УДК 004.934

К.М. Ялова, К.В. Яшина

АЛГОРИТМ ДИНАМІЧНОГО ТРАНСФОРМУВАННЯ ЧАСУ В ЗАДАЧІ РОЗПІЗНАВАННЯ МОВНОГО СИГНАЛУ

Анотація. В роботі представлено результати підвищення ефективності застосування алгоритму динамічного трансформування часу в задачі розпізнавання мовного сигналу. Авторами запропоновано використати в якості вхідного вектору ознак нормалізовані мел-частотні кепстральні коефіцієнти. Для підвищення продуктивності системи розпізнавання мовного сигналу здійснюється предрозрахунок гребінки фільтрів мел-коефіцієнтів та застосовується мультипоточність. З метою оцінки якості отриманих результатів, розроблено командну дикторозалежну систему розпізнавання мовного сигналу, реалізовану у вигляді плагіну для текстового редактору. Адекватність представлених рішень доведено шляхом порівняння отриманих результатів розпізнавання вхідного мовного сигналу із результатами робіт інших авторів. Адекватність представлених рішень доведено шляхом порівняння отриманих результатів із результатами робіт інших авторів.

Ключові слова: Алгоритм динамічного трансформування часу, розпізнавання мовного сигналу, мел-коефіцієнти.

Постановка проблеми. Найбільш ефективними засобами взаємодії людини з машиною є ті, що реалізуються природним для неї чином: через візуальні образи і мову [1]. Створення мовних інтерфейсів може знайти застосування в системах різного призначення: голосове управління для людей з обмеженими можливостями, автовідповідачі, оброблення дзвінків в автоматичному режимі тощо. Однак, не дивлячись на стрімко зростаючі обчислювальні потужності, створення систем розпізнавання мови залишається надзвичайно складною проблемою. Це обумовлюється як її міждисциплінарним характером (необхідно володіти знаннями з лінгвістиці, цифровій обробці сигналів, акустиці, розпізнаванні образів тощо), так і високою обчислювальною складністю розроблених алгоритмів. Останнє накладає суттєві обмеження на системи автоматичного розпізнавання

мови – на обсяг оброблюваного словника, швидкість отримання відповіді і його точність.

Аналіз публікацій по темі дослідження. Алгоритми, що застосовується при розпізнаванні вхідного мовного сигналу, різноманітні: приховані моделі Маркова (ПММ), динамічне програмування, нейронні мережі тощо. Застосуванню алгоритму динамічного трансформування часу (Dynamic Time Warping – DTW, ДТЧ) присвятили свої роботи такі автори, як: Каре С.В. та Навале В.С. [2]. Автори роботи [3] запропонували застосувати алгоритм ДТЧ в рамках мовно-залежної та дикторозалежної системи розпізнавання мовного сигналу. В роботі [4] здійснено порівняльна характеристика ПММ та алгоритму ДТЧ в задачі розпізнавання мовного сигналу. Автори Малькова Е. С. та Шабалина О. А. в роботі [5] застосували алгоритм ДТЧ в системі автоматизованого виявлення дефектів вимови. Класичний опис алгоритму ДТЧ та пов'язаних із ним методів розпізнавання образів наведено в роботі [6]. Не зважаючи на значну кількість наукових робіт, що присвячені проблемі розпізнавання мовного сигналу, задачі підвищення якості розпізнавання мови та розробки нових підходів до реалізації систем автоматичного розпізнавання мовного сигналу, залишаються актуальними науково-практичними завданнями.

Формулювання мети статті. Метою статті є розробка підходу до застосування алгоритму ДТЧ, швидкого перетворення Фур'є, Мел-частотних кепстральних коефіцієнтів, отриманих зі спектру кожного вікна до розв'язання задачі розпізнавання мовного сигналу. В ході досягнення поставленої мети авторами були здійснені наступні етапи роботи:

- проаналізовано особливості застосування алгоритму ДТЧ до розпізнавання мовного сигналу. Застосовано швидке перетворення Фур'є (ШПФ) до аналізу вхідного сигналу та Мел-частотні кепстральні коефіцієнти (МЧКК) для побудови вхідного вектору ознак;

- спроектовано модель системи автоматичного розпізнавання мови (САРМ) із характеристиками: розроблювана система є командною, залежною від диктора із типом структурної одиниці – фразою-командою диктора, який задає команду для роботи з текстом в рамках текстового редактору;

- розроблено програмний модуль розпізнавання мови для текстового редактори у вигляді плагіну, що дозволяє оцінити якість запропонованих рішень та встановити похибку розпізнавання.

