Вибір методики формування словника еталонів

Розпізнавання мови шляхом виділення окремих фонем на практиці не принесло істотних результатів. Якщо повернутися до проблеми сприйняття мови людиною, то виявляється, що навіть досвідчені фонетисти із працею справляються із завданням розчленовування злитої мови на короткі сегменти. Найчастіше щоб розпізнати окрему фонему, слухачеві необхідно почути слово цілком або навіть трохи рядом вартих слів.

Відомо, що чим триваліше мовна одиниця, тим краще вона сприймається на слух. Виходячи із цього, для системи розпізнавання мовних команд як еталони найбільше доцільно використовувати цілі слова.

На рис 5 і 6 показані два слова, записані від різних дикторів. Слова представлені у вигляді точок у підпросторі двох ЛСК. Очевидно, що окремі фонеми досить важко виділити із цілого слова. Сполучні лінії (траєкторії точок ЛСК) відображають перебудування голосового тракту людини в процесі проголошення звуків. Для тих самих слів траєкторії візуально схожі. Ця властивість дозволяє використовувати набори векторів ЛСК, з обліком їхньої часової послідовності, як елементи навчальних словників.

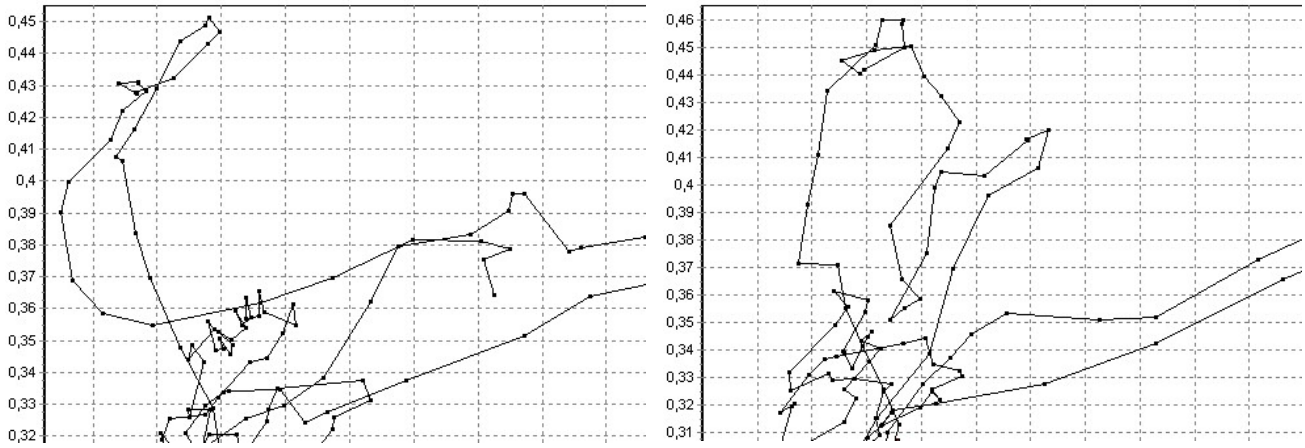

Рисунок 5 – Образи слова «повідомлення» для двох різних дикторів

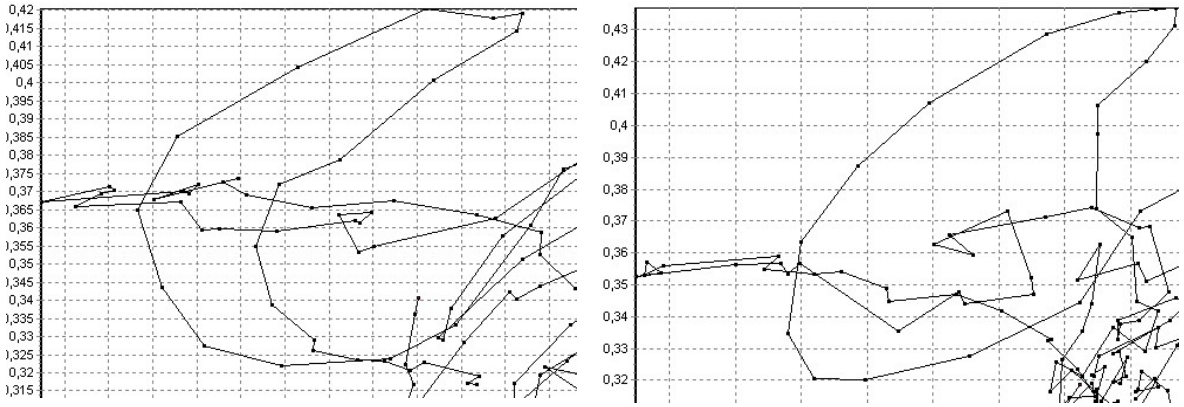

Рисунок 6 – Образи слова «наналаштування» для двох різних дикторів

Оцінка міри близькості між вхідним МС і еталоном виробляється за допомогою методу нелінійного часового вирівнювання (динамічного програмування). Це один з найбільш потужних і широко відомих математичних методів сучасної теорії керування, був запропонований наприкінці 50-х років американським математиком Р. Беллманом для рішення оптимізаційних завдань. Метод дозволяє порівнювати різні по тривалості зразки. Застосовно до мовних сигналів це означає, що порівняння з еталонами можливо практично незалежно від темпу мови.

Нехай рівняється два зразки сигналів, представлених у вигляді масиву векторів (для МС це набори ЛСК):

$$X = \left\{\bar{x}_0, \bar{x}_1, ..., \bar{x}_i, ..., \bar{x}_N\right\} \text{ та } Y = \left\{\bar{y}_0, \bar{y}_1, ..., \bar{y}_i, ..., \bar{y}_M\right\} \qquad (16)$$

Розходження між векторами двох образів визначається послідовністю станів C_K і позначається:

$$F() = C_0, C_1, ..., C_k, ..., C_K, \qquad (17)$$

де C_0 й C_K – початковий і кінцевий стани, $F()$ – функція часового вирівнювання, що проектує часову область одного образа на часову область іншого образа.

