DOI 10.31891/2307-5732-2019-269-1-165-170
УДК 681.327.12

А.Д. ГАФУРОВА, В.В. КОВТУН
Вінницький національний технічний університет

НЕЙРОМЕРЕЖЕВА АДАПТАЦІЯ PLDA ДЛЯ ВИКОРИСТАННЯ У АВТОМАТИЗОВАНІЙ СИСТЕМІ РОЗПІЗНАВАННЯ МОВЦЯ КРИТИЧНОГО ЗАСТОСУВАННЯ

Автори пропонують актуальні системи розпізнавання мовців, де застосовується і-векторне/PLDA моделювання для опису фонограм, синтезують узагальнену PLDA модель із усередненими параметрами по всій базі фонограм без їх сегрегації за рівнем шумів. В результаті такі системи забезпечують прийнятний рівень надійності лише за наявності великої навчальної вибірки як за кількістю, так і за тривалістю фонограм. У роботі автори синтезували окремі PLDA моделі для опису фонограм із детермінованими рівнями відношення сигнал/шум (ВСШ), в результаті чого фактори, які характеризують індивідуальність мовців, зосереджено у найбільш мінливих областях і-векторного простору. Автори запропоновували використовувати нейромережу для прецизійного детектування діапазонів рівнів ВСШ вхідних фонограм із подальшим використанням одержаних даних при синтезі універсальних фонових моделей, які оптимально описують вплив сторонніх акустичних шумів у фонограмах, що дозволяє як підвищити якісні показники роботи автоматизованої системи розпізнавання мовців критичного застосування так і встановлювати факт непридатності для подальшої обробки запропонованої системі в якості вхідних даних фонограми, що також підвищує надійність роботи системи загалом.

Ключові слова: *автоматизована система розпізнавання мовців критичного застосування, і-вектори, нейромережа, суміш PLDA*

A.D. HAFUROVA, V.V. KOVTUN
Vinnytsia National Technical University

NEURAL NETWORK ADAPTATION OF PLDA FOR THE AUTOMATIC SPEAKER RECOGNITION SYSTEM OF CRITICAL USE

Authors offer speaker recognition systems that use i-vector/PLDA modelling to describe phonograms synthesize a generalized PLDA model with averaged parameters throughout the phonogram database without their noise segregation. As a result, such systems provide an acceptable level of reliability only in the presence of a large training sample, both in quantity and duration of phonograms. The authors synthesized individual PLDA models for describing phonograms with deterministic levels of signal-to-noise ratio (SNR), resulting in factors that characterize the individuality of speakers, concentrated in the most volatile regions of the i-vector space. Authors proposed to use a neural network for precise detection of ranges of SNR levels of incoming phonograms with subsequent use of the data obtained in the synthesis of universal background models that optimally describe the influence of extraneous acoustic noises in phonograms, which allows not only to improve the performance of the automatic speaker recognition system of critical use, but also to establish the fact of unfitness for further processing of the input phonogram, which also increases the reliability of the system at all. The authors proposed an improved method for the adaptation of mixtures of PLDA-models to the presence of speech signals in phonograms, on which ACCRM performs voice recognition, dynamic level of VSS. The method based on the use of GNM for training UFM, namely, the GNM in the learning process changes the weight of interneuronal bonds to optimally determine the ranges of changes in the level of BCS in the i-vectors extracted from educational phonograms.

Keywords: *automatic speaker recognition system of critical use, i-vectors, neural network, PLDA mixture.*

