ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ТА КОМП'ЮТЕРНА ТЕХНІКА

681.327.12

Т. В. Грищук[1]
В. В. Ковтун[1]

ПІДВИЩЕННЯ ШУМОСТІЙКОСТІ АВТОМАТИЗОВАНОЇ СИСТЕМИ РОЗПІЗНАВАННЯ МОВЦЯ КРИТИЧНОГО ЗАСТОСУВАННЯ

[1]Вінницький національний технічний університет

Актуальні системи розпізнавання мовців, де застосовується і-векторне/PLDA моделювання для опису фонограм, синтезують узагальнену PLDA модель з 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. Автори пропонують синтезувати окремі PLDA моделі для опису фонограм з детермінованими рівнями відношення сигнал/шум (ВСШ), в результаті чого фактори, які характеризують індивідуальність мовців, будуть зосереджені у наймінливіших зонах і-векторного простору. Статистичний аналіз параметрів таких зон мінливості для фонограм з детермінованим рівнем ВСШ дозволив визначити шумостійкі і інформативні для розпізнавання особи мовця фактори. Для розв'язання цієї задачі отримано аналітичний вираз для PLDA моделі, параметри якої визначаються виключно значеннями і-векторів, у яку введено показники, що описують рівні ВСШ. Також синтезовано цільові функції та етапи ЕМ-алгоритму навчання ВСШ-залежних PLDA сумішей, здійснено перевірку ефективності запропонованих моделей, порівнявши їх з результатами, які показують ВСШ-незалежні суміші для визначеної бази фонограм мовців.

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

Вступ

Автоматизована система розпізнавання мовця критичного застосування повинна розпізнавати особу мовця із визначеною якістю за наданим системі мовним матеріалом, або ж визнати запропоновану системі фонограму непридатною для процедури розпізнавання. Розглянемо заходи, направлені на мінімізацію кількості випадків визнання фонограми із записом мовного сигналу непридатною для подальшого використання за рахунок закладених на етапі проектування механізмів адаптації до шумів оточуючого акустичного середовища. Існують різні підходи [1] до реалізації таких механізмів як на етапі попередньої обробки мовного сигналу, так і на етапах моделювання індивідуальних особливостей фонограм та методів класифікації мовців за їх значеннями. Зокрема здійснюється покращення якості звучання мовних сигналів застосуванням методів шумопригнічення [2], виділення із мовного сигналу стійких до присутності шумів інформативних ознак [3], застосування методів трансформації інформативних ознак з метою підвищення їх надійності [4], [5]. Для виділення інформативних ознак (факторів) із фонограм у актуальних системах розпізнавання мовців найчастіше використовують сумісний факторний аналіз (joint factor analysis, JFA) [6] та і-векторне/PLDA представлення [7]. Зазначимо також, що стійкість систем автоматизованого розпізнавання мовців до присутності у вхідних фонограмах акустичних шумів зростає при використанні інтелектуальних системи класифікації із коректно формалізованою процедурою навчання відповідно до обраного факторного представлення. В контексті задачі розпізнавання мовця, під і-вектором розуміють малорозмірний вектор, який кодує відмінність щільності розподілу імовірностей акустичних ознак, оцінених з фонограми, від еталонних. В такому представленні індивідуальні особливості мовлення особи і шуми навколишнього середовища моделюються у єдиному

просторі акустичних ознак. Отже, виникає необхідність у кластеризації факторного простору і-векторів на класи «інформативна ознака»/«шум», для чого звичайно використовують класичні статистичні методи, зокрема, лінійний дискримінантний аналіз (linear discriminant analysis, LDA) [8] та нормалізацію коваріації всередині класу (within-class covariance normalization,WCCN) [9]. Компенсацію спотворень, спричинених каналом передавання мовної інформації, здійснюють виконуючи додаткову операцію на основі імовірнісного лінійного дискримінантного аналізу (Probabilistic Linear Discriminative Analysis, PLDA) [10]. При PLDA моделюванні кожна фонограма представляється малорозмірним вектором в просторі з базисом, описаним матрицею повної мінливості (на відміну від методу JFA, де кожен запис мовного сигналу описується високорозмірними статистиками Баума-Велша). У методі PLDA опис апріорних розподілів індивідуальних ознак, екстрагованих із мовних сигналів, здійснюється із використанням розподілу Стьюдента із «важкими хвостами» (heavy tailed priors, HTP), що дозволяє отримати стійкі до викидів оцінки параметрів моделі опису індивідуальності мовлення. Припускаючи гаусів характер апріорних розподілів індивідуальних ознак і шумів, за допомогою PLDA-аналізу можна отримати достатньо точні описи мовців застосувавши ML-навчання [11] (maximum likelihood, ML), що дозволяє оптимізувати параметри моделі.

Іншим підходом компенсації шумової складової мовного сигналу, що подається на вхід системи розпізнавання, в і-векторному/PLDA факторному описі, є використання великих мультикомпонентних навчальних баз, у яких по-різному комбінуються фонограми з шумом та без [12]. Також відомі дослідження, коли вплив шумів моделюють безпосереднім втручанням у і-векторний факторний простір [13]. Наприклад у роботі [14], імовірнісну PCA-суміш представлено у факторному просторі так, щоб замінити значення MFCC акустичних векторів при обчисленні репрезентативних статистичних даних на апостеріорні середні залежних від суміші акустичних факторів. Такий підхід спростив нормалізацію акустичних факторів і підвищував надійність класифікації мовців на основі значень і-векторів. У [10] зазначений підхід використали, замінивши UBM сумішшю аналізаторів акустичних факторів, для екстракції і-векторів.

У роботі [11] для адаптації універсальної фонової моделі (universal background model, UBM) до шумного акустичного середовища застосовано векторний ряд Тейлора (vector Taylor series, VTS), з подальшим використанням навченої фонової моделі для екстрагування і-векторів. Втім, у роботі [15] для апроксимації нелінійностей між моделями, що описували шумне середовище, та моделями, що описували середовище без шумів, у кепстральному просторі використано «перетворення без запаху» (Unscented transform, UT) і наведені емпіричні результати показують, що UT перетворення є чутливішим за VTS за суттєвої нелінійності спотворень у описуваному сигналі. У дослідженні [16] базові статистичні параметри екстрактора і-векторів замінено апостеріорними імовірностями зв'язаних станів або сенонів (senones) — станів трифонів, об'єднаних у групи, наприклад, за допомогою дерев рішень, кожна з яких отримує спільну множину параметрів гаусових сумішей, які оцінювалися згортальною нейромережею (convolutional neural network, CNN), яка за своєю природою стійка до присутності шумів у вхідному образі. Результати дослідження показали перспективність та ефективність CNN/*i*-векторного підходу до розпізнавання мовців.

Постановка задачі дослідження

Нехай є множини фонограм з детермінованими рівнями ВСШ X_k, де k — детемінований рівень ВСШ, дБ. Інформацію з фонограм із множин X_k використано для визначення усереднених векторів τ_k і коваріаційних матриць Γ_k для отримання і-векторів, обчислених для кожної з k множин фонограм окремо. У актуальних системах розпізнавання мовців, де застосовується і-векторне/PLDA моделювання для опису фонограм, синтезують узагальнену PLDA модель з 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. Автори пропонують синтезувати окремі PLDA моделі для опису фонограм з детермінованими рівнями ВСШ, в результаті чого фактори, які характеризують індивідуальність мовців, будуть зосереджені у найміливіших областях і-векторного простору. Передбачається, що статистичний аналіз параметрів таких областей мінливості для фонограм з детермінованим рівнем відношення сигнал/шум (signal-to-noise ratio, SNR) дозволить визначити фактори, стійкі до рівня ВСШ, і інформативні для розпізнавання особи мовця. Для розв'язання цієї задачі необхідно отримати аналітичний вираз для PLDA моделі, параметри якої цілком визнача-

ються значеннями і-векторів, ввести у модель показники, що визначають рівні ВСШ, синтезувати цільові функції та етапи ЕМ-алгоритму навчання ВСШ-залежних сумішей та здійснити перевірку ефективності запропонованих моделей, порівнявши їх з результатами, які покажуть ВСШ-незалежні суміші для визначеної бази фонограм мовців.

