НЕЙРОІНФОРМАТИКА ТА ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ

НЕЙРОИНФОРМАТИКА И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

NEUROINFORMATICS AND INTELLIGENT SYSTEMS

УДК 681.327.12

Бісікало О. В.[1], Грищук Т. В.[2], Ковтун В. В.[3]

[1]Д-р техн. наук, професор, декан факультету комп'ютерних систем і автоматики Вінницького національного технічного університету, Вінниця, Україна

[2]Канд. техн. наук, доцент, доцент кафедри комп'ютерних систем управління Вінницького національного технічного університету, Вінниця, Україна

[3]Канд. техн. наук, доцент, доцент кафедри комп'ютерних систем управління Вінницького національного технічного університету, Вінниця, Україна

ОПТИМІЗАЦІЯ КЛАСИФІКАТОРА АВТОМАТИЗОВАНОЇ СИСТЕМИ РОЗПІЗНАВАННЯ МОВЦЯ КРИТИЧНОГО ЗАСТОСУВАННЯ

Актуальність. Розглянуто питання адаптації згортального нейромережевого класифікатора для використання у автоматизовані системі розпізнавання мовців критичного застосування (АСРМКЗ). Об'єктом дослідження є індивідуальні особливості мовного процесу людини.

Мета роботи. Розроблення заходів по виділенню з мовного сигналу індивідуальних для мовця ознак, підвищення їх інформативності в результаті виконання факторного аналізу, їх візуальне представлення для використання згортального нейромережевого класифікатора та оптимізація його архітектури для потреб АСРМКЗ.

Метод. Запропоновано заходи по оптимізації процедури класифікації мовців АСРМКЗ, для чого теоретично обґрунтовано оптимальний спосіб представлення інформативних ознак і метод підвищення їх інформативності, обґрунтовано вид топологію і заходи для підвищення ефективності процесу розпізнавання мовців. Зокрема, обґрунтовано доцільність використання нормалізованих за потужністю кепстральних коефіцієнтів PNCC для опису фонограм, записаних в умовах шумного оточення, запропоновано використовувати фільтри Габора для представлення інформації, що аналізуватиметься згортальною нейромережею, вибрано оптимальний метод факторного аналізу, а саме, розріджений метод аналізу головних компонент, для зменшення розмірності вектору ознак із збереженням його інформативності, запропоновано удосконалену топологію згортальної нейромережі для АСРМКЗ, у якій фільтри Габора інтегровано у згортальний шар, що дозволяє оптимізувати їх параметри при навчанні нейромережі, і на повнозв'язному шарі використано глибоку нейромережу із bottleneck-шаром, ваги якого після навчання використано як вхідні дані для контрольного GMM/HMM-класифікатора.

Результати. Методи представлення та оптимізації інформативних для розпізнавання мовця ознак, методи їх візуального представлення та удосконалення топології згортальної нейромережі для прийняття рішень на їх основі.

Висновки. Отримані теоретичні результати знайшли емпіричного підтвердження. Зокрема, доведено стійкість удосконаленої згортальної нейромережі до присутності шумів оточення у вхідних фонограмах, яка виявилася вищою за показники звичайної згортальної нейромережі та глибокої нейромережі. При зростанні ВСШ>10 дБ контрольний GMM/HMM-класифікатор виявився ефективнішим за нейромережеві щодо імовірності прального розпізнавання мовців, що можна пояснити ефективністю використаної UBM-моделі, але він є і суттєво ресурсоємнішим. Також емпірично виявлено вікна банку фільтрів Габора, які надавали найбільш варіативну щодо індивідуальних особливостей мовлення, інформацію.

Ключові слова: автоматизована система розпізнавання мовців критичного застосування, обробка сигналів, нейромережа, факторний аналіз.

НОМЕНКЛАТУРА

CNN – згортальна нейронна мережа;
DNN – глибока нейронна мережа;
HMM – приховані марковські моделі;
SPCA – розріджений метод главних компонент;
UBM – універсальна фонова модель;
VAD – детектор мовної активності;
WCCN – операції внутрікласової коваріаційної нормалізації;
АСРМКЗ – автоматизована система розпізнавання мовця критичного застосування;
ВСШ – відношення сигнал/шум;
ІПР – середня імовірність правильного розпізнавання;

DOI 10.15588/1607-3274-2018-2-4

$STRF[n_t,k]$, $STRF$ – спектрально-темпоральні рецептивні поля;

LPC – коефіцієнтами лінійного прогнозу;

$MF[m,l]$ – спектральне представлення фонограм;

$MFCC[n]$, $MFCC$ – мел-кепстральні коефіцієнти;

$PLP[n]$, PLP – коефіцієнти перцептивного лінійного прогнозу;

$PNCC[m,l]$, $PNCC$ – нормовані за потужністю кепстральні коефіцієнти;

Σ_x – емпірична коваріаційна матриця;

Λ – множина гаусових сумішей;

φ – фаза;

λ_a, λ_b, λ_t – коефіцієнти забування;

v – домінуючий розріджений головний вектор;

ρ – невід'ємний параметр управління розрідженістю головного вектора;

ω_n, ω_k – темпоральні і спектральні частоти модуляції на різних амплітудних рівнях ΔA; $b(i)$ – коефіцієнти лінійного прогнозу;

D – розмірність простору ознак;

$Gb[x,y]$ – імпульсна передатна характеристика фільтра Габора;

f_0 – частота;

F_d – частота дискретизації;

$F(\ldots)$ – перетворення Фур'є;

$F(Gb)$ – гаусовський сигнал, екстремальне значення якого розташовано на центральній частоті фільтра;

$h[n_t,k]$ – віконна функція Хеннінга;

$h_l[\ldots]$ – віконна функція згладжування;

$h_m[n_t,k]$ – нейрон m-го вектора ознак, який відповідає певному фільтру Габора, рецептивне поле якого утворює матрицю $2K_m+1$ (смуги) на $2N_m+1$ (відліки), орієнтовану на поточну смугу фрейму $x(n,k)$;

im – зображення;

k – частота;

L – кількість фільтрів у банку;

l_1 – штрафна норма;

m, l – номери фрейму та частотної смуги фільтру відповідно;

n – кількість коефіцієнтів;

n_t – момент часу;

$p(x|\lambda)$, GMM – суміш гаусових розподілів;

$O[n_t,k]$ – передатні функції нейронів слухової кори голосного мозку;

S – множина мовців;

$s[n_t,k]$ – комплексна синусоїда;

$Q_f[m,l]$ – значення нижньої обвідної після напвіперіодного випрямляча;

$Q_0[m,l]$ – сигнал після напівперіодного випрямлення;

$Q_{tm}[m,l]$ – сигналу після процедури часового маскування;

$Q_{\text{вх}}[m,l]$ – результат ковзного усереднення $MF[m,l]$ на протязі M фреймів;

$Q_{\text{вих}}[m,l]$ – усереднена за часом потужність;

$\{x_i\}$ – результатами спостережень;

$x[\ldots]$ – короткочасний спектр сигналу.

ВСТУП

Серед існуючих класів автоматизованих систем окреме місце займають так звані критичні системи, які функціонують із високою надійністю і зберігають прогнозований її рівень на протязі всього життєвого циклу автоматизованої системи не залежно від будь-яких зовнішніх обставин. При створені критичних систем віддають перевагу перевіреним та знаним методам та технологіям перед новітніми розробками, які не пройшли всебічної емпіричної перевірки. Ресурсозатратні технології, використання яких для розробки поточних автоматизованих систем є економічно невигідним, допустимі при створенні критичних систем, для яких головним є надійність функціонування.

Актуальна класифікація критичних систем розділяє їх за реакцією на виникнення надзвичайних ситуацій та за галуззю експлуатації. За реакцією на виникнення надзвичайних ситуацій критичні системи розділяють на критичні системи, які мають продовжувати виконання функціональних операцій при виникненні відмов чи впливі непередбачуваних факторів, наприклад літак повинен продовжувати летіти за будь-яких обставин, та критичні системи, які повинні обов'язково безпечно завершувати функціонування не залежно від дії оточуючих факторів, наприклад, потрібно загальмувати потяг щоб перевести його у «безпечний» стан. За галуззю експлуатації критичні системи розділяють на:

– критичні системи збереження життя – це автоматизовані системи, збої у функціонуванні яких можуть привести до загибелі людей, суттєвих пошкоджень навколишнього середовища – це, наприклад, автоматизовані системи управління хімічним виробництвом, літаками, потягами метро, атомних електростанцій;

– критичні системи гарантованого функціонування – ці автоматизовані системи, які створюються із передбаченням гарантованого завершення виконуваної операції, наприклад, системи навігації, управління багажем у аеропортах;

– критичні системи економічного застосування – ці автоматизовані системи, створені з метою уникнення значних матеріальних або репутаційних витрат, що зазвичай забезпечується вчасним припиненням виконання певних операцій інтегрованою критичною системою, наприклад, у системах роботи із клієнтами у банках, інтернет-пошукові системи, ERP системи, системи роботи із біржовими операціями тощо;

– критичні системи інформаційної безпеки – ці автоматизовані системи унеможливлюють втрату конференційної інформації за будь-яких умов.

Автори проводять послідовні дослідження щодо синтезу теоретичних засад та методів для практичного впровадження автоматизованих систем розпізнавання мовця в критичні системи. Запропонований матеріал стосується оптимізації роботи підсистеми класифікації автоматизованої системи розпізнавання мовців критичного застосування (АСРМКЗ).

Об'єктом дослідження є індивідуальні особливості процесу мовотворення людини.

Предметом дослідження – методи виділення інформативних для розпізнавання особи мовця ознак із фонограм мовних сигналів, методи компактифікації їх представлення у факторному просторі, формулювання виду подачі інформативних ознак у відповідності із моделями слухового сприйняття людини і завадостійкі автоматизовані методи прийняття рішень для розпізнавання особи мовця.

Мета дослідження – підвищення стійкості АСРМКЗ до впливу різних видів шумів у вхідних фонограмах.

