СИСТЕМИ ТЕХНІЧНОГО ЗОРУ І ШТУЧНОГО ІНТЕЛЕКТУ З ОБРОБКОЮ ТА РОЗПІЗНАВАННЯМ ЗОБРАЖЕНЬ

УДК 681.327.12

М. М. БИКОВ, В. В. КОВТУН

ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ ОЗНАК РОЗПІЗНАВАННЯ МОВЦІВ ПРИ ВИКОРИСТАННІ ЗАГОРТАЛЬНИХ НЕЙРОМЕРЕЖ

Вінницький національний технічний університет,
21021, вул. Хмельницьке шосе, 95, м. Вінниця, Україна,
E-mail: kovtun_v_v@vntu.edu.ua

Анотація. У статті автори навели результати дослідження ефективності спектральних ознак для прийняття рішень автоматизованою системою розпізнавання мовців критичного застосування із згортальним нейромережевим класифікатором глибокого навчання, використання якого зумовило представлення інформативних ознак у графічному вигляді.

Ключові слова: автоматизована система розпізнавання мовців критичного застосування, розпізнавання образів, цифрова обробка сигналів, кепстральний аналіз, згортальна нейромережа

Аннотация. В статье авторы привели результаты исследования эффективности спектральных признаков для принятия решений автоматизированной системой распознавания дикторов критического применения с сверточным нейросетевым классификатором глубокого обучения, использование которого обусловило представление информативных признаков в графическом виде.

Ключевые слова: автоматизированная система распознавания говорящих критического применения, распознавание образов, цифровая обработка сигналов, кепстральных анализ, свертывающая нейросеть

Abstract. The study represents the results of the research of efficiency of the spectral features of speech signal for automated decision-making critical system for speaker recognition with convolutional neural network deep learning classifier, the use of which caused the submission of informative features in graphical view.

Keywords: automated recognition system speakers critical application, pattern recognition, digital signal processing, cepstral analysis, neural network coagulation

ВСТУП

Індивідуальність акустичних характеристик голосу визначається трьома факторами: механікою коливань мовних низок, анатомією мовного тракту і системою управління артикуляцією. Фізіологічні особливості мовних низок забезпечують індивідуальність процесу їх автоколивань при проходженні через повітряного потоку з легень у процесі мовотворення. Частота коливань низок і форма імпульсів об'ємної швидкості потоку, який протікає артикуляційним трактом, впливають на форму обвідної спектра мовного сигналу і його часові параметри. Геометричні розміри різних відділів мовного тракту, а також механічні властивості його тканин визначають резонансні частоти і швидкість загасання коливань на резонансних частотах, які можна використовувати для розпізнавання особи мовця. В спектрі мовного сигналу ці ознаки представлені енергетичними сплесками (піками) на певних частотах. Інформацію несе як існування піків на певних частотах, так і їх параметри, як то, абсолютне значення, ширина, форма і т.ін. Система управління артикуляцією формує просодичні характеристики: динаміку частоти основного тону, тривалість фонетичних сегментів, швидкість руху артикуляторів, а також, ефекти коартикуляції, які по різному проявляються у різних мовців. Індивідуальність стилю мовлення проявляється на досить тривалих висловах, і може враховуватися при задачах сегментації мовців в потоці мовних сигналів, які містять записи мовлення кількох осіб. Акустично індивідуальний стиль мовлення відтворюється у вигляді контуру частоти основного тону, тривалості слів і його сегментів, ритміки ударних сегментів, тривалості пауз, гучності. Ці ознаки специфічні і носять, в основному, допоміжний характер. Простір

ознак, в якому приймається рішення про особу мовця, має формуватися з урахуванням всіх факторів процесу мовотворення: параметрів джерела голосу, резонансних частот мовного тракту і їх затухання, а також, динамікою управління артикуляцією. Зокрема, найчастіше розглядаються такі параметри голосового джерела як середня частота основного тону. Спектральні характеристики мовного тракту описуються обвідною спектра і його середнім нахилом, формантними частотами і їх смугами та, найкраще, кепстром. Слід відзначити, що критичні системи [1] і автоматизовані системи розпізнавання мовців критичного застосування зокрема, передбачають збереження високої якості функціонування на протязі життєвого циклу функціонування системи що можна забезпечити лише комплексним підходом до вибору інформативних для розпізнавання мовця ознак та застосуванням завадостійкої інтелектуальної системи класифікації.

ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

Автори пропонують дослідити якость розпізнавання мовців за виголошеними парольними фразами згортальною нейронмережею глибокого навчання, яка, на відміну від нейромереж прямого поширення сигналу, ефективно узагальнює подану на її вхід графічну інформацію використовуючи комплекс операцій згортання та субдискретизації. Для забезпечення цілісного представлення мовного сигналу із екстрагуванням індивідуальних його особливостей автори застосовуватимуть результати кепстрального RASTA-аналізу (враховує індивідуальність артикулятоного тракту) та залежності зміни частоти основного тону у часі (враховує індивідуальність голосового джерела), які представлятимуть у вигляді зображень, які подаватимуться на вхід нейромережевого класифікатора.

ВИДІЛЕННЯ ІНФОРМАТИВНИХ ОЗНАК ДЛЯ АВТОМАТИЗОВАНОЇ СИСТЕМИ РОЗПІЗНАВАННЯ МОВЦІВ КРИТИЧНОГО ЗАСТОСУВАННЯ

Для виділення частоти основного тону автори застосували метод виділення та оброблення основного тону на основі модифікованої математичної моделі слухової системи людини [2].

Запропонований авторами спосіб здійснення RASTA-аналізу у задачі розпізнавання мовців наведено далі. У роботах по розпізнаванню мовців [3] домінує метод кепстральних перетворення спектра мовних сигналів, в якому на інтервалі часу в 10—20 мс обчислюється спектр потужності мовного сигналу, потім застосовується зворотне перетворення Фур'є від логарифму цього спектру (кепстра) [4—6] і знаходяться коефіцієнти кепстра:

$$c_n = \frac{1}{\Theta}\int_0^{\Theta} \log\left|S(j\omega,t)\right|^2 e^{-jn\Omega\omega} d\omega \qquad (1)$$

де $\Omega = 2\pi / \Theta$, Θ — верхня частота в спектрі мовного сигналу, $\left|S(j\omega,t)\right|^2$ — спектр потужності.

Кількість кепстральних коефіцієнтів n залежить від бажаного згладжування спектра, і знаходиться в межах від 20 до 40. Якщо використовується гребінка смугових фільтрів, то коефіцієнти дискретного кепстральних перетворення обчислюються зазвичай як

$$c_n = \sum_{m=1}^{M}\left[\log Y(m)\right]\cos\left[\frac{\pi n}{M}\left(m-\frac{1}{2}\right)\right], \qquad (2)$$

де $Y(m)$ — вихідний сигнал m-го фільтра, c_n — n-й коефіцієнт кепстра.

Властивості слухової системи людини враховуються нелінійним перетворенням шкали частот, зазвичай за шкалою Мел, яка утворюється виходячи із припущення про існування в слуховій системі так званих критичних частотних смуг, в межах яких сигнали нероздільні за частотою. Шкала Мел обчислюється як

$$M(f) = 1125 \cdot \ln(1 + f / 700), \qquad (3)$$

де f — частота в Гц, M — частота в Гаммелах.

Існує інша шкала — Барк, при формуванні якої різниця між двома граничними частотами, що обмежують границі критичної смуги, дорівнює 1 барк. Частота в Барк шкалі обчислюється як

$$B = 13\text{arctg}(0,00076 f) + 3,5 arctg(f / 7500). \qquad (4)$$

Коефіцієнти кепстрального перетворення формують простір, в якому відбувається процес розпізнавання мовця. Ці коефіцієнти скорочено позначаються як MFCC — Mel Frequency Cepstral Coefficients або BFCC — Bark Frequency Cepstral Coefficients, в залежності від обраної шкали. Окрім кепстральних коефіцієнтів при розпізнаванні часто використовуються перші і другі різниці кепстральних коефіцієнтів у часі (так звані похідні), що втричі збільшує розмірність простору прийняття рішень, але й

підвищує ефективність розпізнавання мовців. Кепстр описує форму обвідної спектра сигналу, як інтегральної ознаки, яка враховує індивідуальні характеристики джерела збудження мовних сигналів (голосового, турбулентного і імпульсного) та мовного тракту особи.