Синтез універсальної PLDA-моделі

Мовному сигналу, проданому системі розпізнавання мовця у вигляді фонограми, властива динамічна природа у частотному та часовому просторах, що вимагає здійснення процедури нормалізації на етапі передоброблення, який передує етапу екстрагування інформативних ознак. В актуальних системах розпізнавання мовців використовується метод приведення даних різної розмірності у єдиний вектор, що описує мовний сигнал, який називають і-вектором. GMM мовця може розглядатися як і-вектор, компонентами якого є значення математичних очікувань суміші GMM. І-векторне представлення фонограм дозволяє компенсувати варіативність виголошення парольних мовних сигналів під час сеансів розпізнавання особи мовця і доповнюється факторним аналізом, в якому, у випадку використання GMM представлення, гаусівський і-вектор розглядається як лінійна комбінація компонент, що залежать від особи мовця і від впливу каналу поширення мовного сигналу, які вважаються статистично незалежними. Така нормалізація дозволяє враховувати ознаки, притаманні мовцеві, які не були виявлення на етапі навчання системи. Адекватний опис мовного сигналу згаданим вище методом можливий за коректної відповіді на два питання — як створити і-вектор фонограми, і як оцінити і застосувати компенсацію варіабельності сеансів розпізнавання у *i*-векторному просторі. Відповіді на ці питання викладено далі.

Для оцінювання якості процедури кластеризації на k класи фонограм за значеннями і-векторів використано коефіцієнти розбиття (partition coefficients, PC) і ентропійні коефіцієнти розбиття (entropy coefficients, PE):

$$PC = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\xi_{ik}^2; \tag{1}$$

$$PE = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\xi_{ik}\log\xi_{ik}, \tag{2}$$

де N — кількість і-векторів, K — кількість класів, а параметр $\xi_{ik} = \dfrac{\mathrm{N}\left(x_i\middle|\tau_k,\Gamma_k\right)}{\sum\limits_{r=1}^{K}\mathrm{N}\left(x_i\middle|\tau_r,\Gamma_r\right)}$ описує апостеріорну імовірність належності вектора x_i до кластера k. Діапазони зміни значень параметрів PC та PE складають $\left[1/K,1\right]$ і $\left[0,\log K\right]$, відповідно. Значення параметра PC, наближене до 1, або значення параметра PE, наближене до 0, означає однозначну кластеризацію, а значення параметра PC, близьке до $1/K$, або значення параметра PE, близьке до $\log K$ означає неможливість кластеризації.

На відміну від методу спільного факторного аналізу JFA, де простори представлення інформації про індивідуальні особливості мовлення і параметри каналу передавання мовної інформації аналізуються окремо, представлення фонограм з допомогою і-векторів відбувається у спільному факторному просторі. Зокрема, враховуючи MFCC-представлення t-ї фонограми, супервектор μ_t, у якому враховано індивідуальні особливості мовлення і акустичні умови навколишнього середовища, можна описати як

$$\mu_t = \mu + Tx_t, \tag{3}$$

де μ — GMM-супервектор (gaussian mixture model, GMM), у якому враховано індивідуальні особливості мовлення і акустичні умови навколишнього середовища, отриманий в результаті узагальнення усереднених векторів універсальної фонової моделі (universal background model, UBM) [17], T — матриця загальної варіативності малого порядку, а x_t — апостеріорна оцінка малорозмірного і-вектора.

Якщо є множина D-розмірних, нормованих за довжиною і-векторів

$$X = \left\{x_{ij}; i = 1,\ldots,N; j = 1,\ldots,H_i\right\},$$

отриманих для N мовців і H_i фонограм для кожного з мовців, то індивідуальні особливості мовлення $Z=\{z_i; i=1,\ldots,N\}$ і додаткові параметри факторного аналізатора $\omega=\{m,V,\Sigma\}$ оцінюватимемо так:

$$x_{ij}=m+Vz_i+E_{ij}, \tag{4}$$

де $V\in R^{D\times M}$ є факторною матрицею $(M<D)$, M — кількість факторів, $m\in R^D$ — глобальне середнє для X, $z_i\in R^M$ — фактор, що враховує індивідуальні особливості мовлення, з розподілом $N(0,I)$, E_{ij} — залишковий шум з нормальним розподілом $N(0,\Sigma)$. Враховуючи, що і-векторам кожного мовця має відповідати окреме значення фактора z_i, у рівнянні (4) можна згорнути і-вектори *i*-го мовця до такого вигляду:

$$\tilde{x}_i=\tilde{m}+\tilde{V}z_i+\tilde{\varepsilon}_i, \tag{5}$$

де $\tilde{x}_i=\left[x_{i1}^T,\ldots,x_{iH_i}^T\right]^T\in R^{DH_i}$; $\tilde{m}=\left[m^T,\ldots,m^T\right]^T\in R^{DH_i}$; $\tilde{V}=\left[V^T,\ldots,V^T\right]^T\in R^{DH_i}$ і $\tilde{\varepsilon}_i=\left[\varepsilon_{i1}^T,\ldots,\varepsilon_{iH_i}^T\right]^T\in R^{DH_i}$. Параметри факторного аналізатора (5) можна оцінювати за допомогою ЕМ-алгоритму [5], а саме, якщо відомий еталонний і-вектор мовця x_s і тестовий і-вектор x_t, то оцінити ступінь їх корельованості можна відношенням:

$$S_{PLDA}(x_s,x_t)=\frac{p(x_s,x_t|s=t)}{p(x_s|Spk\,s)\,p(x_t|Spk\,t)}=\frac{N\left(\left[x_s^T\ x_t^T\right]^T\left|\left[m^T\ m^T\right]^T,\hat{V}\hat{V}^T+\hat{\Sigma}\right.\right)}{N\left(x_s\left|m,VV^T+\Sigma\right.\right)N\left(x_t\left|m,VV^T+\Sigma\right.\right)}, \tag{6}$$

де $\hat{V}=\left[V^T\,V^T\right]^T$ і $\hat{\Sigma}=diag\{\Sigma,\Sigma\}$.

Отримання ВСШ-незалежної суміші PLDA (ВСШНЗ-PLDA)

Отримаємо PLDA суміш, у якій апостеріорні імовірності не залежать від рівня ВСШ вхідних фонограм. Така модель фактично описує процес контрольованого навчання факторних аналізаторів і є базовою для подальших досліджень. Наведена у виразі (4) PLDA-модель передбачає, що нормалізовані за довжиною і-вектори відповідають Гаусовому розподілу. Однак таке припущення є суттєвим узагальненням при моделюванні міжканальних зв'язків та динамічних рівнів акустичних шумів, яке може призвести до неможливості застосування таких моделей у критичних системах, для яких і-вектори краще описувати сумішшю K факторних аналізаторів з параметрами $\hat{\omega}=\{\phi_k,m_k,\Sigma_k,V_k\}_{k=1}^K$, де ϕ_k — ваги сумішей, тобто і-вектори утворюватимуться лінійно зваженою сумою K щільностей Гаусіан, кожен з яких має власний середній вектор m_k, коваріаційну матрицю Σ_k та підпростір мовців V_k. Надалі символ «$\hat{}$» ми використовуватимемо для представлення множини гіперпараметрів суміші.

Маючи еталонний і-вектор мовця x_s та і-вектор з тестовим вектором x_t, оцінити маргінальну правдоподібність для того ж мовця можна відношенням

$$p(x_s,x_t|s=t)=\sum_{k_s=1}^K\sum_{k_t=1}^K\int p(x_s,x_t,y_{k_s}=1,y_{k_t}=1,z|\hat{\omega})dz=\sum_{k_s=1}^K\sum_{k_t=1}^K P(y_{k_s}=1,y_{k_t}=1|\hat{\omega})\times$$

$$\times\int p(x_s,x_t|y_{k_s}=1,y_{k_t}=1,z,\hat{\omega})\,p(z)\,dz=\sum_{k_s=1}^K\sum_{k_t=1}^K\phi_{k_s}\phi_{k_t}\int p(x_s,x_t|y_{k_s}=1,y_{k_t}=1,z,\hat{\omega})\,p(z)\,dz= \tag{7}$$

$$=\sum_{k_s=1}^K\sum_{k_t=1}^K\phi_{k_s}\phi_{k_t}N\left(\left[x_s^T\ x_t^T\right]^T\left[m_{k_s}^T\ m_{k_t}^T\right]^T,\hat{V}_{k_sk_t}\hat{V}_{k_sk_t}^T+\hat{\Sigma}_{k_sk_t}\right),$$

де $\hat{\Sigma}_{k_sk_t}=diag\left\{\Sigma_{k_s},\Sigma_{k_s}\right\}$, $\hat{V}_{k_sk_t}=\left[V_{k_s}^T V_{k_t}^T\right]^T$ і y_{k_s} та y_{k_t} — індикаторні змінні, які показують яка з

K сумішей описує x_s та x_t репрезентативніше. Аналогічно, оцінити маргінальну правдоподібність для різних мовців можна відношенням

$$p\left(x_s, x_t \middle| x_s \neq x_t\right) = p\left(x_s \middle| Spk\, s\right) p\left(x_t \middle| Spk\, t\right),$$

де $p\left(x_s \middle| Spk\, s\right) = \sum_{k_s=1}^{K} P\left(y_{k_s}=1 \middle| \hat{\omega}\right) \int p\left(x_s \middle| y_{k_s}=1, z, \hat{\omega}\right) dz = \sum_{k_s=1}^{K} \phi_{k_s} N\left(x_s \middle| m_{k_s}, V_{k_s} V_{k_s}^T + \Sigma_{k_s}\right)$.