1 ПОСТАНОВКА ЗАДАЧІ

Формалізуємо постановку задачі розпізнавання мовців так. Нехай $S = \{s_1, s_2, \ldots, s_m\}$ – простір образів, які має розпізнавати АСРМКЗ, а $s_i \in S$ – об'єкт розпізнавання. Використовуючи певні правила сформуємо факторний простір F, причому функція $f(s_i): S \to X$ ставить у відповідність кожній фонограмі із записом мовного сигналу мовця s_i точку $f(s_i)$ у факторному просторі. Кожному i-му мовцеві у факторному просторі відповідає множина точок, кожна з яких відповідає опрацьованій фонограмі, утворюючи кластер $C_i \in X$, при чому кластери різних мовців мають не перетинатися. Вирішувалене правило $r(x): X \to M$ дозволяє із певною імовірністю P_{s_i} стверджувати, що точка $f(s_i)$ у факторному просторі належить s_i мовцеві. Задача авторів – сформувати факторний простір F із чіткими границями кластерів мовців і створити вирішувальне правило R, яке за інформацією про мовців із факторного простору максимізуватиме імовірності правильного розпізнавання P_{s_i} для всіх мовців з множини S.

2 ОГЛЯД ЛІТЕРАТУРИ

Основною специфікою АСРМКЗ є обґрунтовано висока імовірність його правильного розпізнавання при визначених рівнях ВСШ. Існуючі методи підвищення надійності розпізнавання мовців у шумному акустичному середовищі базуються на компенсації відмінностей між фонограмами без шуму («чистими» фонограмами) і фонограмами із шумом за рахунок математичних моделей [1, 2] або ведення додаткових факторів [3–5]. Перший підхід передбачає адаптацію моделі мовців до застосування у шумному середовищі, а другий – на використанні алгоритмів теорії цифрової обробки сигналів для фільтрації шумів із збереженням індивідуальності звучання фонограм або винайденні стійких до шумів та інформативних для розпізнавання особи мовця факторах.

Актуальні роботи, які можна віднести до першого варіанту базуються на моделюванні природної здатності людини якісно виконувати задачу розпізнавання мовця у шумному середовищі описуючи фізіологічні аспекти слухової системи та моделюючи когнітивні функції відповідних відділів кори головного мозку у вигляді нейромереж різної топології. Зокрема, дослідження [6, 7] демонструють можливість застосування параметрів прихованого шару (ваг вхідних зв'язків нейронів обраного прихованого шару) навченої глибокої нейромережі в якості факторів для розпізнавання мовців, і доводять їх більшу інформативність порівняно із традиційними факторами, як то, MFCC, LPC, короткочасною енергією і т. ін. Проте, ці результати отримано для значно схожих навчальних та тестувальних наборів фонограм мовців і не досліджувалися в умовах реального акустичного оточення, характерним для якого є присутність не лише періодичних шумів, притаманних каналам зв'язку, а і стохастичних природних чи техногенних шумів.

Роботи, які можна віднести до другого підходу, демонструють близькі уявлення щодо методів оброблювання мовних сигналів у задачах розпізнавання мови/мовця із подальшою відмінністю у інтерпретації отриманих результатів, яка виконується не автоматично, із незмінним порогом чутливості, що знижує ступінь адаптивності створюваних автоматизованих систем розпізнавання на їх основі, а отже, знижує робастість цих систем. Перспективним є синтез адаптивних методів аналізу мовних сигналів для виділенням індивідуальних особливостей мовлення, які поєднують інформативність із універсальністю та обчислювальною ефективністю.

3 МАТЕРІАЛИ І МЕТОДИ

Застосування згортального нейромережевого класифікатора у АСРМКЗ вимагає інформативного візуального представлення факторів, що описують мовний сигнал. Очевидно, що для цього найкраще підходить спектральне представлення, отримане на виході банку фільтрів

$$MF[m,l] = \sum_{k=0}^{1/2K-1} \left| x\left[m, e^{j\omega_k}\right] h_l\left[e^{j\omega_k}\right] \right|^2, \quad (1)$$

де $\omega_k = 2\pi k / F_d$.

Виділимо з фонограми мовного сигналу множини мел-частотних кепстральних коефіцієнтів [8] і коефіцієнтів лінейного прогнозу [9]. Для кожного фрейму на виході банку фільтрів (1) обрахуємо n MFCC-коефіцієнтів

$$MFCC[n] = L^{-1} \sum_{l=1}^{L} \log\left(MF[m,l] \cos\left[2\pi L^{-1}(l+0{,}5)n\right]\right), \quad (2)$$

де n PLP-коефіцієнтів на основі коефіцієнтів лінійного прогнозу $b(i)$ за рекурсій ним відношенням

$$PLP[n] = -b(n) + n^{-1} \sum_{i=1}^{n-1} (n-i) b(i) PLP[n-i]. \quad (3)$$

MFCC та PLP кепстральні коефіцієнти за рахунок розташування фільтрів у банку за Мел-шкалою є стійки-

ми до лінійних спектральних спотворень і відповідають базовим властивостям моделей слухового сприйняття С. Снефа, О. Гітза, Р Лайона [10], але є усередненним представленням відповідних спектральних характеристик і не дозволяють корегування спектральних характеристик мовного сигналу на частотах, які не входять у критичні смуги слухового сприйняття [11]. Для компенсації цих недоліків Х. Занг, М. Хейтц, И. Брюс и Л. Кані [12] сформулювали модель, яка описує відгук активності слухового нерва на мовний сигнал, супроводжуваний шумом, яка дозволила сформулювати метод представлення мовних сигналів множиною нормалізованих за потужністю кепстральних коефіцієнтів PNCC. Таке представлення дозволяє виконати компенсацію шуму за оцінками середньої у часі потужності, отриманої усередненням на протязі кількох фреймів короткочасної оцінки спектру потужності на виході банку фільтрів (1). Оцінюючи змінний у часі поріг шуму і віднімаючи його від короткочасної оцінки спектра потужності отримаємо чисту мовну складову фонограми:

$$Q_{\text{вих}}[m,l]=\begin{cases}\lambda_a Q_{\text{вих}}[m-1,l]+(1-\lambda_a)Q_{\text{вх}}[m,l], \\ \text{якщо } Q_{\text{вх}}[m,l]\ge Q_{\text{вих}}[m-1,l], \\ \lambda_b Q_{\text{вих}}[m-1,l]+(1-\lambda_b)Q_{\text{вх}}[m-1,l], \\ \text{якщо } Q_{\text{вх}}[m,l]< Q_{\text{вих}}[m-1,l].\end{cases} \quad (4)$$

Зауважимо, що точність обчислення коефіцієнтів PNCC цілком визначається точністю процедури детектування мовної активності (для визначення енергії шуму пауз). Отже, після отримання відкоригованого мовного сигналу до нього застосовують ідеальний лінійний напівперіодний випрямляч, після чого мовний сигнал обробляється у першому каналі із повторним застосуванням фільтру (4) для визначення порогового рівня потужності тільки для фреймів пауз. Одночасно у другому каналі до мовного сигналу застосовується процедура часового маскування

$$Q_{tm}[m,l]=\begin{cases}Q_0[m,l], \\ \text{якщо } Q_0[m,l]\ge \lambda_t Q_P[m-1,l], \\ \mu_t Q_P[m-1,l], \\ \text{якщо } Q_0[m,l]< \lambda_t Q_P[m-1,l].\end{cases} \quad (5)$$

Сформований таким чином пороговий енергетичний детектор мовної активності вибирає для формування вектора PNCC значення з першого (4) або другого (5) каналу за правилом

$$PNCC_n[m,l]=\begin{cases}\max\left(Q_{tm}[m,l],Q_f[m,l]\right), & \text{"мова"}, \\ Q_f[m,l], & \text{"пауза"}.\end{cases} \quad (6)$$

Наглядно операції для визначення векторів факторів за (2), (3) і (6) представлено на рис. 1.

Зауважимо, що у процесі отримання PNCC застосовувався банк гаматонних фільтрів [12], центральні частоти яких лінійно розподілені у частотному діапазоні 200– 8000 Гц відповідно до шкали ERB, яку використовують у психоакустиці для моделювання акустичних фільтрів. Перехід від частотної шкали у ERB виконується за формулою $ERBs = 21{,}41\log_{10}(1+f/229)$. До переваг шкали ERB окрім того, що вона адекватно відтворює такі властивості сприйняття як псевдологарифмічне зростання ширини критичної смуги із зростанням частоти і логарифмічний закон сприйняття інтервалів частот, можна віднести нечутливість до биття і інтермодуляціям між сигналом і фоновим шумом. Приклади MFC- та PN-спектрограм при різному відношенні сигнал/шум наведено на рис. 2.

Фильтри Габора [13] відносять до смугових фільтрів, які, переважно, використовують у задачах визначення крайових ефектів зображень, і дозволяють виявити діапазон частот сигнала у визначеному проміжку x і напрямку y. Імпульсна передатна характеристика фільтра Габора є добутком гаусової функції на гармонічну:

$$Gb[x,y]=K\frac{1}{\sqrt{2\pi\sigma_x}}e^{\frac{-x^2}{2\pi\sigma_x^2}}\frac{1}{\sqrt{2\pi\sigma_y}}e^{\frac{-y^2}{2\pi\sigma_y^2}}\cdot\cos(2\pi f_0 x+\varphi). \quad (7)$$

В процесі фільтрації відбувається згортання вхідного сигналу, в якості якого може виступати будь-яке зображення im, наприклад MFC- або PN-спектрограма, і фільтра Gb у просторовій області. Процес згортання у просторовій області при переході до частотного простору замінюють на множення. В результаті $F(im\cdot Gb)=F(im)\cdot F(Gb)$ – множення в частотній області амплітуди частот вхідного сигналу, близькі до частоти відповідного синусоїдального сигналу, підсилюються, а решта – затухають.

На ефективність фільтрів Габора критично впливають значення висоти σ_x й ширини σ_y гаусової компоненти (7). При обробці зображень емпірично виявлено, що оптимальна чіткість фільтру досягається при $\sigma_x=\frac{1}{\sqrt{2\pi f_0}}$ і $\sigma_y=\frac{3}{2}\sigma_x$, отже у подальший дослідженнях як початкові використовуватимуться ці величини.