Метод ДП полягає в тім, що шукається така функція $F()$, при якій шлях зі стану C_0 в стан C_K, є оптимальним, тобто буде отримана мінімальна накопичена відстань між двома образами.

При побудові оптимального шляху, на кожному кроці алгоритму використовується основна формула ДП:

$$d_{i,j} = \min\begin{Bmatrix} d_{i,j-1} + r(\bar{x}_i, \bar{y}_j) \\ d_{i-1,j-1} + r(\bar{x}_i, \bar{y}_j) \\ d_{i-1,j} + r(\bar{x}_i, \bar{y}_j) \end{Bmatrix}, \text{ де } i = 0...N, j = 0...M. \quad (18)$$

Як відстань між векторами використовується зважена евклідова метрика:

$$r(\bar{x}, \bar{y}) = \sum_{k=0}^{N\_SEC-1} (x_k - y_k)^2, \quad (19)$$

де N_SEC – розмірність векторів ознак.

На виході процедури порівняння виходить деяке число (міра близькості), що представляє собою величину, зворотну ступені близькості між сигналами.

Процедура пошуку по словнику полягає в послідовному порівнянні вхідного сигналу з кожним з еталонів мовних команд. У табл. 1 показаний результат пошуку команди «повідомлення» у словнику із чотирьох командних слів. У результаті вхідний сигнал правильно розпізнаний системою. На рис 9 відображаються траєкторії найкоротших переходів по кадрам від еталонних сигналів до розпізнаваного. Дані по осі ординат нормовані по тривалості еталонних сигналів. По осі абсцис ідуть номера кадрів вхідного сигналу. Ділянки із крутими переходами між точками відображають автоматичне часове масштабування сигналів. Це відбувається, наприклад, якщо при проголошенні диктором розтягується голосний звук.

Таблиця 1 – Результат пошуку команди «повідомлення» у словнику із чотирьох командних слів

Еталон командного слова	Міра близькості
Повідомлення	1,85
Журнал	5,13
Диспетчер	6,82
Календар	4,33

Ідеальний випадок, коли розпізнаваний сигнал збігається з еталонним, являє собою діагональну східчасту траєкторію з лівого нижнього кута у верхній правий. На рис 9 для еталонів «журнал» і «календар» спостерігається істотне відхилення від діагоналі, що може бути додатковим критерієм для ухвалення рішення при розпізнаванні слів.

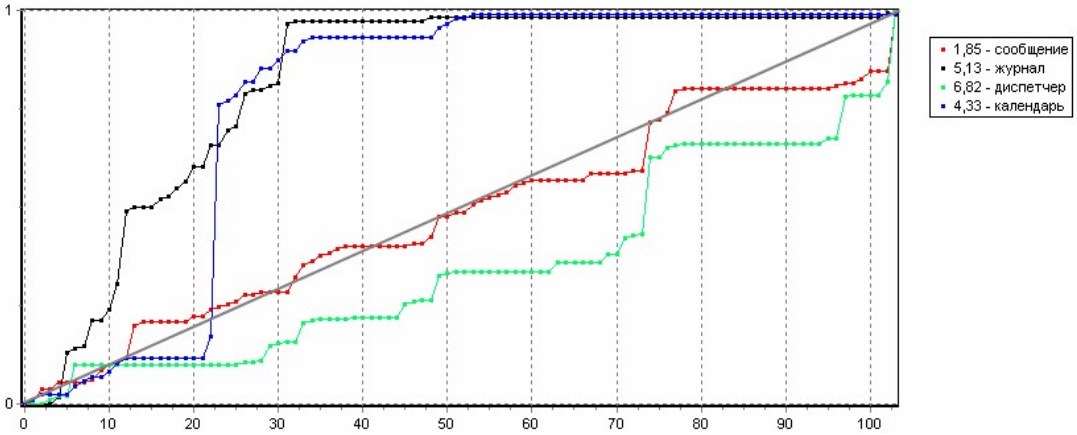

Рисунок 7 – Оптимальні траєкторії при порівнянні з еталонами

Перше ніж буде розпізнане ціле командне слово, на базі запропонованої моделі можливе **розпізнавання більше дрібних мовних одиниць**. Це дозволить скоротити область

пошуку в словнику й підвищити точність алгоритму. На рис 11 представлений результат розпізнавання цілого слова «режими» на словнику, що складається з набору складів. У якості одного з елементів словника використовується «еталон тиші» (позначений як «_»), що дозволяє без застосування додаткових алгоритмів виділяти паузи в мовних сигналах.

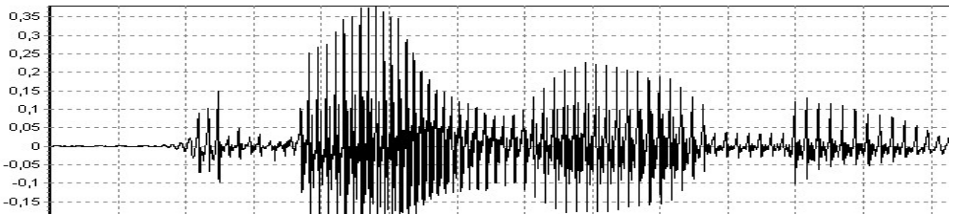

Рисунок 8 – Часова діаграма слова « ре-жи-ми»

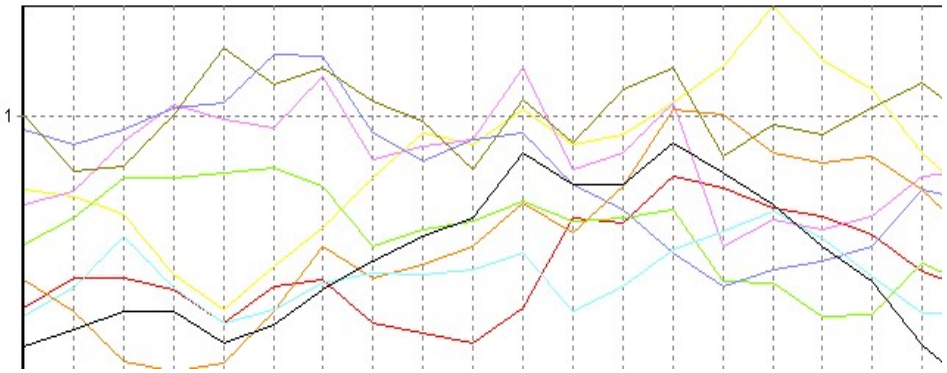

Рисунок 9 – Результат пошуку складів: «__ререре__жижижижи_мимимими____»

Вхідний сигнал розбивається на кадри по середній довжині еталонів. На графіку показані діаграми міри близькості до кожного з еталонів для всіх кадрів мовного сигналу. У результаті одержуємо послідовність розпізнаних складів. Шляхом згортки й подальша семантична обробка можлива одержання цілого слова. Дана методика може використовуватися для побудови СРМ на словниках більших обсягів.

