DOI 10.31891/2307-5732-2018-267-6(1)-143-150
УДК 681.327.12

А.О. БЕРЕЗА, Т.В. ГРИЩУК, В.В. КОВТУН
Вінницький національний технічний університет

ОЦІНЮВАННЯ НАДІЙНОСТІ СЕАНСУ РОЗПІЗНАВАННЯ ОСОБИ АВТОМАТИЗОВАНОЮ СИСТЕМОЮ РОЗПІЗНАВАННЯ МОВЦЯ КРИТИЧНОГО ЗАСТОСУВАННЯ

Автори пропонують метод оцінювання надійності сеансу розпізнавання особи автоматизованою системою розпізнавання мовця критичного застосування (АСРМКЗ), який, на відміну від існуючих, використовує Баєсівську мережу, яка описує зв'язки між оцінкою сеансу розпізнавання, згенерованою класифікатором АСРМКЗ, встановленим у системі значенням порогу правдоподібності, оцінками надійності сеансу розпізнавання та оцінками впливу основних збурюючих факторів на мовний сигнал, що дозволяє за рахунок притаманних Байєсівській мережі (БМ) властивостей проводити оцінювання надійності сеансу розпізнавання в умовах часткової невизначеності згаданих параметрів та досягти заданого значення показника якості системи розпізнавання, ідентифікувавши ненадійні сеанси розпізнавання. Автори синтезували базову та уточнену Байєсівські мережі для оцінювання надійності сеансів розпізнавання мовців за даними тестувальної вибірки, які відрізняються тим, що в уточненій БМ параметр надійності виражено залежною від оцінок рівня присутніх у мовному сигналі збурень змінною, тоді як базова БМ є більш універсальною та потребує меншої кількості обчислень.

Ключові слова: *автоматизована система розпізнавання мовців критичного застосування, надійність, імовірнісна оцінка, сеанс розпізнавання.*

A.O. BEREZA, T.V. GRISHCHUK, V.V. KOVTUN
Vinnytsia National Technical University

ESTIMATION OF THE RELIABILITY OF A PERSON RECOGNITION SESSION BY AN THE AUTOMATIC SPEAKER RECOGNITION SYSTEM OF CRITICAL USE

The authors propose a method for an estimation of the reliability of a person recognition session by an automatic speaker recognition system of critical use (ASRSCU) that, unlike existing ones, uses the Bayesian network (BN), which describes the relationships between the evaluation of the recognition session generated by the ASRSCU classifier, system likelihood threshold value , estimates of the reliability of the recognition session, and estimates of the influence factors with are in the speech signal, which, due to the inherent Bayesian se properties, allow an estimation of the reliability of test session under partial uncertainties of said parameters and achieve a predetermined quality indicator to identify unreliable recognition sessions. The authors synthesized the basic and refined Bayesian networks for an estimation of the reliability of speaker recognition sessions based on test sample data, which differ in that in the refined BN the reliability parameter is expressed by the variable-dependent influence factors, while the base BN is more universal and requires less the number of calculations.

Keywords: *automatic speaker recognition system of critical use, reliability, probabilistic estimation, person recognition session.*

ВСТУП

Незважаючи на велику кількість методів обробки мовних сигналів, які застосовуються у сучасних системах розпізнавання мовця взагалі та використовувалися авторами при створенні автоматизованої системи розпізнавання мовця критичного застосування (АСРМКЗ) [1] зокрема, імовірною є ситуація коли ступінь збурення парольного мовного сигналу у фонограмі виявляється завеликим для розпізнавання особи мовця із прийнятною якістю. Неможливість детектування таких наперед ненадійних спроб розпізнавання призводить до різкого падіння якісних показників роботи АСРМКЗ і в цілому унеможливлює її критичне застосування. Існування методу адекватної оцінки надійності сеансу розпізнавання мовця дозволило б визначити необхідність проходження мовцем повторного сеансу розпізнавання, підвищуючи таким чином надійність системи розпізнавання в цілому.

На даний час проведено ряд досліджень [2–4] щодо комбінування різних джерел помилок систем розпізнавання образів у вигляді узагальненої якісної міри, використовуючи яку можна було б автоматизувати процес визначення порогу достовірності рішень, які приймаються класифікатором системи розпізнавання. Один з перспективних підходів, який застосовується для оцінювання надійності систем розпізнавання мови [5,6], оснований на використанні Баєсової мережі (Bayesian network, БМ) для імовірнісного оцінювання надійності сеансу розпізнавання за визначеними якісними показниками. Відзначимо, що згадані дослідження проведені для оцінювання надійності систем розпізнавання мови, які принципово відрізняються від систем розпізнавання мовця. Останні для прийняття рішень використовують інформацію яка персоніфікує мовні сигнали. Також відкритим є питання оптимальної конфігурації Баєсової мережі та достатньої множини факторів, що впливають на надійність сеансу розпізнавання мовця.

ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕННЯ

Метою наведених у статті досліджень є створення методу оцінювання надійності сеансу розпізнавання мовця на основі аналізу якісних оцінок, нормативні значення яких визначаються як на етапі навчання систем, так і при аналізі отриманого під час сеансу розпізнавання мовного сигналу, представленого у вигляді фонограмами. Складність такої задачі оцінювання надійності полягає у комплексній структурі об'єкту дослідження, який включає в себе мовні сигнали, результати їх оцінювання класифікатором системи розпізнавання щодо належності до класу кожного із зареєстрований у системі осіб-мовців, остаточний результат, прийнятий системою розпізнавання, та тип і ступінь збурень, присутніх у

мовному сигналі. Отже, задачею дослідників є створення моделі для оцінювання надійності сеансу розпізнавання мовця, здатної підтримувати результативність процесу оцінювання надійності в умовах часткової невизначеності враховуваних параметрів, вибір та введення у модель можливості врахування збурюючих факторів, присутніх у мовному сигналі, та емпіричне дослідження адекватності створеного математичного апарату.

ВИБІР ПОКАЗНИКІВ ОЦІНЮВАННЯ НАДІЙНОСТІ АСРМКЗ

Згідно із стандартом ISO/IEC 19795-6:2012 [7] працеспроможність біометричних систем розпізнавання, до яких можна віднести і АСРМКЗ, оцінюється насамперед за такими показниками як імовірність помилкового допуску (False Acceptance Rate, ІПД) – імовірність несанкціонованого допуску (помилка першого роду), виражена у відсотках кількість допусків системою неавторизованих осіб, і імовірність помилкового недопуску (False Rejection Rate, ІПНД) – імовірність помилкового затримання (помилка другого роду), виражена у відсотках кількість відмов у допуску системою авторизованим особам. Параметри ІПД і ІПНД взаємопов'язані і змінюючи пороги ІПД і ІПНД можна варіювати чутливість біометричної системи, щоб вона з більшою імовірністю пропускала зареєстрованих осіб, але при цьому зростатиме і імовірність пропуску системою і незареєстрованих осіб. Результати тестування системи розпізнавання згідно [8] можуть представлятися у вигляді графіка робочої характеристики, яка є параметрично заданою кривою порога прийняття рішень. На такому графіку по вісі абсцис відкладаються оцінки імовірностей помилкового допуску ІПД, а по вісі ординат – оцінки імовірностей помилкового недопуску ІПНД. Для більшої наочності для осей використовують шкалу нормального відхилення [9] (іноді логарифмічну або іншу шкалу). Такий спосіб представлення результатів у задачі біометричної ідентифікації одержав назву кривої компромісного визначення помилки (Detection Error Trade-off curve, КВП крива). Кожна точка КВП кривої відповідає отриманим емпірично помилкам 1-го і 2-го роду при фіксованому порозі прийняття рішень готової до експлуатації системи розпізнавання. Для представлення результатів оцінок системи розпізнавання у вигляді єдиного параметра використовується коефіцієнт рівного рівня помилок (Equal Error Rate, РРП), рівний значенню імовірності помилок при такому рівні порогу прийняття рішень, при якому імовірності помилок помилкового допуску ІПД і помилкового недопуску ІПНД збігаються або найбільш близькі за значенням. Також в статті використовується такий специфічний для систем розпізнавання мовця параметр як функція помилки детектування (Detection cost function, ФПД) з вкрай високою вартістю помилкового пропуску нецільового мовця: $\text{ФПД} = \text{ІПНД} + 549 \cdot \text{ІПД}$. Надалі ми будемо оперувати мінімальним (minФПД) і фактичним (actФПД) значеннями функції помилки детектування в точці прийняття рішення, для визначення яких будується КВП крива, на якій параметру minФПД відповідає значення функції вартості помилки з оптимальним порогом, а параметру actФПД – значення функції вартості помилки з порогом, встановленим вручну.

Основним джерелом помилок АСРМКЗ за умови достатньої розмірності факторного простору для персоналізації фонограм мовців є збурення мовного сигналу. Отже, визначимо множину показників, які б дозволили врахувати ці збурення при оцінюванні надійності сеансу розпізнавання мовця АСРМКЗ. Очевидно, що присутній у фонограмі із записом мовного сигналу шум негативно впливає на ефективність розпізнавання мовця. В даній роботі для оцінювання рівня відношення сигнал/шум (ВСШ) ми використовуватимемо метод основний на аналізі періодичних властивості інтервалів мовлення. Основна енергія мовного сигналу концентрується в околі частоти основного тону мовця, тоді як енергія шуму не детермінована у частотному просторі. Цей факт дозволяє оцінювати рівень ВСШ сигналу використовуючи банки смугових фільтрів H_s та H_n відповідно:

$$H_s(z,t) = \frac{0.5z^{T_p(t)} + 1 + 0.5z^{-T_p(t)}}{1 - a_s z^{-T_p(t)}},\ H_n(z,t) = \frac{0.5z^{T_p(t)} + 1 + 0.5z^{-T_p(t)}}{1 - a_n z^{-T_p(t)}}, \tag{1}$$

де $T_p(t)$ – період основного тону в момент часу t, а a_s і a_n – модифікатори пропускної спроможності фільтрів. Емпірично виявлено найбільш універсальні значення модифікаторів $a_s = 0.25$ і $a_n = 0.7$. Для оцінювання параметрів основного тону використовується авторський метод [10]. Отже, використовуючи (1) ми оцінюємо рівень ВСШ аналізуючи розподіл енергії сигналу у частотному просторі із подальшим усередненням значення рівня ВСШ для всіх мовних фрагментів вхідної фонограми.