Щоб врахувати геометричні та фізіологічні характеристики артикуляції тракту, авторами був розроблений алгоритм обчислення RASTA-PLP коефіцієнтів. Аналіз мовних сигналів RASTA-PLP складається з двох частин — PLP (Perceptual linear prediction) [7] — лінійний прогноз із урахуванням особливостей слухового сприйняття людини і RASTA (RelAtive SpecTrA) — оброблення мовного сигналу з метою видалення спектральних компонент, швидкість зміни яких відмінна від швидкості зміни вибраних компонент мови, що дозволяє зосередити увагу на областях спектру, що відповідають найбільш інформативним для розпізнавання мовця ознакам, або менш піддавалися впливу шумів. Пропонується такий алгоритм обчислення RASTA-коефіцієнтів:

1. Запис мовного сигналу розбивається на фрейми — відрізки довжиною 10-30 мс з перекриттям 5—10 мс;

2. На кожному фреймі обчислюється квадрат модуля перетворення Фур'є, що еквівалентно процедурі отримання спектрограми фрейму;

3. Частотний діапазон $[0, f_s / 2]$ розбивається на N критичних смуг (critical bands). Ці смуги відповідають рівномірному розподіленню частотного діапазону в bark-шкалі. Підраховуються логарифми енергій $\log E_i$ у всіх критичних смугах;

4. Виконується RASTA-фільтрація фреймів, при чому дискретна передавальна функція RASTA-фільтру має вигляд

$$R(B) = 0{,}1B^4 \frac{2 + B^{-1} - B^{-3} - 2B^{-4}}{1 - 0{,}94B^{-1}}, \tag{5}$$

де B обчислюється із співвідношення (4),

5. Через фільтр $R(B)$ пропускається кожна з N спектральних траєкторій, отриманих на попередньому етапі. RASTA-фільтрація прибирає сталі складові логарифмів спектральних компонент;

6. Згладжений логарифмічний спектр, отриманий в результаті RASTA-фільтрації, повертається в лінійний масштаб. Потім на кожному фреймі він множиться на криву рівної гучності [4], яка визначається співвідношенням

$$H(f) = \frac{f^4}{\left(f^2 + 1{,}6 \cdot 10^5\right)^2} \cdot \frac{f^2 + 1{,}44 \cdot 10^6}{f^2 + 9{,}61 \cdot 10^6}, \tag{6}$$

де f — частота в лінійному масштабі;

7. Отримані на попередньому етапі спектри для кожного фрейма підносяться до ступеня 0,33;

8. Від спектра береться зворотне перетворення Фур'є, результатом якого буде автокореляційна функція $R(k)$, $k = 0, \ldots, L_{FFT}$;

9. Обчислюються коефіцієнтів лінійного прогнозу порядку p за допомогою рекурсії Левінсона-Дарбина [7];

10. Кепстральні коефіцієнти c_n обчислюються через рекурентні співвідношення

$$c_n = -a_n - \sum_{k=1}^{n-1} \frac{k}{n} c_k a_{n-k}, \; n = 1, \ldots, p, \tag{7}$$

11. Проводиться «ліфтинг» коефіцієнтів c_n:

$$c_n^{'} = n^{0.6} c_n, \; n = 1, \ldots, p. \tag{8}$$

На рисунках 1—4 наведено результати роботи запропонованого алгоритму RASTA-аналізу мовного сигналу, спектрограму якого, отриману із застосуванням швидкого перетворення Фур'є наведено на рисунку 1. На рисунку 2 переставлено спектрограму, смугові фільтри для отримання якої були організовані за Барк-шкалою. Рисунок 3 містить візуалізацію обчислених для фрейму 13 коефіцієнтів кепстрального аналізу мовного сигналу, а рисунок 4 — містить результати RASTA-аналізу мовного сигналу.