Вступ

При створенні автоматизованої системи розпізнавання мовців критичного використання (АСРМКЗ) першочергову увагу слід приділити оцінюванню надійності функціонування системи в умовах шумного навколишнього акустичного середовища. Зазначимо, що в першому наближенні ці шуми можна розділити на детерміновані (спричинені умовами каналу передавання акустичної інформації, наприклад) та стохастичні (спричинені втручанням непередбачуваних факторів, що може призвести до виникнення критичної ситуації). Відзначимо, що компенсувати вплив шумів першого типу у сучасних системах розпізнавання мовців до певного рівня вдається як застосуванням методів цифрової обробки сигналів так і на рівні екстрагування інформативних для розпізнавання особи мовця факторів. Шуми другого типу за своєю природою вимагають створення імовірнісного математичного апарату, що, враховуючи комплексність задачі розпізнавання мовця, є нетривіальною задачею, розв'язання якої починається із вибору способу опису мовного сигналу у просторі ознак. У сучасних класичних системах розпізнавання мовця в якості комплексної інформативної ознаки (фактора) активно використовують *i*-вектори [1–3], екстраговані з фонограм із записами мовних сигналів різної тривалості. У просторі загальної мінливості *i*-вектори представляють апостеріорними середніми значеннями прихованих змінних факторного аналізатора [1]. Відзначимо, що *i*-вектор є комплексним способом опису мовного сигналу і виділення в ньому актуальної для розпізнавання мовця інформації є окремою задачею, яку зараз найчастіше вирішують методом імовірнісного лінійного дискримінантного аналізу (Probabilistic Linear Discriminative Analysis, PLDA) [4], який аналізуючи множини і-векторів формує їх індивідуальне представлення відповідно до класів мовців у факторному просторі. Останнім часом для сегрегації залежної від особи мовця інформації з і-векторів застосовують, зокрема, глибокі нейромережі (Deep neural networks, DNN) [5–7]. Наприклад, у дослідженні [7] при створенні системи розпізнавання мовця універсальну фонову модель (Universal background model,

UBM), яка є незалежною від особи мовця моделлю гаусових сумішей (Gaussian mixture model, GMM), замінили фонетично навченою ГНМ для обчислення апостеріорних імовірностей фреймів фонограм і емпірично довели більшу ефективність такої модифікації у порівнянні із стандартним УФМ/*i*-вектор-підходом. ГНМ/*i*-вектор-підхід дозволив якісно класифікувати фрейми із мовним сигналом за категоріями сенонів, що покращило показники якості розпізнавання мовців за визначеною парольною фразою.

У роботі [8, 9], на основі спостереження, що *i*-вектори, отримані з фонограм, які мали подібні рівні відношення сигнал/шум (ВСШ), схильні групуватися у *i*-векторному просторі, запропоновано метод представлення ВСШ-залежних просторів мовців у ВСШ-незалежній PLDA, а у роботах [10, 11] апостеріорні імовірності значень індикаторних змінних зв'язано із рівня ВСШ фонограм. Узагальнюючи ці підходи можна зробити висновок, що рівень ВСШ є важливим критерієм, ведення якого у створювану PLDA-модель дозволяє застосовувати її для розпізнавання мовця взагалі і робити це за фонограмами у яких мовний сигнал супроводжується шумами навколишнього середовищі зокрема.

ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

При реалізації АСРМКЗ актуальною є задача створення стійкого до присутнього у *i*-векторі шуму класифікатора. У роботі [1] автори запропонували навчати множини PLDA-моделей на основі груп факторів, екстрагованих із фонограм із визначеними рівнями ВСШ, із подальшою класифікацією мовців на основі значень апостеріорних імовірностей належності аналізованих факторів до однієї з навчених PLDA-моделей. Емпіричні дослідження виявили залежність якісних показників роботи АСРМКЗ від точності детектування рівнів ВСШ вхідних фонограм, але використовуючи класичні підходи покращити якісні характеристики цього процесу виявилося неможливим. Отже, автори пропонують дослідити можливість інтеграції штучної нейромережі глибокого навчання у АСРМКЗ для автоматизації процесу детектування рівнів ВСШ у фонограмах у процесі навчання нейромережі.

СТВОРЕННЯ PLDA-СУМІШІ ІЗ ІНТЕГРОВАНОЮ ЗАЛЕЖНОЮ ВІД РІВНЯ ВСШ ГНМ

Враховуючи позитивний досвід застосування ГНМ/*i*-вектор-апарату для GMM-навчання, ми пропонуємо удосконалений метод навчання PLDA-суміші за допомогою ГНМ для розпізнавання мовців за фонограмами із динамічним рівнем ВСШ. Передбачається, що апостеріорні імовірності рівня ВСШ, отримані з *i*-вектора, використовуються як індикаторні змінні у функціях правдоподібності суміші PLDA для управління процесом навчання моделі суміші, а апостеріорні рівні ВСШ отримуються від навченої на детектування відповідного рівня ВСШ ГНМ. І, нарешті, *i*-вектори, отримані з фонограм мовців, які розпізнаються, і мовців, на розпізнавання яких навчено систему, та апостеріорні рівні ВСШ, отримані від ГНМ, використовуються для лінійного комбінування маргінальних функцій правдоподібності різних сумішей PLDA. Отримана в результаті суміш, на відміну від залежної від рівня ВСШ суміші PLDA, не залежатиме від фактичного рівня ВСШ тестової та еталонної фонограм, а використовуватиме лише їх апостеріорні імовірності рівня ВСШ.