Іншим використаним авторами індикатором збурень мовних сигналів є коефіцієнт модуляції у момент часу t визначається із співвідношення

$$KM(t) = \frac{v_{\max}(t) - v_{\min}(t)}{v_{\max}(t) + v_{\min}(t)}, \tag{2}$$

де $v(t)$ – це обвідна сигналу, а $v_{\max}(t)$ і $v_{\min}(t)$ – її локальні максимум і мінімум в околі часу t. Обвідна апроксимує абсолютні значення мовного сигналу, після зниження частоти його дискретизації до 60 Гц, як виконувалося у [11]. Передбачається, що із зростанням рівня ВСШ у фонограмі зростатиме і кількість локальних мінімумів її обвідної, знижуючи відповідно до (2) значення коефіцієнта модуляції. Далі ми усереднюватимемо значення коефіцієнта модуляції для кожної аналізованої фонограми.

Комплектним індикатором наявності у мовному сигналі збурень слугує спектральна ентропія, яка визначається шляхом інтерпретування короткочасного спектру як розподілу імовірності однієї дискретної випадкової величини X із подальшим обрахунком ентропії розподілу. Спектральний розподіл визначається шляхом нормалізації значень короткочасного спектра:

$$p_X\left(f\right) = \frac{e\left(f\right)}{\sum_{k=1}^{N} e\left(k\right)}, \tag{3}$$

де $e\left(f\right)$ – спектральна енергії для частоти f, p_X – це спектральний розподіл. Розширюючи вираз (3) на t-й фрейм фонограми отримаємо такий вираз для обчислення спектральної ентропії:

$$H\left(t\right) = -\sum_{k=1}^{N} p_{X_t}\left(k\right)\log\left(p_{X_t}\left(k\right)\right). \tag{4}$$

Логічним виглядає міркування, що гармонійна структура мовного сигналу знижуватиме значення показника (4), а зростання рівня ВСШ призводитиме до збурення структури мовного сигналу і зростання значення показника (4) відповідно. Далі ми усереднюватимемо значення спектральної ентропії для кожної аналізованої фонограми.

Також оцінювати рівень збурень у мовному сигналі передбачається за допомогою моделей гаусових сумішей (Gaussian Mixture Model, МГС) і універсальної фонової моделі (Universal Background Model, УФМ) зокрема. УФМ – це модель гаусових сумішей, яка представляє імовірнісний розподіл особливостей процесу мовотворення осіб-мовців, фонограми яких містилися у навчальний вибірці. Фактично УФМ містить еталони мовців, на розпізнавання яких було навчено АСРМКЗ. Фонограми із вищім рівнем ВСШ матимуть суттєвішу відмінність від еталонів, збережених у УФМ, отже, рівень правдоподібності, визначений класифікатором при порівнянні аналізованої фонограми із даними УФМ може використовуватися для оцінювання якості цієї фонограми.

МОДЕЛЮВАННЯ НАДІЙНОСТІ АСРМКЗ ЗА ДОПОМОГОЮ БАЄСОВОЇ МЕРЕЖІ

Для оцінювання підсумкової міри надійності тестування АСРМКЗ із врахуванням якісних оцінок фонограм тестувальної вибірки автори пропонують використовувати Баєсівську мережу (Bayesian network, БМ), яка моделюватиме взаємодію між випадковими величинами, задіяними у процесі розпізнавання мовців. Баєсівська мережа – це імовірнісна графічна модель, яка представляє множину випадкових змінних та множину умовних залежностей, з ними пов'язаних, у вигляді орієнтованого ациклічного графу. Вершини графу, який відповідає створюваній БМ, представляють випадкові змінні у Байєсовому сенсі. Це можуть бути спостережувані величини, латентні змінні, невідомі параметри або гіпотези тощо. Ребра графу представляють умовні залежності об'єкту моделювання, зокрема, відсутність з'єднання між вершинами графу означає умовну незалежність змінних, які цим вершинам відповідають. Кожній вершині відповідає функція імовірності, яка перетворює отримані на входи вершини значення від батьківських вершин у імовірність (або розподіл імовірності) змінної, яку представляє ця вершина, що і є виходом цієї вершини.

Автори обрали саме математичний апарат БМ для моделювання надійності тестування АСРМКЗ тому, що оскільки БМ є замкненою моделлю змінних та їх взаємозв'язків, то її можна використовувати для отримання відповідей на імовірнісні запити стосовно них, тобто мережу можна використовувати для уточнення даних про стан певної підмножини змінних при наявності даних спостережень (evidence) щодо інших змінних. Такий процес обчислення апостеріорного розподілу змінних для заданого свідчення називають імовірнісним висновком (probabilistic inference), який дає універсальну достатню статистику для висновку, чи потрібно підбирати значення досліджуваної підмножини змінних, які мінімізують певну цільову функцію, що моделює імовірність помилково прийнятих АСРМКЗ рішень.