Розрахувавши кепстральні коефіцієнти можна виконати обернену до аналізу мовного сигналу задачу — його синтез, так синтезована парольна фраза, візуалізована у вигляді спектрограми і спектрограми, смугові фільтри для отримання якої були організовані за барк-шкалою, наведені на рисунках 5 та 6 відповідно. Можна помітити, що сигнал зазнав суттєвого згладжування в області низьких

частот, що спричинятиме додаткові похибки при визначенні інформативних для розпізнавання мовця ознак, таких, наприклад, як частота основного тону.

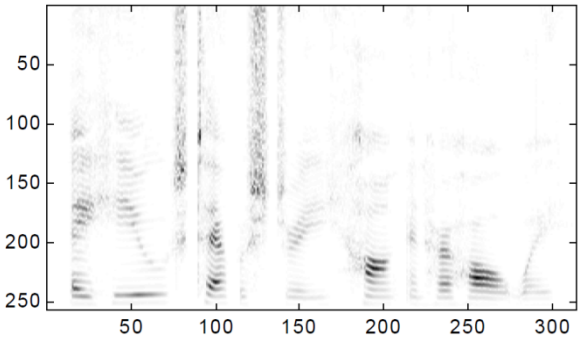

Рис. 1. FFT-спектрограма тестового запису

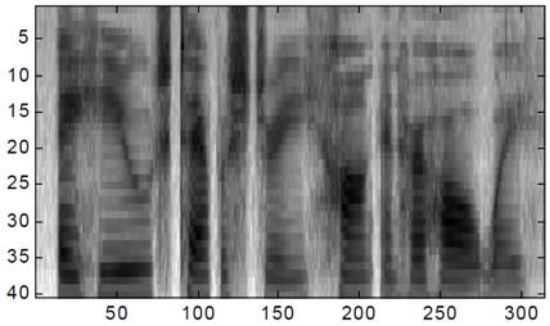

Рис. 2. BARK-спектрограма тестового запису

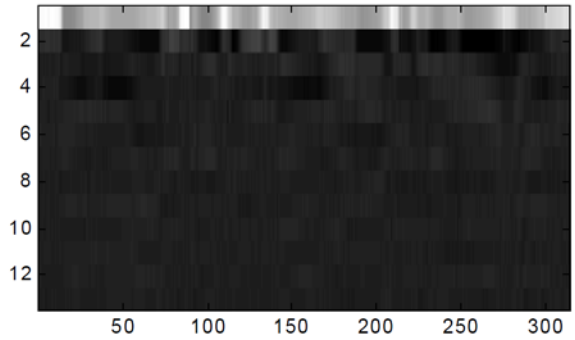

Рис. 3. Візуалізація 13 кепстральних коефіцієнтів після дискретного косинусного перетворення

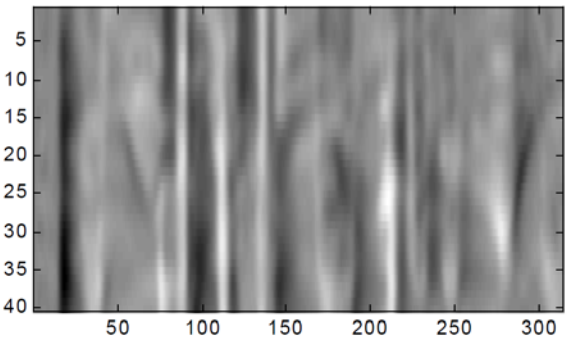

Рис. 4. RASTA-аналіз тестового запису

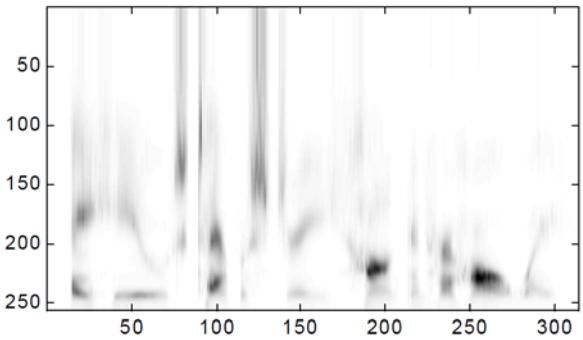

Рис. 5. Відтворення оригінальної FFT-спектрограми за значеннями кепстральних коефіцієнтів