Відношення для обчислення імовірності $p\left(x_t \middle| Spk\, t\right)$ виглядає аналогічно вищеотриманому. Отже, підсумкова міра правдоподібності для процесу розпізнавання мовців за інформацією з ВСШ-незалежних PLDA-сумішей описуватиметься таким відношенням:

$$S_{ВСШНЗ-PLDA}\left(x_s, x_t\right) = \frac{\sum_{k_s=1}^{K}\sum_{k_t=1}^{K} \phi_{k_s}\phi_{k_t} N\left(\left[x_s^T\ x_t^T\right]^T, \left[m_{k_s}^T\ m_{k_t}^T\right]^T, \hat{V}_{k_s k_t}\hat{V}_{k_s k_t}^T + \hat{\Sigma}_{k_s k_t}\right)}{\left[\sum_{k_s=1}^{K} \phi_{k_s} N\left(x_s \middle| m_{k_s}, V_{k_s}V_{k_s}^T + \Sigma_{k_s}\right)\right]\left[\sum_{k_t=1}^{K} \phi_{k_t} N\left(x_t \middle| m_{k_t}, V_{k_t}V_{k_t}^T + \Sigma_{k_t}\right)\right]}. \quad (8)$$

Позначимо $Y = \left\{y_{ijk}\right\}_{k=1}^{K}$ як множину прихованих скритих індикаторних змінних, що вказують який з K-факторних аналізаторів $\hat{\omega} = \left\{\phi_k, m_k, \Sigma_k, V_k\right\}_{k=1}^{K}$ достовірно ідентифікує x_{ij}. Практично, $y_{ijk}=1$ якщо k-й факторний аналізатор ідентифікує x_{ij}, і $y_{ijk}=0$ у зворотному випадку. Тоді цільова функція ЕМ-алгоритму матиме такий вигляд:

$$\begin{aligned} Q\left(\hat{\omega}' \middle| \hat{\omega}\right) &= E_{y,z}\left\{\ln p\left(X, Y, Z \middle| \omega'\right) \middle| X, \hat{\omega}\right\} = E_{y,z}\left\{\sum_{ijk} y_{ijk} \ln\left[p\left(y_{ijk} \middle| \omega'\right) p\left(x_{ij} \middle| z_i, \omega_k'\right) p\left(z_i \middle| \hat{\omega}'\right)\right] \middle| X, \hat{\omega}\right\} = \\ &= \sum_{ijk} E_{y,z}\left\{y_{ijk} \ln\left[\phi_k' N\left(x_{ij} \middle| m_k' + V_k' z_i, \Sigma_k'\right) N\left(z_i \middle| 0, I\right)\right] \middle| X, \hat{\omega}\right\}. \end{aligned} \quad (9)$$

Зауважимо, що реальні апріорні імовірності z і y взаємозалежні, що робить процедуру обчислення (9) суттєво ресурсоємною або ж взагалі неможливою, тому на практиці для обчислення (9) використовують варіаційну Байєсову процедуру виводу (variational Bayesian inference procedure, VB), яка дозволяє виконати факторизований варіаційний розподіл за прихованими змінними z_i і y_{ijk}, припускаючи $q\left(z_i, y_{ijk}\right) = q\left(z_i\right) q\left(y_{ijk}\right)$. Процедура VB оцінює факторизований варіаційний розподіл, який прямує до справжніх зв'язаних апріорних імовірнісних розподілів двох взаємозалежних прихованих змінних $p\left(z_i, y_{ijk} \middle| X\right)$. Отже, на VB-E етапі автори оцінюють оптимальний варіаційний розподіл або варіаційні параметри з найвищою нижньою границею функції правдоподібності (варіаційну нижню межу). З урахуванням оновленого варіаційного розподілу на VB-M етапі параметри суміші $\left\{\phi_k, m_k, V_k, \Sigma_k\right\}_{k=1}^{K}$ оцінюються шляхом подальшої оптимізації варіаційної нижньої межі. Втім, для ще більшого спрощення розрахунків, замість використання VB-методу, ми зроблено припущення, що прихована змінна z_i апріорно незалежна від y_{ijk}, тобто $p\left(z_i, y_{ijk} \middle| X_i\right) = p\left(z_i \middle| X_i\right) p\left(y_{ijk} \middle| x_{ij}\right)$. Правомірність такого припущення у задачі розпізнавання мовців випливає з аналізу апробованого у задачі розпізнавання мови математичного апарату Марковських моделей скінченної щільності (hidden Markov models continuous-density, CDHMM) [18], де HMM-стани і Гаусові суміші також вважаються апріорно незалежними. З урахуванням вищенаведених міркувань, визначення ЕМ-критерію для вибраного способу факторного представлення розділимо на Е- та М-етапи, відповідно:

Е-етап:

$$\left\langle y_{ijk}\left|x_{ij}\right.\right\rangle \equiv Ey\left\{y_{ijk}\left|x_{ij},\widehat{\omega}\right.\right\}=\frac{\phi_k N\left(x_{ij}\left|m_k,V_kV_k^T+\Sigma_k\right.\right)}{\sum\limits_{k'=1}^{K}\phi_{k'}N\left(x_{ij}\left|m_{k'},V_{k'}V_{k'}^T+\Sigma_{k'}\right.\right)};\quad L_i=I+\sum_{k=1}^{K}H_{ik}V_k^T\Sigma_k^{-1}V_k;$$

$$\left\langle y_{ijk}z_i\left|X_i\right.\right\rangle \equiv E_{Y,Z}\left\{y_{ijk}z_i\left|X_i,\widehat{\omega}\right.\right\}=\left\langle y_{ijk}\left|X_i\right.\right\rangle\left\langle z_i\left|X_i\right.\right\rangle;\left\langle z_i\left|X_i\right.\right\rangle=L_i^{-1}\sum_{k=1}^{K}V_k^T\Sigma_k^{-1}\sum_{j\in H_{ik}}\left(x_{ij}-m_k\right);\quad(10)$$

$$\left\langle y_{ijk}z_iz_i^T\left|X_i\right.\right\rangle=\left\langle y_{ijk}\left|x_{ij}\right.\right\rangle\left\langle z_iz_i^T\left|X_i\right.\right\rangle;\ \left\langle z_iz_i^T\left|X_i\right.\right\rangle=L_i^{-1}+\left\langle z_i\left|X_i\right.\right\rangle\left\langle z_i\left|X_i\right.\right\rangle^T$$

М-етап:

$$m_k'=\frac{\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{H_i}\left\langle y_{ijk}\left|x_{ij}\right.\right\rangle x_{ij}}{\sum\limits_{i=1}^{N}\sum\limits_{j=1}^{H_i}\left\langle y_{ijk}\left|x_{ij}\right.\right\rangle};\ \phi_k'=\frac{\sum\limits_{ij}\left\langle y_{ijk}\left|x_{ij}\right.\right\rangle}{\sum\limits_{ijl}\left\langle y_{ijl}\left|x_{ij}\right.\right\rangle};\ V_k'=\left[\sum_{ij}\left(x_{ij}-m_k'\right)\left\langle y_{ijk}z_i\left|X_i\right.\right\rangle^T\right]\left[\sum_{ij}\left\langle y_{ijk}z_iz_i^T\left|X_i\right.\right\rangle\right]^{-1};\quad(11)$$

$$\Sigma_k'=\frac{\sum\limits_{ij}\left[\left\langle y_{ij}\left|x_{ij}\right.\right\rangle f_{ijk}'f_{ijk}'^T-V_k'\left\langle y_{ijk}z_i\left|X_i\right.\right\rangle f_{ijk}'^T\right]}{\sum\limits_{ij}\left\langle y_{ijk}\left|x_{ij}\right.\right\rangle};\quad f_{ijk}'=x_{ij}-m_k',$$

де H_{ik} містить індекси і-векторів *i*-го мовця, які належать суміші k, а H_{ik} — це кількість елементів у H_{ik}.