На рисунку 1, а) показано графічне представлення базової БМ, створеної для оцінювання надійності прийнятих АСРМКЗ рішень на основі фонограм тестувальної вибірки. Незафарбовані вершини відповідають латентним змінним, а зафарбовані – змінним, які спостерігаються. Зафарбованим вершинам меншого розміру відповідають детермінованим параметрам. Напрямлені дуги між вершинами вказують на наявність умовної залежності між змінним, які відповідають зв'язним вершинам. Опишемо змінні, що відповідають вершинам БМ. Нехай $\hat{s}_i$ – оцінка АСРМКЗ i-ї фонограми із тестувальної вибірки, а Q_i – результати якісної оцінки цієї ж фонограми. $\theta_i = \{T, N\}$ – індикаторна змінна сеансу розпізнавання мовця за i-ю фонограмою, де T – гіпотеза про те, що i-ї фонограма тестувальної вибірки відповідає певній фонограмі навчальної вибірки, задовольняючи порогу правдоподібності АСРМКЗ, а N – гіпотеза, супротивна вищеописаній. $\hat{\theta}_i$ – це прийняте АСРМКЗ рішення по i-ї фонограмі із врахуванням значення порогу правдоподібності ξ_θ. Змінна $R_i \in \{R, U\}$ описує надійність сеансу розпізнавання за i-ю фонограмою, де R – гіпотеза про те, що прийняте АСРМКЗ рішення надійне, а U – гіпотеза, супротивна вищеописаній. Значення $\pi_\theta = (P_T, P_N)$ описують апріорні імовірності P_T і P_N, коли виконуються гіпотези T і N стосовно змінної θ відповідно, при чому $P_N = 1 - P_T$. Змінні $\pi_R = (P_R, P_U)$ описують апріорні імовірності P_R і P_U, коли виконуються

гіпотези R і U стосовно змінної R відповідно, при чому $P_R = 1 - P_U$.

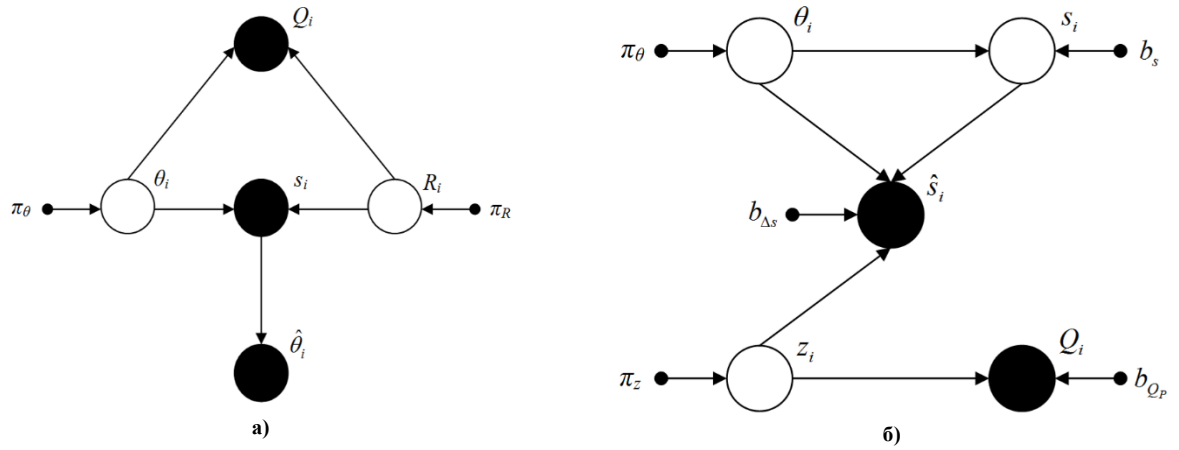

Рис. 1. Архітектура Баєсовських мереж оцінювання надійності АСРМКЗ: а) базова БМ, б) уточнена БМ

За своїм призначенням БМ дозволяє сформулювати узагальнений імовірнісний розподіл вищеописаних змінних, що відповідають її вершинам, у вигляді рівняння

$$P\left(\hat{s}, Q, R, \theta, \hat{\theta} \middle| \pi_\theta, \pi_R\right) = P\left(\hat{s} \middle| R, \theta\right) P\left(Q \middle| R, \theta\right) P\left(\hat{\theta} \middle| R, \theta\right) P\left(\theta \middle| \pi_\theta\right) P\left(R \middle| \pi_R\right), \quad (5)$$

використовуючи яке можна отримати апостеріорну імовірність розподілу R в залежності від значень змінних, які спостерігаються, так:

$$P\left(R \middle| \hat{s}, Q, \hat{\theta}, \pi_\theta, \pi_R\right) = \frac{\sum_{\theta \in \{T,\ N\}} P\left(\hat{s}, Q, R, \theta, \hat{\theta} \middle| \pi_\theta, \pi_R\right)}{\sum_{R} \sum_{\theta \in \{T,\ N\}} P\left(\hat{s}, Q, R, \theta, \hat{\theta} \middle| \pi_\theta, \pi_R\right)}. \quad (6)$$

Розподіл імовірності $P\left(\hat{s} \middle| R, \theta, \hat{\theta}\right)$ описується УФМ, а $P\left(Q \middle| R, \theta\right)$ – МГС, а імовірність $P\left(\hat{\theta} \middle| \theta, R\right)$ визначається як $P\left(\hat{\theta} \middle| \theta, R\right) = \begin{cases} 1, & якщо\ \hat{\theta} = \theta \vee R_i = R \wedge \hat{\theta} = \theta \vee R_i = U, \\ 0, & у\ іншому\ випадку. \end{cases}$

У сформульованій вище структурі БМ зв'язок між θ і Q описує припущення що на фонограми навчальної і тестувальної вибірок збурюючи фактори впливали по різному. Запропонована у вигляді БМ модель оцінювання надійності АСРМКЗ за якістю тестувальних фонограм здійснює оцінювання надійності на основі показників оцінки АСРМКЗ і індикатора якості фонограм. Якщо ми бажаємо виконувати оцінювання надійності без урахування показника s_i , то можна видалити відповідну вершину із БМ та видалити з (5) імовірність $P\left(s \middle| R, Q\right)$.