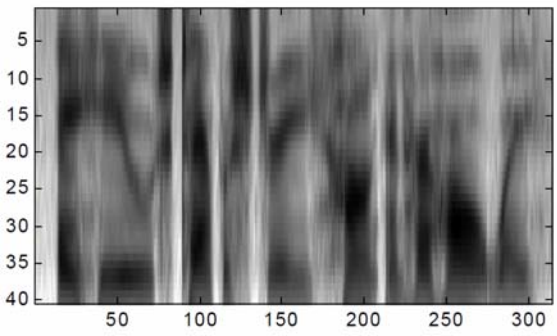

Рис. 6. Відтворення оригінальної Барк-спектрограми за значеннями кепстральних коефіцієнтів

ФОРМУВАННЯ ЗАГОРТАЛЬНОЇ НЕЙРОМЕРЕЖІ ДЛЯ АВТОМАТИЗОВАНОЇ СИСТЕМИ РОЗПІЗНАВАННЯ МОВЦІВ КРИТИЧНОГО ЗАСТОСУВАННЯ

Для розпізнання мовців за результатами RASTA-аналізу пропонується використовувати згортальну нейромережу [8], особливістю якої є самостійний вибір інформативних для розпізнавання ознак з вхідного зображення. Основна ідея згортальної нейронної мережі полягає в чергуванні згортальних (C-layers) і субдискретизуючих (S-layers) шарів, які фіналізуються повнозв'язним (F-layer) шаром. Діяльність згортальної нейромережі базується та трьох засадах: локальність полів сприйняття, поділюваність ваг і просторова субдискретизація. Під локальним розуміється сприйняття, коли на вхід одного нейрона подається не все зображення, а окремі його області. При такому підході топологія зображення зберігається від шару до шару. Концепція поділюваних ваг полягає в тім, що для великої кількості зв'язків використовується невелика кількість ваг. Наприклад, якщо на вхід подається

зображення розмірами 50 х 50 пікселів, то на вхід кожного з нейронів наступного шару подається невелика область цього зображення, наприклад, 5 х 5, і для обробки кожного з фрагментів буде використаний той самий набір ваг (ядер). Просторова субдискретизація за допомогою S-шарів в згортальних нейромережах передбачає зменшення просторової розмірності зображення. Чергування шарів дозволяє складати карти ознак з карт попереднього шару, що сприяє на практиці розпізнаванню складних ієрархій ознак.

У процесі функціонування згортальних шарів кожен фрагмент зображення поелементно множиться на невелику матрицю ваг, а отримані добутки підсумовуються. Отримана сума визначає один піксель вихідного зображення — карти ознак. На кожному шарі відбувається формування кількох таких карт. Спочатку вихідне зображення розділяється на області, з яких будуються карти ознак. При цьому використовується один той самий набір вагових коефіцієнтів, а на виході отримується значення, яке відповідає одному елементу вихідний карти. Цей алгоритм виконується кілька разів для кожного елемента, потім завантажуються інші коефіцієнти і формується наступна карта ознак. На шарах субдискретизації виходами є не одне, а кілька значень, приблизно в два рази менше кількості входів, що дозволяє знижувати розмірність зображення і виявляти в подальшому більш загальні, інваріантні до масштабування ознаки. Повнозв'язні шари є класичними персептронами, які добре розпізнають прості образи. Архітектура мережі, створеної для розпізнання в автоматизованій системі розпізнавання мовців критичного застосування за результатами RASTA-аналізу, наведена на рисунку 7.