Запропоновано рішення завдання пошуку слів у безперервному мовному потоці. Як елементи словника використовуються цілі слова. На вхід системи подається тривала ділянка мовного сигналу. У даному прикладі, фраза: «Чорна тойота номер три два один убік Київа» (рис 10).

Пошук іде без попередньої сегментації фрази на окремі слова. На рис 10 і 11 спостерігаються локальні мінімуми в області шуканих еталонних одиниць. На рис 12 яскраво вираженого мінімуму ні, тому що шукане слово («зелена») не було вимовлено в пропозиції. Співвідношення значення середньої міри близькості по всім кадрам МС і значення міри близькості на локальному мінімумі є критерієм, що дозволяє автоматично визначати, є присутнім чи взагалі шукане слово в аналізованій фразі.

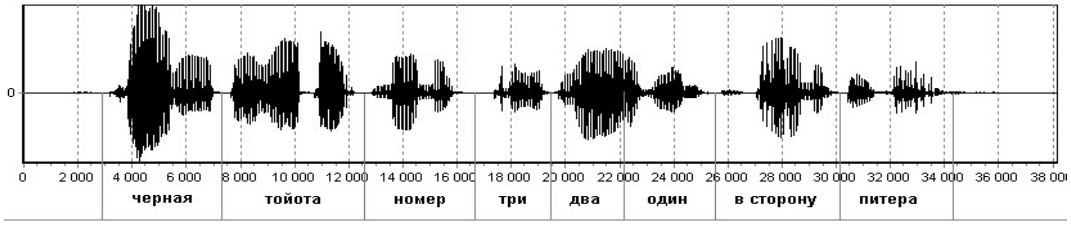

Рисунок 10 – Часова діаграма цілої фрази

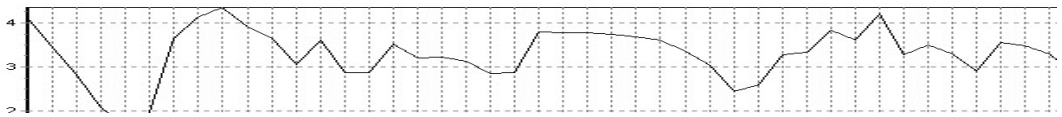

Рисунок 11 – Пошук слова «чорна» (співвідношення міри близькості = 0,5)

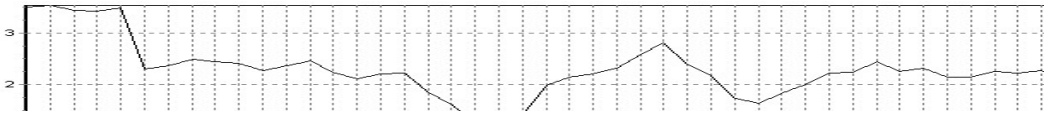

Рисунок 12 – Пошук слова «номер» (співвідношення міри близькості = 0,5)

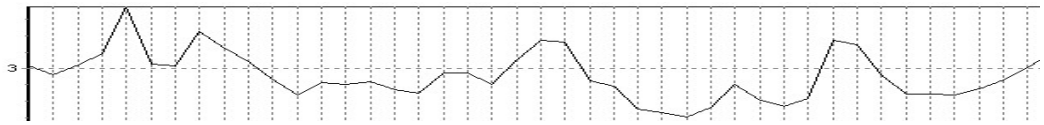

Рисунок 13 – Пошук слова «зелена» (співвідношення міри близькості = 0,8)

Критерій для оцінки вірогідності розпізнавання слів

При розпізнаванні мовних команд на базі словника з набору цілих слів, виходить таблиця зі значеннями міри близькості до елементів словника. Еталон з мінімальним значенням є шуканим – розпізнаним. Навіть якщо на вхід системи буде подане слово, що не входить у словник, у кожному разі буде отриманий результат – один з еталонів. Що приведе до помилки розпізнавання.

Запропоновано рішення завдання автоматичного відсівання помилкових спрацьовувань системи. Таблиця результатів розпізнавання нормується (табл. 2). Далі підраховується різниця в значенні міри близькості між першим і другим еталоном. У даному прикладі це 0,73. Якщо ця різниця не перевищує граничне значення 0,5, то слово буде вважатися нерозпізнаним і системою буде виданий запит на повторне уведення команди. Запропонований критерій дозволяє оцінювати вірогідність розпізнавання поточного слова.

Таблиця 2 – Таблиця результатів розпізнавання

Еталон	Міра близькості	Після нормування
Повідомлення	1,74	1,00
Пам'ять	3,01	1,73
Наналаштування	3,06	1,75
Годинники	3,45	1,98
Офіс	3,53	2,03
Режими	3,68	2,11
Засоби	4,06	2,33
Контакти	4,13	2,37
Теми	4,15	2,38
Журнал	4,25	2,44
Зв'язок	4,58	2,63
Календар	4,70	2,70

Оцінка впливу параметрів моделі ЛП на вірогідність розпізнавання

У ході досвідів, на словнику з 42 командних слів від 4 дикторів, варіювався розмір кадрів МС і ступінь апроксимуючого полінома (порядок моделі). На рис 14 і 15 наведені графіки відповідних залежностей. Найкраща вірогідність розпізнавання досягається, коли розмір вікна збігається з періодами основного тону МС. При зміні порядку моделі, максимум досягається на 10 коріннях, далі спостерігається пологий графік кривої.