Апостеріорні очікування $\left\langle z_i,X_i\right\rangle$ у рівняннях (10) показують, що з множини H_i і-векторів мовця i тільки H_{ik} з них синтезуються *k*-м компонентом суміші. Ця інформація виявляється особливо важливою для задачі розпізнавання мовців, оскільки дозволяє, використовуючи дані фонограм з різним рівнем ВСШ, чіткіше розділити кластери мовців у і-векторному факторному просторі.

Отримання ВСШ-залежної суміші PLDA (ВСШЗ-PLDA)

У вищенаведених моделях апріорні імовірності сумішей визначаються значеннями і-векторів, які не розділяючи описують як корисний сигнал так і шумову складову фонограми. Автори пропонують вдосконалити вищеописані моделі, так, щоб оцінювати апріорні імовірності за рівнем ВСШ фонограм, що дозволить ЕМ-алгоритму чіткіше розділити кластери мовців у і-векторному просторі. Спираючись на результати емпіричних досліджень [10], які свідчать, що різні рівні шумів у фонограмах зміщують відповідні і-вектори у просторі ознак, припускаємо, що і-вектори в задачі розпізнавання мовця краще моделювати комбінацією ВСШ-залежних PLDA сумішей (ВСШЗ-PLDA) з параметрами $\widehat{\theta}=\left\{\widehat{\lambda},\widehat{\omega}\right\}=\left\{\lambda_k,\omega_k\right\}_{k=1}^{K}=\left\{\pi_k,\mu_k,\sigma_k,m_k,V_k,\Sigma_k\right\}_{k=1}^{K}$, де значення $\lambda_k=\left\{\pi_k,\mu_k,\sigma_k\right\}$ містить апріорну імовірність, середнє арифметичне і стандартне відхилення ВСШ у *k*-й групі. У цій моделі кластеризація і-векторів здійснюється за рівнем ВСШ фонограм навчальної вибірки, а і-вектори формуються лінійною комбінацією гаусових щільностей, в яких комбіновані ваги є апостеріорними імовірностями ВСШ фонограм.

Позначимо l — рівень ВСШ фонограми, і-вектор якої — x. Позначимо y_k індикаторні змінні, які визначать котрий з факторних аналізаторів відповідає за синтез x, тоді апостеріорні імовірності y_k визначимо як

$$\gamma_l\left(y_k\right)\equiv P\left(y_k=1\left|l,\widehat{\lambda}\right.\right)=\frac{\pi_k N\left(l\left|\mu_k,\sigma_k^2\right.\right)}{\sum\limits_{k'=1}^{K}\pi_{k'}N\left(l\left|\mu_{k'},\sigma_{k'}^2\right.\right)}.\quad(12)$$

З (12) випливає, що на відміну від описаного вище ВСШНЗ-PLDA, вирівнювання і-векторів у

ВСШЗ-PLDA базується на апостеріорних імовірностях ВСШ, а не на апостеріорних імовірностях і-векторів. У моделі ВСШ-залежної суміші PLDA нижні індекси i, j, k визначають мовця, сесію та суміш, відповідно. Індикаторна змінна y_{ijk} з'єднує і-вектор x_{ij} з відповідним ВСШ l_{ij} і вказує, який з факторних аналізаторів ω_k синтезував x_{ij}. На відміну від ВСШНЗ-PLDA, y_{ijk} визначає не лише x_{ij}, а і ВСШ l_{ij}, який описується гаусовою сумішшю з параметрами $\widehat{\lambda} = \{\lambda_k\}_{k=1}^{K} = \{\pi_k, \mu_k, \sigma_k\}_{k=1}^{K}$.

Враховуючи еталонний і-вектор мовця x_s, тестовий і-вектор x_t і відповідні рівні ВСШ l_s та l_t (у дБ) відповідної фонограми, маргінальна правдоподібність розпізнавання того ж мовця визначається як

$$p(x_s, x_t, l_s, l_t | s = t) = p(l_s)\, p(l_t)\, p(x_s, x_t | l_s, l_t, s = t) = p_{st} \sum_{k_s=1}^{K} \sum_{k_t=1}^{K} \int p\left(x_s, x_t, y_{k_s} = 1, y_{k_t} = 1, z \middle| \widehat{\theta}, l_s, l_t\right) dz =$$

$$= p_{st} \sum_{k_s=1}^{K} \sum_{k_t=1}^{K} \gamma_{l_s, l_t}\left(y_{k_s} y_{k_t}\right) \cdot \int p\left(x_s, x_t \middle| y_{k_s} = 1, y_{k_t} = 1, z, \widehat{\omega}\right) p(z)\, dz =$$

$$= p_{st} \sum_{k_s=1}^{K} \sum_{k_t=1}^{K} \gamma_{l_s, l_t}\left(y_{k_s} y_{k_t}\right) \cdot N\left(\left[x_s^T\ x_t T\right]^T \middle| \left[m_{k_s}^T\ m_{k_t}^T\right]^T, \hat{V}_{k_s k_t} \hat{V}_{k_s k_t}^T + \hat{\Sigma}_k\right),$$

де $p_{st} = p(l_s)\, p(l_t)$, $\hat{V}_{k_s k_t} = \left[V_{k_s}^T\ V_{k_t}^T\right]^T$, $\hat{\Sigma}_k = diag\left\{\Sigma_{k_s}, \Sigma_{k_t}\right\}$ і

$$\gamma_{l_s, l_t}\left(y_{k_s}, y_{k_t}\right) \equiv P\left(y_{k_s} = 1, y_{k_t} = 1 \middle| l_s, l_t, \widehat{\lambda}\right) = \frac{\pi_{k_s} \pi_{k_t} N\left(\left[l_s\ l_t\right]^T \middle| \left[\mu_{k_s}\ \mu_{k_t}\right]^T, diag\left\{\sigma_{k_s}^2, \sigma_{k_t}^2\right\}\right)}{\sum_{k_s'=1}^{K} \sum_{k_t'=1}^{K} \pi_{k_s'} \pi_{k_t'} N\left(\left[l_s\ l_t\right]^T \middle| \left[\mu_{k_s'}\ \mu_{k_t'}\right]^T, diag\left\{\sigma_{k_s'}^2, \sigma_{k_t'}^2\right\}\right)}.$$

Відповідно маргінальна правдоподібність розпізнавання різних мовців визначається як $p(x_s, x_t, l_s, l_t | s \neq t) = p(x_s, l_s | Spk\, s)\, p(x_t, l_t | Spk\, t)$, де імовірність

$$p(x_s, l_s | Spk\, s) = p(l_s) \sum_{k_s=1}^{K} \int p\left(x_s, y_{k_s} = 1, z \middle| \widehat{\theta}, l_s\right) dz = p(l_s) \sum_{k_s=1}^{K} \gamma_{l_s}\left(y_{k_s}\right) N\left(x_s \middle| m_{k_s}, V_{k_s}, V_{k_s}^T + \Sigma_{k_s}\right)$$

і ймовірність $p(x_t, l_t | Spk\, t)$ описується аналогічно.

