УДК 004.891

д.т.н., проф. **Мясіщєв О.А.** (ХмНУ)
к.т.н. **Лєнков Є.С.** (ВІТІ)
Ожаровський С.Г. (ХмНУ)

МЕТОД РОЗПІЗНАВАННЯ МОВЛЕННЯ ПО КОРОТКОМУ СЛОВНИКУ З ВИКОРИСТАННЯМ MEL-КЕПСТРАЛЬНИХ КОЕФІЦІЄНТІВ

У статті приведено основні принципи розпізнавання мовлення, особливості використання для цієї цілі технології короткого словника.

В результаті проведення досліджень, визначено збільшення швидкості розпізнавання при знаходженні ефективного алгоритму для визначення команди з потоку вхідного звукового сигналу, яка заздалегідь визначена набором команд. Для цього потрібно виділити з вхідного звукового сигналу інформативні характеристики, які повноцінно його описують, але потребують менше обчислень при обробці.

Ключові слова: розпізнавання мовлення, короткий словник, алгоритми, Mel-кепстральні коефіцієнти.

Вступ. На сьогоднішній день на ринку існує ряд програм розпізнавання мови, які можна використовувати в домашніх умовах або на роботі. Кожна програма пропонує своєму користувачеві ряд можливостей, наприклад, диктувати якийсь текст безпосередньо комп'ютеру, який у якості секретаря записує кожне слово. Таким чином можна швидко написати листа по електронній пошті або написати звіт для роботи. За допомогою голосових команд можна також отримати доступ до командних функцій, наприклад, відкрити файл або

меню виклику. Деякі програми призначені тільки для певних цілей, наприклад, для використання в медичній або юридичній практиці.

Системи розпізнавання мови також використовують люди з фізичними відхиленнями. Наприклад, люди, які втратили обидві свої руки, або позбулися зору і поки не звикли використовувати брайлівський друк [1] . Такі програми дозволяють голосом управляти роботою комп'ютера або набирати будь-який текст. Деякі такі програми після кожної сесії зберігають голосові дані користувача, щоб той потім міг почати роботу з того місця, де зупинився.

Всі програми розпізнавання мови діляться на дві категорії: Програми з невеликим словниковим запасом, призначені для більшості користувачів

Такі системи ідеально підходять для автоматизованого телефонного відповіді. Ці програми здатні розпізнавати декілька видів голосів, розуміти акцент і розбирати мовні зразки користувачів. Однак, управління цими програмами обмежена всього декількома зумовленими командами, наприклад, роботою з меню та управлінням з цифрами [2].

Програми з великим словниковим запасом, розраховані на обмежену кількість користувачів.

Ці системи найбільше підходять для невеликих компаній, де з програмою буде працювати тільки персонал. Але, не дивлячись на те, що ці програми працюють дуже чітко і містять кілька десятків тисяч словників, їх необхідно «підлаштувати» під кожного користувача або під певну групу користувачів, оскільки ступінь точності може значно впасти, у випадку, якщо програмою буде користуватися «Не представленою »їй особою.

Системи розпізнавання мови, створені кілька років тому, також поділялися ще за одним критерієм - по сприйняттю мови: мова з паузами і безперервна мова. Програмі набагато легше зрозуміти окремі слова з постійною паузою між ними. Однак, більшість користувачів воліє говорити зі звичайною швидкістю і не переривати свою мову постійними паузами. Тому практично всі сучасні системи здатні розуміти безперервну мову [3].

Постановка проблеми. Метою роботи є знаходження ефективного алгоритму для розпізнавання команди з потоку вхідного звукового сигналу, яка заздалегідь визначена набором команд; виконання практичної реалізації алгоритму.

Для досягнення мети необхідно зробити виділення з вхідного мовного сигналу інформативних характеристик, які повноцінно описують вхідний сигнал, але вимагають менших обчислювальних витрат при обробці [4]. Далі проводиться порівняння масивів інформативних ознак аналізованої команди і зразка вимови. На підставі зробленого аналізу, вибирається текстове представлення найбільш схожого зразка і виводиться користувачеві.