Протягом останнього десятиліття ряд фізіологічних експериментів на різних видах ссавців показав, що нейрони у первинній слуховій корі чутливі до певних спектрально-темпоральних структур, названих спектрально-темпоральними рецептивними полями STRF [14], які є функціональними дескрипторами лінійної обробки змінних акустичних спектрів слуховою системою. Отримані на основі результатів цих досліджень спектрально-темпоральні ознаки увійшли до моделі STRF, яку почали використовувати у системах розпізнавання мови і мовців. Ряд досліджень [14, 15] дозволили визначити множину фільтрів із спектральною, темпоральною та спектрально-темпоральною модуляцією, які у першому наближенні моделюють шаблони збудження нейронів для типових спектрально- темпоральних складових вхідних сигналів. Зокрема, дослідження роботи первинної слухової кори головного мозку виявили множину одиночних пульсацій із синусоїдально-модульованими спектральними профілями з усталеними частотами у просторах часу і логарифмічної частоти, яка може ефективно описувати рецептивні поля та передатні

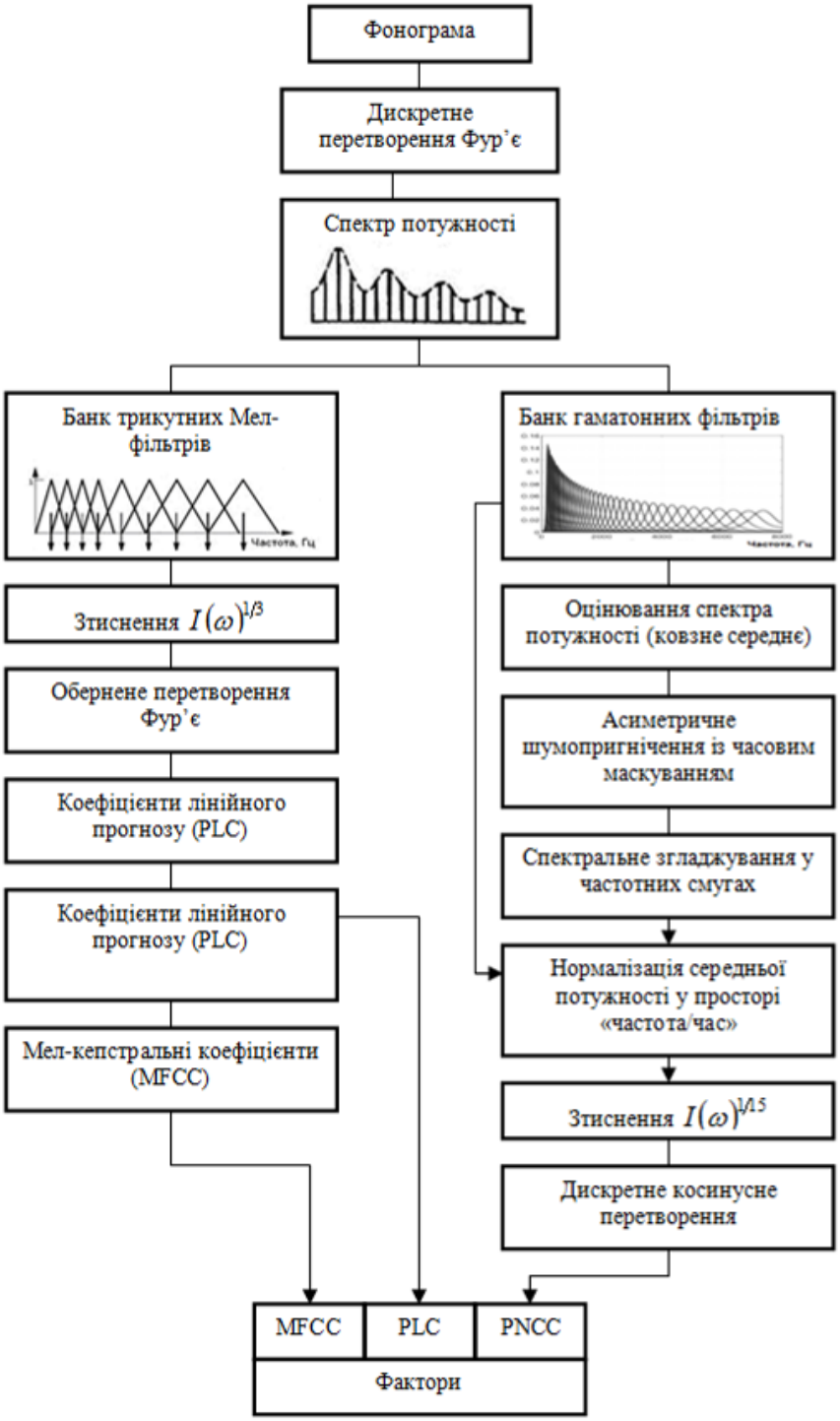

Рисунок 1 – Алгоритм отримання факторів з фонограм

функції нейронів слухової кори голосного мозку $O[n_t,k]=\Delta A\sin\left(2\pi\omega_{n_t}n_t+2\pi\omega_k k+\Phi\right)$. STRF-обробка описується процедурою фільтрації виду $O[n_t,k]=STRF[n_t,k]*x[n_t,k]$, тобто у довільний момент часу n_t і на частоті k реакція нейронів $O[n_t,k]$ описується згортанням STRF та динамічного спектру подразника в околі моменту часу і частоти $x[n_t,k]$. Відповідно, STRF діє як фільтр, який видає піки як реакцію на вхідні сигнали, що за своїми характеристиками наближаються до спектрально-темпоральних ознак, що описують запам'ятовані образи.

Для практичної обробки мовних сигналів зазвичай використовується апроксимація STRF 2D функцією Габора. Для синтезу фільтру Габора, який моделюватимуть STRF, помножимо комплексну синусоїду на віконну функцію Хеннінга:

$$STRF[n_t,k]\cong Gb[n_t,k]=s[n_t,k]\cdot h[n_t,k], \qquad (8)$$

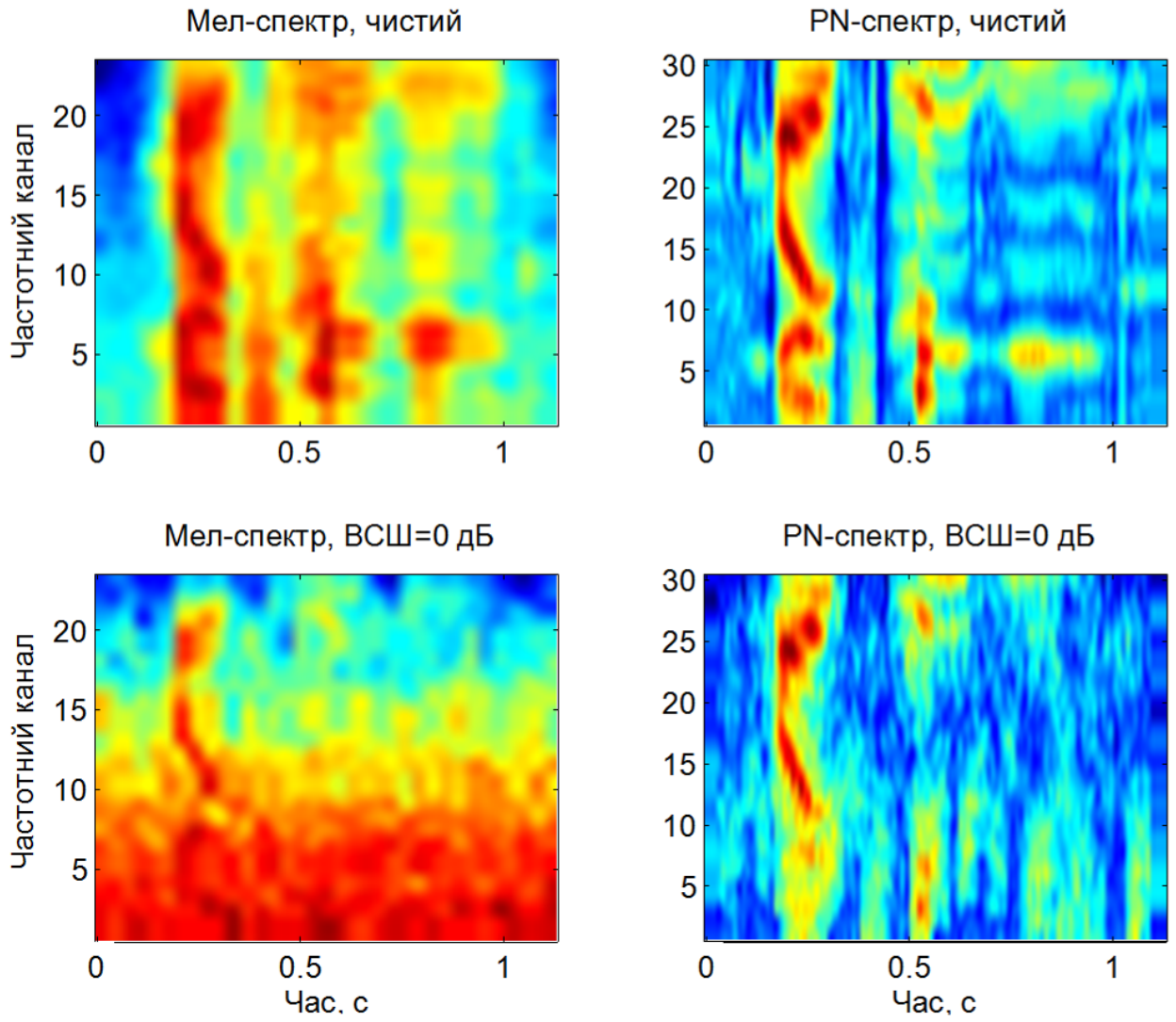

Рисунок 2 – Приклади візуалізація Мел- та PN-спектрів без/з шумом

де комплексну синусоїду, з темпоральною частотою модуляції ω_n і спектральною частотою модуляції ω_k, представлено як $s[n_t,k]=e^{i\omega_n n+i\omega_k k}$, а віконну функцію Хеннінга – як $h[n_t,k]=\left(\frac{1}{2}\left(1-\cos\frac{2\pi n_t}{W_{n_t}+1}\right)\right)\left(\frac{1}{2}\left(1-\cos\frac{2\pi k}{W_k+1}\right)\right)$, де параметри W_n і W_k визначаються як 1.75 цикли відповідної частоти модуляції: $W_{n_t}=1{,}75\frac{2\pi}{|\omega_{n_t}|}$, $W_k=1{,}75\frac{2\pi}{|\omega_k|}$. Для чисто темпоральних або спектральних фільтрів (8) дає нескінченну допоміжну функцію, отже, обмежимося 40 частотними каналами або 99 часовими фреймами, що відповідає максимальній довжині інших фільтрів у відповідних вимірах.