Процес навчання звичайної PLDA-суміші (незалежної від рівня ВСШ) [12] еквівалентно процедурі самокластеризації *i*-векторів у множину Гаусіан, по одній для кожного з мовців, якого розпізнаватиме система (PLDA-модель і класифікатор). Оскільки в процесі навчання інформація про рівень ВСШ ігнорується, *i*-вектори, отримані із «шумних» та «чистих», відповідно до присутності у записі мовного сигналу сторонніх акустичних шумів, фонограм узагальнюються разом, що розмиває межі кластерів мовців тим більше, чим більше «шумних» фонограм використовувалося при навчанні. Автори припускають, що ГНМ може виковувати функцію вчителя у процесі навчання моделей PLDA-сумішей, тобто управляти процесом кластеризації *i*-векторів у залежні від рівня ВСШ групи фонограм і зробити це прецизійніше відносно рівня ВСШ, порівняно із стандартним ЕМ-алгоритмом. Відповідно, подавши на вхід ГНМ *i*-вектори на виході мережі ми повинні отримати апостеріорні імовірності рівнів ВСШ (ядра активності нейронів вихідного шару визначають присутність відвідного рівня ВСШ у фонограмі, з якої екстраговано вхідний *i*-вектор).

Отже, з кожної фонограми начальної вибірки екстрагується *i*-вектор, множина яких згодом подається на вхід ГНМ. Значення виходів навченої ГНМ (апостеріорні імовірності наявності рівнів ВСШ у фонограмах) потім використовуються як апостеріорні імовірності значень індикаторних змінних моделі PLDA-суміші, що робить кластери мовців у *i*-векторному факторному просторі залежними від рівнів ВСШ. Для вхідного *i*-вектора x_{ij}, екстрагованого з j-ї фонограми i-го мовця, апостеріорна імовірність присутності k-го рівня ВСШ у j-й фонограмі, визначена ГНМ, описується відношенням

$$\gamma_{x_{ij}}\left(y_{ijk}\right) \equiv P\left(y_{ijk}=1 \middle| x_{ij}, \underline{w}\right), \tag{1}$$

де y_{ijk} – індикаторна змінна, яка вказує на компонент суміші, що описує спостереження x_{ij}, а $\underline{w}$ – ваги ГНМ, яка детектує рівні ВСШ. Розповсюдимо (1) на суміш K PLDA-моделей:

$$p\left(x_{ij}\right) = \sum_{k=1}^{K} \int P\left(y_{ijk}=1 \middle| x_{ij}, \underline{w}\right) p\left(x_{ij} \middle| z, y_{ijk}=1, \theta_k\right) p(z) dz =$$

$$= \sum_{k=1}^{K} \gamma_{x_{ij}}\left(y_{ijk}\right) N\left(x_{ij} \middle| m_k, V_k V_k^T + \Sigma_k\right), \tag{2}$$

де z – фактор мовця, який зв'язаний із всіма компонентами суміші, m_k, V_k і Σ_k описують математичне сподівання, підпростір мовців та коваріаційну матрицю k-ї групи за рівнем ВСШ відповідно. Параметри моделі у відношенні (2) згорнуто до виду $\theta = \{m_k, V_k, \Sigma_k\}_{k=1}^{K}$. Варіативність індивідуальних ознак мовців описується добутком $V_k V_k^T$, а варіативність фонограм коефіцієнтом Σ_k, $k = 1,...,K$.

