УДК 004.05
DOI https://doi.org/10.32838/2663-5941/2022.1/19

***Климчук І.О.***
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

***Потапова К.Р.***
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

***Тарасенко-Клятченко О.В.***
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

# ПРО ОСОБЛИВОСТІ ОРГАНІЗАЦІЇ ЗВУКОВОГО ІНТЕРФЕЙСУ ДЛЯ ЛЮДЕЙ З ПОРУШЕННЯМИ МОВНОГО АПАРАТУ

*У статті досліджено методи розпізнавання мови людей з порушенням мовного апарату по короткому словнику з використанням mel-кепстральних коефіцієнтів. Визначено, що однією з головних форм взаємодії для людини є мовлення. Мова є носієм інформації, що використовується людиною для передачі повідомлень сигналом. За фізичною природою доведено, що це акустичний сигнал, який безперервно змінюється в часі. Визначено, що стрімко зростаючі обчислювальні потужності, створення систем розпізнавання мови залишається надзвичайно складною проблемою. Комерційні програми з розпізнавання мовлення з'явилися на початку дев'яностих років. Доведено, що ці програми використовують люди, які через травми рук не в змозі набирати велику кількість тексту або мають порушення мовного апарату. Програми переводять голос користувача в текст. Імовірність точного перекладу у таких програм не дуже висока, але з часом вона поступово покращується. У наш час на ринку є низка програм розпізнавання мови, які можна використовувати в домашніх умовах або на роботі. Важливі частини визначено у сучасних стандартних уже заснованих алгоритмах з розпізнавання мовлення – це моделювання мови та акустичне моделювання. Доведено, що розпізнавання у подібних наявних методах відбувається окремими словами в обмеженому словнику, а зі збільшенням словника збільшується час розпізнавання, що є суттєвим недоліком. Визначено, що найефективніший метод розпізнавання по короткому словнику – з використанням mel-кепстральних коефіцієнтів, які часто використовуються як характеристики мовних сигналів. Метод має дуже невеликий набір значень, який у разі розпізнавання успішно замінює тисячі відліків мовного сигналу. У цьому методі набагато менший обсяг даних, ніж спектрограма або тимчасове уявлення сигналу. Для кращого результату розглянуто спосіб розбиття вихідного слова на відрізки невеликої тривалості і як обчислювати коефіцієнти для кожного з них.*

***Ключові слова:*** *розпізнавання мови, мовний сигнал, короткий словник, mel-кепстральні коефіцієнти, програмний додаток, порушення мовного апарату.*

**Постановка проблеми.** Однією з головних форм взаємодії для людини є мовлення. Зазвичай методи розпізнавання мовлення використовують люди, які мають порушення мовного апарату.

Головна мета розпізнавання мовлення – це отримання різної інформації через вхідний мовленнєвий сигнал: про що говориться, хто саме говорить, якою мовою, в якому фізичному стані перебуває мовець тощо. Проблеми, що вирішує метод:

1) автоматичне перетворення мовленнєвого сигналу на текст;

2) введення інформації голосом, диктувальна машина;

3) пошук ключових слів і фраз у потоці мовлення;

4) смислова інтерпретація голосових повідомлень;

5) ідентифікація та верифікація диктора;

6) адаптація до голосу диктора та акустичного каналу;

7) розпізнавання мови, якою говорить диктор, його акценту;

8) усний переклад з однієї мови на іншу.

**Аналіз наявних рішень.** У наш час системи розпізнавання мови для загального призначення ґрунтуються на прихованих моделях Маркова. Приховані моделі Маркова користуються попу-

лярністю тому, що їх можна навчити автоматично. У використанні нейронні мережі роблять менше помітних припущень щодо статистичних властивостей ознак, ніж HMM. Користуючись для оцінки ймовірностей мовний сегмент, нейронні мережі дозволяють дискримінаційне навчання природним та ефективним способом. Допускаючи, що джерела збудження і форма голосового тракту можуть бути повністю незалежні, мовний апарат людини можемо представити як вигляд сукупностей генераторів тональних сигналів і шумів, а також фільтрів [8]. Схематично це можна представити так:

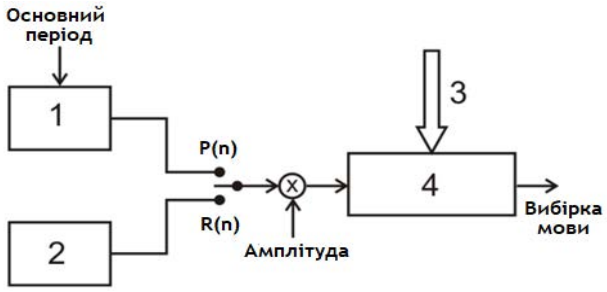


**Рис. 1. Сукупність генераторів, сигналів і шумів**
1. Генератор імпульсної послідовності (тонів).
2. Генератор випадкових чисел (шумів).
3. Коефіцієнти цифрового фільтра (параметри голосового тракту).
4. Нестаціонарний цифровий фільтр.