Виклад основного матеріалу досліджень. Мел – це одиниця висоти звуку, заснована на сприйнятті цього звуку нашими органами слуху. Як відомо, амплітудно-частотна характеристика людського вуха навіть віддалено не відповідає прямій, та амплітуда не зовсім точна міра гучності звуку. Тому і ввели емпірично підібрані одиниці гучності, наприклад, фон.

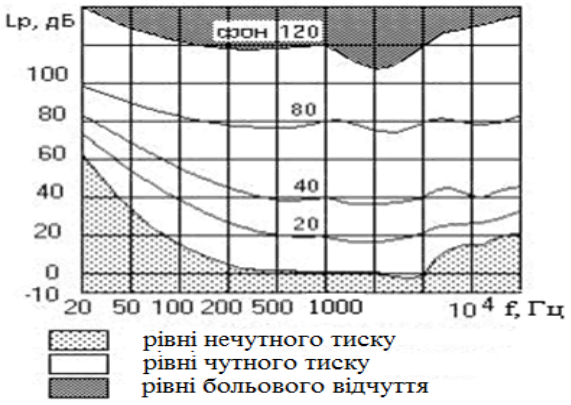

Рис. 1. Фон

Аналогічно, сприйнята людським слухом висота звуку не зовсім лінійно залежить від його частоти.

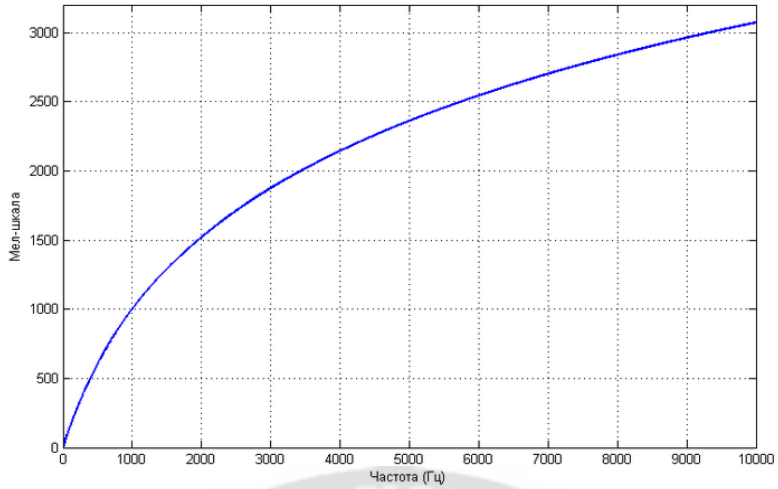

Рис. 2. Залежність частоти від висоти звуку

Ця залежність не зовсім точна, але описується простою формулою:

$$m = 1125 \ln(1 + f / 700).$$

Подібні одиниці виміру часто використовують при вирішенні задач розпізнавання, так як вони дозволяють наблизитися до механізмів людського сприйняття, яке поки що лідирує серед відомих систем розпізнавання мови [1].

У відповідності з теорією мовостворення мова являє собою акустичну хвилю, яка випромінюється системою органів: легкими, бронхами і трахеєю, а потім перетвориться в голосовому тракті. Якщо припустити, що джерела збудження і форма голосового тракту відносно незалежні, мовний апарат людини можна представити у вигляді сукупності генераторів тонових сигналів і шумів, а також фільтрів. Схематично це можна представити так:

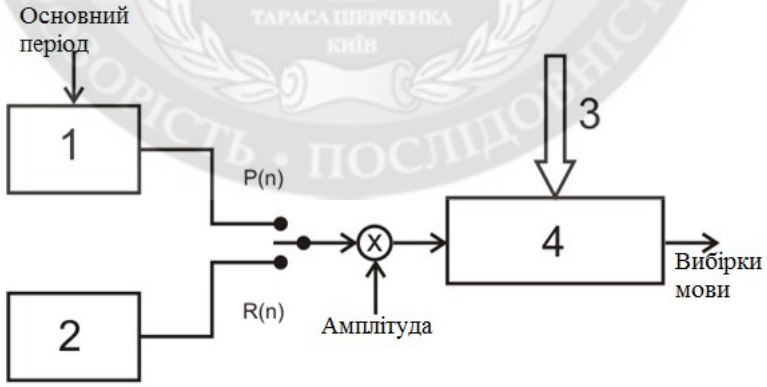

Рис. 3. Схема створення звуків, мовним апаратом людини

1. Генератор імпульсної послідовності (тонів)
2. Генератор випадкових чисел (шумів)
3. Коефіцієнти цифрового фільтра (параметри голосового тракту)
4. Нестаціонарний цифровий фільтр

Сигнал на виході фільтра (4) можна представити у вигляді згортки:

$$f(t) = s(t) \otimes h(t),$$

де s(t) - початковий вигляд акустичної хвилі, а h(t) - характеристика фільтра (залежить від параметрів голосового тракту).