Позначимо $Y = \{y_{ijk}\}_{k=1}^{K}$ множину прихованих індикаторних змінних, за значенням яких вибиратимемо один з K факторних аналізаторів спираючись на рівень ВСШ вхідної фонограми. Зокрема, $y_{ijk} = 1$ якщо k - та PLDA-модель вказує на x_{ij} та $y_{ijk} = 0$ у іншому випадку. Маючи набір D-розмірних нормованих за довжиною i-векторів $X = \{x_{ij}, i = \overline{1,S}, j = \overline{1,H_i}\}$, параметри θ можна отримати з вхідних даних набору оцінюванням методом максимальної правдоподібності. Спираючись на початкове значення θ, ми прагнемо віднайти нову оцінку θ', яка максимізуватиме цільову функцію

$$Q(\theta', \theta) = E_{Y,Z}\{\ln p(X, Y, Z|\theta')|X, \theta\} = E_{Y,Z}\left\{\sum_{ijk} y_{ijk} \ln\left(p(y_{ijk} = 1|\theta')p(x_{ij}|z_i, \theta')p(z_i)\right)|X, \theta\right\}. \quad (3)$$

Аналітичне викладення ЕМ-алгоритму, якій максимізуватиме (3), аналогічне отриманим авторами у відношеннях (8) і (13) у [1], за умови заміни апріорних сподівань значень індикаторних змінних y_{ijk} з урахуванням рівнів ВСШ $L = \{l_{ij}\}$ $\langle y_{ijk}|L\rangle$ на значення з виходів ГНМ $\gamma_{x_{ij}}(y_{ijk})$. Також аналітичний вигляд логарифмічних функцій визначення оцінки правдоподібності для вхідного x_s та еталонного x_t і-векторів аналогічні наведеним у (10), (11) і (15), (16) у [1] із урахувань таких замін:

$$\gamma_{x_s}(y_{k_s})\gamma_{x_t}(y_{k_t}) \to \gamma_{l_s, l_t}(y_{k_s}, y_{k_t}),\ \gamma_{x_s}(y_{k_s}) \to \gamma_{l_s}(y_{k_s}),\ \gamma_{x_t}(y_{k_t}) \to \gamma_{l_t}(y_{k_t}), \quad (4)$$

де l_s та l_t позначають рівень ВСШ вхідної та еталонної фонограм відповідно.

В результаті функція визначення оцінки правдоподібності набуде такого виду

$$S_{ГНМ-PLDA}(x_s, x_t) = \frac{\sum_{k_s=1}^{K}\sum_{k_t=1}^{K}\gamma_{x_s}(y_{k_s})\gamma_{x_t}(y_{k_t})e^{-0.5\log(\alpha\hat{\Lambda}_{k_s k_t})-0.5D\left(\left[x_s^T\ x_t^T\right]^T\left\|\left[m_{k_s}^T\ m_{k_t}^T\right]^T\right.\right)}}{\left(\sum_{k_s=1}^{K}\gamma_{x_s}(y_{k_s})e^{-0.5\log(\alpha\Lambda_{k_s})-0.5D(x_s\|m_{k_s})}\right)\left(\sum_{k_t=1}^{K}\gamma_{x_t}(y_{k_t})e^{-0.5\log(\alpha\Lambda_{k_t})-0.5D(x_t\|m_{k_t})}\right)}, \quad (5)$$

де α є скаляром для уникнення експоненціювання дуже великих від'ємних чисел (подальші результати отримано при $\alpha = 5$), $\hat{\Lambda}_{k_s k_t} = \hat{V}_{k_s}\hat{V}_{k_t}^T + \hat{\Sigma}_{k_s k_t}$, $\Lambda_{k_s} = V_{k_s}V_{k_s}^T + \Sigma_{k_s}$, $\hat{\Sigma}_{k_s k_t} = diag\{\Sigma_{k_s}, \Sigma_{k_t}\}$ та $D(x\|y)$ – відстань Махаланобіса між x і y.

ПОСТАНОВКА ЕКСПЕРИМЕНТУ ТА АНАЛІЗ РЕЗУЛЬТАТІВ

В якості бази фонограм для навчання та тестування створеної із застосуванням вищеописаного математичного апарату АСРМКЗ використано базу записів із безкоштовної бази даних NOIZEUS [13] – 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 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. 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 5 дБ, 10 дБ або 15 дБ.