За рахунок налаштовуваних параметрів спектральної та темпоральної частоти модуляції функції Габора матимуть різну інтенсивність і нахил в залежності від кількості коливань у вхідному сигналі. Банк фільтрів Габора, використаний авторами, містить 59 фільтрів, налаштованих на різні темпоральні та спектральні частоти модуляції (темпоральні частоти модуляції ω_n, Гц: 0; 1,9; 3,9; 6,2; 9,9; 15,7; 25; спектральні частоти модуляції ω_k, циклів/октаву: –0,25; –0,1224; –0,06; –0,0293; 0; 0,0293; 0,06; 0,1224; 0,25), які візуалізовано на рис. 3.

При фільтрації PN-спектру створеним банком фільтрів Габора (див. приклад на рис. 4) фільтри із високою частотою модуляції (вузькі фільтри) фільтрують швидкозмінювану частину спектра, а фільтри із низькою модуляційною частотою (широкі фільтри) фільтрують низькочастотну область мовного сигналу. Комплекс вузьких та широких фільтрів, що відповідають різним спектральним частотам модуляції, генерує ознаки, які описують спектральну динаміку вхідного сигналу.

Описаний вище математичний апарат виділення з фонограм факторів та їх представлення після фільтрації Габора є досить ресурсоємним, отже, важливою задачею є здійснення факторного аналізу отриманих з фільтрів ознак для підвищення їх інформативності і зменшення кількості обчислень. Зазвичай у системах розпізнавання мови/мовця застосовують метод главних ком-

понент PCA для отримання лінійної комбінації базових факторів із максимальною варіативністю. В основному, головні вектори є щільними (тобто містять ненульові значення), що ускладнює інтерпретацію результатів PCA. Отже, пропонується використати для оптимізації простору інформативних ознак розріджений метод главних компонент SPCA [16], позбавлений цього недоліку. Цільова функція SPCA отримується із класичної PCA-функції введенням l_1-штрафної норми і набуває виду:

$$\max v^T \Sigma_x v - \rho\left(\|v\|_1^2\right),\ \|v\|_2 = 1. \qquad (8)$$

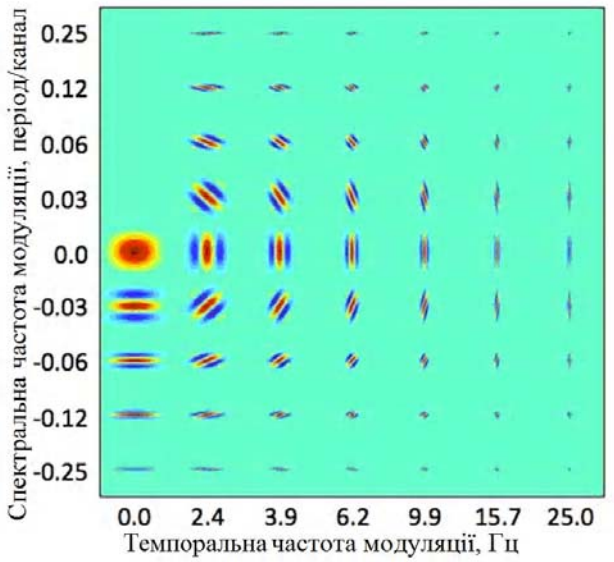

Рисунок 3 – Частотна роздільна здатність створеного банку фільтрів Габора

Перший доданок (8) є цільовою функцією класичного PCA. Вираз (8) можна перетворити на

$$\max Tr\left(\Sigma_x, vv^T\right) - \rho\left(1^T \left|vv^T\right| 1\right),\ \|v\|_2 = 1, \qquad (9)$$

Для підвищення надійності аналізу представимо l_1-штрафну норму у вигляді

$$\rho\left(1^T \left|vv^T\right| 1\right) = \min_U Tr\left(Uvv^T\right),\ -\rho \le U_{ij} \le \rho \qquad (10)$$

і узагальнимо вирази (8)–(10) у вигляді виразу (11)

$$\max \min Tr\left(\left(\Sigma_x + U\right)vv^T\right),\ \|v\|_2 = 1,\ -\rho \le U_{ij} \le \rho, \qquad (11)$$

який, із врахуванням кінцевого результату, представимо як

$$\max \min v^T\left(\Sigma_x + U\right)v,\ \|v\|_2 = 1,\ -\rho \le U_{ij} \le \rho \qquad (12)$$

або

$$\min_U \lambda_{\max}\left(\Sigma_x + U\right),\ -\rho \le U_{ij} \le \rho, \qquad (13)$$

де $\lambda_{\max}\left(\Sigma_x + U\right)$ – максимальне власне значення $\left(\Sigma_x + U\right)$.

Із виразу (13) видно, що для реалізації SPCA необхідно шукати мінімальне з найбільших власних значень коваріаційної і шумової матриць. Для прискорення пошуку U застосуємо алгоритм розширеного метода Лагранжа ALM [16] і обмежимося обчисленням лише головного розрідженого власного вектора, нульовим значенням у якому відповідатимуть ненадійним ознакам.

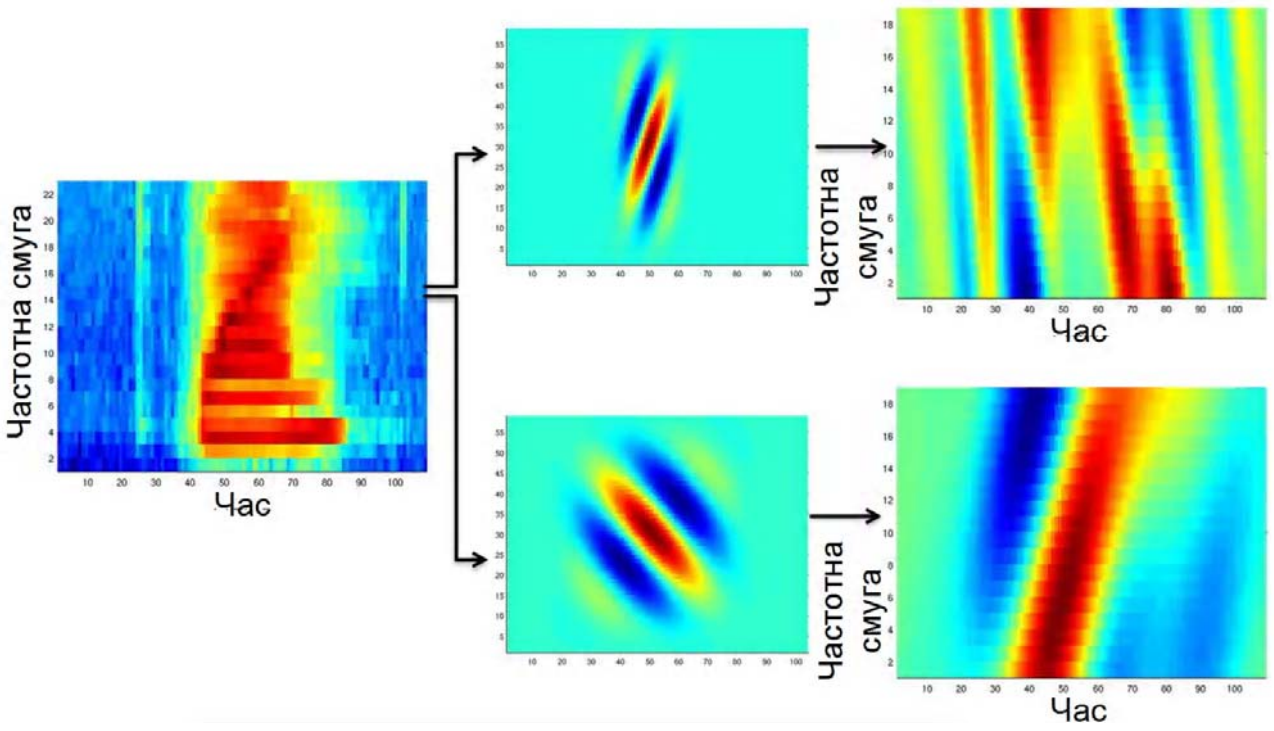

Рисунок 4 – Вище: фільтр із високою темпоральною модуляцією ($\omega_n = 6{,}2$ Гц, $\omega_k = 0{,}03$ цикли/канал) та результат фільтрації; нижче: фільтр із низькою темпоральною модуляцією ($\omega_n = 2{,}4$ Гц, $\omega_k = 0{,}03$ цикли/канал) та результат фільтрації

Отже, для подальшого використання обмежимося лише ознаками, яким відповідають ненульові значення у розрідженому головному векторі.

У створеній АСРМКЗ MFC- та PN-спектрограми, отримані при аналізі вхідних фонограм, після проходження банку фільтрів Габора та оптимізації за допомогою SPCA використовуються в якості вхідних даних для глибокого нейромережевого класифікатора (див. рис. 5), у топологію якого введено вузьке горло, параметри якого після навчання використовуються у HMM/GMM-системі розпізнавання мовця, яку буде описаній згодом.

Маючи візуальні зображення MFC- та PN-спектрограм очевидним є можливість застосування для розпізнавання мовця згортальної нейромережі. Першим (вхідним) шаром згортальної нейромережі є згортальний шар, який виконує процедуру фільтрації вхідного зображення використовуючи фільтр (ядро згортання). Аналіз роботи згортального шару дозволяє передбачити можливість інтеграції 2D фільтрів Габора у топологію згортальної нейромережі, замінивши ядра згортання фільтрами Габора із відповідними часовими і частотними характеристиками. Модифікована таким чином згортальна нейромережа, зображена на рис. 6, використовує коефіцієнти фільтрів Габора як параметри ядра згортання і виконує їх оптимізацію під час навчання, що робить таку нейромережу потенційно більш ефективною за звичайні CNN або DNN нейромережі, де передфільтрація Габора викнується із незмінними параметрами, адже винесена за межі класифікатора.

Автори передбачають, що описана модифікація згортального шару забезпечується застосуванням лінійних функцій активації нейронів шару замість сигмоїдних та встановленням параметру перекривання ядер рівним нулю, що робить згортальний шар передатною функцією фільтрів рецептивних полів. Розміри рецептивних полів для отримання однакових із фільтрами Габора частотно-часових параметрів встановлюємо рівними $2K_m+1$ (смуг) на $2N_m+1$ (відліків) для відповідного фільтру m. Коефіцієнти фільтра Габора інтегруються у вхідні ваги нейромережі W_m згідно відношення

$$h_m[n_t,k]=lin\left(\sum_{i=-N_m}^{N_m}\sum_{j=-K_m}^{K_m}Gb_m(-i,-j)\cdot x(n_t+i,k+j)\right)=$$

$$=Gb_m[n_t,k]*x[n_t,k]. \qquad (14)$$

Ваги зв'язків W_m виконують фільтрацію у рецептивному полі, де індекси коефіцієнтів фільтрів обернені індексам ваг як по вертикалі так і по горизонталі відповідно рівняння (14).