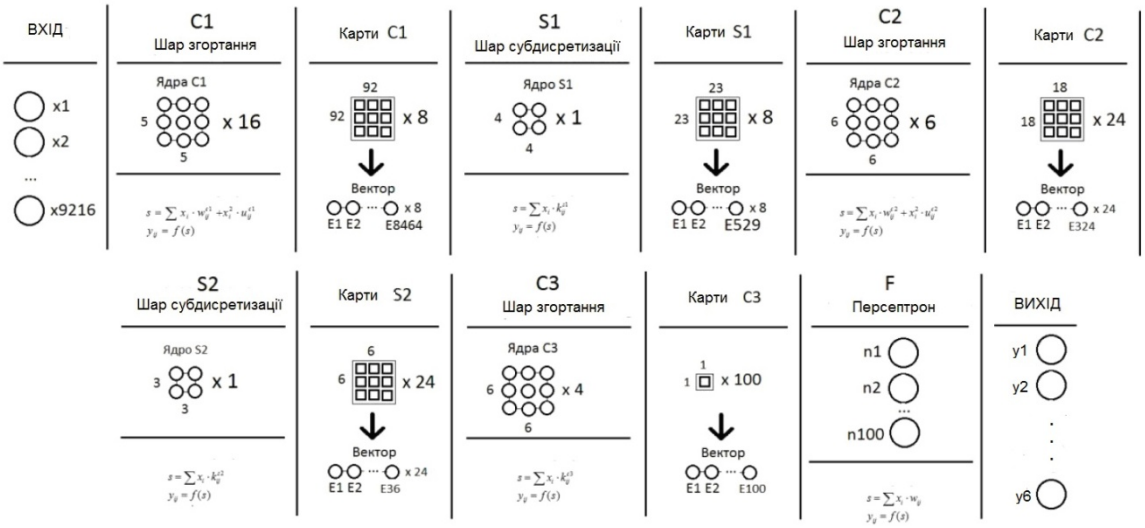

Рис. 7. Архітектура згортальної нейромережі глибокого навчання для розпізнавання мовців за даними RASTA-аналізу та частоти основного тону

Для проведення процедури розпізнавання ми створили загортальні нейромережі глибокого навчання з використанням кросплатформних бібліотек Caffe [9] з відкритим програмним кодом. На розпізнавання мовців нейромережу було навчено із використанням алгоритму стохастичного градієнтного спуску (Stochastic Gradient Descent Algorithm) [10].

ПОСТАНОВКА ЕКСПЕРИМЕНТУ ТА АНАЛІЗ РЕЗУЛЬТАТІВ

В якості бази еталонних записів, які піддавалися RASTA-аналізу та виділенню частоти основного тону, використано записи із безкоштовної бази даних NOIZEUS [11] — спеціалізованої бази даних Школи інжинірингу та комп'ютерних наук Еріка Джонсона при Університеті Техасу в Далласі, США, яка використовується для дослідження алгоритмів покращення звуку і складається з 30 речень англійської розмовної мови, вимовлених трьома чоловіками та трьома жінками (по 5 на кожного диктора, частота дискретизації записів складає 25 кГц, але задля додавання шуму була зменшена до 8 кГц) та записів типових побутових та техногенних шумів. В ході експерименту автоматизовану систему розпізнавання мовців критичного застосування навчали як записами чистих парольних фраз, так і парольними фразами із додаванням шумів. В першому випадку навчальна вибірка містила 18 парольних фраз, у другому — 576, де до чистого сигналу додавався штучний шум з рівнями шум/сигнал 0 дБ, 5 дБ, 10 дБ, 15 дБ відповідно.

Для навчання нейромережі використано 60 % обсягу бази аудіозаписів, в яку увійшли екземпляри записів без шумів та із різним рівнем шум/сигнал (5, 10. 15 дБ) відповідно. Тестуючи вибірка

склала решту 40 % аудіозаписів. Узагальнені результати експерименту представлено на рисунку 8, де імовірність правильного розпізнавання розраховувалася за формулою

$$P = \frac{\sum_i (Np_i)}{N}, \tag{9}$$

де Np_i — кількість правильних результатів розпізнавання i -го мовця, N — загальна кількість експериментів.