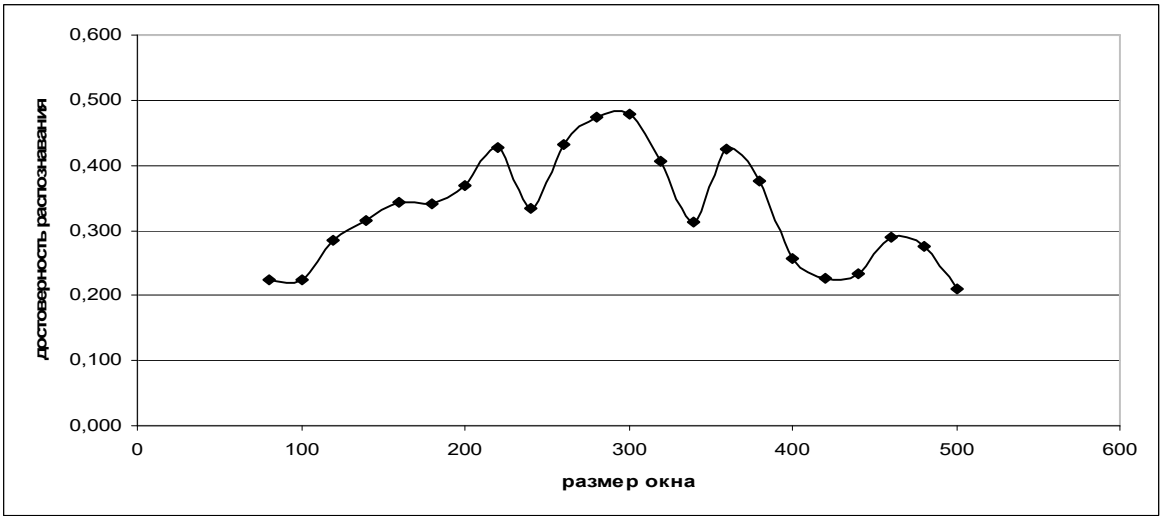

Рисунок 14 – Вплив розміру вікна на вірогідність розпізнавання

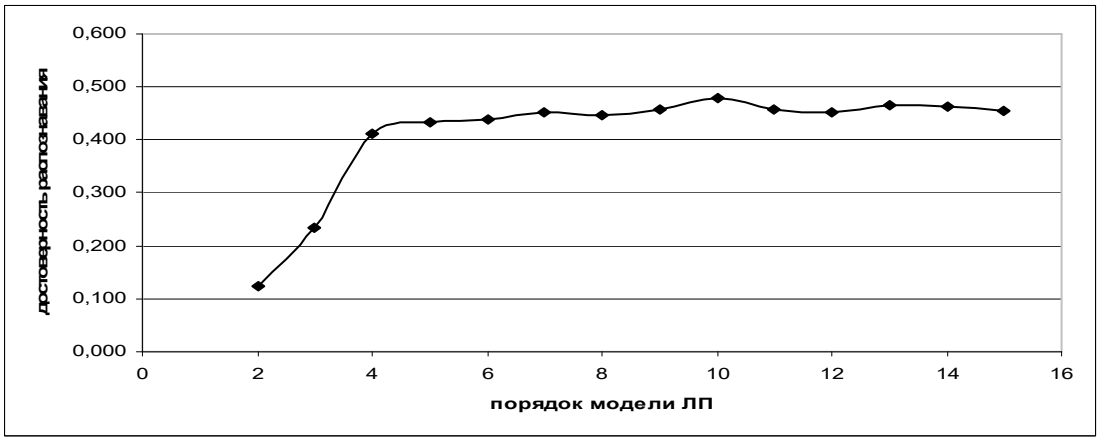

Рисунок 15 – Вплив порядку моделі на вірогідність розпізнавання

Результати досліджень погодяться із загальновідомими оцінками оптимальних параметрів моделі ЛП. Що підтверджує адекватність запропонованого критерію оцінки вірогідності розпізнавання мовних команд.

Оцінка якості сформованого словника еталонів

При використанні системи розпізнавання мовних команд в умовах підвищеної зашумленості або на вузькополосних каналах зв'язку, навіть на словниках малих обсягів (до 50 слів) можливо велика кількість помилок. Для збільшення надійності запропоновано використовувати корекцію словника еталонів.

Таблиця 3 – Аналіз того, наскільки елементи відрізняються друг від друга

Еталони	Календар	Контакти	Наналашту вання	Офіс	Пам'ять	Режими	Зв'язок	Повідомле ння	Засоби	Теми	Годинники
Журнал	2,64	2,59	3,11	2,45	3,47	2,73	3,34	4,73	2,8	2,39	3,54
Календар		2,54	3	2,78	2,96	2,4	2,72	4,24	3,15	2,28	3,47
Контакти			2,77	2,87	3,32	3,09	3,17	4,35	3,1	2,52	3,49
Налалашт				2,24	2,53	3,15	3,41	3,07	3,46	2,82	3,33

ування											
Офіс					2,36	2,71	2,59	3,48	3,31	2,46	3,2
Пам'ять						3,51	2,66	2,6	4	3,06	3,7
Режими							2,59	3,44	2,72	*1,95*	3,23
Зв'язок								4,29	3,01	2,4	2,98
Повідомлення									4,16	3,9	3,23
Засоби										2,4	3,2
Теми											2,88
Середнє	2,83	3,07	2,99	2,77	3,11	2,87	3,01	3,77	3,21	2,64	3,30

Після формування словника виробляється аналіз того, наскільки елементи відрізняються друг від друга (табл. 3). Підраховується середнє значення (у даному прикладі 3,04). Якщо деякі елементи словника занадто схожі один на одного (міра близькості менше порога, рівного 2), то пропонується замінити один з еталонів, наприклад, синонімом. Після цього виробляється повторний аналіз словника.

У даному прикладі (табл. 3), після заміни одного зі схожої пари слів «теми» або «режими», відсоток правильно розпізнаних команд збільшився на 9,8%.

