УДК 321.396, 004.934.81

Д.В. Феньов,
к.т.н., доц. (Інститут служби зовнішньої розвідки України)

НАПРЯМ ІДЕНТИФІКАЦІЇ МОВНИХ СИГНАЛІВ ВІДПОВІДНО ДО ПОТОЧНИХ ПОРЯДКОВИХ СТАТИСТИК СПЕКТРІВ

У статі розглянуто принципово новий метод спектральної ідентифікації мовних сигналів окремих слів (словоформ) на основі використання поточних вибіркових порядкових статистик поточних амплітудних спектрів відповідних мовних сигналів, які характеризують зміни в часі їх індивідуальних частотно-часових та статистичних характеристик. Як еталони при цьому використовуються набори (словники) векторів впорядкованих значень вибіркових поточних порядкових статистик амплітудних спектрів на інтервалах тривалості слів (словоформ).

Ключові слова: *мовні сигнали, порядкові статистики амплітудних спектрів, ідентифікація мовних одиниць.*

В статье рассмотрен принципиально новый метод спектральной идентификации речевых сигналов отдельных слов (словоформ) на основе использования текущих выборочных порядковых статистик текущих амплитудных спектров соответствующих речевых сигналов, которые характеризуют изменения во времени их индивидуальных частотно-временных и статистических характеристик. В качестве эталонов при этом используются наборы (словари) векторов упорядоченных значений выборочных текущих порядковых статистик амплитудных спектров на интервалах длительности слов (словоформ).

Ключевые слова: *речевые сигналы, порядковые статистики амплитудных спектров, идентификация речевых единиц.*

The fundamentally new method of spectral identification of speech signals of separate words (word forms) is considered in the paper on the basis of using current selective ordinal statistics of current amplitude spectra of corresponding speech signals that characterize changes in time of their individual time-frequency and statistical characteristics. As the pattern, we use sets (dictionaries) of the vectors of ordered values of selective current order statistics of amplitude spectra on the intervals of word duration (word forms).

Keywords: *speech signals, ordinal statistics of amplitude spectra, identification of speech units.*

Постановка проблеми

Автоматичне розпізнавання ізольованих голосових команд у системах комп'ютерного інтерфейсу; ключових мовних комбінацій у системах верифікації, ідентифікації та охоронної сигналізації; визначення мови та тематики потоку суцільного мовлення за тематичними словниками в системах розвідки та контррозвідки – це досить короткий перелік основних важливих практичних задач

для технічних систем ідентифікації, верифікації та розпізнавання мовних сигналів [1; 2]. Одним із важливих завдань, що впливають на показники якості ідентифікації та розпізнавання, є обґрунтований вибір фіксованого набору (вектору) інформаційних параметрів, які дозволяють достатньо надійно в умовах коартикуляції (взаємного впливу звуків та переходів між окремими звуками), мінливості та нелінійності темпу промовляння описати елементарні компоненти мовних сигналів (контекстно-незалежних фонем (монофонів), контекстно-залежних фонем (дифонів, трифонів) та алофонів) та їх максимально правдоподібні конкатенації (слоги, слова, фрази). Найбільш поширеним у практиці технічної ідентифікації мовних сигналів є використання сукупностей векторів коефіцієнтів MFCC, LPC, PLP, які отримують у часовій, частотній та кепстральній областях. Однак перелічені параметри характеризуються досить складними негаусівськими багатомірними розподіленнями імовірностей їх значень та не забезпечують робастності розпізнавання мови в умовах перекручень та завад різного походження.

Аналіз останніх досліджень і публікацій

Для подолання наслідків впливу перекручень та завад різного походження на результати спектральної ідентифікації диктора вперше запропонована оригінальна методика ідентифікації з використанням векторів вибіркових порядкових статистик (вибіркових квантилів) всієї сукупності поточних реалізацій амплітудних спектрів по сегментам часового аналізу [2; 3]. Послідовність часових відліків реалізації спонтанного мовного сигналу після видалення пауз розбивається на сегменти стаціонарності мовного сигналу 10–30 мс з перекриттям сегментів на 5–15 мс. Для кожного сегменту аналізу всієї сукупності 500–1000 сегментів розраховується амплітудний спектр сигналу на основі дискретного перетворення Фур'є, а як інформаційний вектор для всіх сегментів аналізу використовуються вектори вибіркових процентілів у діапазоні 50 %...95 % всіх реалізацій спектрів. Завадостійка спектральна ідентифікація диктора за порядковими статистиками забезпечує 99 % імовірність правильної ідентифікації для 1 % помилки першого роду, не залежить від фонетичного складу промови та переходів з одної мови на іншу [3; 4].