Основна частина. Під розпізнанням мови розуміють процес трансформації мовного сигналу в цифрову інформацію. САРМ – це інформаційна система, що перетворює вхідний мовний сигнал в розпізнане повідомлення. При цьому повідомлення може бути представлено як у формі тексту цього повідомлення, так і одразу перетворено в зручну для його подальшої обробки форму. САРМ класифікуються за такими ознаками: розмір словника (обмежений набір слів або великий словник), залежність від диктора (дикторозалежні або дикторонезалежні), тип мови (злита, роздільна), призначення (системи диктування, командні системи), алгоритм розпізнавання, що використовується, тип структурної одиниці (фрази, слова, фонеми тощо), принципи виділення структурних одиниць (ШПФ, МЧКК) [1].

Загальна схема процесу розпізнавання мовного сигналу наступна: отримання оцифрованого звукового сигналу, який поступає з мікрофону користувача, отримання спектру вікна сигналу, розрахунком МЧКК та порівняння вектору ознак вхідного сигналу з шаблонами.

Для аналізу звукової хвилі, скористаємося теоремою Фур'є, де зазначено, що будь-яке складне періодичне коливання можна розділити на суму простих гармонійних коливань [7]. В результаті отримаємо набір амплітуд, фаз і частот для кожного синусоїдального компонента хвилі:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi}{N}kn},$$

де n – кількість значень сигналу, к – кількість частот, xn – значення сигналу в певних точках часу, x_k – комплексні амплітуди синусоїдальних сигналів, з яких складається початковий сигнал, k=0,...K-1 – індекс частоти, n=0,...,N-1 – дискретні часові точки, в яких проводилося вимірювання сигналу. Для відновлення дискретного сигналу зі спектру використаємо зворотне перетворення Фур'є:

$$x_n = \frac{1}{K}\sum_{k=0}^{K-1} X_k e^{-\frac{2\pi}{K}kn}.$$

Для побудови спектрограми, що надає змогу прослідкувати за змінами спектру сигналу в часі на всьому звуковому відрізку застосуємо віконне перетворення Фур'є – спектр розраховується від послідовних вікон сигналу, кожне з яких перекриває частину попереднього вікна. Отримані спектри є стовпцями у спектрограмі. Для значного прискорення процесу, було застосовано алгоритм швидкого перетворення Фур'є (ШПФ), що працює з комплексними числами і розмірами перетворень, які є ступенями двійки [7]. Якщо частота тону збігається з однією з частот сітки ШПФ, то спектр буде мати єдиний гострий пік, що вкаже на частоту і амплітуду тону. Якщо ж частота тону не збігається ні з однією з частот сітки, то ШПФ сформує тон із наявних в сітці частот, скомбінованих із різними вагами. Графік спектру при цьому розмивається по частоті, що є небажаним, оскільки воно може закрити собою слабші звуки на сусідніх частотах. Для зменшення ефекту розмиття спектру, сигнал перед обчисленням ШПФ множиться на вагові вікна – функції, що спадають до країв інтервалу – це зменшує розмиття спектру за рахунок деякого погіршення частотного розділення. В даній роботі було застосовано вікно Хеммінга (рис. 1), яке можна визначити із наступного рівняння:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right).$$

де N – ширина вікна, рівень бокових пелюсток: -42дБ.