Отже, враховуючи вищенаведене, підсумкова оцінка маргінальної правдоподібності $S_{SD-mPLDA}$ ВСШ-залежної суміші PLDA визначається як

$$S_{ВСШЗ-PLDA}(x_s, x_t) = \frac{\sum_{k_s=1}^{K} \sum_{k_t=1}^{K} \gamma_{l_s, l_t}\left(y_{k_s} y_{k_t}\right) N\left(\left[x_s^T\ x_t T\right]^T \middle| \left[m_{k_s}^T\ m_{k_t}^T\right]^T, \hat{V}_{k_s k_t} \hat{V}_{k_s k_t}^T + \hat{\Sigma}_{k_s k_t}\right)}{\left[\sum_{k_s=1}^{K} \gamma_{l_s}\left(y_{k_s}\right) N\left(x_s \middle| m_{k_s}, V_{k_s}, V_{k_s}^T + \Sigma_{k_s}\right)\right]\left[\sum_{k_t=1}^{K} \gamma_{l_t}\left(y_{k_t}\right) N\left(x_t \middle| m_{k_t}, V_{k_t}, V_{k_t}^T + \Sigma_{k_t}\right)\right]} \quad (13)$$

Позначимо $Y = \left\{y_{ijk}\right\}_{k=1}^{K}$ — множину прихованих індикаторних змінних, що визначають який з K факторних аналізаторів має обиратися за рівнем ВСШ вхідної фонограми. Також позначимо $L = \left\{l_{ij}\right\}$, $i = 1, ..., N$, $j = 1, ..., H_i$ ВСШ вхідної фонограми. Очевидно, $y_{ijk} = 1$, якщо k-й факторний аналізатор синтезує x_{ij}, і $y_{ijk} = 0$, якщо ні. Отже, допоміжна функцій ЕМ-алгоритму буде такою:

$$Q\left(\hat{\theta}',\hat{\theta}\right)=E_{Y,Z}\left\{\log p\left(X,L,Y,Z\left|\hat{\theta}'\right.\right)\left|X,L,\hat{\theta}\right.\right\}=$$

$$=E_{Y,Z}\left\{\sum_{ijk}y_{ijk}\log\left[p\left(l_{ij}\left|y_{ijk}=1\right.\right)p\left(y_{ijk}\right)\times p\left(x_{ij}\left|z_i,y_{ijk}=1,\omega_k'\right.\right)p\left(z_i\right)\right]\left|X,L,\hat{\theta}\right.\right\}= \quad (14)$$

$$=\sum_{i=1}^{N}\sum_{j=1}^{H_i}\sum_{k=1}^{K}E_{Y,Z}\left\{y_{ijk}\log\left[N\left(l_{ij}\left|\mu_k',\sigma_k'\right.\right)\pi_k'\times N\left(x_{ij}\left|m_k'+V_k'\,z_i,\Sigma_k'\right.\right)N\left(z_i\left|0,I\right.\right)\right]\left|X,L,\hat{\theta}\right.\right\},$$

де $\pi_k'\equiv P\left(y_{ijk}=1\right)$ — апріорна імовірність k-го факторного аналізатора. Максимізація (14) дозволяє сформулювати ЕМ-критерій так:

Е-етап:

$$\left\langle y_{\ ijk}\left|L\right.\right\rangle\equiv E_Y\left\{y_{ijk}\left|L,\hat{\lambda}\right.\right\}=\frac{\pi_k N\left(l_{ij}\left|\mu_k,\sigma_k^2\right.\right)}{\sum_{r=1}^{K}\pi_r N\left(l_{ij}\left|\mu_r,\sigma_r^2\right.\right)};\ \ L_i=I+\sum_{k=1}^{K}H_{ik}V_k^T\,\Sigma_k^{-1}V_k\ ;$$

$$\left\langle y_{ijk}\left|X,L\right.\right\rangle=\left\langle y_{ijk}\left|L\right.\right\rangle\left\langle z_i\left|X\right.\right\rangle;\ \left\langle z_i,X\right\rangle=L_i^{-1}\sum_{k=1}^{K}\sum_{j\in H_{ik}}V_k^T\,\Sigma_k^{-1}\left(x_{ij}-m_{\,k}\right); \quad (15)$$

$$\left\langle y_{ijk}z_i\left|X,L\right.\right\rangle=\left\langle y_{ijk}\left|L\right.\right\rangle\left\langle z_iz_i^T\left|X\right.\right\rangle;\ \ \left\langle z_iz_i^T\left|X\right.\right\rangle=L_i^{-1}+\left\langle z_i\left|X\right.\right\rangle\left\langle z_{\ i}\left|X\right.\right\rangle^T.$$

М-етап:

$$m_k'=\frac{\sum_{i=1}^{N}\sum_{j=1}^{H_i}\left\langle y_{ijk}\left|L\right.\right\rangle x_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{H_i}\left\langle y_{ijk}\left|L\right.\right\rangle};\ \ \pi_k'=\frac{\sum_{ij}\left\langle y_{ijk}\left|L\right.\right\rangle}{\sum_{ijl}\left\langle y_{ijl}\left|L\right.\right\rangle};\ \ \mu_k'=\frac{\sum_{ij}\left\langle y_{ijk}\left|L\right.\right\rangle l_{ij}}{\sum_{ij}\left\langle y_{ijk}\left|L\right.\right\rangle};$$

$$\sigma_k^{2'}=\frac{\sum_{ij}\left\langle y_{ijk}\left|L\right.\right\rangle\left(l_{ij}-\mu_k'\right)^2}{\sum_{ij}\left\langle y_{ijk}\left|L\right.\right\rangle};\ V_k'=\left[\sum_{ij}f_{ijk}'\left\langle y_{ijk}z_i\left|X,L\right.\right\rangle^T\right]\left[\sum_{ij}\left\langle y_{ijk}z_iz_i^T\left|X,L\right.\right\rangle\right]^{-1}; \quad (16)$$

$$\Sigma_k'=\frac{\sum_{ij}\left[\left\langle y_{ijk}\left|L\right.\right\rangle f_{ijk}'f_{ijk}'^T-V_k'\left\langle y_{ijk}z_i\left|X,L\right.\right\rangle f_{ijk}'^T\right]}{\sum_{ij}\left\langle y_{ijk}\left|L\right.\right\rangle};\ f_{ijk}'=x_{ij}-m_k'.$$

Зауважимо, що для більшої точності апостеріорне середнє $\left\langle y_{ijk}\left|L\right.\right\rangle$ має бути $\left\langle y_{ijk}\left|x_{ij},l_{ij}\right.\right\rangle$. Для гарантування, що кластеризація відбувається у відповідності до ВСШ рівня фонограм, а не значень і-векторів, припускаємо, що y_{ijk} апостеріорно не залежить від x_{ij}. Також, за аналогією із вищеописаним, щоб не застосовувати VB-алгоритм для апроксимації реальних апостеріорних значень y_{ijk} і z_{ij}, припускаємо, що y_{ijk} і z_{ij} апостеріорно незалежні.

Отже, отримані авторами оцінювання міри правдоподібності для універсального PLDA, ВСШ-незалежного PLDA і ВСШ-залежного PLDA математично описані відношеннями (6), (8) та (13) відповідно, а сформульовані у рівняннях (10), (11) і (15), (16) етапи формування ЕМ-клаифікаторів дозволять емпірично оцінити адекватність запропонованих теоретичних результатів.

Постановка експерименту та аналіз результатів

Як базу фонограм для навчання та тестування створеної із застосуванням вищеописаного математичного апарату автоматизованої системи розпізнаванням мовців критичного застосування (АСРМКЗ) використано базу записів з безкоштовної бази даних NOIZEUS [19] — 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 з 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 з мовців була хоча б одна фонограма з ВСШ = 0 дБ. Фонограми навчальної вибірки використовувалися як вхідні дані для синтезу залежних від статі мовця UBM моделей, повних матриць варіативності, LDA-WCCN и PLDA моделей. Для детектування інтервалів мовної активності у фонограмах застосовувався двоканальний VAD алгоритм (Voice activity detection, VAD) [20]. Інтервали мовної активності тривалістю 3 секунди розбивалися на фрейми тривалістю 30 мс з 15 мс зсувом, із даних яких екстрагувалися 19 MFCC коефіцієнтів, їх енергія, перша і друга їхні похідні. Для компенсації ефекта Гиббса виконувалося зважування сигналу вікном Хемінга. Ефекти канальних спотворень на факторному рівні компенсувалися розрахунком кепстрального середнього (cepstral mean subtraction) та, враховуючи достатню тривалість фреймів аналізу, здійсненням операції факторного вирівнювання (feature warping) [7]. До кожної чистої фонограми (з рівнем ВСШ = 0) навчальної вибірки підмішувався запис акустичних шумів, вид та рівень ВСШ яких вибирався випадково із мовної бази. В результаті на одну чисту фонограму припадало десять з рівнем ВСШ 5 дБ, 10 дБ або 15 дБ.

Системи і-векторів створеної АСРМКЗ базуються на залежних від статі мовця UBM моделях з 1024 сумішами, навчених на мовному матеріалі бази NOIZEUS, і матрицях повної варіативності із 500 факторами, до яких застосовувалися операції внутрікласової коваріаційної нормалізації (within-class covariance normalization, WCCN) [9] і нормалізації довжини і-векторів [21]. Потім застосовувався лінійний дискримінантний аналіз (linear discriminant analysis, LDA) [8] та ще раз WCCN для зменшення розмірності факторного простору до 200 перед навчанням PLDA і PLDA сумішей з 150 прихованих змінних.