Для детектування інтервалів мовної активності у фонограмах використано двоканальний детектор (Voice activity detector, VAD) [14]. До детектованих інтервалів мовної активності застосовувалася нормалізація кепстрального середнього (Cepstral mean normalization, CMN) [1] і процедура вирівнювання ознак (Feature warping, FW) [15] із фреймом тривалістю 3 с, після чого сигнал розбивався на фрейми тривалістю 25 мс із перекриванням 5 мс із застосуванням віконної функції Хеммінга і з кожного фрейму екстрагувався 60-мірний вектор інформативних ознак, який включав в себе 19 Мел-кепстральних коефіцієнтів, їх енергію, першу та другу похідні.

Для екстрагування i-векторів автори створили гендерно-залежні універсальні фонові моделі із 1024

сумішами та матрицями загальної мінливості із 500 факторами. Далі *i*-вектори розділялися на K груп (діапазони рівнів ВСШ для віднесення *i*-вектора до k-ї групи $(k = 1,...,K)$ наведено у таблиці 1) відповідно до оціненого за методикою [1] рівня ВСШ і формували навчальну вибірку для ГНМ. Кількість *i*-векторів у кожній групі буда однакова. Вхідний шар ГНМ містив 500 нейронів вхідних вузлів і три прихованих шари ГНМ містили обмежені машини Больцмана (Restricted Boltzmann machine, RBM) [16], для навчання яких застосовувався алгоритм контрастної розбіжності (Contrastive divergence algorithm, CDA) [16]. Для переднавчання ГНМ використовувався міні-пакетний метод зв'язаних градієнтів (Mini-batch conjugate gradient descent, mBCGD) із 100 прикладами у пакеті для оцінювання градієнту. Навчений на етапі переднавчання для кращого узагальнення навчальних даних softmax-вихідний шар фіналізував створену ГНМ. Подальший етап тонкого налаштування ГНМ тривав 30 епох із застосуванням методу зв'язаних градієнтів для мінімізації перехресної ентропії між бажаними і реальними значеннями виходів мережі. Після навчання ГНМ набула здатності оцінювати апріорні імовірності груп SNR, що спостерігаються у вхідному *i*-векторі.

Таблиця 1

Границі діапазонів рівнів ВСШ, (дБ) для різної кількості груп за рівнем ВСШ (*K*)

K	Група 1	Група 2	Група 3	Група 4	Група 5
2	(-∞,20]	(20, ∞)	-	-	-
3	(-∞,10]	(10,20]	(20, ∞)	-	-
4	(-∞,10]	(10,15]	(15,20]	(20, ∞)	-
5	(-∞,5]	(5,10]	(10,15]	(15,20]	(20, ∞)

Перед навчанням PLDA-моделей із подібною до реалізованої у [1] структурою відбувалося нормування довжини *i*-векторів до 500 компонент і застосувався метод внутрікласового нормування коваріації (Within-class covariance normalization, WCCN) [1] для їх очищення. Далі застосовувався лінійний дискримінантний аналіз (Linear Discriminative Analysis, LDA) для зменшення їх корельованості в межах класу одного мовця, що дозволило зменшити довжини *i*-векторів до 200 компонент із збереженням адекватності даних навчальної вибірки. Потім здійснювалося навчання різних типів PLDA-моделей із 150 прихованими характеристичними для мовця факторами, а саме, звичайних гаусових PLDA; незалежних від рівня ВСШ PLDA [1] із заміною $\gamma_{x_{ij}}(y_{ijk})$ у (5) апріорною імовірністю k-суміші; залежних від рівня ВСШ PLDA [1] із заміною $\gamma_{x_{ij}}(y_{ijk})$ у (5) одновимірною GMM, яка описує розподіл рівня ВСШ; запропонованою у (5) ГНМ-PLDA із заміною $\gamma_{x_{ij}}(y_{ijk})$ у (5) значеннями з виходів налаштованої на детектування рівня ВСШ у вхідному *i*-векторі ГНМ.