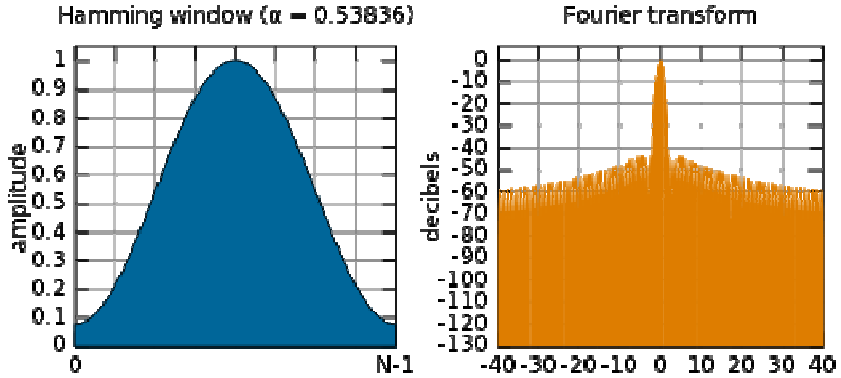

Рисунок 1 – Схема вікна Хеммінга

Застосування вікна Хеммінкга зменшує рівень розмиття спектру приблизно на 40 дБ щодо головного піку. Вхідний сигнал розділявся на проміжки по 20-40мс, оскільки розмір такого проміжку є достатнім для отримання надійної спектральної оцінки. Початковий стан спектральної оцінки містить забагато інформації

для автоматичного розпізнавання мовного сигналу. Тому доцільно розділити увесь спектр на ділянки, які стануть проекціями частот у відповідному діапазоні на Мел-шкалу:

$$M(f) = 1127 * \ln\left(1 + \frac{f}{700}\right),$$

де f – частота, яка проектується на Мел-шкалу.

$$f(i) = Floor\left((n+1) * \frac{h(i)}{R}\right),$$

де n – розмір вікна ШПФ, R – частота дискретного сигналу.

Отримані значення переводяться назад до частотного вигляду за допомогою наступного рівняння:

$$h(m) = 700 * \left(\exp\left(\frac{m}{1127}\right) - 1\right),$$

де m – проекція частоти на Мел-шкалі.

Після чого було сформувано банк фільтрів, щоб отримати представлення про те, скільки енергії існує в різних частостних областях сигналу:

$$H_m(k) = \begin{cases} 0, \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)}, \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)}, \\ 0, \end{cases} H_m(k) = \begin{cases} 0, & k < f(m-1). \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) \le k \le f(m). \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)}, & f(m) \le k \le f(m+1). \\ 0, & k > f(m+1). \end{cases}$$

де m – кількість МЧКК, k – поточна частота.

Дискретне косинусне перетворення (ДКП) енергії логарифмів банку фільтрів розраховується наступним чином:

$$X_k = \sum_{n=0}^{N-1} x_n * \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right].$$

Потрібно зазначити, що лише половина ДКП коефіцієнтів використовуються як характеристики сигналу. Це пояснюється тим, що вищі ДКП коефіцієнти є швидкими змінами в енергії фільтрувального банку, і виявляється, що ці швидкі зміни фактично знижують продуктивність системи розпізнавання, тому можна отримати невелике поліпшення, відкинувши їх [7].

В процесі розпізнавання мови найскладніше полягає в здійсненні процедури порівняння двох мовних елементів, які характеризуються ще і протяжністю в часі, тому існує досить багато таких

процедур і методів. Алгоритм ДТЧ – це алгоритм визначення оптимальної послідовності трансформацій (перетворень) часу між двома часовими рядами шляхом розрахунку значень деформації між ними. Припустимо, що існує дві числові послідовності (a1, a2, …, an) та (b1, b2, …, bm). Для розрахунку локальних відхилень між елементами двох послідовностей можна розрахувати абсолютне відхилення значень двох елементів (Евклідова відстань). У результаті чого буде отримано матрицю відхилень d:

$$d(x,y)=\sqrt{\sum_{i=1}^{n}\left(x_i-y_i\right)^2}$$

Елементи матриці мінімальних відстаней між послідовностями визначаються за формулою:

$$md_{ij}=d_{ij}+\min(md_{i-1,j-1},md_{i-1},md_{i,j-1}),$$

де md_{ij} – мінімальна дистанція між і та j елементами послідовності a і b.

Алгоритм знаходження мінімального шляху в обраній матриці, за яким можна пройти від елементу mdnm до md00 складається з наступних кроків:

1. Шлях прокладається тільки вперед – індекси і та j ніколи не збільшуються.

2. Індекси зменшуються лише на одиницю за одну ітерацію.

3. Шлях починається в правому нижньому куті і закінчується в верхньому лівому куті матриці.

На основі отриманого шляху, можна оцінити глобальну деформацію наступним чином:

$$GC=\frac{1}{p}\sum_{i=1}^{p}w_i,$$

де wi – елементи мінімального шляху деформації, з – кількість елементів шляху деформації.