Однак безпосереднє використання вибіркових порядкових статистик усієї сукупності поточних реалізацій амплітудних спектрів по сегментам часового аналізу для спектральної ідентифікації мовних одиниць (фонем, слів, словоформ, фраз) не є ефективним. Це зумовлено тим, що відокремлене слово в безперервному мовному потоці суттєво змінює свої акусто-фонетичні характеристики залежно від фонетичного контексту та місцеположення в потоці мовлення (початок, кінець, середина). Особливості фонетичного контексту суттєво виявляються на стиках слів за рахунок коартикуляції. Крім того, значна нестабільність темпу промовляння суттєво загострює проблему нормалізації мовного сигналу в часі. Це зумовлює необхідність додаткового дослідження як порядкових статистик спектрів мовних одиниць (фонем, слів, фраз), так і особливостей їх зміни на стиках елементів слова, тобто дослідження характеристик поточних порядкових статистик поточних спектрів мовних сигналів слів (словоформ). Тому для спектральної ідентифікації словоформ обмеженої тривалості доцільно використовувати робастні поточні порядкові статистики поточних амплітудних спектрів за сегментами аналізу мовного сигналу, які змінюються у часі та відображають для кожного слова індивідуальні особливості зміни значень амплітуд поточних спектрів у часі відповідно порядку (послідовності) слідування слогів (фонем, алофонів).

Мета і завдання статті

Метою статті є викладення теоретичних основ нового методу спектральної ідентифікації слів (словоформ) на основі зіставлення сукупностей поточних порядкових статистик поточних амплітудних спектрів мовних сигналів словоформ з еталонними (зразковими). Визначення основних особливостей та переваг використання поточних порядкових статистик для спектральної ідентифікації мовних сигналів є основним завданням статті.

Виклад основного матеріалу

Розглянемо принципово новий метод спектральної ідентифікації мовних сигналів на основі зіставлення поточних вибіркових порядкових статистик амплітудних спектрів мовних сигналів словоформ з відповідними статистиками словоформ-еталонів (словником). Використання робастних поточних порядкових статистик поточних амплітудних спектрів мовних сигналів суттєво зменшує вплив дестабілізуючих факторів на ймовірність правильної ідентифікації мовних одиниць. Це витікає з асимптотичної нормальності та робастності оцінок квантилів за значеннями вибіркових порядкових статистик обмежених вибірок випадкових параметрів мовного сигналу з неперервними щільностями розподілу їх миттєвих значень [5].

Для формування вибіркових поточних порядкових статистик послідовність часових відліків реалізації мовного сигналу розбивається на сегменти з однаковою кількістю відліків, що зазвичай відповідає сегменту стаціонарності мовного сигналу 10–30 мс (для частоти дискретизації 8 кГц відповідно 80–240 відліків). Після виконання спектрального перетворення сегментів мовного сигналу, для кожного дискретного значення частоти f_j поточного спектра Фур'є може бути отримана сукупність значень модулів $|X(f_j)|$, по усіх n сегментів формування поточних порядкових статистик. При достатньому значенні об'єму вибірки (для частот дискретизації 8000 Гц та вище) визначаються порядкові статистики (вибіркові квантилі) для кожного дискретного значення частоти поточного спектра Фур'є [4]. Порядкові статистики $X_{\text{p}}(f_j)$ можна розглядати як впорядковану сукупність нових випадкових величин (випадковий вектор). Статистичні характеристики порядкових статистик $X_{\text{p}}(f_j)$ варіаційного ряду $\tilde{X}_{1/n}(f_j),...,\tilde{X}_{n/n}(f_j)$ істотно відрізняються від статистичних характеристик початкової вибірки спектральних амплітуд мовного сигналу, що зумовлено тим, що операція формування елементів варіаційного ряду не є лінійною операцією [3]. Взаємне перекриття сегментів аналізу становить 0,025...0,05 тривалості сегменту аналізу. Поточні порядкові статистики мовних одиниць характеризують залежністю $X_{\text{p}}(f_j)$ від часу існування мовної одиниці.

Сукупність поточних порядкових статистик, значення яких змінюються в часі відповідно до індивідуальних змін амплітудного спектра, є інформаційним еквівалентом сигналів мовних одиниць. Навіть для різного промовляння одного слова залежності поточних порядкових статистик подібні. Як ілюстрація на рис. 1 показані залежності значень поточних квантилів 50 % для двох однакових слів "one" різного промовляння. Наведені залежності наочно демонструють високу подібність.