Графічно уточнену Баєсівську мережу для оцінювання надійності АСРМКЗ представлено на рисунку 1, б). На відміну від базової БМ у ній присутні вершини s_i та z_i і здійснено описані далі перепозначення, зокрема, $\hat{s}_i$ – це оцінка АСРМКЗ i-ї фонограми, отриманої за реальних умов експлуатації (тобто за присутності у мовному сигналі збурень, спричинених впливом навколишнього середовища), а s_i – це оцінка АСРМКЗ i-ї фонограми, у якій збурення відсутні. Змінні зв'язані між собою залежністю $\hat{s}_i = s_i + \Delta s$. Зазвичай значення змінної s_i невідоме, але якщо у навчальній вибірці є інформація про вид та ступінь збурень присутніх фонограм, то значення s_i можна визначити на етапі навчання БМ. Змінна z_i описує якісний стан. Це вектор, який містить елементи z_{ik} , $k = 1,...,K$, які описують ступінь збурення мовного сигналу у тестовій фонограмі відповідно до вказаних вище типів збурень. Звичайно, змінні z_i і Δs зв'язані. Множини вхідних параметрів $b_S = \{\mu_S, \Lambda_S\}$, $b_{\Delta S} = \{\mu_{\Delta S}, \Lambda_{\Delta S}\}$ і $b_{Q_P} = \{\mu_{Q_P}, \Lambda_{Q_P}\}$ пов'язані із використанням математичного апарату МГС для представлення мовних сигналів у просторі ознак і описують параметри функцій щільності розподілу (вектор математичного сподівання μ і коваріаційну матрицю Λ) представлення об'єктів, які описуються змінними s , Δs і Q_P відповідно. Отже, апостеріорну імовірність s в контексті МГС можна описати так:

$$P(s|\hat{s},Q)=\sum_{\theta\in\{T,N\}}\sum_{k=1}^{K}P(\theta,z_k=1|\hat{s},Q)N\left(s\middle|\mu'_{s_{k\theta}},\Lambda^{-1}_{s_{k\theta}}\right), \quad (7)$$

де

$$\Lambda'_{s_{k\theta}}=\Lambda_{\Delta s_{k\theta}}+\Lambda_{s_\theta}, \quad (8)$$

$$\mu'_{s_{k\theta}}=\Lambda^{-1}_{s_{k\theta}}\left(\Lambda_{\Delta s_{k\theta}}\left(\hat{S}-\mu_{\Delta s_{k\theta}}\right)+\Lambda_{s_\theta}\mu_{s_\theta}\right). \quad (9)$$

Ваги суміші визначаються як

$$P(\theta,z_k=1|\hat{s},Q)=\frac{P(\hat{s}|\theta,z_k=1)P(Q|z_k=1)P(\theta)\pi_{z_k}}{\sum_{\theta\in\{T,N\}}\sum_{k=1}^{K}P(\hat{s}|\theta,z_k=1)P(Q|z_k=1)P(\theta)\pi_{z_k}}, \quad (10)$$

$$P(\hat{s}|\theta,z_k=1)=N\left(\hat{s}\middle|\mu'_{\hat{s}_{k\theta}},\Lambda^{-1}_{\hat{s}_{k\theta}}\right), \quad (11)$$

$$\Lambda'_{\hat{s}_{k\theta}}=\Lambda_{s_\theta}\Lambda^{-1}_{s_{k\theta}}\Lambda_{\Delta s_{k\theta}}, \quad (12)$$

$$\mu'_{\hat{s}_{k\theta}}=\mu_{s_\theta}+\mu_{\Delta s_{k\theta}}. \quad (13)$$

Зв'язки між змінними удосконаленої БМ описується такими співвідношеннями:

$$P(s|\theta)=N\left(s\middle|\mu_{s_\theta},\Lambda^{-1}_{s_\theta}\right), \quad (14)$$

$$P(\hat{s}|s,z_k=1,\theta)=N\left(\hat{s}\middle|s+\mu_{\Delta s_{k\theta}},\Lambda^{-1}_{\Delta s_{k\theta}}\right), \quad (15)$$