У табл. 4 показані результати розпізнавання для чотирьох варіантів первинних ознак:

- LSP – лінійних спектральних корінь (пари)
- LPC – коефіцієнти лінійного пророкування
- PLP – коефіцієнти перцептивного пророкування
- MFCC – мел-кепстральні коефіцієнти

Таблиця 4 – Результати розпізнавання

Диктори	Час розрахунку, мс				Вірогідність				% помилок			
	LSP	LPC	PLP	MFCC	LSP	LPC	PLP	MFCC	LSP	LPC	PLP	MFCC
Чоловік1	9,05	3,05	13,09	13,02	1,91	1,30	1,49	1,13	0,00	4,76	2,38	2,38
Чоловік2	8,45	2,85	11,91	12,30	0,78	0,41	0,78	0,69	4,76	23,80	23,80	19,05
Жінка1	8,31	2,24	12,20	11,50	1,46	0,90	1,07	1,06	3,84	19,23	7,69	7,69
Жінка2	7,95	2,15	10,30	10,23	1,35	0,75	0,99	0,95	2,86	8,57	5,71	5,71
Середнє	8,44	2,57	11,88	11,76	1,38	0,84	1,08	0,96	2,87	14,09	9,90	8,71

Видно, що для ЛСК спостерігається мінімальний відсоток помилок 2,87% і максимальний ступінь вірогідності 1,38. При цьому час розрахунку порівнянний з іншими методами. Що дозволяє говорити про можливість успішного застосування даних ознак у більше складних системах розпізнавання мови.

Для реалізації алгоритму розпізнавання мови, був обраний алгоритм MFCC.

Опис алгоритму MFCC

Існують безліч методів розпізнавання мови, у переважній більшості випадків вони засновані на методах статистичного аналізу й теорії ймовірностей (Hidden Markov Model, Gaussian Mixture Model і т.п.). Як відомо, компанія google надає безкоштовний сервіс по розпізнаванню коротких мовних повідомлень. На основі цього сервісу було навіть запропоноване розпізнавання мови за допомогою мікроконтролера. Однак, виникає питання: є чи можливість зробити свою систему розпізнавання мови, нехай навіть на досить обмеженому за розміром словнику, без використання «зовнішніх» сервісів, при цьому щоб вона працювала швидко й із прийнятною якістю?

Основна ідея

Отже, для розпізнавання мови будемо використовувати MFCC. Щоб не вдаватися в подробиці скажу, що відноситися до них варто лише як до деякого фільтра, на вході якого – фонограма, на виході – набір векторів (коефіцієнти), що ми й будемо розпізнавати як деяке слово або набір слів.

Справедливості заради варто відзначити, що існують безліч інших акустичних ознак, що використовуються для розпізнавання мови: Perceptual linear predictive (PLP), Linear prediction cepstral coefficient (LPCC), Linear frequency cepstral coefficients (LFCC).

Основна ідея полягає у використанні лінійного дискрімінантного аналізу для ідентифікації слова. Однак, він застосовний лише для векторів однакової розмірності. Т.к. слова можуть бути різної довжини, виникає питання: яким образом перетворити послідовність довільного числа MFCC-векторів у вектор фіксованої розмірності?

Можна надійти в такий спосіб: знаходити місця «згущення» розподілу цих векторів і в якості результуючого вектора брати конкатенацію векторів, що є центрами «згущень».

Такий конкатенований вектор будемо називати супервектором середніх, а самі центри – середніми значеннями. При цьому в якості «відправної точки» будемо використовувати супервектор середніх, отриманий на всіх MFCC-векторах всієї бази навчання.

Перетворивши в такий спосіб послідовність MFCC-векторів в один супервектор середніх фіксованої розмірності, ми можемо застосовувати різні методи класифікації.

Очевидний принциповий недолік такого підходу: не враховується динаміка розподілу MFCC-ознак за часом, отже, система апріорі не здатна розрізняти, приміром, слова «головриба» і «абирвалг», тому що загальний розподіл MFCC-векторів таких слів буде приблизно однаковим (відповідно, центри «згущень» будуть збігатися).

Опис алгоритму

Як базу навчання будемо використовувати безліч файлів, кожний з яких являє собою набір MFCC-векторів, отриманих з фонограми із записом того або іншого слова. При цьому файли із записом того самого слова повинні бути об'єднані в одну групу.

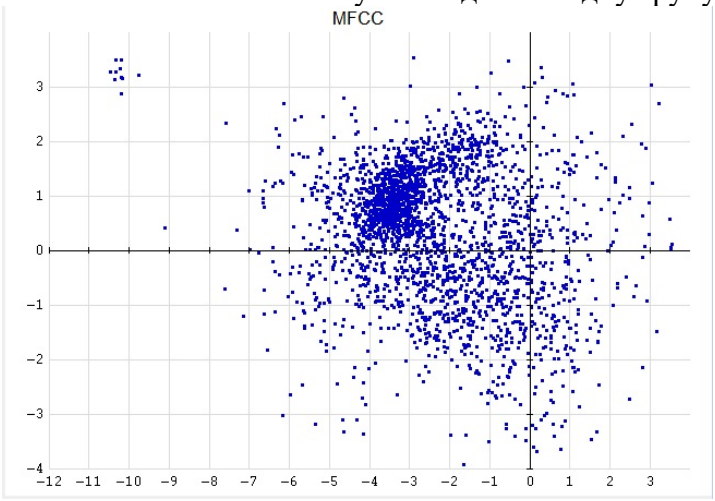

Рисунок 16 – Розподіл перших двох компонентів MFCC-векторів всієї бази навчання

Алгоритм складається з наступних етапів:

1. Знаходимо супервектор середніх для всієї бази навчання за допомогою алгоритму K-середніх.