Визначним критерієм для створюваної системи є надійність розпізнавання мовців, яку можна суттєво підвищити використавши поряд із нейромережевою класифікацією інший ефективний метод прийняття рішень для співставлення отриманих результатів, як це показано на рис. 5.

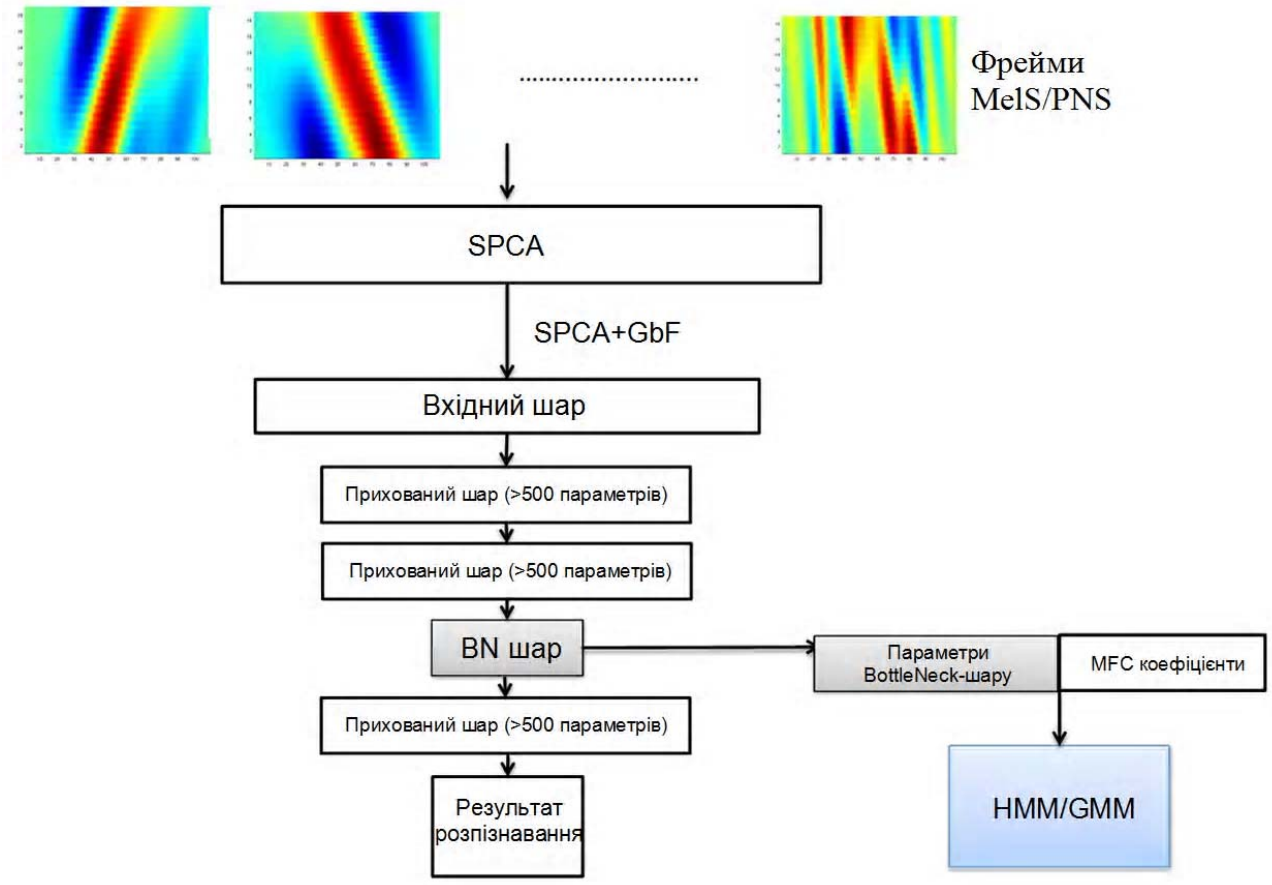

Рисунок 5 – Архітектура АСРМКЗ із глобоким нейромережевим класифікатором і можливістю HMM/GMM аналізу

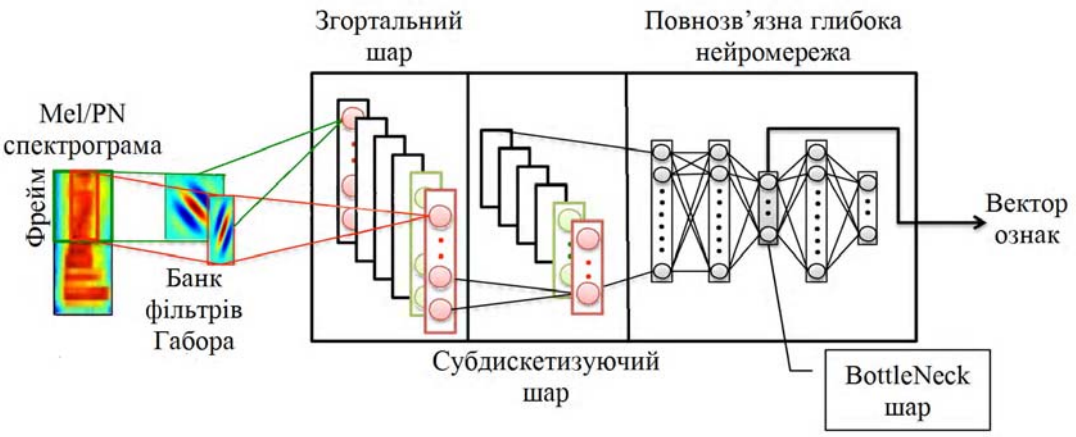

Рисунок 6 – Архітектура модифікованого згортального нейромережевого класифікатора GbCNN АСРМКЗ

У сучасних системах розпізнавання мовця використовують метод апроксимації щільності імовірності в просторі ознак GMM і метод HMM [17]. Гаусова суміш є зваженою сумою M щільності імовірності, яка описується відношенням $p(x|\lambda)=\sum_{i=1}^{M} p_i b_i(x)$, де x – D-вимірний вектор випадкових величин, $b_i(x)$ – функції щільності розподілу складових моделі, p_i – ваги компонентів суміші. Остаточно модель гаусової суміші описується у вигляді вектора $\lambda_i=\{\mu_i,\Sigma_i,p_i\}$, $i=1,\ldots,M$. У нашому випадку, кожному i-му мовцеві відповідає унікальна модель λ_i. Практично процес знаходження оптимальних параметрів λ_i вирішується алгоритмом оцінювання-максимізації [17], робота якого аналогічна алгоритму Баумана-Велча для оцінюванню параметрів прихованих марковських моделей [17]. Так якщо відомий початковий вектор ознак λ_i і обчислений за певними правилами вектор $\overline{\lambda}_i$, то за умови, що $p(x|\overline{\lambda}_i)\geq p(x|\lambda_i)$, вектор $\overline{\lambda}_i$ вважається базовим для повторної ітерації обчислень, які повторюються поки зберігається позитивна динаміка покращення параметрів моделі гаусової суміші i-го мовця. Розширюючи ці міркування на процес розпізнавання мовця, задачу GMM-HMM класифікації можна сформулювати так. Нехай множину мовців $S=\{s_1,s_2,\ldots,s_n\}$ описано множиною гаусових сумішей $\Lambda=\{\lambda_1,\lambda_2,\ldots,\lambda_n\}$, тоді задача класифікації полягає у знаходженні моделі мовця із найбільшим значенням апостеріорної імовірності для визначеної парольної фрази: $S=\arg\max_{1\leq k\leq n}\Pr(\lambda_k|X)=\arg\max_{1\leq k\leq n}\frac{p(x|\lambda_k)\Pr(\lambda_k)}{\Pr(x)}$. Цей вираз за умови рівноімовірної появи мовців $\left(\Pr(\lambda_k)=1/n\right)$, однакового значення $\Pr(x)$ для всіх моделей та за умови незалежності процедур розпізнавання можна переписати так: $S=\arg\max_{1\leq k\leq n}\sum_{i=1}^{M}\log p(x_i|\lambda_k)$, де $p(x_i|\lambda_k)$ – є гаусовою сумішшю, із описаними вище характеристиками.

У сучасних системах розпізнавання мовця використовують метод апроксимації щільності імовірності в просторі ознак GMM і метод HMM [17]. Гаусова суміш є зваженою сумою M щільності імовірності, яка описується відношенням $p(x|\lambda)=\sum_{i=1}^{M} p_i b_i(x)$, де x – D-вимірний вектор випадкових величин, $b_i(x)$ – функції щільності розподілу складових моделі, p_i – ваги компонентів суміші. Остаточно модель гаусової суміші описується у вигляді вектора $\lambda_i=\{\mu_i,\Sigma_i,p_i\}$, $i=1,\ldots,M$. У нашому випадку, кожному i-му мовцеві відповідає унікальна модель λ_i. Практично процес знаходження оптимальних параметрів λ_i вирішується алгоритмом оцінювання-максимізації [17], робота якого аналогічна алгоритму Баумана-Велча для оцінюванню параметрів прихованих марковських моделей [17]. Так якщо відомий початковий вектор ознак λ_i і обчислений за певними правилами вектор $\overline{\lambda}_i$, то за умови, що $p(x|\overline{\lambda}_i)\geq p(x|\lambda_i)$, вектор $\overline{\lambda}_i$ вважається базовим для повторної ітерації обчислень, які повторюються поки зберігається позитивна динаміка покращення параметрів моделі гаусової суміші i-го мовця. Розширюючи ці міркування на процес розпізнавання мовця, задачу GMM-HMM класифікації можна сформулювати так. Нехай множину мовців $S=\{s_1,s_2,\ldots,s_n\}$ описано множиною гаусових сумішей $\Lambda=\{\lambda_1,\lambda_2,\ldots,\lambda_n\}$, тоді задача класифікації полягає у знаходженні моделі мовця із найбільшим значенням апостеріорної імовірності для визначеної парольної фрази: $S=\arg\max_{1\leq k\leq n}\Pr(\lambda_k|X)=\arg\max_{1\leq k\leq n}\frac{p(x|\lambda_k)\Pr(\lambda_k)}{\Pr(x)}$. Цей ви-

раз за умови рівноімовірної появи мовців $\left(\Pr(\lambda_k) = 1/n\right)$, однакового значення $\Pr(x)$ для всіх моделей та за умови незалежності процедур розпізнавання можна переписати так: $S = \arg\max_{1\le k\le n} \sum_{i=1}^{M} \log p(x_i|\lambda_k)$, де $p(x_i|\lambda_k)$ – є гаусовою сумішшю, із описаними вище характеристиками.