Таблиця 2

Залежність критеріїв якості роботи АСРМКЗ від наборів навчальних та тестових даних і методу моделювання факторів

Набір навчальних даних	Метод моделювання		Мовці-чоловіки				Мовці-жінки			
			Тест. набір 1		Тест. набір 2		Тест. набір 1		Тест. набір 2	
			P_+, %	minDCF	P_+, %	minDCF	P_+, %	minDCF	P_+, %	minDCF
Набір 1	PLDA		100,00	0,32	97,14	0,30	96,87	0,36	97,53	0,35
	PLDA (і-век.+ВСШ)		100,00	0,32	97,07	0,30	96,90	0,35	97,62	0,34
	ВСШ незалежна PLDA	2 суміші	100,00	0,31	97,01	0,33	96,92	0,36	97,66	0,36
		3 суміші	100,00	0,32	97,06	0,32	97,02	0,36	97,45	0,37
		4 суміші	100,00	0,30	97,09	0,31	96,84	0,37	97,64	0,35
	ВСШ залежна PLDA	2 суміші	100,00	0,32	97,03	0,32	96,90	0,37	97,41	0,37
		3 суміші	100,00	0,33	97,14	0,31	97,40	0,34	97,41	0,34
		4 суміші	100,00	0,32	97,10	0,32	97,16	0,34	97,26	0,36
	ВСШ залежна PLDA із ГНМ	2 суміші	100,00	0,27	98,60	0,28	98,12	0,29	98,05	0,28
		3 суміші	100,00	0,26	98,63	0,27	98,14	0,29	98,11	0,29
		4 суміші	100,00	0,26	98,65	0,26	98,27	0,27	98,13	0,29
Набір 2	PLDA		100,00	0,33	96,91	0,33	97,06	0,36	97,36	0,37
	PLDA (і-век.+ВСШ)		100,00	0,33	96,77	0,32	97,02	0,36	97,28	0,34
	ВСШ незалежна PLDA	2 суміші	100,00	0,33	96,91	0,35	96,94	0,37	97,33	0,36
		3 суміші	100,00	0,33	96,79	0,32	97,18	0,36	97,41	0,35
		4 суміші	100,00	0,31	96,87	0,31	97,14	0,36	97,48	0,35
	ВСШ залежна PLDA	2 суміші	100,00	0,32	97,00	0,33	97,10	0,36	97,36	0,36
		3 суміші	100,00	0,32	97,00	0,33	97,10	0,36	97,33	0,36
		4 суміші	100,00	0,32	96,92	0,32	96,91	0,37	97,14	0,38
	ВСШ залежна PLDA із ГНМ	2 суміші	100,00	0,29	97,56	0,31	98,01	0,33	98,08	0,33
		3 суміші	100,00	0,31	97,87	0,30	98,04	0,30	98,17	0,31
		4 суміші	100,00	0,31	98,03	0,29	98,31	0,29	98,26	0,31

Результати поведених експериментів з розпізнавання мовців системою на базі відповідних PLDA моделей наведено у таблиці 2. Для оцінювання якості роботи АСРМКЗ використано два критерії: імовірність правильного розпізнавання P_+ та мінімальна функція вартості виявлення (minimum detection cost

function, minDCF) [1]. Функція вартості виявлення DFC обчислюється як зважена сума імовірності відмови мовцеві, який має право доступу, P_{fa} і імовірності надання доступу мовцеві, який такого права не має, P_{miss}: $DFC = 0.1P_{miss} + 0.01P_{fa}$. Відповідно, мінімум функції DFC визначається за отриманими оцінками результатів розпізнавання.

Результати у таблиці 2 демонструють перевагу ВСШЗ-PLDA і ГНМ-PLDA-моделей над PLDA і ВСШНЗ-PLDA-моделями. PLDA-модель показує найгірші результати, що можна пояснити відсутністю у математичному представленні моделі можливості урахування рівня ВСШ у фонограмах мовних сигналів. Решта моделей включають таку адаптацію, проте прослідковується тенденція до зростання показників ефективності, демонстрованих ГНМ-PLDA-моделлю із зростанням кількості детектованих рівнів ВСШ, адже можливість навчання ГНМ робить модель чутливішою, тоді як у ВСШЗ-PLDA-моделі межі рівнів ВСШ встановлюються дослідником емпірично. ВСШЗ-PLDA і ГНМ-PLDA-моделі мають високу адаптивність до рівня ВСШ у вхідних фонограмах тому, що при їх використанні результати розпізнавання мовців отримуються шляхом узагальнення оцінок PLDA, які базуються на апостеріорних імовірностях індикаторних змінних, що залежать від вхідних фонограм. Проте, ваги суміші ВСШНЗ-PLDA-моделі обчислюються лише на основі навчальних *i*-векторів і більше не зазнають змін що зменшує адаптивність цієї моделі суміші. Результати, показані ВСШЗ-PLDA і ГНМ-PLDA близькі тому, що у першій моделі апріорні імовірності y_{ijk} обчислюються із використанням 1-D GMM, яка описує розподіл рівня ВСШ, а у другій моделі апостеріорні імовірності обчислює чутлива до рівня ВСШ ГНМ. Проте, ВСШЗ-PLDA потребує інформації про рівень ВСШ тестових фонограм, а ГНМ-PLDA – ні, що робить останню універсальнішою.