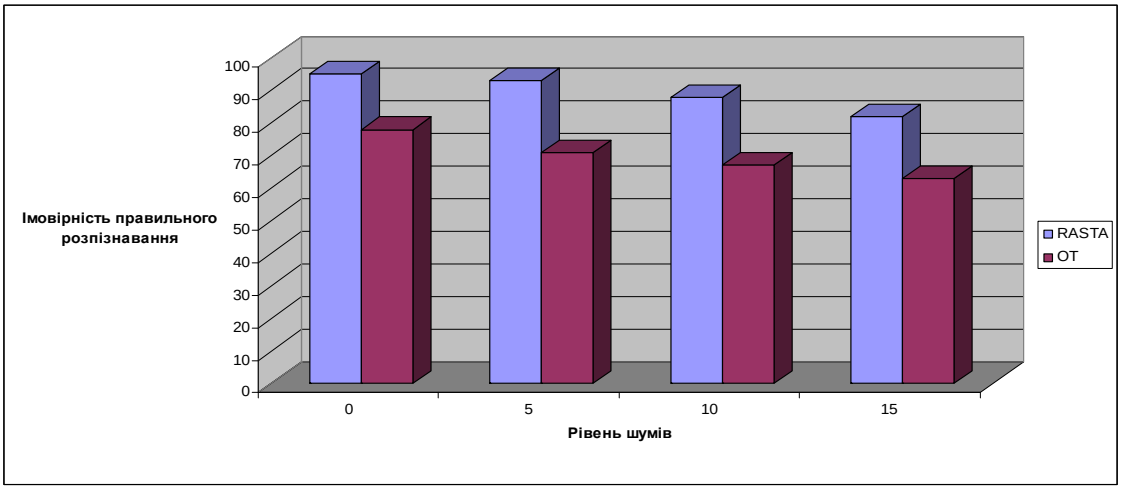

Рис. 8. Результати експерименту із розпізнавання дикторів

ВИСНОВКИ

Отже, авторами розглянуто проблему дослідження ефективності інформативних ознак для автоматизованої системи розпізнавання мовців критичного застосування. Для цього створили адаптований для специфіки критичних систем алгоритм кепстрального RASTA-аналізу (враховує індивідуальність артикулятоного тракту) та врахували залежність зміни частоти основного тону у часі (враховує індивідуальність голосового джерела). Отримані значення інформативних ознак подавалися на загортальну нейромережу для розпізнавання осіб мовців. Результати роботи системи дозволяють стверджувати, що загортальна нейромережа глибокого навчання може ефективно використовуватися в задачі створення автоматизованих систем розпізнавання мовців критичного застосування за даними RASTA-аналізу парольних записів (імовірність правильного розпізнавання на рівні 87 % для 6 дикторів при співвідношенні шум/сигнал 15 дБ). Проте, якість прийнятих нейромережею рішень залежить від поданої на її вхід інформації, так використання для розпізнавання дикторів зображень частоти основного тону, показало нижчу ефективність ніж використання RASTA-аналізу, які описують мовний сигнал у частотній області, що зумовлює доцільність проведення додаткових досліджень для вибору найкращого візуального представлення мовного сигналу для класифікації загортальною нейромережею.