Для задачі розпізнавання слів можна взяти перші 13 з 24 обчислених коефіцієнтів [1], але скільки-небудь придатні результати в такому випадку починалися з 16. У будь-якому разі це набагато менший обсяг даних, ніж спектрограма або тимчасове уявлення сигналу.

Для кращого результату можна розбити вихідне слово на відрізки невеликої тривалості й обчислювати коефіцієнти для кожного з них [9].

**Постановка завдання.** Метою роботи є дослідження можливості використання методу mel-кепстральних коефіцієнтів для розпізнавання мови людей з порушенням мовного апарату.

Для досягнення мети розв'язуються такі наукові завдання:

– розглянути структуру мовного сигналу та виконати його опис;

– розглянути загальну структуру системи автоматичного розпізнавання мови;

– розглянути побудову блоку виділення ознак;

– дослідити метод mel-кепстральних коефіцієнтів.

**Виклад основного матеріалу дослідження.** Питання людино-машинної взаємодії є одними з найважливіших у разі створення нових комп'ютерів. Найбільш ефективними засобами взаємодії людини з машиною були б ті, які є природними для неї: через візуальні образи і мову.

Створення мовних інтерфейсів могло б знайти застосування у системах самого різного призначення [2]:

1) голосове управління для людей з обмеженими можливостями;

2) надійне управління бойовими машинами, «розуміючими» тільки голос командира;

3) автовідповідачі, що опрацьовують в автоматичному режимі сотні тисяч дзвінків на добу (наприклад, у системі продажу авіаквитків);

4) та інше.

При цьому мовний інтерфейс повинен включати у себе два компоненти: систему автоматичного розпізнавання мови для прийому мовного сигналу і перетворення його в текст або команду, і систему синтезу мовлення, що виконує протилежну функцію – конвертацію повідомлення від машини в мову.

Це зумовлюється як її міждисциплінарним характером (необхідно володіти знаннями у філології, лінгвістиці, цифровій обробці сигналів, акустиці, статистиці, розпізнаванні образів і т. д.), 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.

Всі описані вище проблеми об'єднує необхідність створення компактного, надійного, самостійного і максимально швидкодіючого пристрою. Таким чином, пошук нових архітектурних рішень щодо розпізнавання мови є актуальною темою. Одним з перспективних напрямів є дослідження і використання методу mel-кепстральних коефіцієнтів по короткому словнику.

***Мовний сигнал і його опис.*** Мова – це історично сформована форма спілкування людей за допомогою мовних конструкцій, створюваних на основі певних правил [3]. Тоді як усне мовлення – звукове коливання, що характеризується частотою і амплітудою.

Більшість сигналів (мовних у тому числі) мають схожість, тому для обробки їх на цифрових комп'ютерах вони перетворюються на дискретні сигнали за допомогою аналого-цифрового перетворення (АЦП).

***Загальна структура системи автоматичного розпізнавання мови.*** У разі автоматичного розпізнавання мови (САРМ) можна виділити такі етапи: виділення ознак, навчання і розпізнавання

(рис. 2). Спершу з вихідного сигналу отримують вектор ознак – дуже короткий опис мовного сигналу, в якому присутня тільки корисна інформація для розпізнавання. Для цього використовуються методи, що працюють як у частотній сфері, так і в тимчасовій, але у разі використання цього методу проблема подання мови не вирішується. Тому дослідження ведуться дотепер [4].

Послідовність векторів ознак довжиною T називають акустичною або послідовністю, що спостерігається, $O=(o_1, o_2, ..., o_T)$. За допомогою цієї послідовності людина передає ланцюжок слів $W=(w_1, w_2, ..., w_N)$. Сама задача розпізнавання мови ставиться таким чином: необхідно відшукати ланцюжок слів $W$, який відповідає акустичній послідовності $O$ [5].

Для правильного розпізнавання потрібно повністю перевірити всі ланцюжки слів $W$, але це нереально в реальному житті та практиці. Для того щоб полегшити завдання, можна вводити різні обмеження. Можна розпізнавати тільки ізольовані слова за допомогою граматики мови або звуження завдання.