Для навчання PLDA моделі використовувались два набори даних. У перший набір увійшли фонограми з рівнями ВСШ = 5 дБ, ВСШ = 15 дБ та змішаним рівнем ВСШ ≈ 10 дБ. Розподіл ВСШ для першого набору показано на лівій частині рис. 1. Помітно, що рівні ВСШ на рисунку в основному не рівні 5 чи 15 дБ, а лише наближаються до цих значень, так як на рисунку показано фактичні рівні ВСШ, отримані рішенням VAD рівнянь [13]. У процедурі отримання другого набору фонограми з шумом зливаються у загальний файл, який згодом розбивається на 1800 фонограм з випадково обраним рівнем ВСШ. Розподіл ВСШ для другого набору показано у правій частині рис. 1. Зауважимо, що дані з другого набору є ближчими до реальних, так як не несуть змістовної інформації.

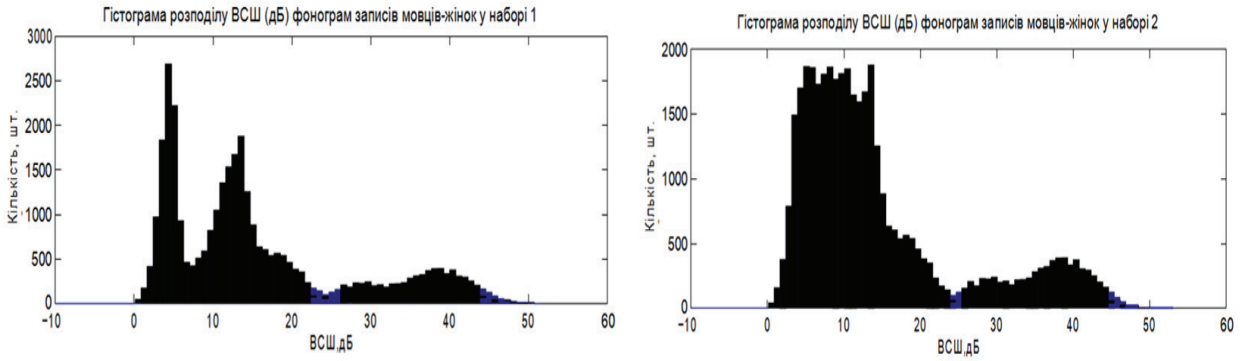

Рис. 1. Гістограми розподілу ВСШ (дБ) фонограм записів мовців-жінок у наборах 1 і 2

Отже, всі три вищеописані математичні PLDA моделі (PLDA, ВСШЗ-PLDA і ВСШНЗ-PLDA) навчалися на фонограмах з першого та другого наборів за допомогою EM-алгоритмів, формалізованих у рівняннях (10), (11) і (15), (16), коли $K = \{2, 3, 4\}$. Результати поведених експериментів з розпізнавання мовців системою на базі відповідних PLDA моделей наведено у табл. 1. Для оцінювання якості роботи АСРМКЗ використано два критерії: імовірність правильного розпізнавання P_+ та мінімальна функція вартості виявлення (minimum detection cost function, minDCF) [22]. 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.

З табл. 1 випливає, що з позиції критерію minDCF ВСШНЗ-PLDA є кращим за PLDA у разі розпізнавання мовців-чоловіків, а у разі розпізнавання мовців-жінок картина протилежна. У ВСШНЗ-PLDA апостеріорна імовірність ϕ_k у рівнянні (8) визначається апостеріорною імовірністю індикаторної змінної навчальних даних (див. (11). Ці апріорні імовірності використовуються як основа для K сумішей. Відповідно, ваги суміші для об'єднання PLDA оцінок у рівнянні (8) не залежать від тестових фонограм, тобто та сама комбінація ваг буде використовуватися для класифікації незалежно від значень факторів, екстрагованих з тестової фонограми. Така обставина робить PLDA суміш грубою, в той час як у ВСШЗ-PLDA суміші апостеріорні імовірності індикаторних змінних y_{ijk} враховують ВСШ тестової фонограми для визначення комбінації ваг у рівнянні (13).

Таблиця 1

Залежність критеріїв якості роботи АСРМКЗ від наборів навчальних та тестових даних і методу моделювання факторів 78

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		96,51	0,32	97,14	0,30	96,87	0,36	97,53	0,35
	PLDA (і-век.+ВСШ)		96,81	0,32	97,07	0,30	96,90	0,35	97,62	0,34
	ВСШ незалежна PLDA	2 суміші	96,59	0,31	97,01	0,33	96,92	0,36	97,66	0,36
		3 суміші	96,76	0,32	97,06	0,32	97,02	0,36	97,45	0,37
		4 суміші	96,81	0,30	97,09	0,31	96,84	0,37	97,64	0,35
	ВСШ залежна PLDA	2 суміші	96,72	0,32	97,03	0,32	96,90	0,37	97,41	0,37
		3 суміші	97,06	0,33	97,14	0,31	97,40	0,34	97,41	0,34
		4 суміші	96,89	0,32	97,10	0,32	97,16	0,34	97,26	0,36
Набір 2	PLDA		96,68	0,33	96,91	0,33	97,06	0,36	97,36	0,37
	PLDA (і-век.+ВСШ)		96,76	0,33	96,77	0,32	97,02	0,36	97,28	0,34
	ВСШ незалежна PLDA	2 суміші	96,68	0,33	96,91	0,35	96,94	0,37	97,33	0,36
		3 суміші	96,87	0,33	96,79	0,32	97,18	0,36	97,41	0,35
		4 суміші	96,63	0,31	96,87	0,31	97,14	0,36	97,48	0,35
	ВСШ залежна PLDA	2 суміші	96,67	0,32	97,00	0,33	97,10	0,36	97,36	0,36
		3 суміші	96,62	0,32	97,00	0,33	97,10	0,36	97,33	0,36
		4 суміші	96,65	0,32	96,92	0,32	96,91	0,37	97,14	0,38

На рис. 2 візуалізовано і-вектори для ВСШНЗ-PLDA і ВСШЗ-PLDA. Кожній точці на рисунку відповідає і-вектор, екстрагований із множини фонограм першого набору. З рисунку видно, що для ВСШЗ-PLDA і-вектори з однаковими компонентами суміші групуються відносно майже однакових ВСШ, тоді як для ВСШНЗ-PLDA і-вектори зі значною різницею ВСШ відносять до однієї суміші. Такий результат доводить доцільність використання інформації про рівень ВСШ при розпізнаванні особи мовця, зокрема, ВСШЗ-PLDA модель підсилює зв'язок між компонентами суміші і тестовими і-векторами, що збільшує ймовірність правильного розпізнавання мовців.

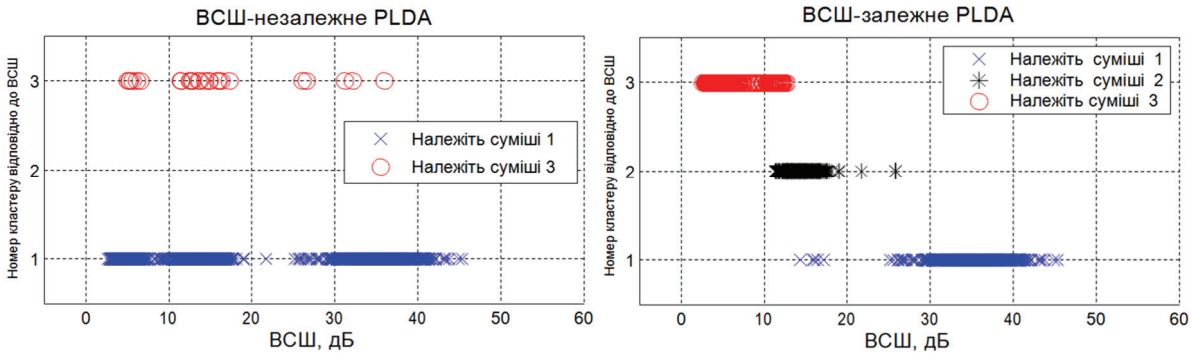

Рис. 2. Розподіл і-векторів з тестового набору 2 до відповідних кластерів за рівнем ВСШ в залежності від методу моделювання факторів, якщо $K = 3$

Наведені у табл. 1 результати експериментів показують, що ВСШ ВСШЗ-PLDA модель дозволяє отримати кращі результати ніж ВСШНЗ-PLDA модель майже для всіх варіантів тестових даних, коли для навчання моделей використовувалися фонограми із першого набору. Проте, коли навчання моделей відбувалося за даними з другого набору ситуація виявляється протилежною. Це можна пояснити тим, що використання фонограм з трьома рівнями ВСШ у формуванні першого набору навчальних даних забезпечує більшу інформативність порівняно з другим способом отримання навчальних даних. Але, якщо даних про рівень ВСШ у навчальних даних невідомий, то краще використовувати для навчання ВСШНЗ-PLDA модель, так як при формуванні другого набору навчальних даних рівень ВСШ у межах навчальних даних динамічно змінювався.