4 ЕКСПЕРИМЕНТИ

У наведеному матеріалі автори обгрунтували ряд удосконалень АСРМКЗ, як то, використання факторів на основі PN-представлення мовних сигналів для опису індивідуальності мовлення; використання банку фільтрів Габора у складі АСРМКЗ; застосування SPCA для аналізу факторів опису індивідуальності мовлення; можливість інтеграції фільтрів Габора у топологію згортальної нейромережі; застосування GMM/HMM класифікації на основі ознак bottleneck-шару навченого глибокого нейромережевого класифікатора у складі АСРМКЗ. Далі приведемо поставку емпіричних досліджень адекватності запропонованих удосконалень та оцінювання ефекту від їх впровадження.

В якості бази фонограм для навчання та тестування створеної із застосуванням вищеописаних удосконалень АСРМКЗ використано базу записів із безкоштовної бази даних NOIZEUS [2] – 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 АСРМКЗ навчали як фонограмами без додавання шумів, так і фонограмами із додаванням шуму. Навчальна вибірка містила 594 фонограми, де до чистого сигналу додавався штучний шум з рівнями шум/сигнал 0 дБ, 5 дБ, 10 дБ, 15 дБ відповідно. Навчання створеної системи проводилося на фонограмах всіх чотирьох типів відповідно до рівня ВСШ, за умови, що серед навчальної вибірки для кожного із мовців була хоча б одна фонограма із ВСШ = 0 дБ. Фонограми навчальної вибірки використовувалися як вхідні дані для синтезу залежних від статі мовця UBM моделей, повних матриць варіативності, моделей. Для детектування інтервалів мовної активності у фонограмах застосовувався двохканальний VAD алгоритм [18]. Інтервали мовної активності тривалістю 3 секунди розбивалися на фрейми тривалістю 30 мс із 15 мс зсувом, із даних яких екстрагувалися 19 MFCC та PNCC коефіцієнтів, їх енергія, перша і друга їх похідні. До кожної чистої фонограми (із рівнем ВСШ=0) навчальної вибірки підмішувався запис акустичних шумів, вид та рівень ВСШ яких обирався випадково із мовної бази. В результаті на одну чисту фонограму припадало десять із рівнем ВСШ 0, 5, 10 або 15 дБ.

Системи *i*-векторів створеної АСРМКЗ базуються на залежних від статі мовця UBM моделях із 1024 сумішами, навчених на мовному матеріалі бази NOIZEUS, і матрицях повної варіативності із 500 факторами, до яких застосовувалися операції внутрікласової коваріаційної нормалізації WCCN [18] і нормалізації довжини i-векторів [19].

Для реалізації повнозв'язної DNN із чотирма прихованими шарами, серед яких третій – bottleneck-шар, містив 25 нейронів, був використаний фреймворк Caffe [19]. Загальна кількість керованих параметрів нейромережі становила близько $3{,}5\cdot10^6$. Для прискорення навчання нейромережі початкові її параметри отримано за допомогою переднавчання із використанням обмежених машин Больцмана [19], а подальше навчання відбувалося за алгоритмом зворотного розповсюдження помилки із коефіцієнтом швидкості навчання рівним 0,008 і використанням значення перехресної ентропії в якості функції втрат. Цей же класифікатор в подальшому інтегрувався у повнозв'язний шар згортальних нейромереж.

При синтезі CNN на згортальному шарі утворювалося 120 фільтрів, кожен з яких накривав 9 частотних смуг і 15 послідовних відліків вхідних зображень. При створеній згортальної нейромережі із інтегрованими фільтрами Габора GCNN розміри ядер фільтрів у часовому просторі знаходитися у межах від 7 до 99 відліків, а для частотного простору – від 7 до 40 смуг.

5 РЕЗУЛЬТАТИ

У таблиці 1 наведено набори інформативних ознак, які далі подавалися на класифікатори для прийняття рішень. Набори 1 і 2 утворені значеннями відповідного виду кепстральних коефіцієнтів без додаткової обробки, у наборах 3 і 4 MF- та PN-кепстральні коефіцієнти фільтрувалися банком фільтрів Габора, які накривали частотний діапазон 0-8000 Гц та піддавалися логарифмічній компресії LC. Набори 5 і 6 утворювалися аналогічно наборам 3 і 4, але замість LC-компресії використовувалася нелінійнай компресія енергії PNC, а у наборах 7 і 8 – розріджений метод главних компонент SPCA. Всі набори представляли собою як набори векторів так і спектрограм для подальшого передавання на вхідні

Таблиця 1 – Базові набори інформативних ознак для розпізнавання мовців

№ п/п	Синтезована інформативна ознака	Вид банку фільтрів	Віднімання енергії пауз	Вид компресії	Фільтрація Габора
1	MFCC	Мел	–	–	–
2	PNCC	Гаматон	+	–	–
3	MFC-Gb	Мел	–	LC	+
4	PNC-Gb	Гаматон	+	LC	+
5	MFC-Gb	Мел	–	PNC	+
6	PNC-Gb	Гаматон	+	ONC	+
7	MFC-Gb	Мел	–	SPCA	+
8	PNC-Gb	Гаматон	+	SPCA	+

шари глибокої нейромережі або згортальної нейромережі відповідно. Також за результатами SPCA досліджено інформативність окремих фільтрів Габора на основі оцінювання варіативності отриманих від них інформативних ознак, результати якого представлено на рисунку 7, де темніші області визначають ступінь важливості фільтрів (з позиції зростання варіативності отриманих ознак). Ознаки, отримані з виходів фільтрів з темпоральною частотою модуляції 0, 2,4 і 3,9 Гц (що відповідає приблизно 1, 0,7 та 0,5 с довжини фільтрів), складають понад 90% від загальної розмірності ітогового вектора ознак.

Для навчання HMM/GMM моделей застосовувалися ваги bottleneck-шару відповідних навчених нейромереж. На рис. 8 наведено середні ІПР в залежності від виду класифікатора і рівня ВСШ.

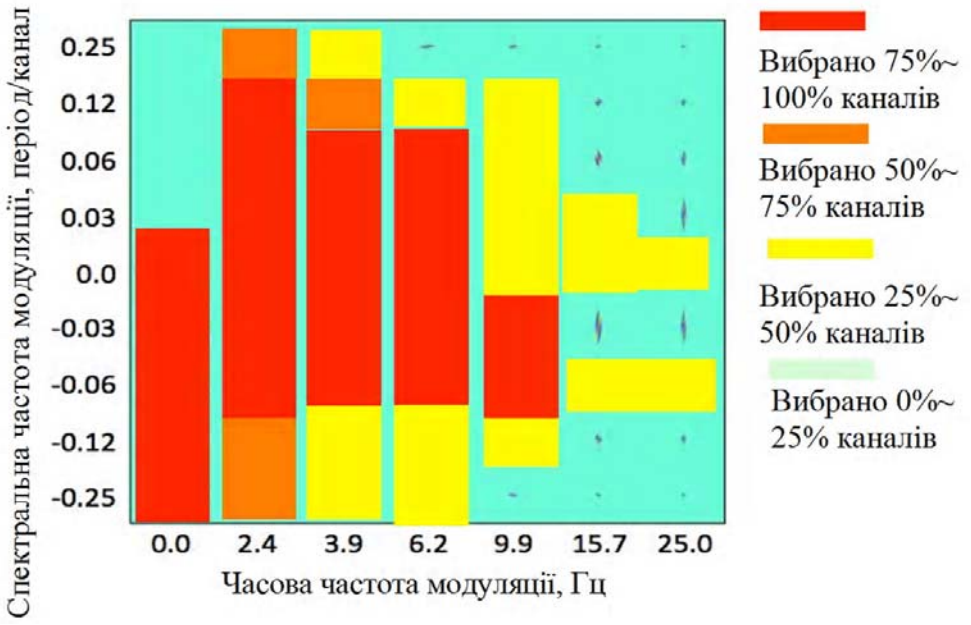

Рисунок 7 – Ефективність фільтрів Габора, оцінювана за результатами SPCA-компресії

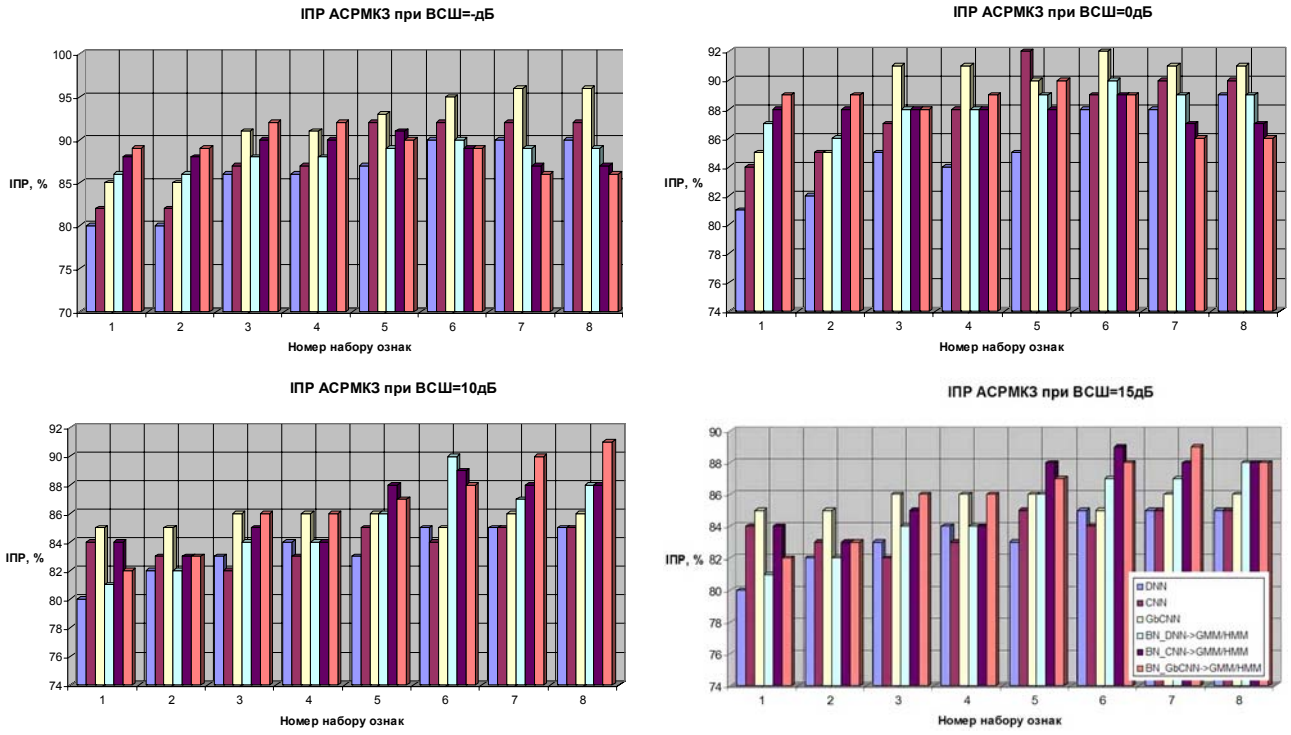

Рисунок 8 – ІПР АСРМКЗ в залежності від виду класифікатора, набору ознак і ВСШ фонограм