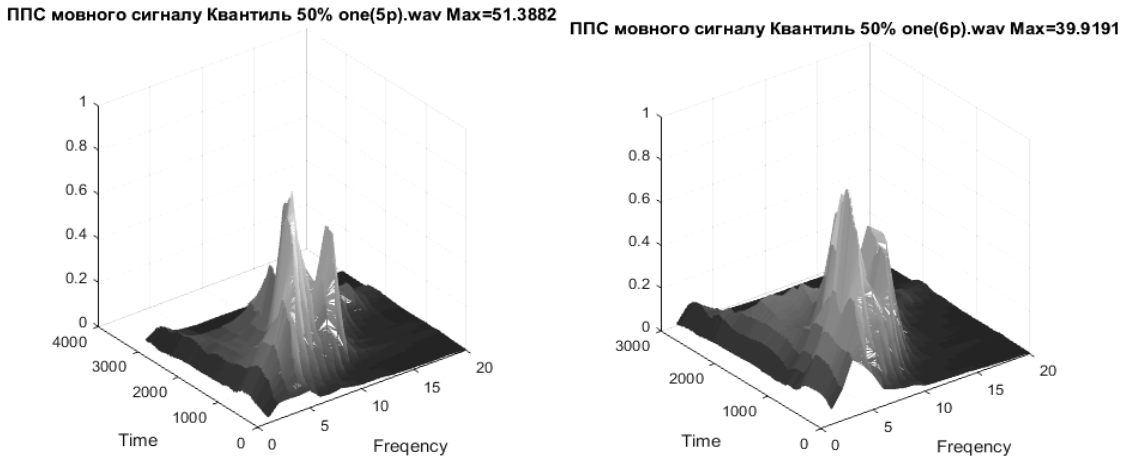

Рис. 1. Поточні порядкові статистики слів “one” різного промовляння

С другого боку залежності поточних порядкових статистик поточних амплітудних спектрів мовних сигналів словоформ з різним фонетичним складом суттєво відрізняються. Для прикладу на рис. 2 показані залежності поточних квантилів 50 % для двох різних слів “seven” та “nine”. Наведені залежності наочно демонструють суттєві відмінності характеристик для різних слів.

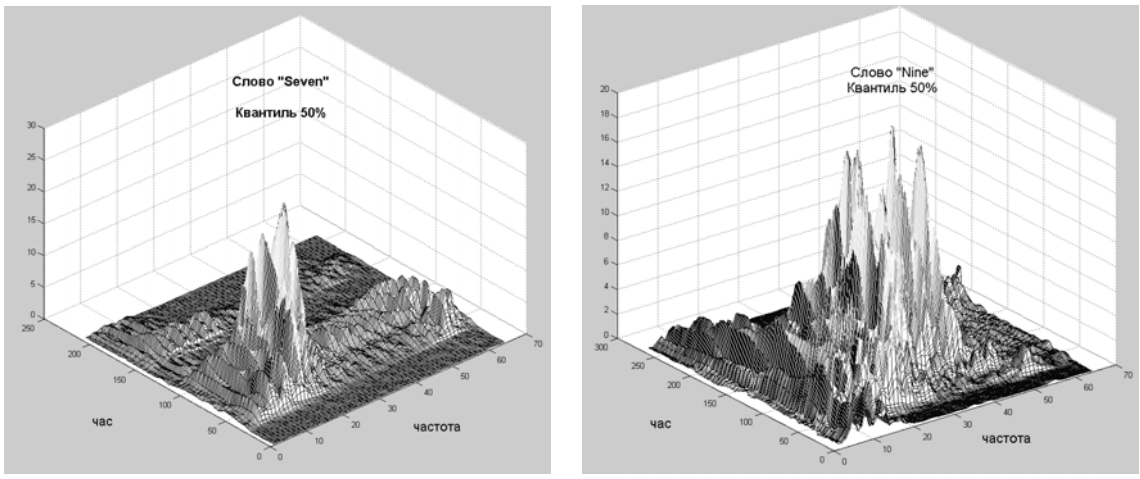

Рис. 2. Поточні порядкові статистики слів “seven” та “nine”

Таким чином, поточні порядкові статистики поточних амплітудних спектрів мовних сигналів відображають індивідуальні зміни в часі частотно-часових характеристик мовних одиниць. Тому еталоном (зразком) мовної одиниці при такому підході до спектральної ідентифікації є матриця (набір векторів) значень поточних порядкових статистик.