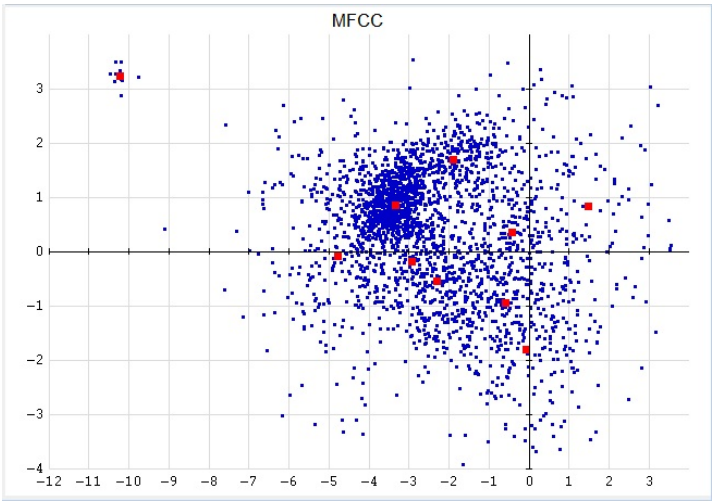

Рисунок 17 – Приклад роботи алгоритму K-середніх для K=10, де більші червоні квадрати і є шукані середні значення

2. Для кожного файлу бази знаходимо власні середні значення за формулою:

Mk = a * Mk0 + (1 – a) * Mk', k = 1:K

де Mk0 – середнє значення, знайдене в п.1, Mk' – середнє значення, отримане в результаті застосування однієї ітерації алгоритму K-середніх для MFCC-векторів файлу з використанням як початкове значення Mk0,

a = R/(R + Nk),

де R – коефіцієнт «чутливості», Nk – число MFCC-векторів, що відповідають середньому значенню Mk'.

3. Знайдені в такий спосіб середні значення будемо називати адапритованими середніми значеннями.

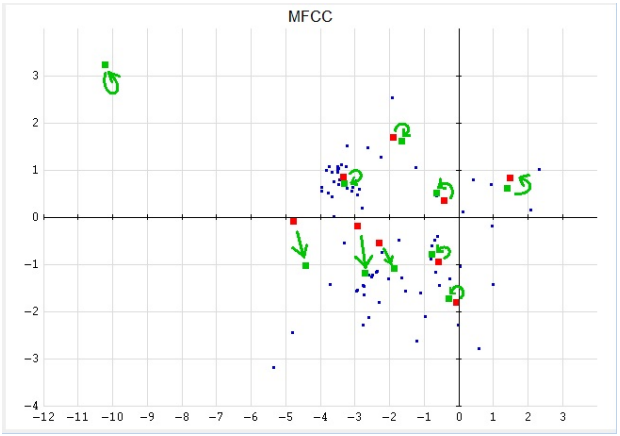

Рисунок 18 – Приклад адаптованих середніх значень для файлу

4. Маючи тепер замість вихідних фонограм адаптовані супервектора середніх, проводимо LDA для N класів (кожний клас відповідає одному слову).

5. У результаті ми повинні одержати матрицю, що складається з векторів нового базису, при проекції на який вихідні адаптовані супервектора середніх повинні досить добре розділятися.

6. Проектуємо всі адаптовані супервектора середніх на новий базис і знаходимо середні значення й СКВ (середнє квадратичне відхилення) проекцій для кожного класу.

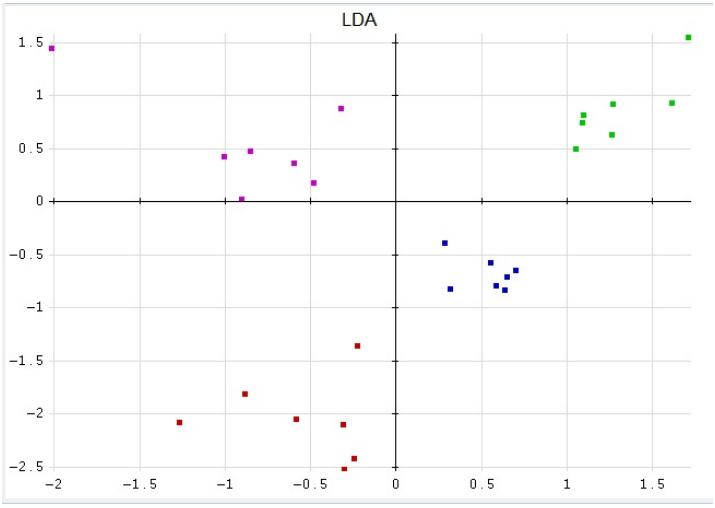

Рисунок 19 – Приклад для N=4

7. Для визначення приналежності тестової фонограми тому або іншому класу (тобто розпізнавання), виконуємо для неї пп. 2 і 4, далі знаходимо відстані отриманої проекції до середніх значень всіх класів (можна додатково нормувати їх на відповідне СКВ). Мінімальна відстань і буде відповідати класу, до якого належить тестова фонограма.

Реалізація

Створення власної системи розпізнавання слів складається з наступних етапів:

1. Запис фонограм для навчання й тестування.

Для запису можна скористатися будь-якою програмою, що вміє записувати звук і зберігати його у форматі WAVE. Я рекомендую використовувати безкоштовну програму Audacity.

Розроблена система не вміє виділяти мовні сегменти, тому при записі потрібно намагатися, щоб у фонограмі була присутня тільки мова. Чим якісніше використовується мікрофон, тим якіснішою виходить система. Записувати необхідно в моно-режимі із частотою дискретизації 16000.

2. Побудова MFCC-векторів.

Для побудови MFCC-векторів можна використовувати безкоштовну бібліотеку SPro 5.0. Я взяв на себе відповідальність, небагато перебрав цю бібліотеку, виправив парочку помилок і зробив складання програми sfbcep.exe під windows (див. папку ../ spro-5.0). 32-розрядна версія цієї програми лежить у папці ../tools. Для побудови MFCC-векторів я використовував наступні параметри:

sfbcep.exe -format=wave – sample-rate=16000 -mel – freq-min=0 – freq-max=8000 – fft-length=256 -length=16.0 -shift=10.0 – num-ceps=13 [вхідний WAVE-Файл] [вихідний файл із MFCC-векторами]

Навчання й тестування системи

Для навчання й тестування системи я написав програму wrsystem мовою visual C++. Реалізація алгоритму LDA була запозичена з бібліотеки ALGLIB.