Для оцінки якості запропонованих рішень відносно процесу розпізнавання мовного сигналу було спроектовано командну, залежну від диктора систему автоматичного розпізнавання мовного сигналу, що перетворює вхідний мовний сигналу диктора на команду форматування тексту в рамках текстового редактору. Розроблена система забезпечена функціями: формування словника команд диктора, про-

ведення навчання системи під конкретного диктора та виконання розпізнаних команд.

З метою встановлення адекватності запропонованих рішень та оцінки похибки розпізнавання мовного сигналу був розроблений журнал транзакцій, що отримував опис кожної операції розпізнавання команди диктора. Аналіз даних журналу транзакцій дав змогу встановити усереднене значення точності розпізнавання для кожної команди та узагальнене значення похибки розпізнавання в цілому, що не перевищує 6%. Здійснивши порівняння значення отриманої точності із даними досліджень авторів [2-5,8] (рис.2) автори дійшли висновку про непоганий результат розпізнавання вхідного сигналу.

Рисунок 2 – Порівняльна характеристика отриманих результатів

Висновки. В роботі представлено результати підвищення ефективності застосування алгоритму ДТЧ в задачі розпізнавання мовного сигналу. З метою оцінки якості отриманих результатів, розроблено командну дикторозалежну систему розпізнавання мовного сигналу, реалізовану у вигляді плагіну для текстового редактору, який дозволяє вирішувати наступні задачі: виявлення початку та кінця команди, що вимовляється, розпізнавання та порівняння вимовленої команди з набором шаблонів, виконання розпізнаних команд програмним додатком. Запропонований підхід до застосування алгоритму ДТЧ, використання в якості вхідного вектору ознак нормалізовані МЧКК, предрозрахунку гребінки фільтрів мелкоефіцієнтів та використання мультипоточності при розрахунках дало змогу отримати середню якість розпізнавання 94%, що на відміну від існуючих результатів авторів [2-5,8] є кращими на 1-14%. Похибка розпізнавання команд складає в середньому 6%, що в порівнянні з

результатами інших авторів, свідчить про високу ефективність запропонованих рішень.

ЛІТЕРАТУРА

1. Аграновский А. В. Теоретические аспекты алгоритмов и классификации речевых сигналов /А. В. Аграновский, Д. А. Леднов. – М.: Радио и связь, 2004. – 164 с.
2. Kare C.B. Speech recognition by Dynamic Time Warping [Електронний ресурс] / C. B. Kare, V. S. Navale. – Режим доступу: http://www.iosrjournals.org/ iosrjece/ papers/ NCIEST/ Volume %202/3.% 2012-16.pdf.
3. Xianglilan Z. One-against-All Weighted Dynamic Time Warping for Language-Independent and Speaker-Dependent Speech Recognition in Adverse Conditions [Електронний ресурс] / Z. Xianglilan, S. Jiping, L. Zhigang. – Режим доступу: http://journals.plos.org/ plosone/ article?id=10.1371 /journal.pone.0085458.
4. Wielgat R. Comparison of HMM and DTW methods in automatic recognition of pathological phoneme pronunciation [Електронний ресурс] / R. Wielgat, T. Zielinski, P. Swiętojanski. – Режим доступу: http://www.isca-speech.org/archive/archive_papers/ interspeech_2007/ i07_1705.pdf.
5. Малькова Е. С. Методы распознавания речи в задаче автоматизированного выявления дефектов произношения / Е. С. Малькова, О.А.Шабалина // Известия Волгоградского государственного технического университета. – 2015. – №2.– С. 65-71.
6. Akila A. Slope Finder – A Distance Measure for DTW based Isolated Word Speech Recognition. [Електронний ресурс]. – Режим доступу: http://www.ijecs.in/issue/v2-i12/ 13% 20ijecs.pdf.
7. Кристалинский Р. Е. Преобразование Фурье и Лапласа в системах компьютерной математики / Р. Е Кристалинский, В. Р. Кристалинский. – М.: Стереотип,2012. – 216 с.
8. Запрягаев С.А. Распознавание речевых сигналов / С. А.Запрягаев, А.Ю.Коновалов // Вестник Воронежского государственного университета.– 2009.– №2.– С. 37-46.

 ISSN 1562-9945