Основні етапи спектральної ідентифікації мовних сигналів по поточним порядковим статистикам поточних амплітудних спектрів можна представити у вигляді послідовності операцій. На етапі попередньої обробки мовних сигналів проводиться перетворення для зменшення впливу форми сигналу голосового тракту диктору, наприклад, здійснюється цифрова фільтрація з характеристикою

$1\text{-}0{,}97z^{-1}$. Профільтрований сегмент мови обробляється часовим вікном, наприклад Хемінга.

Основною процедурою, що дозволяє істотно знизити вплив варіативності значень амплітуд спектра, є нормування мовних сигналів (забезпечення постійної потужності на сегментах коротко часового спектрального аналізу мовних сигналів). Для нормування потужності на кожному сегменті аналізу безперервного речового мовного сигналу $x(t)$, на основі стандартного дискретного перетворення Гілберта відліків $x[n], n = \overline{1,T}$, формуються відліки $y[n]$ сигналу $y(t)$, ортогонального початковому сигналу. По сукупності формованих комплексних відліків $\tilde{z}[n] = x[n] + jy[n] = U[n]\exp(j\tilde{Q}[n])$, відповідних безперервному аналітичному сигналу $\tilde{z}(t) = x(t) + j \cdot y(t) = U(t)\exp[j\tilde{Q}(t)]$, після логарифмічного перетворення отримують відліки комплексного процесу $\ln(\tilde{z}[n]) = U[n] + j\tilde{Q}[n]$. Значення реальної компоненти $U[n] = \mathrm{Re}(\tilde{z}[n]) = |\tilde{z}[n]|$ цього процесу є відліками $U[n] = |\tilde{z}[n]| = \sqrt{x^2[n] + y^2[n]}$ огинаючої $U(t) = \sqrt{x^2(t) + y^2(t)}$ безперервного аналітичного сигналу $\tilde{z}(t)$. Значення уявної компоненти $\tilde{Q}[n] = \mathrm{Im}(\tilde{z}[n])$ є головними значеннями відліків безперервної повної фази $Q(t)$ аналітичного сигналу $\tilde{z}(t)$ в інтервалі $(-\pi/2, \pi/2)$.

Повна фаза $Q(t)$ аналітичного мовного сигналу $\tilde{z}(t)$ з відліками $\tilde{z}[n]$ має в цілому лінійний тренд, який визначається значенням середньої частоти спектра мовного сигналу. Варіації (без розривів) значень повної фази $Q(t)$ на сегменті аналізу від лінійного тренду зумовлені варіаціями миттєвих частот мовного сигналу на сегменті від середнього значення. Тому на основі відомих співвідношень $\cos(Q[n]) = x[n]/U[n]$, $\sin(Q[n]) = y[n]/U[n]$ можна відновити за головними значеннями $Q[n]$ послідовність миттєвих відліків $Q[n]$ безперервної повної фази $Q(t)$ з лінійним трендом і нульовим початковим значенням. Послідовність отримуваних згладжених відліків повної фази $Q[n]$ містить в собі усю інформацію про високочастотну структуру початкового мовного сигналу і його низькочастотної складової (що огинає). Це зумовлено тим, що відліки $F[n]$ миттєвої циклічної частоти $F(t)$ аналітичного сигналу (похідній повної фази, що є $Q(t)$ також пов'язані з огинаючою $U(t)$ (по Гілберту) початкового мовного сигналу:

$$F(t) = \partial Q(t)/\partial t = \frac{W[x(t), y(t)]}{\sqrt{x^2(t) + y^2(t)}},$$

де $W[x(t), y(t)]$ – визначник Вронського процесів $x(t), y(t)$, що характеризує міру їх лінійної незалежності. Наступним кроком є формування послідовності відліків нового комплексного аналітичного сигналу $\bar{z}[n] = A\exp(jQ[n])$ з постійним значенням амплітуди A (що нормованою огинає), відліки повної фази $Q[n]$ якого подібні до фазової структури початкового мовного сигналу на сегменті аналізу. При цьому середня потужність (енергія) аналітичного сигналу $\bar{z}[n] = A\exp(jQ[n])$ однакова для усіх сегментів аналізу. Ця обставина істотно спрощує рішення наступної задачі статистичної спектральної ідентифікації мовних сигналів. Експериментально встановлено, що після ЦАП (відновлення) дійсної компоненти $\mathrm{Re}(\bar{z}[n]) = A\cos(Q[n])$ нормованого аналітичного мовного сигналу $\tilde{z}[n]$ по усіх сегментах, вона сприймається на слух практично так само, як і відновлений по відліках $x[n]$ мовний сигнал без пауз.