Програма wrsystem має два режими роботи: навчання (у випадку наявності параметра -learn) і тестування. Ця програма приймає на вхід три основних параметри:

– Шлях до файлу з описом бази навчання (тестування) (параметр -base). Приклад файлу з описом бази лежить у папці ../base, також опис формату можна подивитися, запустивши програму з параметром -help.

– Шлях до бінарного файлу, що зберігає результат навчання системи (параметр -system). У режимі навчання цей файл створюється, у режимі тестування – зчитується.

– Шлях до файлу, у який записуються результати тестування системи на зазначеній базі: матриця переплутування й значення WER (Word Error Rate) (параметр -test_results**).**

Результати експериментів

Як експеримент я створив систему, що вміє розпізнавати 14 слів, записаних моїм голосом. Для навчання системи я записав кожне слово 4-5 разів, а для тестування – 7 разів. Разом база навчання містить 63 файлу, а база тестування – 98. Використовувалися наступні параметри при навчанні:

- Кількість середніх значень: 10.
- Коефіцієнт «чутливості» при адаптації: 20.
- Розмірність проекції: 20.
- Використання нормалізації на СКВ: відсутній.

Результат тестування на базі навчання показав рівень помилки розпізнавання слів (WER) 1,6%, а на базі тестування 5,1%.

На що варто звернути увагу

Для того, щоб будь-яка система (включаючи описану тут) могла якісно розпізнавати мову будь-якої людини, необхідно мати величезну базу навчання із записом всіх слів, вимовлених різними людьми в різному емоційному стані з використанням різних записуючих пристроїв (телефон, мікрофон, що підслухує пристрій і т.п.). Тобто система, що ви навчите, використовуючи тільки свій голос і тільки вашу домашню гарнітуру, напевно не буде працювати для ваших знайомих і навіть для вас, якщо ви будете використовувати який-небудь інший мікрофон.

Структурна схема розробленої системи зображена на рисунку 20. На ній показано структуру системи дистанційного голосового керування робототехнічним комплексом.

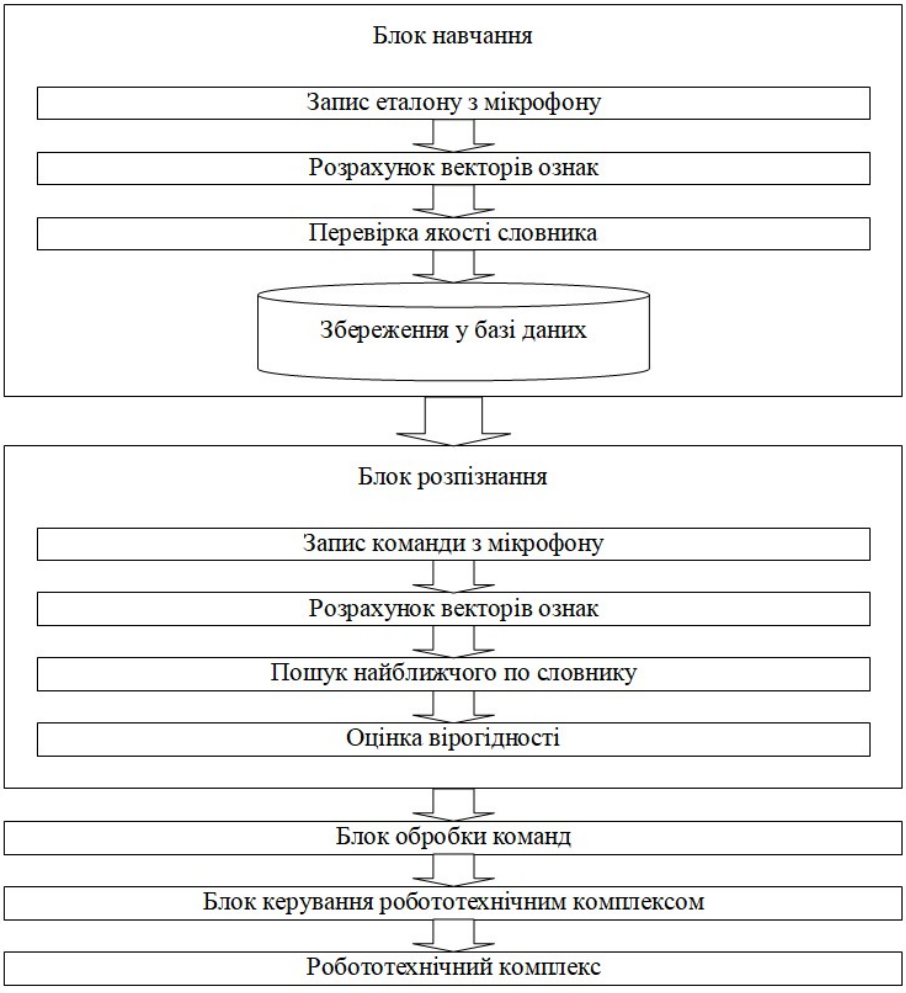

Рисунок 20 – Структурна схема системи

Висновки. У статті наведені теоретичне узагальнення й рішення наукового завдання дослідження методів дистанційного голосового керування робототехнічним комплексом.Рішення даного завдання полягало у вирішенні наступних задач: Був проведений огляд існуючих систем дистанційного голосового керування робототехнічним комплексом. Досліджена система дистанційного голосового керування робототехнічним комплексом;На основі отриманих результатів досліджень створена програмна реалізація системи дистанційного голосового керування робототехнічним комплексом. Розроблені під час виконання випускної кваліфікаційної роботи за другим (магістерським) рівнем вищої освіти алгоритми дозволяють успішно вирішувати завдання дистанційного голосового керування робототехнічним комплексом.