Результати досліджень дозволяють стверджувати, що інформація про рівень ВСШ у фонограмах може використовуватися як додатковий фактор до і-векторів при PLDA моделюванні та оцінюванні. Щоб забезпечити рівням ВСШ ту ж розмірність, що і в інших компонентів і-векторів виконаємо процедуру Z-нормалізації ВСШ перед додаванням їх до і-векторів. Результати ефективності роботи АСРМКЗ з таким набором факторів показано у стовпцях «PLDA (і-век.+ВСШ)» табл. 1. Очевидно, що коли для навчання системи використовується перший набір, така модифікація факторного простору є доцільною, втім, при навчанні системи другим набором даних така інформація не приводить до зростання ефективності розпізнавання, так як зміна значень ВСШ у навчальних і тестових даних носить випадковий характер, і може навіть погрішити результати, так як ВСШ — незалежний від мовця фактор, а значення і-векторів навпаки.

Забезпечення надійності процедури розпізнавання мовця під час експлуатації системи розпізнавання у реальному акустичному середовищі є основою для віднесення створеної системи до класу критичних систем. Отже, дані бібліотеки фонограм першого і другого набору розбито на множину навчальних та тестових даних до яких додано різних видів шумів на рівнях ВСШ = 5 дБ і 15 дБ, так, щоб ці параметри для тестової та навчальної вибірки не збігались. Візуально результат таких перетворень показано на рис. 3 і якщо порівняти ці гістограми з гістограмами навчальних вибірок, показаних на рис. 1, видно суттєво різний розподіл рівня ВСШ. У табл. 2 наведено результати тестування АСРМКЗ з визначеними вище умовами, які підтверджують, що навіть за таких несприятливих умов ВСШЗ-PLDA та ВСШНЗ-PLDA працюють краще за PLDA моделі, а показники ВСШЗ-PLDA взагалі виявляються найкращими.

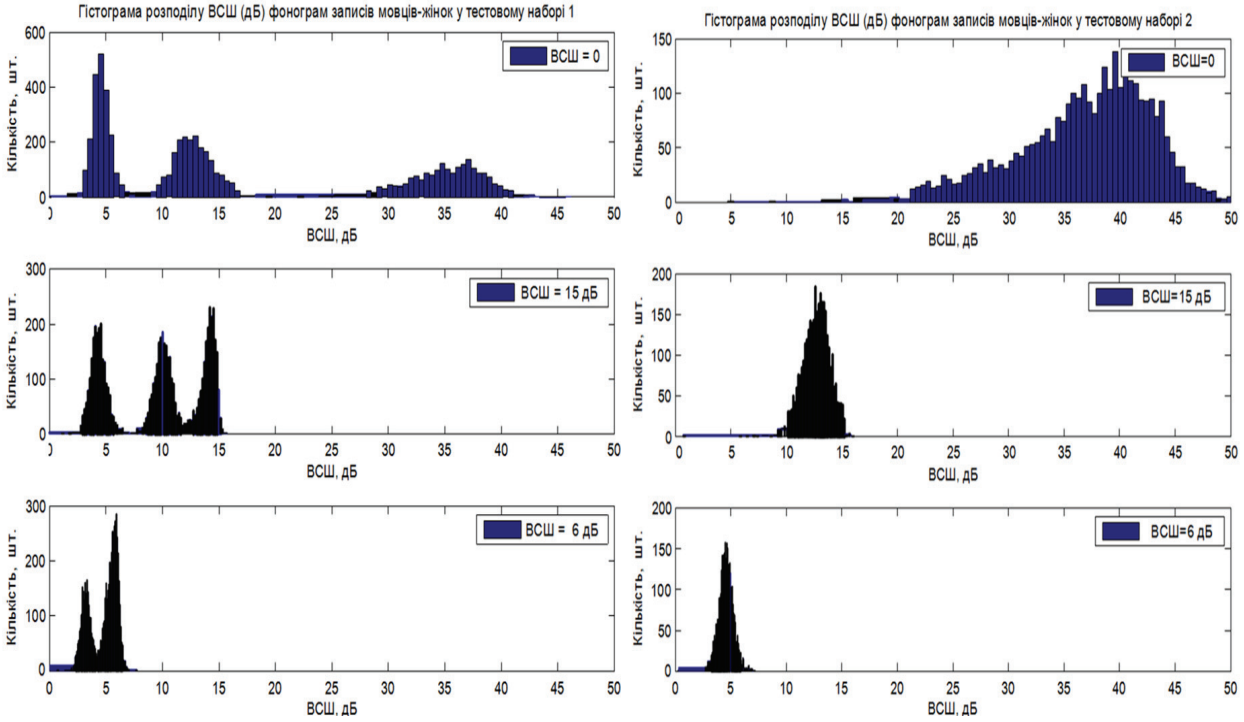

Рис. 3. Гістограми розподілу ВСШ (дБ) фонограм записів мовців-жінок у тестових наборах 1 і 2

Таблиця 2

Залежність критеріїв якості роботи АСРМКЗ від наборів навчальних та тестових даних і методу моделювання факторів при використанні ВСШ як додаткового фактора

Набір навчальних даних	Метод моделювання	Тест.набір 1 + ВСШ 15дБ		Тест.набір 1 + ВСШ 6дБ		Тест.набір 2 + ВСШ 15дБ		Тест.набір 2 + ВСШ 6дБ	
		P+, %	minDCF	P+, %	minDCF	P+, %	minDCF	P+, %	minDCF
Набір 1	PLDA	97,21	0,37	96,88	0,39	96,62	0,42	94,27	0,58
	ВСШ-незалежна PLDA	97,30	0,37	97,02	0,39	96,72	0,40	94,37	0,58
	ВСШ-залежна PLDA	97,45	0,36	96,98	0,42	96,63	0,42	93,89	0,60
Набір 2	PLDA	97,19	0,37	96,82	0,41	96,50	0,42	93,92	0,59
	ВСШ-незалежна PLDA	97,31	0,37	96,92	0,40	96,51	0,42	94,06	0,59
	ВСШ-залежна PLDA	97,09	0,39	96,94	0,41	96,56	0,43	94,02	0,59

Висновки

Отже, основним джерелом помилок в роботі автоматизованої системи розпізнавання мовців критичного застосування є присутність шумів у вхідних фонограмах. З метою підвищення стійкості таких систем до впливу шумів автори розробили ВСШ-залежну модель PLDA. Для тестування запропонованих вдосконалень підготовлено два набори тестових даних: з різними детермінованими рівнями ВСШ та з динамічно змінюваним ВСШ у межах фонограми. Також авторами запропоновано використовувати інформацію про рівень ВСШ в якості додаткового фактора для розпізнавання. Результати тестів довели адекватність запропонованих авторами моделей ВСШ-залежних моделей PLDA сумішей для всіх варіантів тестових даних.

Зокрема, результати експериментів, наведені у табл. 1 і 2, показують, що ВСШ ВСШЗ-PLDA модель дозволяє отримати кращі результати ніж ВСШНЗ-PLDA модель майже для всіх варіантів тестових даних, коли для навчання моделей використовувалися фонограми з першого набору. Проте, коли навчання моделей відбувалося за даними з другого набору ситуація виявлялась протилежною. Це можна пояснити тим, що використання фонограм з трьома рівнями ВСШ для формування першого набору навчальних даних забезпечує більшу інформативність порівняно з другим способом отримання навчальних даних.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] М. М. Биков, та В. В. Ковтун, «Оцінювання надійності автоматизованих систем розпізнавання мовців критичного застосування,» *Вісник Вінницького політехнічного інституту*, № 2, с. 70-76, 2017.

[2] R. Saeidi, and D. A. van Leeuwen, “The Radboud University Nijmegen submission to NIST SRE-2012,” [Online]. Available: http://repository.ubn.ru.nl/bitstream/handle/2066/116114/116114.pdf?sequence=1. Accessed on: February 14. 2018.

[3] Y. Shao, and D. Wang, “Robust speaker identification using auditory features and computational auditory scene analysis.” [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.151.4921&rep=rep1&type=pdf. Accessed on: February 14. 2018.