***Побудова блоку виділення ознак.*** Завдання блоку виділення ознак – побудувати ланцюжок векторів ознак $O=(O_1, O_2, ..., O_T)$ вихідного сигналу. Як було зазначено раніше, мова – нестаціонарний сигнал. Однак через інертність мовного тракту в межах досить короткого проміжку часу від 10 до 40 мс його характеристики не змінюються, тобто його можна вважати стаціонарним [3]. Тому блок виділення ознак сканує вхідний сигнал короткочасним ковзаючим вікном, у межах якого і складається один вектор ознак O*t* (рис. 3).

***Mel-кепстральний метод виділення ознак.*** Дослідження показали, що найкраще мова представляється ознаками, отриманими в частотній сфері. До таких ознак належать коефіцієнти лінійного передбачення (Linear Predictive Codes – LPC), перцепційні коефіцієнти лінійного передбачення (Perceptual Linear Prediction – PLP), mel-кепстральні коефіцієнти (Mel-Frequency Cepstral Coefficients – MFCC). Ці три ознаки ґрунтуються на акустичній моделі мовостворення, згідно з якою мовний сигнал можна представити у вигляді сигналу на виході лінійної системи з мінливими в часі параметрами, збудженої квазіперіодичними імпульсами (у разі проголошення вокалізованого звуку) або випадковим шумом (у разі невокалізованих звуків).

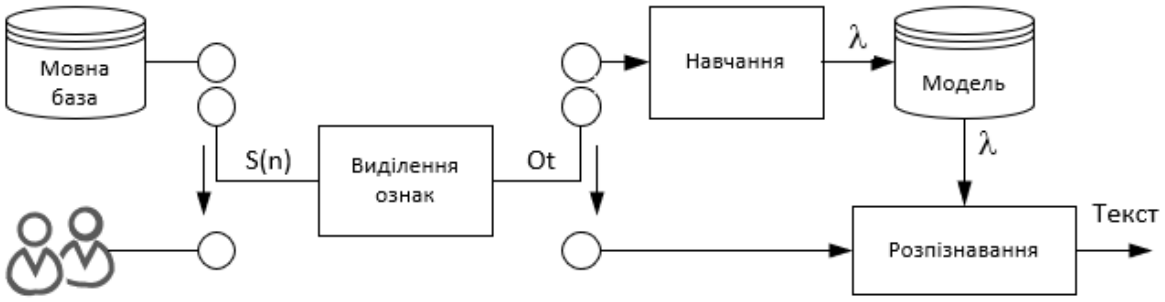


**Рис. 2. Загальна схема САРМ**


**Рис. 3. Виділення ознак**

Оцінка параметрів цієї лінійної системи і є завданням методу знаходження ознак у частотній сфері. Головною метою методів лінійного передбачення є можливість апроксимації поточного відліку мовного сигналу [6]:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k), \quad (1)$$

де $\{a_k\}$ – коефіцієнти лінійного передбачення.

Метод MFCC може відокремити сигнал збудження від параметрів мовного тракту, використовуючи гомоморфні перетворення. Для цього за допомогою дискретного перетворення Фур'є (ДПФ) переходять у частотну область, де обчислюють логарифм спектра вхідного сигналу, а потім виконують обчислення ДПФ (ОДПФ) або дискретне косинусне перетворення. Як і в PLP, для моделювання сприйняття мови людиною перед знаходженням логарифма на спектр вхідного сигналу накладають набір mel-фільтрів, смуги пропускання яких вибираються по mel-шкалі, рухаючись у сторону високих частот.

Перевагою MFCC порівняно з LPC і PLP є простота реалізації за схожих показників. Підвищення швидкодії методу обгрунтовується ефективністю процедури знаходження ДПФ і ОДПФ – швидкого перетворення Фур'є (ШПФ). Аналізуючи всі стани галузі розпізнавання мови, можна побачити, що натепер MFCC застосовується найбільш широко.

Знаходження MFCC здійснюється в декілька етапів (рис. 4).

1. Нормалізація початкового сигналу, що дозволить вирівняти його амплітуду і посилення високих частот. Форманти низьких частот мають велику амплітуду порівняно з формантами більш високих частот, хоч останні теж несуть важливу для розпізнавання інформацію. Тому до вхідного сигналу застосовують фільтр:

$$s'(n) = s(n) - 0.95 \cdot s(n-1), \quad (2)$$

2. Виділення короткочасної ділянки сигналу (~32 мс.) і накладання віконної функції $w(k)$ для мінімізації витоку спектра. Одержаний сегмент називають фреймом довжиною $K$ відліків і подальшу роботу ведуть у межах фрейму:

$$x_t(k)=s'(k+t \cdot K) \cdot w(k),\ 0 \le k \le K-1 \quad (3)$$

Таким чином, вектор ознак отримують для кожного $0 \le t \le T$ фрейму вхідного сигналу $s'(n)$. Як віконну функцію зазвичай використовують функцію Хемінга:

$$w(k) = 0.54 - 0.46 \cdot \cos\left(\frac{2\pi k}{K-1}\right), \quad (4)$$

3. Обчислення ДПФ для фрейму:

$$X_t(m) = \sum_{k=0}^{K-1} x_t(k) \cdot e^{\frac{-j2\pi mk}{K}} \quad (5)$$

Для обчислення ДПФ використовується алгоритм швидкого перетворення Фур'є (ШПФ).

4. Накладення набору Q мел-фільтрів на фрейм у частотній області так, що на виході кожного q-го фільтра знімають енергію $Yt(q)$ у смузі пропускання цього фільтра. Таким чином моделюють сприйняття мови людиною: роздільна здатність слуху зростає у разі руху по спектру від низьких частот до високих. Центральні частоти $Fq$ мел-фільтрів вибираються за так званою мел-шкалою, яка залежить від звичайної по логарифмічному закону (рис. 5):

$$F_{mel} 2095 \cdot log_{10}\left(1 + \frac{F_{Hz}}{700}\right), \quad (6)$$

5. Логарифмування $Yt(q)$. На цьому кроці виконуються гомоморфні перетворення. Для того щоб відокремити сигнал збудження від характеристики фільтра, обчислюється логарифм модуля $Yt(q)$:

$$L_t(q) = \log\left|Y_t(q)\right|, \quad (7)$$

6. Обчислення ОДПФ. На останньому етапі отримують мел-кепстральні коефіцієнти шляхом обчислення ОДПФ для $Lt$(q). При цьому, оскільки $Lt$(q) дійсний і симетричний, ОДПФ буде еквівалентно дискретному косинусному перетворенню:

$$Q_t(p) = \sum_{q=1}^{Q} L_t(q) \cdot \cos\left(p \cdot \left(q - \frac{1}{2}\right) \cdot \frac{\pi}{Q}\right), p = 0, \ldots, P. \quad (8)$$

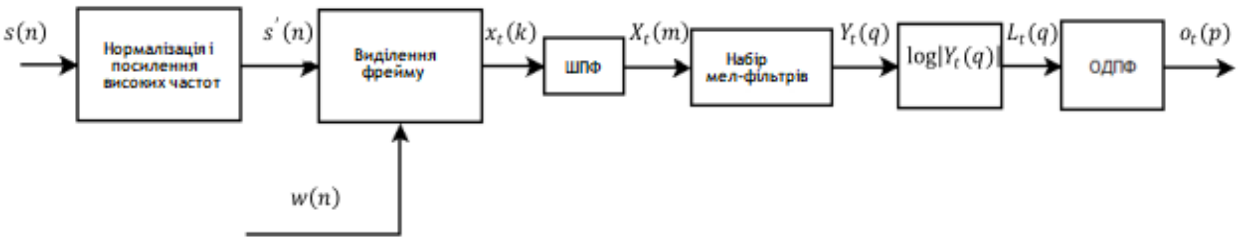


**Рис. 4. Знаходження mel-кепстральних коефіцієнтів**

У результаті отримують *P* коефіцієнтів, при цьому хоча *P* може дорівнювати *Q* зазвичай беруть лише половину значень MFCC. Це пояснюється тим, що кепстра сигналу збудження зазвичай «правіше» кепстра мовного тракту.

Таким чином, на розпізнавач надходить p-мірний вектор *ot*, що містить мел-кепстральні коефіцієнти для *t*-го фрейму [7].

**Висновки.** У цій роботі детально розглянуто метод використання mel-кепстральних коефіцієнтів з точки зору організації звукового інтерфейсу для людей із дефектами мовлення.