$$P(Q|z_k=1)=N\left(Q\middle|\mu_{Q_k},\Lambda^{-1}_{Q_k}\right), \quad (16)$$

$$P(z)=\prod_{k=1}^{K}\pi_{z_k}^{z_k}. \quad (17)$$

У базовій БМ надійність сеансу розпізнавання за i-ю фонограмою описано окремою змінною R_i, тоді як в уточненій БМ цей показник у явному вигляді відсутній і визначається опосередковано, способом, описаним далі. У типових системах розпізнавання мовця рішення щодо особи мовця $\hat{\theta}$ приймається системою із урахуванням значення порогу правдоподібності ξ_θ у вигляді оцінки $\hat{s}$. Щоб визначитися із достовірністю прийнятого АСРМКЗ щодо особи мовця рішення використаємо БМ для оцінювання апостеріорного розподілу s із урахуванням прийнятої системою оцінки та результату якісної оцінки вхідної фонограми у вигляді $P(s|\hat{s},Q)$, що дозволяє оцінити надійність прийнятого АСРМКЗ рішення так:

$$P(R_i=R|\hat{s},Q)=\begin{cases}P(s>\xi_\theta|\hat{s},Q), & якщо\ \hat{\theta}=T,\\ P(s<\xi_\theta|\hat{s},Q), & якщо\ \hat{\theta}=N.\end{cases} \quad (18)$$

На рис. 2 наведено приклади розподілу результатів оцінки АСРМКЗ, які потім використовувалися для обчислення апостеріорної імовірності показника надійності R_i при відомому значенні якісного стану z. Сині криві описують розподіл оцінки для тестових фонограм без збурень, червоні криві описують розподіл оцінки для тестових фонограм із збуреннями, а зафарбований фрагмент під зеленою кривою обмежує частку надійних тестових випробувань із урахуванням сформульованого вище математичного апарату. Розміри зафарбованого фрагменту дозволяють однозначно стверджувати, що уточнена БМ потенційно набагато чутливіша за базову, що, проте, необхідно підтвердити емпірично.

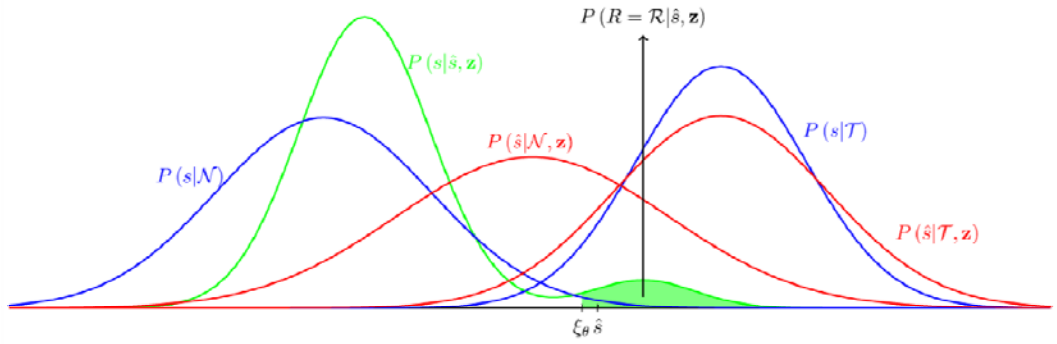

Рис. 2. Розподіл оцінки АСРМКЗ при обчисленні $P\left(R_i\in R\middle|\hat{s},z_k=1\right)$

ПОСТАНОВКА ЕКСПЕРИМЕНТУ ТА АНАЛІЗ РЕЗУЛЬТАТІВ

Для емпіричних досліджень запропонованих БМ оцінювання надійності сеансів автоматизованого розпізнавання мовців автори використали створену у [11] АСРМКЗ, у які застосовувалося і-векторне представлення фонограм із нейромережевим класифікатором. Фонограми представлялися у просторі ознак 400-мірними і-векторами, на основі 20 MF-кепстральних коефіцієнтів та їх перших та других похідних і 2048-елементної діагональної коваріаційної УФМ матриці. Передобробка фонограм включала центрування, передфільтрацію та нормування тривалості звучання фонограм.

В якості бази фонограм для формування навченої і тестувальної вибірок використовувалася база записів із безкоштовної бази даних NOIZEUS [10] – 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 мовця, частота дискретизації записів складає 25 кГц, але задля додавання шуму була зменшена до 8 кГц) та записів типових побутових та техногенних шумів, які можна підмішувати до незбурених мовних сигналів у довільний спосіб. При формуванні вибірки для навчання БМ до її складу включено фонограми із різним рівнем ВСШ та різними видами техногенних шумів для кожного мовця.

Спочатку параметр оцінки АСРМКЗ калібрувався методом лінійної логарифмічної регресії із використанням Bosaris Toolkit [4] на основі лише фонограм без збурень із навчальної вибірки. Далі відкалібрована АСРМКЗ виконувала операцію розпізнавання мовців із тестувальної вибірки, у якій були присутні фонограми із рівнем ВСШ у 0, 5, 10, 15 дБ, при встановленому рівні Байєсового порогу прийняття рішень у 2.29. Загалом результативність розпізнавання мовців за фонограмами із рівнем ВСШ у 0 дБ становила РРП = 2.2%, minФПД = 0.14 і actФПД = 0.17, а середні значення цих показників при розпізнаванні мовців за всіма фонограмами із тестувальної вибірки склали minФПД = 0.99 і actФПД = 2.96 при тому, що при значенні показника actФПД більше одиниці система розпізнавання вважається ненадійною. Отже, покращити результативність досліджуваної АСРМКЗ пропонується шляхом виявлення фонограм, які найбільш впливають на значення показника actФПД, і проаналізувавши умови, за яких вони були записані, визнавати в подальшому ненадійними сеанси розпізнавання, у фонограмах яких спостерігаються ідентифіковані спотворення, значення яких перевищує встановлений рівень.