У результаті виконання розглянутих операцій аналого-цифрового перетворення отримуємо нормований по потужності комплексний аналітичний сигнал і його спектр Фур'є для кожного дискретного значення частоти f_f (сукупність значень модулів $X_i(f_j), i=\overline{1,N}$) по усіх N сегментах аналізу. При цьому можна порівняно просто (наприклад, використовуючи алгоритми швидкого сортування і формування варіаційного ряду $\tilde{X}_{1/n}(f_j),...,\tilde{X}_{n/n}(f_j)$), визначити порядкові статистики (вибіркові квантилі) $X_p(f_j)$ для кожного дискретного значення частоти f_f поточного спектра Фур'є [4; 5].

На рис. 3 показані квантилі 50 % та 95 % на частоті 38 для ненормованого та нормованого сигналу.

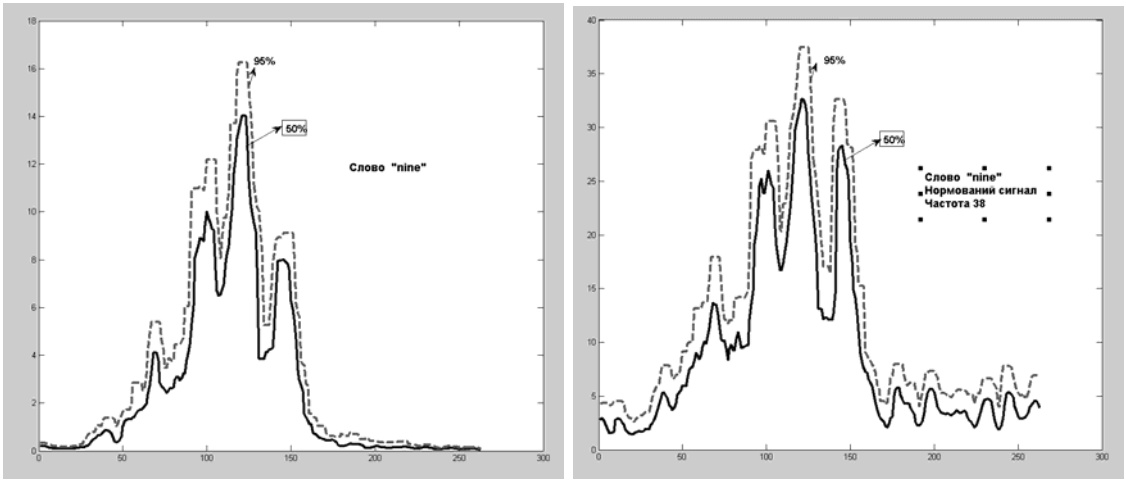

Рис. 3. Квантилі 50 % та 95 % для нормованого та ненормованого мовного сигналу

Для вирішення наступного завдання статистичної ідентифікації джерел мовних сигналів найбільш ефективними є сукупність вибіркових процентилів $X_{k/100}(f_j), k=\overline{1,99}$, які ділять увесь діапазон значень варіаційного ряду $\tilde{X}_{1/n}(f_j),...,\tilde{X}_{n/n}(f_j)$ на 100 нерівних за величиною підінтервалів, в які потрапляє однакова (не менше десяти-двадцяти) кількість елементів ряду.

Аналізовані порядкові статистики $X_p(f_j)$ можна розглядати як впорядковану сукупність нових випадкових величин (випадковий вектор). У межі для $n\to\infty$, функція щільності розподілу ймовірності (ПРВ) $g[X_p(f_j)]$ кожної порядкової статистики $X_p(f_j)$ асимптотично нормальна:

У межах, для $n\to\infty$, значення порядкових статистик $X_p(f_j)$ вибірки прагнуть до відповідних значень квантилів $X_p(f_j)$ функцій розподілу $G[X(f_j)]$ генеральної сукупності.