СПИСОК ЛІТЕРАТУРИ

1. Critical system — Wikipedia [Електронний ресурс] — Режим доступу : https://en.wikipedia.org/wiki/Critical_system.
2. Биков М. М. Аналіз ефективності ідентифікації мовця за частотою основного тону / М. М. Биков, В. В. Ковтун. — Вісник Хмельницького національного університету. — 2004. — № 2. — Ч.1. — Т. 2 (60). — С. 20—23.
3. Рабинер Л. Цифровая обработка речевых сигналов / Л. Рабинер, Р. Шафер. — М. : Радио и связь, 1981. — 496 с.
4. Hermansky H. RASTA processing of speech / H. Hermansky, N. Morgan. — IEEE Trans. Speech and Audio Processing. — 1994. — 2, N 6. — P. 578—589.
5. Hermansky H. Perceptual Linear Prediction (PLP) analysis of speech / H. Hermansky. — J. Acoust. Soc. America. — 1990. — 87. — P. 1738—1753.
6. rasta-plp speech analysis — ICSI [Електронний ресурс] — Режим доступу : http://www.icsi.berkeley.edu/pubs/techreports/tr-91-069.pdf.
7. Perceptual Linear Predictive (PLP) Analysis of Speech [Електронний ресурс] — Режим доступу : http://seed.ucsd.edu/mediawiki/images/5/5c/PLP.pdf

8. CS231n: Convolutional Neural Networks for Visual Recognition [Електронний ресурс] — Режим доступу: http://cs231n.github.io/convolutional-networks/
9. Caffe | Deep Learning Framework [Електронний ресурс] — Режим доступу: http://caffe.berkeleyvision.org/.
10. An overview of gradient descent optimization algorithms [Електронний ресурс] — Режим доступу: http://sebastianruder.com/optimizing-gradient-descent/.
11. NOIZEUS: Noisy speech corpus - Univ. Texas-Dallas [Електронний ресурс] — Режим доступу: http://ecs.utdallas.edu/loizou/speech/noizeus/.

SPYSOK LITERATURY

1. Critical system — Wikipedia [Yelektronniy resurs] — Rezhim dostupu: https://en.wikipedia.org/wiki/Critical_system.
2. Bikov M. M. Analíz yefektivností ídentifíkatsíí movtsya za chastotoyu osnovnogo tonu / M. M. Bikov, V. V. Kovtun. — Vísnik Khmel'nits'kogo natsíonal'nogo unívesitetu. — 2004. — № 2. — CH.1. — T.2(60). — S. 20—23.
3. Rabiner L. Tsifrovaya obrabotka rechevykh signalov / L. Rabiner, R. Shafer. — M. : Radio i svyaz', 1981. — 496 s.
4. Hermansky H. RASTA processing of speech / H. Hermansky, N. Morgan. — IEEE Trans. Speech and Audio Processing. — 1994. — 2, N 6. — P. 578—589.
5. Hermansky H. Perceptual Linear Prediction (PLP) analysis of speech / H. Hermansky. — J. Acoust. Soc. America. — 1990. — 87. — P. 1738—1753.
6. rasta-plp speech analysis — ICSI [Yelektronniy resurs] — Rezhim dostupu: http://www.icsi.berkeley.edu/pubs/techreports/tr-91-069.pdf.
7. Perceptual Linear Predictive (PLP) Analysis of Speech [Yelektronniy resurs] — Rezhim dostupu: http://seed.ucsd.edu/mediawiki/images/5/5c/PLP.pdf
8. CS231n: Convolutional Neural Networks for Visual Recognition [Yelektronniy resurs] — Rezhim dostupu: http://cs231n.github.io/convolutional-networks/
9. Caffe | Deep Learning Framework [Yelektronniy resurs] — Rezhim dostupu: http://caffe.berkeleyvision.org/.
10. An overview of gradient descent optimization algorithms [Yelektronniy resurs] — Rezhim dostupu: http://sebastianruder.com/optimizing-gradient-descent/.
11. NOIZEUS: Noisy speech corpus — Univ. Texas-Dallas [Yelektronniy resurs] — Rezhim dostupu: http://ecs.utdallas.edu/loizou/speech/noizeus/.

Надійшла до редакції 16.12.2016 р.

БИКОВ МИКОЛА МАКСИМОВИЧ — к. т. н., доцент, професор кафедри комп'ютерних систем управління, Вінницький національний технічний університет, м. Вінниця, Україна.

КОВТУН В'ЯЧЕСЛАВ ВАСИЛЬОВИЧ — к. т. н., доцент, доцент кафедри комп'ютерних систем управління, Вінницький національний технічний університет, м. Вінниця, Україна.