ВИСНОВКИ

Автори запропонували удосконалений метод адаптації сумішей PLDA-моделей до присутності у фонограмах мовних сигналів, за якими АСРМКЗ здійснює розпізнавання мовців, динамічного рівня ВСШ. Метод заснований на використанні ГНМ для навчання УФМ, а саме, ГНМ у процесі навчання змінює ваги міжнейронних зв'язків для оптимального визначення діапазонів зміни рівня ВСШ у *i*-векторах, екстрагованих із навчальних фонограм. Результати емпіричних досліджень довели ефективність запропонованого удосконалення порівняно із класичними PLDA-моделями і запропонованими авторами у [1] ВСШЗ-PLDA і ВСШНЗ-PLDA. Окрім прецизійної точності детектування діапазонів рівнів ВСШ вхідних фонограм запропонований нейромережевий компонент PLDA-моделі може довільно масштабуватися за потребою дослідника і адаптуватися під умови експлуатації АСРМКЗ. До недоліків запропонованих удосконалень можна віднести обчислювальну складність процесу створення і навчання ГШМ, проте, математичний вираз (5) можна застосувати і для класичного персептрону, використання якого є суттєво ресурсоефективнішою.

Література

1. Ковтун В.В. Підвищення шумостійкості автоматизованої системи розпізнавання мовця критичного застосування / Т.В. Грищук, В.В. Ковтун // Вісник Вінницького політехнічного інституту. – 2018. – № 1. – С. 98–111.

2. Reynolds D. A. Speaker verification using adapted Gaussian mixture models / D. A. Reynolds, T. F. Quatieri, R. B. Dunn // Digital Signal Processing. – 2000. – Vol. 10. – № 1. – P. 19–41.

3. Shao Y. Robust speaker identification using auditory features and computational auditory scene analysis / Y. Shao, D. Wang // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2008. – P. 1589–1592.

4. Elder J.H. Probabilistic linear discriminant analysis for inferences about identity / S.J.D. Prince, J.H. Elder // ICCV. – 2007. – P. 1–8.

5. Yaman S. Bottleneck features for speaker recognition / S. Yaman, J.W. Pelecanos, R. Sarikaya // Odyssey. – 2012. – Vol. 12. – P. 105–108.

6. Ghahabi O. Deep belief networks for i-vector based speaker recognition / O. Ghahabi, J. Hernando // ICASSP. – 2014. – P. 1700–1704.

7. Variani E. Deep neural networks for small footprint text-dependent speaker verification / E. Variani, X. Lei, E. McDermott, I. Lopez M., J. Gonzalez-Dominguez // ICASSP. – 2014. – P. 4052–4056.

8. Zhao X.J. Deep neural networks for cochannel speaker identification / X.J. Zhao, Y.X. Wang, D.L. Wang / ICASSP. – 2015. – P. 4824–4828.

9. Garcia-Romero D. Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition / D. Garcia-Romero, X. Zhou, C.Y. Espy-Wilson // ICASSP. – 2012. – P. 4257–4260.

10. McLachlan G. Mixtures of factor analyzers / G. McLachlan, D. Peel // Finite Mixture Models. – 2000. – P. 238–256.

11. Ghahramani Z. The EM algorithm for mixtures of factor analyzers / Z. Ghahramani, G.E. Hinton // Technical Report CRGTR-96-1, University of Toronto. – 1996.

12. Garcia-Romero D. Analysis of i-vector length normalization in speaker recognition systems / D. Garcia-Romero, C.Y. Espy-Wilson // Interspeech. – 2011. – P. 249–252.

13. Ковтун В.В. 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. – 2017. – № 3(249). – С. 222–228.

14. Ковтун В.В. Використання множини мікрофонів у автоматизованій системі розпізнавання мовця критичного застосування / М.М. Биков, В.В. Ковтун // Вісник Вінницького політехнічного інституту, Вінниця. – 2017. – № 3. – С. 84–91.