У частотній області це виглядає так:

$$F(\omega) = S(\omega)H(\omega).$$

Добуток можна прологарифмувати, щоб отримати замість нього суму:

$$\ln[S^2(\omega)\cdot H^2(\omega)] = \ln S^2(\omega) + \ln H^2(\omega).$$

Тепер нам потрібно перетворити цю суму так, щоб отримати непересічні набори характеристик вихідного сигналу і фільтра. Для цього є кілька варіантів, , наприклад якщо ми використаємо зворотне перетворення Фур'є, то ось що отримаємо:

$$C(q) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \ln[F(\omega)^2 e^{i\omega q}\, d\omega].$$

Також залежно від цілей можна використовувати пряме перетворення Фур'є або дискретне косинусное перетворення[2].

Тепер розберемось як перетворити мовний сигнал в набір коефіцієнтів MFCC. В якості прикладу візьмемо слово температура, ось його часове представлення:

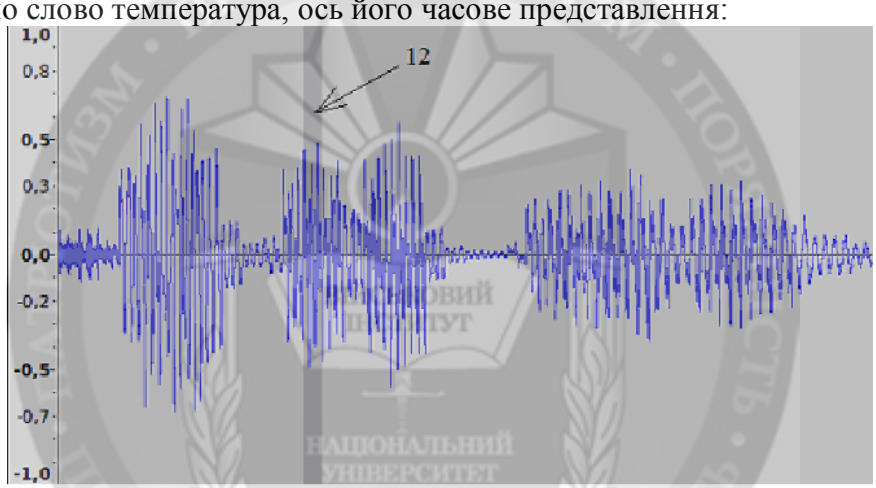

Рис. 4. Осцилограма слова «Температура»

Насамперед нам потрібен спектрограма вихідного сигналу, яку ми отримуємо за допомогою перетворення Фур'є [5]. Для простоти прикладу беремо спектрограму з інтервалу №12:

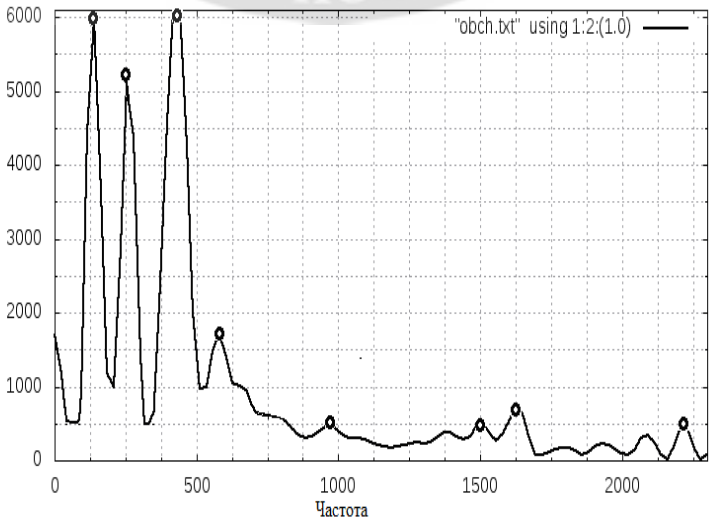

Рис. 5. Спектрограма слова «Температура»

Отриману спектрограму нам потрібно розташувати на мел-шкалі. Для цього ми використовуємо вікна, рівномірно розташовані на мел-осі.