Враховуючи контрольованість процесу навчання БМ із відомими характеристиками фонограм навчальної вибірки, можна точно встановити кількість фонограм без збурень, визначити значення параметру s для кожної фонограми навчальної вибірки із відомим ступенем збурень, і встановити значення параметру z в залежності від даних про рівень ВСШ навчальної та тестувальної фонограм для відповідних мовців: $z = \left(SNR_{trn}, SNR_{tst}\right)$. Встановивши значення цих параметрів можна обчислити розподіли Δs для кожної пари рівнів ВСШ і θ, $P\left(\Delta s \mid \theta, SNR_{trn}, SNR_{tst}\right)$. Проведені обчислення $P\left(\Delta s \mid \theta, SNR_{trn}, SNR_{tst}\right)$ виявили ряд залежностей, а саме, якщо із двох фонограм одна з яких не містить збурень, а у другій вони присутні, і якщо сеанс розпізнавання визнано надійним, то середнє значення Δs спадає із зростанням ступеня збурень, а якщо сеанс розпізнавання визнано ненадійним, то середнє значення Δs повільно зростає із ступенем збурень; якщо із двох фонограм обидві збурені незначно, то якщо сеанс розпізнавання визнано надійним, то середнє значення Δs прямує до 0, а якщо сеанс розпізнавання визнано ненадійним, то прямує до ∞; якщо із двох фонограм обидві збурені суттєво, то якщо сеанс розпізнавання визнано надійним, то середнє значення Δs повільно зменшується, а якщо сеанс розпізнавання визнано ненадійним, то середнє значення Δs стрімко зростає. Скориставшись цими правилами можна емпірично оцінювати значення параметра z для вхідної фонограми, аналізуючи її фрагменти, і по результатам оцінювання робити висновок щодо надійності процедури розпізнавання на її основі.

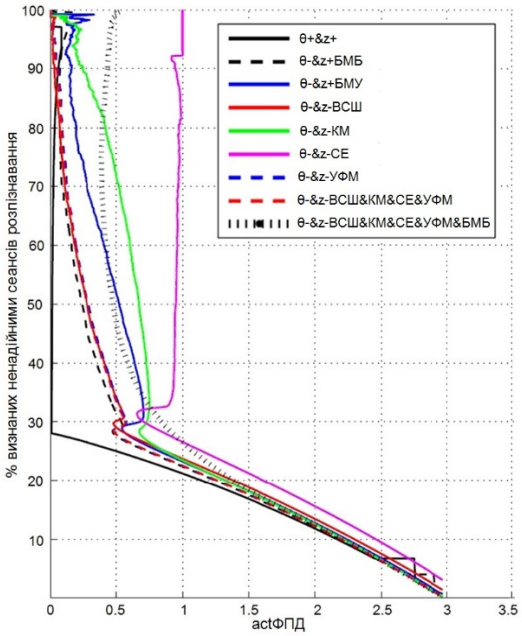

Рис. 3. Залежність між показником *actФПД* і кількістю визнаних базовою і уточненою БМ ненадійними сеансів розпізнавання при різних початкових умовах

Скориставшись сформульованими вище правилами проведено експеримент по виявленню залежності між кількістю визнаних ненадійними сеансів розпізнавання і показником actФПД, результати якого представлені на рисунку 3.

Наведені на рисунку 3 залежності отримано при варіюванні пороговим значенням апостеріорної

імовірності $P\left(R\left|\hat{s},Q\right.\right)$, обчислюваної за відношенням (18). Кращою вважається залежність, яка забезпечує вище значення actФПД відкидаючи меншу кількість випробувань. Рішення щодо визнання сеансу розпізнавання ненадійним приймалося на основі оцінок як базової так і уточненої БМ при різних початкових умовах, вставлюваних варіюванням параметрами θ та z. Спочатку отримаємо оцінки БМ за умови, що значення змінних θ та z відомі. При такому припущенні множина гаусіан, описуваних рівнянням (7), вироджується до єдиної функції – цей гіпотетичний сценарій дозволяє нам оцінити рівень нижньої межі показника actФПД для досліджуваної АСРМКЗ. Як видно з рисунку (крива θ+&z+), за таких початкових умов можна звести значення actФПД майже до нуля, відкидаючи близько 30% сеансів розпізнавання. Далі розглядаємо випадок, коли змінна θ невідома, а значення змінної z є детерміновані, тобто відомі рівні спотворень для фонограм як навчальної, так і тестувальної вибірок. За таких умов суміш (7) міститиме два гаусіани із вагами-імовірностями $P\left(\theta\left|\hat{s},z_k\right.=1\right)$. Отримані при таких початкових умовах результати (криві θ-&z+БМБ для базової і θ-&z+БМУ для уточненої БМ) показують погіршення значення actФПД, але мешу кількість визнаних ненадійними сеансів розпізнавання порівняно із попередніми дослідами. Розглянемо нарешті найхарактерніший для типових систем розпізнавання мовця випадок, для якого величини θ і z невідомі. За такої початкової умови розподіли $P\left(Q\left|z_k\right.=1\right)$ опишемо діагональними та взаємоковаріаційними матрицями комбінації збурюючи факторів, описаних відношеннями (1)-(4). Отримані результати (крива θ-&z-ВСШ для уточненої БМ і при окремо варійованому рівні ВСШ, θ-&z-КМ – для уточненої БМ і при окремо варійованому коефіцієнті модуляції, θ-&z-СЕ – для уточненої БМ і при окремо варійованій спектральній ентропії, θ-&z-УФМ – для уточненої БМ і при окремо варійованому відхиленні від УФМ, θ-&z-ВСШ&КМ&СЕ&УФМ – для уточненої БМ і варіюванні значеннями всіх збурюючи факторів, θ-&z-ВСШ&КМ&СЕ&УФМ&БМБ – для базової БМ і варіюванні значеннями всіх збурюючи факторів) показують, що найбільшу кількість визнаних ненадійними сеансів розпізнавання при збереженні прийнятного рівня actФПД отримано при варіюванні значенням коефіцієнту модуляції, за яким йде рівень ВСШ, а спектральна ентропія і логарифмічний поріг правдоподібності УФМ виявляється малоінформативним. Загалом, для всіх проведених дослідів базова БМ виявилася менш інформативнішою за уточнену, використовуючи інформацію від якої вдалося зменшити значення actФПД з 2.96 до 0.72, 0.27 та 0.09 при визнанні ненадійними 25, 50 або 75% сеансів розпізнавання із врахуванням всіх збурюючи факторів.