Таким чином, при переході від аналізу вибірки випадкових значень амплітудного спектра (вид і параметри досить складних функцій розподілу яких на практиці, зазвичай, невідомі), до аналізу значно меншого об'єму порядкових статистик (вибіркових квантилів) $X_p(f_j)$ амплітудного спектра нормованого за

потужністю аналітичного мовного сигналу, істотно спрощується їх наступна обробка. Це зумовлено тим, що сукупність нових випадкових величин – порядкових статистик є асимптотично нормальним. При цьому вибіркові процентилі робастні до аномальних значень спектральних компонент $X_i(f_j)$, які розташовані на "хвостах" відповідного варіаційного ряду $\tilde{X}_{1/n}(f_j),...,\tilde{X}_{n/n}(f_j)$ [5]. У результаті від аналізу стохастичної сукупності значень $X_i(f_j)$ переходимо до аналізу статистично стійких залежностей порядкових статистик $X_p(f)$ від частоти f спектра істотно меншого об'єму. При достатній кількості сегментів мовного сигналу, коли всі основні фонетичні конструкції мови, залежності порядкових статистик $X_p(f)$ від частоти спектра відбивають основні особливості спектра мовного сигналу. Ці особливості відбиваються як в самому характері поведінки залежностей порядкових статистик, так і в співвідношенні значень залежностей порядкових статистик. Статистична стійкість і асимптотична ефективність порядкових статистик $X_i(f_j)$ амплітудного спектра Фур'є забезпечують подібність залежностей $X_p(f)$ при збільшенні кількості сегментів мовного сигналу (збільшенні об'єму початкової вибірки). Таким чином, для кожного k-го мовної одиниці бази можна сформувати складені значень усіх порядкових статистик амплітудного спектра Фур'є нормованого аналітичного сигналу. Подальша процедура ідентифікації мовних одиниць полягає в зіставленні (попарно) порядкових статистик спектрального образу з порядковими статистиками спектральних образів бази еталонів. Вектор різниці спектральних образів є асимптотично нормальним з нульовим середнім при їх збігу і відмінним від нуля середнім для різних мовних одиниць за однакових умов реєстрації відповідних мовних сигналів. Рішення про ідентичність мовних одиниць з одним із цільових еталонів бази полягає у визначенні номера мовної одиниці за критерієм мінімуму квадрата норми Евкліда різниць спектральних образів (векторів) бази еталонів X_k і вхідного образу мовної одиниці Y.

Значення сформованих поточних порядкових статистик поточних спектрів мовного сигналу слова, що ідентифікується, зіставляється з відповідними значеннями **поточних** порядкових статистик поточних спектрів мовних сигналів слів бази даних (словника).

Подальше підвищення ефективності методу статистичної ідентифікації можна забезпечити за рахунок багатоканальної частотної обробки (наприклад, використання банків фільтрів мел- або барк- шкали), ідентифікації залежностей по кожній поточній порядковій статистиці з мажоритарним правилом прийняття рішення про ідентичність сигналів в цілому та використання двох порогових правил прийняття рішення про ідентичність залежностей поточних порядкових статистик для зниження ймовірності неправильної ідентифікації.

Висновки

Розглянутий новий метод робастної спектральної ідентифікації словоформ на основі зіставлення вибіркових поточних порядкових статистик поточних амплітудних спектрів нормованих по потужності мовних сигналів словоформ є достатньо продуктивним та перспективним. Поточні порядкові статистики поточних амплітудних спектрів мовних сигналів словоформ достатньо повно характеризують зміни у часі індивідуальних частотно-часових характеристик сигналів мовних одиниць, що важливо при ідентифікації голосових команд у системах комп'ютерного інтерфейсу; ключових мовних комбінацій у системах

верифікації, ідентифікації та охоронної сигналізації; визначення мови та тематики потоку суцільного мовлення за тематичними словниками в системах моніторингу.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Dong Yu. Li Deng Automatic Speech Recognition A Deep Learning Approach Springer-Verlag London 2015. 329 p.

2. *Селетков В. Л.* Вариант спектральной идентификации речевых сигналов. Изв. вузов. Радиоэлектроника. 2006. № 4 [ч. 2]. С. 63–69.

3. *Селетков В.Л. Кузнецов М.В.* Идентификация диктора по порядковым статистикам спектров речевых сигналов. Сучасна спеціальна техніка. 2014. № 4(39). С. 33–43.

4. Пат. 112813 Україна. МПК 2016 G10L 15/00, G10L 17/00. Спосіб ідентифікації слів за порядковими статистиками спектрів аналітичних мовних сигналів. Опубл. 25.10.16.

5. *Дейвид Г.* Порядковые статистики. Москва: Наука, 1979. 335 с.

Отримано 11.09.2017
Рецензент Єрохін В.Ф., д.т.н., проф.