15. Hatch A. Within-class covariance normalization for SVM-based speaker recognition / A. Hatch, S. Kajarekar, A. Stolcke // ISCSLP, Pittsburgh. – 2006. – P. 1471–1474.

16. Bengio Y. Learning deep architectures for AI / Y. Bengio // Foundations and trends R in Machine Learning. – 2009. – Vol. 2. – № 1. – P. 1–127.

References

1. Kovtun V.V. Pidvyshchennia shumostiikosti avtomatyzovanoi systemy rozpiznavannia movtsia krytychnoho zastosuvannia / T.V. Hryshchuk, V.V. Kovtun // Visnyk Vinnytskoho politekhnichnoho instytutu. – 2018. – № 1. – S. 98–111.

2. Reynolds D. A. Speaker verification using adapted Gaussian mixture models / D. A. Reynolds, T. F. Quatieri, R. B. Dunn // Digital Signal Processing. – 2000. – Vol. 10. – № 1. – R. 19–41.

3. Shao Y. Robust speaker identification using auditory features and computational auditory scene analysis / Y. Shao, D. Wang // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – 2008. – R. 1589–1592.

4. Elder J.H. Probabilistic linear discriminant analysis for inferences about identity / S.J.D. Prince, J.H. Elder // ICCV. – 2007. – R. 1–8.

5. Yaman S. Bottleneck features for speaker recognition / S. Yaman, J.W. Pelecanos, R. Sarikaya // Odyssey. – 2012. – Vol. 12. – R. 105–108.

6. Ghahabi O. Deep belief networks for i-vector based speaker recognition / O. Ghahabi, J. Hernando // ICASSP. – 2014. – R. 1700–1704.

7. Variani E. Deep neural networks for small footprint text-dependent speaker verification / E. Variani, X. Lei, E. McDermott, I. Lopez M., J. Gonzalez-Dominguez // ICASSP. – 2014. – R. 4052–4056.

8. Zhao X.J. Deep neural networks for cochannel speaker identification / X.J. Zhao, Y.X. Wang, D.L. Wang / ICASSP. – 2015. – R. 4824–4828.

9. Garcia-Romero D. Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition / D. Garcia-Romero, X. Zhou, C.Y. Espy-Wilson // ICASSP. – 2012. – R. 4257–4260.

10. McLachlan G. Mixtures of factor analyzers / G. McLachlan, D. Peel // Finite Mixture Models. – 2000. – R. 238–256.

11. Ghahramani Z. The EM algorithm for mixtures of factor analyzers / Z. Ghahramani, G.E. Hinton // Technical Report CRGTR-96-1, University of Toronto. – 1996.

12. Garcia-Romero D. Analysis of i-vector length normalization in speaker recognition systems / D. Garcia-Romero, C.Y. Espy-Wilson // Interspeech. – 2011. – R. 249–252.

13. Kovtun V.V. Optymizatsiia alfavitu informatyvnykh oznak dlia avtomatyzovanoi systemy rozpiznavannia movtsiv krytychnoho zastosuvannia / A.O. Bereza, M.M. Bykov, A.D. Hafurova, V.V. Kovtun // Visnyk Khmelnytskoho natsionalnoho universytetu, seriia: Tekhnichni nauky. – 2017. – № 3(249). – S. 222–228.

14. Kovtun V.V. Vykorystannia mnozhyny mikrofoniv u avtomatyzovanii systemi rozpiznavannia movtsia krytychnoho zastosuvannia / M.M. Bykov, V.V. Kovtun // Visnyk Vinnytskoho politekhnichnoho instytutu, Vinnytsia. – 2017. – № 3. – S. 84–91.

15. Hatch A. Within-class covariance normalization for SVM-based speaker recognition / A. Hatch, S. Kajarekar, A. Stolcke // ISCSLP, Pittsburgh. – 2006. – R. 1471–1474.

16. Bengio Y. Learning deep architectures for AI / Y. Bengio // Foundations and trends R in Machine Learning. – 2009. – Vol. 2. – № 1. – R. 1–127.

Рецензія/Peer review : 27.1.2019 р. Надрукована/Printed :16.2.2019 р.
Рецензент: д.т.н., проф. Бісікало О.В.