[4] J. Pelecanos, and S. Sridharan, *Feature warping for robust speaker verification.* [Online]. Available: http://www.isca-speech.org/archive_open/archive_papers/odyssey/odys_213.pdf. Accessed on: February 14. 2018.

[5] М. М. Биков, та В. В. Ковтун, «Використання множини мікрофонів у автоматизованій системі розпізнавання мовця критичного застосування,» *Вісник Вінницького політехнічного інституту*, № 3, с. 84-91, 2017.

[6] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A study of inter-speaker variability in speaker verification,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 980-988, 2008.

[7] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788-798, 2011.

[8] C. Bishop, *Pattern Recognition and Machine Learning*. New York, USA: Springer, 2006.

[9] A. Hatch, S. Kajarekar, and A. Stolcke *Within-class covariance normalization for SVM-based speaker recognition* [Online]. Available: http://www.isca-speech.org/archive/archive_papers/interspeech_2006/i06_1874.pdf. Accessed on: February 14, 2018.

[10] T. Hasan, and John H. L. Hansen, “Maximum likelihood acoustic factor analysis models for robust speaker verification in noise,” *IEEE Transactions on Audio, Speech And Language Processing*, vol. 22, no. 2, pp. 381-391, 2014.

[11] Y. Lei, L. Burget, and N. Scheffer, “A noise robust i-vector extractor using vector Taylor series for speaker recognition,” *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6788–6791. 21 October 2013. 2013. DOI: 10.1109/ICASSP.2013.6638976.

[12] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, “Towards noise-robust speaker recognition using probabilistic linear discriminant analysis,” *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*

(ICASSP), p. 4253-4256. 31 August 2012. 2012. DOI: 10.1109/ICASSP.2012.6288858.

[13] N. Li, and M. W. Mak, "SNR-invariant PLDA modeling in nonparametric subspace for robust speaker verification," *IEEE/ACM Trans. on Audio Speech and Language Processing*, vol. 23, no. 10, pp. 1648-1659, 2015.

[14] T. Hasan, and J. Hansen, "Acoustic factor analysis for robust speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 4, pp. 842-853, 2013.

[15] D. Martinez, L. Burget, T. Stafylakis, Y. Lei, P. Kenny, and E. Lleida, "Unscented transform for i-vector-based noisy speaker recognition," *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4070-4074. 14 July 2014. 2014. DOI: 10.1109/ICASSP.2014.6854361.

[16] M. McLaren, Y. Lei, N. Scheffer, and L. Ferrer, *Application of convolutional neural networks to speaker recognition in noisy conditions.* [Online]. Available: https://pdfs.semanticscholar.org/f6b0/984d6289acdb87139f1ca4abc42d31cb24fc.pdf. Accessed on: February 14, 2018.

[17] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41, 2000.

[18] L. Rabiner, and B. H. Juang *Fundamentals of Speech Recognition*. NJ, USA: Prentice-Hall International, Inc., 1993.

[19] А. О. Береза, М. М. Биков, та В. В. Ковтун, «Оптимізація алфавіту інформативних ознак для автоматизованої системи розпізнавання мовців критичного застосування,» *Вісник Хмельницького національного університету, серія: Технічні науки*, № 3 (249), с. 222-228, 2017.

[20] M. W. Mak, and H. B. Yu, "A study of voice activity detection techniques for NIST speaker recognition evaluations," *Computer, Speech and Language*, vol. 28, no. 1, pp. 295-313, 2013.

[21] D. Garcia-Romero, and C. Espy-Wilson, *Analysis of i-vector length normalization in speaker recognition systems*. [Online]. Available: http://www.isr.umd.edu/Labs/SCL/publications/conference/dgromero_is11_lnorm_final.pdf. Accessed on: February 14. 2018.

[22] R. Saeidi, and D. A. van Leeuwen. *The Radboud University Nijmegen submission to NIST SRE-2012.* [Online]. Available: http://repository.ubn.ru.nl/bitstream/handle/2066/116114/116114.pdf?sequence=1. Accessed on: February 14, 2018.

Рекомендована кафедрою комп'ютерних систем управління ВНТУ

Грищук Тетяна Вікторівна — канд. техн. наук, доцент, доцент кафедри комп'ютерних систем управління;

Ковтун В'ячеслав Васильович — канд. техн. наук, доцент, доцент кафедри комп'ютерних систем управління, e-mail: kovtun_v_v@vntu.edu.ua

T. V. Gryshchuk[1]
V. V. Kovtun[1]

Increase Noise Resistance of the Automatic Speaker Recognition System of Critical Use

[1]Vinnytsia National Technical University

The relevant speaker recognition systems in which i-vector/PLDA modeling is applied to the description of soundtracks synthesize the generalized PLDA model with average parameters on all soundtracks base without their segregation on the noise level. As a result such systems provide the acceptable level of reliability only in the presence of the large training selection, both by quantity, and on duration of soundtracks. Authors suggest to synthesize separate PLDA models for the description of soundtracks with the determined levels the relation signal / noise (RSN) therefore factors which characterize specific features of a speaker's voice, will be concentrated in the most changeable areas of i-vector space. It is assumed that statistical analysis of the parameters of such variability regions for phonograms with a signal-to-noise ratio determinants will determine the factors that are stable to the noise level in the signal and informative for the speaker's identity recognition. The statistical analysis of parameters of such areas of variability for soundtracks with the determined RSN level allowed to define noise resistant and informative for speaker recognition factors. For the solution of this task analytical expression for PLDA model which parameters are defined only by values of i-vectors, into which it is entered the parameters describing the RSN levels is received. Criterion functions and stages EM-algorithm of training RSN depended PLDA mixture are also synthesized and check of efficiency of the offered models by their comparison with results which show RSN independed PLDA mixture for a certain base of the speaker's soundtracks is carried out. For complex testing of the proposed theoretical results, the authors formed two test samples of phonograms that differed in the way of making noise into a signal. Experimental results show that the RSN depended PLDA model allows for better results than the RSN independed PLDA model for almost all test data variants, when phonograms from the first set were used for training models. However, when the training of models occurred according to data from the second set, the situation turns out to be the opposite. This can be explained by the fact that the use of phonograms with the three levels of the RSN the formation of the first set of training data provides greater informativity than the second way of obtaining training data.

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

Gryshchuk Tetiana V. — Cand. Sc. (Eng.), Assistant Professor of the Chair of Computer Control Systems;

Kovtun Viacheslav V. — Cand. Sc. (Eng.), Assistant Professor of the Chair of Computer Control Systems, e-mail: kovtun_v_v@vntu.edu.ua

Т. В. Грищук[1]
В. В. Ковтун[1]

Повышение шумоустойчивости автоматизированной системы распознавания диктора критического применения

[1]Винницкий национальный технический университет

Актуальные системы распознавания диктора, в которых применяется і-векторное/PLDA моделирования для описания фонограмм, синтезируют обобщенную PLDA модель с усредненными параметрами по всей базе фонограмм без их сегрегации по уровню шумов. В результате такие системы обеспечивают приемлемый уровень надежности только при наличии большой обучающей выборки, как по количеству, так и по продолжительности фонограмм. Авторы предлагают синтезировать отдельные PLDA модели для описания фонограмм с детерминированными уровнями отношение сигнал/шум (ОСШ), в результате чего факторы, характеризующие индивидуальные особенности голоса диктора, будут сосредоточены в наиболее изменчивых областях і-векторного пространства. Статистический анализ параметров таких областей изменчивости для фонограмм с детерминированным уровнем ОСШ позволил определить шумоустойчивые и информативные для распознавания диктора факторы. Для решения этой задачи получено аналитическое выражение для PLDA модели, параметры которой определяются исключительно значениями і-векторов, в которую введены параметры, описывающие уровни ОСШ. Также синтезированы целевые функции и этапы EM-алгоритма обучения ОСШ-зависимых PLDA смесей и осуществлена проверка эффективности предложенных моделей путем их сравнения с результатами, которые показывают ОСШ-независимые PLDA смеси для определенной базы фонограмм говорящих.

Ключевые слова: автоматизированная система распознавания диктора критического применения, і-векторы, смесь PLDA.

Грищук Татьяна Викторовна — канд. техн. наук, доцент, доцент кафедры компьютерных систем управления;

Ковтун Вячеслав Васильевич — канд. техн. наук, доцент, доцент кафедры компьютерных систем управления, e-mail: kovtun_v_v@vntu.edu.ua