Список літератури

1. Smirnova, T., Gnatyuk, S., Yudin, O., Sydorenko, V., Polozhentsev, A., «The Model for Calculating the Quantitative Criteria for Assessing the Security Level of Information and Telecommunication Systems». CEUR Workshop Proceedings Volume 3156, 2022, Pages 390-399.
2. Smirnova T., Gnatyuk S., Berdibayev R., Avkurova Zh., Iavich M. «Cloud-Based Cyber Incidents Response System and Software Tools». Communications in Computer and Information Science, 2021, vol 1486. Springer, Cham. pp 169-184.
3. Smirnov O., Kuznetsov A., Kiian A., Kuznetsova T. «Non-binary constant weight coding technique». CEUR Workshop Proceedings. Volume 2740, 2020, Pages 102-114.
4. Smirnov O., Alimseitova Zh., Adranova A., Akhmetov B., Lakhno V., Zhilkishbayeva G. «Models and algorithms for ensuring functional stability and cybersecurity of virtual cloud resources». Journal of theoretical and applied information technology Vol.98. No 21, 2020, P. 3334-3346.
5. Smirnov O., Kuznetsov A., Kiian A., Cherep A., Kanabekova M., Chepurko I. «Testing of code-based pseudorandom number generators for post-quantum application». 2020 IEEE 11th International Conference on Dependable Systems, Services and Technologies (DESSERT), Ukraine, Kyiv, May 14-18. 2020. P. 172-177.
6. Smirnov O., Kuznetsov A., Pushkar'ov A., Serhiienko R., Babenko V., Kuznetsova T., «Representation of Cascade Codes in the Frequency Domain». In: Radivilova T., Ageyev D., Kryvinska N. (eds) Data-Centric Business and Applications. Lecture Notes on Data Engineering and Communications Technologies, vol 48. Springer, Cham. 2021. pp 557-587.
7. Smirnov, O., Markovets, O. Vovk, N., Turchyn, Y., «Model of informational support for social network administrators' content creation». CEUR Workshop Proceedings Volume 2616, 2020, Pages 125-136.
8. Smirnov, O., Drieieva, H., Drieiev, O., Polishchuk, Y., Brzhanov, R., Aleksander, M. «Method of fractal traffic generation by a model of generator on the graph». CEUR Workshop Proceedings Volume 2616, 2020, Pages 366-379.
9. Smirnov, O., Drieieva, H., Drieiev, O., Simakhin, V., Bondar, S., Odarchenko, R. «Managing multifractal properties of the binary sequence generated with the Markov chains», CEUR Workshop Proceedings Volume 2608, 2020, Pages 633-645.
10. Smirnov O. Kuznetsov A., Zaicenko Yu., Pastukhov M., Oleshko O., Kuznetsova K., «Formation of Discrete Signals with Special Correlation Properties». International Conference on Information and Telecommunication Technologies and Radio Electronics, UkrMiCo 2019; Odessa; Ukraine; 9-13 September 2019. P.22-28.
11. Smirnov, O., Kuznetsov, A., Kolovanova, I., Kuznetsova, T., «Noise immunity of the algebraic geometric codes». International Journal of Computing; 2019, Volume 18, Issue 4 – Research Institute for Intelligent Computer Systems – 2019. – P. 393-407.
12. Smirnov, O., Kuznetsov, A., Reshetniak, O., Ivko, N., Katkova, T., Kuznetsova, T., «Generators of Pseudorandom Sequence with Multilevel Function of Correlation». 2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T), Kyiv, Ukraine, 8 – 11 October 2019 . P.517-522.
13. Smirnov, O., Odarchenko, R., Abakumova, A., Usik, P., Kundyz, M., «QoE optimization technique for media delivery in 5G networks». 2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T), Kyiv, Ukraine, 8 – 11 October 2019. P.597-601.
14. Smirnov, O., Krasnobayev, V., Yanko, A., Kuznetsova, T. «Methods of nulling numbers in the system of residual classes». CEUR Workshop Proceedings, Vol 2588, P. 90-106, 2019.
15. Smirnov, O., Kuznetsov, A., Kovalchuk, D., Averchev, A., Pastukhov, M., Kuznetsova, K., «Formation of Pseudorandom Sequences with Special Correlation Properties», 2019 3rd International Conference on Advanced Information and Communications Technologies, AICT -2019/ Lviv, Ukraine, 2-6 July, 2019, P. 395-399.
16. Smirnov, O., Kuznetsov, A., Kiian, A., Zamula, A., Rudenko, S., Hryhorenko, V., «Variance Analysis of Networks Traffic for Intrusion Detection in Smart Grids», 2019 IEEE 6th International Conference On Energy Smart Systems (2019 IEEE ESS), Kyiv, Ukraine April 17-19, 2019 P. 353-358.
17. Smirnov, O., Kuznetsov, A., Kavun, S., Babenko, B., Nakisko, O., Kuznetsova, K., «Malware Correlation Monitoring in Computer Networks of Promising Smart Grids», 2019 IEEE 6th International Conference On

Energy Smart Systems (2019 IEEE ESS), Kyiv, Ukraine April 17-19, 2019 P. 347-352.
18. Smirnov, O., Kuznetsov, A., Kovalchuk, D., Pastukhov, M., Kuznetsova, K., Prokopovych-Tkachenko, D., «Discrete Signals with Special Correlation Properties», CEUR Workshop Proceedings Volume 2353, CEUR Workshop Proceedings 2019, Pages 618-629.
19. Smirnov A.A., Kuznetsov A.A., Danilenko D.A., Berezovsky A., «The statistical analysis of a network traffic for the intrusion detection and prevention systems», Telecommunications and Radio Engineering. – Volume 74, Issue 1. – Begel House Inc. – 2015. – P. 61-78.
20. Вінтенко Б.Ю., Смірнов О.А., Коваленко А.С., Смірнов С.А., Буравченко К.О. «Дослідження вимог міжнародних стандартів IEC60880 та IEC62138 з розробки програмного забезпечення інформаційно-керуючих систем АЕС, важливих для безпеки». Системи управління, навігації та зв'язку, 2023, вип. 3(73), С. 155-166.
21. Аль-Мудхафар Акіл Абдулхуссейн М., Смірнова Т.В., Буравченко К.О., Смірнов О.А. «Метод оцінки та підвищення користувальницького досвіду абонентів в програмно-конфігурованих мережах на основі використання машинного навчання». Сучасні інформаційні системи, 2023, том 7, № 2, С. 49-56.