6 ОБГОВОРЕННЯ

Результати експериментів, наведені на рис. 8, показали, що запропонована авторами концепція інтеграції фільтрів Габора у згортальний шар CNN дозволяє підвищити як кількісні показники ефективності роботи АСРМКЗ так і підвищити її стійкість до зростання ВСШ, що досягається притаманній CNN адаптивності до мінливості вхідних даних, яка, проте, недостатньо компенсується при високих значеннях ВСШ, що цілком виправдовує застосування контрольного класифікатора, не зважаючи на зростання ресурсоємності ітогової системи.

Як видно з результатів, наведених на рис. 8, із зростанням рівня ВСШ зростає і ефективність HMM/HMM класифікації, що можна пояснити ефективністю роботою UBM-моделі. Також можна відзначити, що набори ознак 7 і 8 демонструють найвищу інформативність, яка зберігається із зростанням ВСШ, що доводить доцільність і адекватність запропонованої процедури факторного аналізу і застосування фільтрації Габора для представлення інформативних ознак.

Використання PNC-ознак також виявляється доцільним при зростанні ВСШ, що обумовлено закладеному у метод розрахунків нормованих за потужністю кепстральних коефіцієнтів заходів для компенсації шумів. Загалом подавання зображень PN-спектрограм на вхід CNN виявилося ефективнішим з MFC- спектрограми із зростанням ВСШ.

Використання розрідженого методу главних компонент дозволило зменшити час на навчання основного і контрольного класифікаторів на 15–17% в залежності від співвідношення «сигнал»/«пауза» у вхідних фонограмах на користь першого класу.

Також наведені на рис. 7 результати доводять більшу інформативність низькочастотної області сигналу, що дозволяє припустити доцільність введення додаткових факторів, які описують індивідуальні особливості мовного процесу у низькочастотній області мовного сигналу, наприклад, характеристик періоду/частоти основного тону.

ВИСНОВКИ

У роботі запропоновано заходи до підвищення стійкості АСРМКЗ до впливу шумів у вхідних фонограмах.

До наукової новизни отриманих результатів можна віднести те, що вперше запропоновано орієнтований на розпізнавання мовця метод інтеграції фільтрів Габора, які імітують роботу слухової системи людини на основі нормалізованих за потужністю кепстральних коефіцієнтів, у вхідний шар згортальної нейронної мережі, що дозволило автоматизовано варіювати чутливістю представлення інформативних для розпізнавання мовця ознак змінюючи параметри фільтрів Габора на етапі навчання нейромережі. Удосконалено нейромережевий класифікатор АСРМКЗ шляхом введення bottleneck-шару у повнозв'язний сегмент згортальної нейронної мережі, що дозволило використовувати його параметри після навчання в якості вхідних даних для контрольного GMM/HMM-класифікатора. Удосконалено спосіб представлення інформативних для розпізнавання мовця ознак у факторному просторі за рахунок застосування розрідженого методу аналізу головних компонент, що дозволяє зменшити довжину вектора інформативних ознак у 2-3 рази із збереженням його інформативності.

Практична цінність отриманих результатів полягає у створенні програмного забезпечення АСРМКЗ, яке реалізує запропоновані наукові результати і дозволяє перевірити їх адекватність у тому числі за даними спеціалізованої бази даних Школи інжинірингу та комп'ютерних наук Еріка Джонсона при Університеті Техасу NOIZEUS.

Проведені дослідження виявили, що при зростанні ВСШ>10 дБ контрольний GMM/HMM-класифікатор виявився ефективнішим за нейромережеві щодо імовірності прального розпізнавання мовців, що можна пояснити ефективністю використаних UBM-моделей, але він є і суттєво ресурсоємнішим. Отже, можливим напрямом подальших досліджень може стати підвищення обчислювальної ефективності GMM/HMM-класифікатора АСРМКЗ.

ПОДЯКИ

Роботу виконано в рамках кафедральній науково-дослідній роботи № 46К4 «Методи моделювання та оптимізації складних систем на основі інтелектуальних технологій» на кафедрі комп'ютерних систем управління Вінницького національного технічного університету за підтримки колективу кафедри і спорідненої кафедри автоматики та інформаційно-вимірювальної техніки ВНТУ.

СПИСОК ЛІТЕРАТУРИ

1. Kalinli O. Noise adaptive training using a vector Taylor series approach for noise robust automatic speech recognition / O. Kalinli, M. L. Seltzer, A. Acero // [Electronic resource]. – Access mode: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/Ozlem_ICASSP09_final.pdf
2. Ковтун В. В. Оцінювання надійності автоматизованих систем розпізнавання мовців критичного застосування / В. В. Ковтун, М. М. Биков, // Вісник Вінницького політехнічного інституту, Вінниця. – 2017. – № 2. – С. 70–76.
3. Kim C. Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring / C. Kim, R. M. Stern // [Electronic resource]. – Access mode: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.184.9018&rep=rep1&type=pdf
4. Normalized amplitude modulation features for large vocabulary noise-robust speech recognition / [V. Mitra, H. Franco, M. Graciarena, A. Mandal] // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 25–30 March 2012 : proceedings. – Kyoto, Japan: IEEE, 2012. - P. 4117-4120. DOI: 10.1109/ICASSP.2012.6288824.
5. Speech Processing, Transmission and Quality Aspects (STQ). [Electronic resource]. – Access mode: http://www.etsi.org/deliver/etsi_es/201100_201199/201108/01.01.03_60/es_201108v010103p.pdf
6. Graves A. Speech recognition with deep recurrent neural networks / A. Graves, A. R. Mohamed, G. Hinton // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 26–31 May 2013 : proceedings. – Vancouver, BC, Canada : IEEE, 2013. – P. 6645–6649. DOI: 10.1109/ICASSP.2013.6638947
7. Mohamed A. Acoustic modeling using deep belief networks / A. Mohamed, G. Dahl, G. Hinton // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 31 January 2011 : proceedings. – IEEE, 2011. – P. 14–22. DOI: 10.1109/TASL.2011.2109382

8. Davis S. Comparison of parametric representation of monosyllabic word recognition in continuously spoken sentences / S. Davis, P. Mermelstein // [Electronic resource]. – Access mode: http://www.cs.northwestern.edu/~pardo/courses/eecs352/papers/Davis1980-MFCC.pdf
9. Hermansky H. Perceptual Properties of Current Speech Recognition Technology / H. Hermansky, J. Cohen, R. Stern // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 23 July 2013 : proceedings. – IEEE, 2013. – P. 1968–1985. DOI: 10.1109/JPROC.2013.2252316.
10. Virtanen T. Techniques for Noise Robustness in Automatic Speech Recognition / T. Virtanen, R. Singh, B. Raj // John Wiley & Sons, Ltd, Chichester, UK. – 2012. DOI: 10.1002/9781118392683.ch1.
11. Stern R. Hearing is Believing. Biologically inspired methods for robust automatic speech recognition // R. Stern, N. Morgan // [Electronic resource]. – Access mode: https://pdfs.semanticscholar.org/d4a9/a6aa42dcb2011e45a99b0174da6a47777b7a.pdf
12. Kim C. Power-normalized cepstralcoefficients (PNCC) for robust speech recognitions / C. Kim, R. Stern // [Electronic resource]. – Access mode: http://www.cs.cmu.edu/~robust/Papers/OnlinePNCC_V25.pdf
13. Movellan J. Tutorial on Gabor Filters. [Electronic resource] / J. Movellan. – Access mode: http://mplab.ucsd.edu/tutorials/gabor.pdf
14. Mesgarani N. Speech Processing with a Cortical Representation of Audio / N. Mesgarani, S. Shamma // [Electronic resource]. – Access mode: https://pdfs.semanticscholar.org/f1d8/f93cdb64390b3a65f930cee4346c30bd86e4.pdf
15. Morgan N. Using spectro-temporal features to improve AFE feature extraction for automatic speech recognition / N. Morgan, S. Ravuri // [Electronic resource]. – Access mode: https://pdfs.semanticscholar.org/c7c5/04087f2107f0ea9a3cedeeaf5cc0c48c0c92.pdf
16. Berthet Q. Optimal Detection of Sparse Principal Components in High Dimension / Q. Berthet, P. Rigollet / [Electronic resource]. – Access mode: https://arxiv.org/pdf/1202.5070.pdf
17. Оптимізація алфавіту інформативних ознак для автоматизованої системи розпізнавання мовців критичного застосування / [А. О. Береза, М. М. Биков, А. Д. Гафурова, В. В. Ковтун] // Вісник Хмельницького національного університету, серія: Технічні науки, Хмельницький. – 2017. – №3 (249). – С. 222–228.
18. Mak M. W. A study of voice activity detection techniques for NIST speaker recognition evaluations / M. W. Mak, H. B. Yu // [Electronic resource]. – Access mode: https://pdfs.semanticscholar.org/541f/9cfacdac000aadd57cd33b6d86dc96bc3308.pdf
19. Research of neural network classifier in speaker recognition module for automated system of critical use / [Mykola M. Bykov, Viacheslav V. Kovtun, Andrzej Smolarz et al] // SPIE 10445, Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments 2017, 1044521; DOI:10.1117/12.2280930.

Стаття надійшла в редакцію 31.12.2017.
Після доробки 22.01.2018.