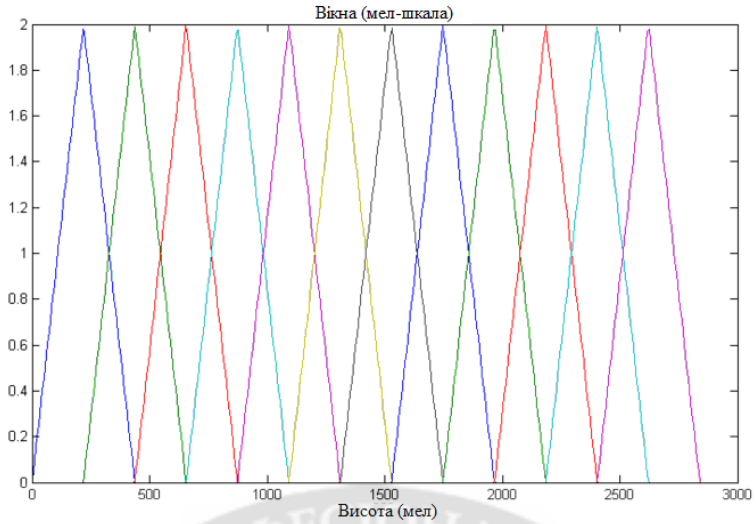

Рис. 6. Мел-шкала

Якщо перевести цей графік в частотну шкалу, то ось що ми отримаємо:

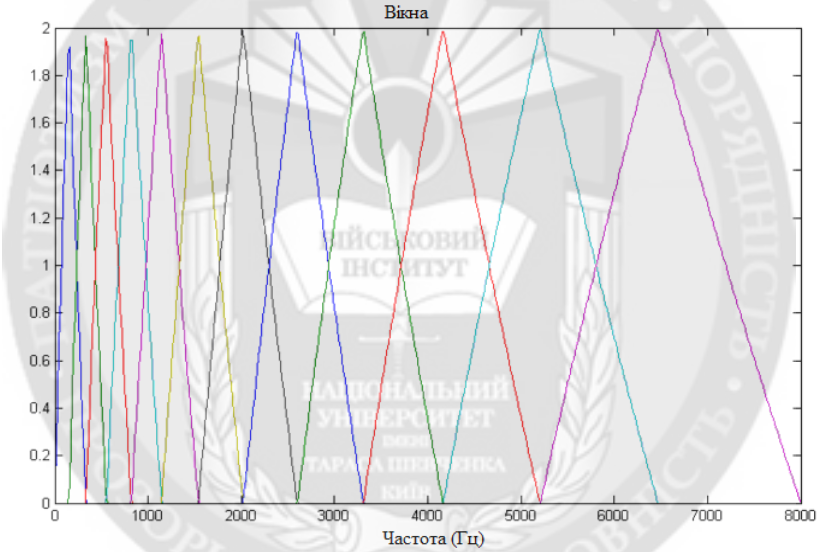

Рис. 7. Мел-шкала переведена в частотну шкалу

На цьому графіку помітно, що вони «збираються» в області низьких частот, забезпечуючи вищу «дозвіл» там, де воно необхідне для розпізнавання.