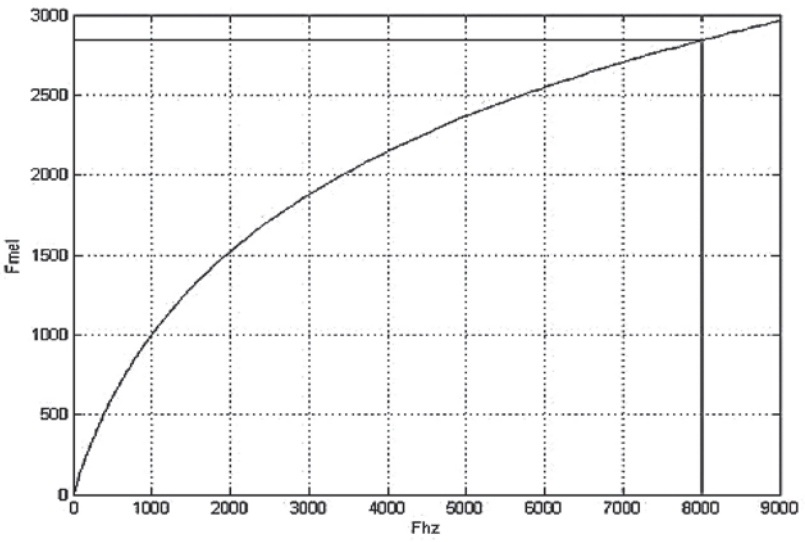


**Рис. 5. Відповідність частот звичайної шкали частотам мел-шкали**

Основні результати роботи полягають у такому:

1. Проведено аналіз задачі представлення і розпізнавання мови, виділені основні компоненти систем автоматичного розпізнавання мови.

2. Розглянуто методи попередньої обробки і виділення ознак мовного сигналу, серед яких вибрано підхід, що оснований на знаходженні мел-кепстральних коефіцієнтів.

**Список літератури:**

1. Davis S. and Mermelstein P. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. In *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 28, No. 4, 1980.

2. Ronzhin A., Karpov A. Russian Voice Interface. *Pattern Recognition and Image Analysis*, 2007. Vol. 17, No. 2, pp. 321–336.

3. Вінцюк Т.К. Аналіз, розпізнавання і інтерпретація мовних сигналів. Київ : Наукова думка, 1987. 264 с.

4. Ognev I.V., Ognev A.I., Paramonov P.A., Sutula N.A. The use of extrema distribution as a feature vector for speech patterns recognition. The 11th International Conference *"Pattern Recognition and Image Analysis: New Information Technologies"*, Vol. 1, 2013. Pp. 114–117.

5. Claudio Becchetti, Lucio Prina Ricotti. Speech Recognition. Theory and C++ Implementation. Wiley. 1999, 428 p.

6. Рабинер Л., Шафер Р. Цифровая обработка речевых сигналов. Москва : Радио и связь, 1981. 496 с.

7. Шарий Т.В. О проблеме параметризации речевого сигнала в современных системах распознавания речи. *Вісник Донецького національного університету*, 2008, Вип. 2, с. 536–54.

8. The 11th International scientific and practical conference *"European scientific discussions"* (September 12–14, 2021). Potere della ragione Editore, Rome, Italy, 2021, 337 p., UDC 001.1, ISBN 978-88-32934-02-1, Pp. 64–68.

9. UDC 001.1 The 7th International scientific and practical conference *"Results of modern scientific research and development"* (September 19–21, 2021). Barca Academy Publishing, Madrid, Spain. 2021. 336 p. ISBN 978-84-15927-33-4, Pp. 105–108.

**Klymchuk I.O., Potapova K.R., Tarasenko-Klyatchenko O.V. FEATURES OF ORGANIZATION OF SOUND INTERFACE FOR PEOPLE WITH SPEECH DEFECTS**

*The article investigates the methods of speech recognition of people with speech disorders in a short dictionary using Mel-cepstral coefficients. It is determined that one of the main forms of human interaction is speech. Language is a carrier of information used by humans to transmit messages – a signal. By physical nature, it is proven that this is an acoustic signal that is constantly changing over time. It is determined that the rapidly growing computing power, the creation of language recognition systems remains an extremely difficult problem. Commercial speech recognition programs appeared in the early 1990s. It has been proven that people who, due to hand injuries, are unable to type a large amount of text or have speech disorders use these programs. Programs translate the user's voice into text. The probability of accurate translation*

*in such programs is definitely not very high, but over time, it gradually improves. There are a number of language recognition programs on the market today that can be used at home or at work. Important parts are defined in modern standard already established algorithms for speech recognition – it's language modeling and acoustic modeling. It is proved that recognition in such existing methods occurs in separate words in a limited dictionary, and with the increase of the dictionary the recognition time increases, which is a significant disadvantage. It is determined that the most effective method of recognition by short vocabulary – using mel-keptral coefficients, which are often used as a characteristic of speech signals. The method has a very small set of values, which in recognition has been proven to successfully replace thousands of samples of speech signal. This method has much less data than the spectrogram or temporal representation of the signal. For the best result, the method of dividing the source word into segments of short duration is considered, and the coefficients for each of them are calculated.*

***Key words:*** *speech recognition, speech signal, short dictionary, mel-keptral coefficients, software application, speech apparatus disorders.*