Бисикало О. В.[1], Грищук Т. В.[2], Ковтун В. В.[3]

[1]Д-р техн. наук, профессор, декан факультета компьютерных систем и автоматики Винницкого национального технического университета, Винница, Украина

[2]Канд. техн. наук, доцент, доцент кафедры компьютерных систем управления Винницкого национального технического университета, Винница, Украина

[3]Канд. техн. наук, доцент, доцент кафедры компьютерных систем управления Винницкого национального технического университета, Винница, Украина

ОПТИМИЗАЦИЯ КЛАССИФИКАТОРА АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ РАСПОЗНАВАНИЯ ДИКТОРА КРИТИЧЕСКОГО ПРИМЕНЕНИЯ

Актуальность. Рассмотрены вопросы адаптации сверточного нейросетевого классификатора для использования в автоматизированные системы распознавания диктора критического применения (АСРДКП). Объектом исследования является индивидуальные особенности речевого процесса человека.

Цель работы. Разработка средств выделения из речевого сигнала индивидуальных для диктора признаков, повышение их информативности в результате выполнения факторного анализа, их визуальное представление для использования сверточного нейросетевого классификатора и оптимизация его архитектуры для нужд АСРДКП.

Метод. Предложены меры по оптимизации процедуры классификации диктора АСРДКП, для чего теоретически обоснован оптимальный способ представления информативных признаков и метод повышения их информативности, обосновано топологию и меры по повышению эффективности процесса распознавания диктора. В частности, обоснована целесообразность использования нормализованных по мощности кепстральных коэффициентов PNCC для описания фонограмм, записанных в условиях шумного окружения, предложено использовать фильтры Габора для представления информации, которая будет анализироваться сверточной нейросетью, выбран оптимальный метод факторного анализа, а именно, разреженный метод анализа главных компонент, для уменьшения размерности вектора признаков с сохранением его информативности, предложено усовершенствованную топологию сверточной нейросети для АСРДКП, в которой фильтры Габора интегрированы в сверточный слой, что позволяет оптимизировать их параметры в процессе обучения нейросети, и в полносвязном слое использована глубокую нейросеть с bottleneck-слоем, веса которого после обучения выступают в качестве входных данных для контрольного GMM / HMM-классификатора.

Результаты. Методы представления и оптимизации информативных для распознавания диктора признаков, методы их визуального представления и усовершенствование топологии сверточной нейросети для принятия решений на их основе.

Выводы. Полученные теоретические результаты нашли эмпирическое подтверждение. В частности, доказано устойчивость усовершенствованной сверточной нейросети к присутствию шумов во входных фонограммах, которая оказалась выше показателей обычной сверточной нейросети и глубокой нейросети. При росте ОСШ > 10 дБ контрольный GMM / HMM-классификатор оказался эффективнее нейросетевого, что можно объяснить эффективностью использованных UBM-моделей, но он является существенно более ресурсозатратным. Также эмпирически определены параметры окон банка фильтров Габора предоставляющих наиболее вариативную относительно индивидуальных особенностей речи информацию.

Ключевые слова: автоматизированная система распознавания диктора критического применения, обработка сигналов, нейросеть, факторный анализ.

Bisikalo O. V.[1], Grischuk T. V.[2], Kovtun V. V.[3]
[1]Dr.Sc., Professor, Dean of Faculty of Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia, Ukraine
[2]Ph.D., Associate Professor of Computer Control Systems Department, Vinnitsia National Technical University, Vinnitsa, Ukraine
[3]Ph.D., Associate Professor of Computer Control Systems Department, Vinnitsia National Technical University, Vinnitsa, Ukraine

THE AUTOMATIC SPEAKER RECOGNITION SYSTEM OF CRITICAL USE CLASSIFIER OPTIMIZATION

Context. The questions of adapting the convolution neural network classifier use in automatic speaker recognition system of critical use (ASRSCU) are considered. The research object is the individual features of the human speech process.

Objective. Development of means for separating individual features from the speaker's speech signal, increasing their informativeness as a result of the factor analysis, their visual representation for the use of the convolution neural network classifier, and optimizing its architecture for the needs of ASRSCU.

Method. Measures are proposed to optimize the speaker recognition procedure of the ASRSCU, for which the optimal way of informative features representation and the method of increasing their informativeness are theoretically justified, the topology and measures for increasing of the speaker recognition process efficiency are justified. In particular, it is justified the use of power normalized cepstral coefficients (PNCC) for the description of phonograms recorded in noisy environment conditions. We propose to use Gabor filters to represent information that will be analyzed by a convolution neural network, an optimal method of factor analysis (a sparse main components analyzing method) to reduce of the features vector length while preserving its informativeness, an improved topology of the convolution neural network in which the Gabor filters are integrated in to the convolution layer, which allows them to optimize their parameters during the neural network training process, and in a fully connected layer a deep neural network with a bottleneck layer is used, whose weights after training are uses as inputs for the GMM/HMM control classifier.

Results. Methods of representation and optimization of the speaker's individual features, methods for their visual presentation and improvement of the topology of a convolution neural network for making speaker recognition on their basis.

Conclusions. The obtained theoretical results have found empirical confirmation. In particular, the stability of an improved convolution neural network to the noisy input phonograms proved to be higher than the results of an ordinary convolution neural network and a deep neural network. With an SNR increase up to 10 dB, the GMM/HMM classifier is more efficient than the neural network, which can be explained by the efficiency of the used UBM models, but it is much more resource-intensive. Also, the parameters of the Gabor filter bank frames that provide the most variable individual features from the speech signal for speaker recognition are determined empirically.

Keywords*:* automated speaker recognition system of critical use, signal processing, neural network, feature analysis.

REFERENCES

1. Kalinli O., Seltzer M. L., Acero A. Noise adaptive training using a vector Taylor series approach for noise robust automatic speech recognition, [Electronic resource], Access mode: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/Ozlem_ICASSP09_final.pdf
2. Kovtun V. V., Bykov M. M. Otsiniuvannia nadiinosti avtomatyzovanykh system rozpiznavannia movtsiv krytychnoho zastosuvannia, *Visnyk Vinnytskoho politekhnichnoho instytutu, Vinnytsia*, 2017, No. 2, pp. 70–76.
3. Kim C., Stern R. M. Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring, [Electronic resource]. Access mode: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.184.9018&rep=rep1&type=pdf
4. Mitra V., Franco H., Graciarena M., Mandal A. Normalized amplitude modulation features for large vocabulary noise-robust speech recognition, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 25–30 March 2012 : proceedings. Kyoto, Japan, IEEE, 2012, pp. 4117–4120. DOI: 10.1109/ICASSP.2012.6288824.
5. Speech Processing, Transmission and Quality Aspects (STQ), [Electronic resource]. Access mode: http://www.etsi.org/deliver/etsi_es/201100_201199/201108/01.01.03_60/es_201108v010103p.pdf
6. Graves A., Mohamed A. R., Hinton G. Speech recognition with deep recurrent neural networks, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 26–31 May 2013, proceedings, Vancouver, BC, Canada, IEEE, 2013, pp. 6645–6649. DOI: 10.1109/ICASSP.2013.6638947
7. Mohamed A., Dahl G., Hinton G. Acoustic modeling using deep belief networks, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, *31 January 2011, proceedings*, IEEE, 2011, pp. 14–22. DOI: 10.1109/TASL.2011.2109382
8. Davis S., Mermelstein P. Comparison of parametric representation of monosyllabic word recognition in continuously spoken sentences, [Electronic resource], Access mode: http://www.cs.northwestern.edu/~pardo/courses/eecs352/papers/Davis1980-MFCC.pdf
9. Hermansky H., Cohen J., Stern R. Perceptual Properties of Current Speech Recognition Technology, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 23 July 2013 : proceedings*, *IEEE,* 2013, pp. 1968–1985. DOI: 10.1109/JPROC.2013.2252316.
10. Virtanen T., Singh R., Raj B. Techniques for Noise Robustness in Automatic Speech Recognition, *John Wiley & Sons, Ltd, Chichester, UK,* 2012. DOI: 10.1002/9781118392683.ch1.
11. Stern R., Morgan N. Hearing is Believing. Biologically inspired methods for robust automatic speech recognition, [Electronic resource]. Access mode: https://pdfs.semanticscholar.org/d4a9/a6aa42dcb2011e45a99b0174da6a47777b7a.pdf
12. Kim C., Stern R. Power-normalized cepstralcoefficients (PNCC) for robust speech recognitions, [Electronic resource]. Access mode: http://www.cs.cmu.edu/~robust/Papers/OnlinePNCC_V25.pdf
13. Movellan J. Tutorial on Gabor Filters. [Electronic resource]. Access mode: http://mplab.ucsd.edu/tutorials/gabor.pdf
14. Mesgarani N., Shamma S. Speech Processing with a Cortical Representation of Audio, [Electronic resource]. Access mode: https://pdfs.semanticscholar.org/f1d8/f93cdb64390b3a65f930cee4346c30bd86e4.pdf
15. Morgan N., Ravuri S. Using spectro-temporal features to improve AFE feature extraction for automatic speech recognition, [Electronic resource]. Access mode: https://pdfs.semanticscholar.org/c7c5/04087f2107f0ea9a3cedeeaf5cc0c48c0c92.pdf
16. Berthet Q., Rigollet P. Optimal Detection of Sparse Principal Components in High Dimension, [Electronic resource]. Access mode: https://arxiv.org/pdf/1202.5070.pdf
17. Bereza A. O., Bykov M. M., Hafurova A. D., Kovtun V. V. Optymizatsiia alfavitu informatyvnykh oznak dlia avtomatyzovanoi systemy rozpiznavannia movtsiv krytychnoho zastosuvannia, *Visnyk Khmelnytskoho natsionalnoho universytetu, seriia: Tekhnichni nauky*, Khmelnytskyi, 2017, No. 3(249), pp. 222–228.
18. Mak M. W., Yu H. B. A study of voice activity detection techniques for NIST speaker recognition evaluations, [Electronic resource]. Access mode: https://pdfs.semanticscholar.org/541f/9cfacdac000aadd57cd33b6d86dc96bc3308.pdf
19. Mykola M. Bykov, Viacheslav V. Kovtun, Andrzej Smolarz, Mukhtar Junisbekov, Aliya Targeusizova, Maksabek Satymbekov Research of neural network classifier in speaker recognition module for automated system of critical use, *SPIE 10445, Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments* 2017, 1044521; DOI: 10.1117/12.2280930.