Простим перемножением векторів спектра сигналу і віконної функції знайдемо енергію сигналу, яка потрапляє в кожне з вікон аналізу. Ми отримали деякий набір коефіцієнтів, але це ще не ті MFCC, які ми шукаємо. Поки їх можна було б назвати Мел-частотними спектральними коефіцієнтами. Зводимо їх в квадрат і логарифмуємо. Нам залишилося тільки отримати з них кепстральних, або «спектр спектра». Для цього ми могли б ще раз застосувати перетворення Фур'є, але краще використовувати дискретне косинусное перетворення. В результаті отримуємо послідовність такого вигляду:

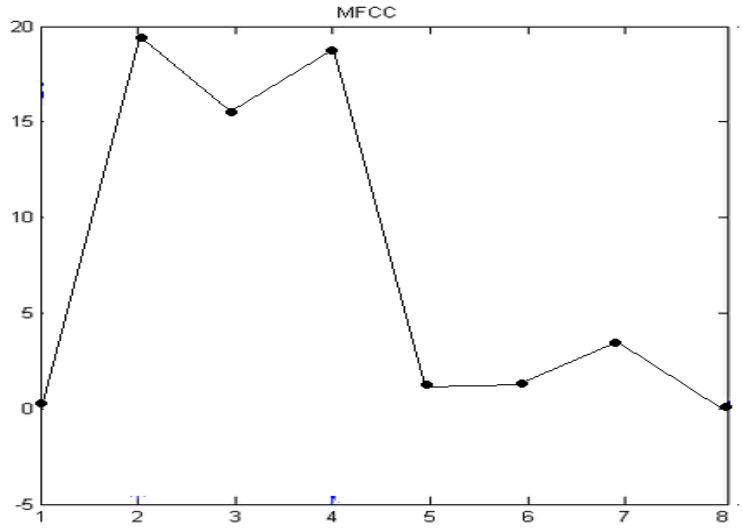

Рис. 8. Мел-кепстральні коефіцієнти

Висновки. Таким чином ми маємо дуже невеликий набір значень, який при розпізнаванні успішно замінює тисячі відліків мовного сигналу. Для задачі розпізнавання слів можна брати перші 7 з 16 обчислених коефіцієнтів, але придатні результати починаються десь з 8. У кожному разі це набагато менший обсяг даних, ніж спектрограмма або тимчасове уявлення сигналу.

ЛІТЕРАТУРА:

1. Вікепедія. [Electronic resource].
– Mode of access: https://en.wikipedia.org/wiki/Speech_recognition.
2. Xuedong Huang, Alex Acero, Hsiao-Wuen Hon, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Prentice Hall, 2001, ISBN:0130226165.
3. Now a Machine That Talks With the Woice of Man. (14 January 1939 r.). Science News Letter.
4. Галунов В.И., Соловьев А. Современные проблемы в области распознавания речи. Информационные технологии и вычислительные системы. – №2. – 2004.
5. Мясищев А.А. Программы распознавания команд с помощью ДПФ и библиотеки FANN. [Electronic resource]. – Mode of access: http://webstm32.sytes.net/obrazec/prog.htm, 2015.

Без рецензії.

д.т.н., проф. Мясищев А.А., к.т.н. Ленков Е.С., Ожаровский С.Г.

МЕТОД РАСПОЗНАВАНИЯ РЕЧИ, ПО КОРОТКОМУ СЛОВАРЮ, С ИСПОЛЬЗОВАНИЕМ MEL-КЕПСТРАЛЬНЫХ КОЭФФИЦИЕНТОВ

В статье приведены основные принципы распознавания речи, особенности использования для этой технологии короткого словаря.

В результате проведения опытов, определено увеличение скорости распознавания при нахождении эффективного алгоритма для определения команды из потока входного звукового сигнала, которая заранее определена набором команд. Для этого нужно выделить из входного звукового сигнала информативные характеристики, которые полноценно его описывают, но требуют меньше вычислений при обработку.

Ключевые слова: распознавание речи, короткий словарь, алгоритмы, Mel-кепстральные коэффициенты.

Prof. Myasischev A.A., Ph.D. Lenkov E.S., Ozharovkiy S.G.

SPEECH RECOGNITION METHOD, IN BRIEF DICTIONARIES, WITH MEL-CEPSTRAL COEFFICIENTS

The article describes the main principles of the speech recognition, especially the use of this technology for a short vocabulary, because it increases the recognition rate.

As a result of experiments to determine the increase recognition speed in finding an efficient algorithm for determining the flow of commands from an input audio signal, which is pre-defined set of commands. To do this, select from the input audio signal informative characteristics that describe it fully, but require less computing for processing.

Keywords: speech recognition, a short vocabulary, algorithms, Mel-cepstral coefficients.