ВИСНОВКИ

У даній статті запропоновано новий метод оцінювання надійності сеансу розпізнавання особи автоматизованою системою розпізнавання мовця критичного використання, який, на відміну від існуючих, використовує Баєсівську мережу, яка описує зв'язки між оцінкою сеансу розпізнавання, згенерованою класифікатором АСРМКЗ, встановленим у системі значенням порогу правдоподібності, оцінками надійності сеансу розпізнавання та оцінками впливу основних збурюючи факторів на мовний сигнал, що дозволяє за рахунок притаманних Байєсівській мережі властивостей проводити оцінювання надійності сеансу розпізнавання в умовах часткової невизначеності згаданих параметрів та досягти заданого значення показника якості системи розпізнавання ідентифікувавши ненадійні сеанси розпізнавання.

Запропоновані авторами теоретичні результати знайшли свого емпіричного підтвердження при аналізі результатів роботи АСРМКЗ, параметри чутливості якої було відкалібровано на даних навчальної вибірки. До складу тестувальної вибірки були включені фонограми із різним видом та ступенем збурень, серед яких враховувалися такі показники, як рівень відношення сигнал/шум, коефіцієнт модуляції, спектральна ентропія та ступінь відхилення параметрів парольного мовного сигналу від значень цих же параметрів, узагальнених на етапі навчання системи у вигляді універсальної фонової моделі. Автори запропонували базову та уточнену Байєсівські мережі для оцінювання надійності сеансів розпізнавання мовців за даними тестувальної вибірки, які відрізняються тим, що в уточненій БМ параметр надійності виражено залежною від оцінок рівня присутніх у мовному сигналі збурень змінною, тоді як базова БМ є більш універсальною та потребує меншої кількості обчислень.

Використовуючи універсальний показник якості систем розпізнавання мовців actФПД проведено комплексне дослідження ефекту від використання отримуваної від БМ інформації при винесені вердикту щодо надійності сеансів розпізнавання, результати якого показали, що базова БМ виявилася менш інформативнішою за уточнену, використовуючи інформацію від якої вдалося зменшити значення actФПД з 2.96 до 0.72, 0.27 та 0.09 при визнанні ненадійними 25, 50 або 75% сеансів розпізнавання із врахуванням всіх збурюючи факторів, при тому, що працеспроможною визнається система розпізнавання, значення показника actФПД якої перевищує одиницю. Також результати експерименту показують, що найбільшу кількість визнаних ненадійними сеансів розпізнавання при збереженні прийнятного рівня actФПД отримано при зростанні значень коефіцієнту модуляції, за яким йде рівень ВСШ, а спектральна ентропія і логарифмічний поріг правдоподібності УФМ виявилися малоінформативними для запропонованого методу оцінювання надійності.

Література

1. Kovtun V.V. Research of neural network classifier in speaker recognition module for automated system of critical use / Mykola M. Bykov, Viacheslav V. Kovtun, Andrzej Smolarz, Mukhtar Junisbekov, Aliya Targeusizova, Maksabek Satymbekov // SPIE 10445, Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments. – 2017. – 1044521. – DOI: 10.1117/12.2280930.

2. Huggins M.C., Grieco J.J. Confidence metrics for speaker identification [Electronic resource]. Access mode: https://pdfs.semanticscholar.org/613d/7b60da94100d152f38611cd1ea9fd42056e3.pdf

3. Campbell W. M. Estimating and evaluating confidence for forensic speaker recognition / W. M. Campbell, D. A. Reynolds, J. P. Campbell // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 23-23 March 2005: proceedings. – Philadelphia, PA, USA: IEEE, 2005. – P. 4117–4120. – DOI: 10.1109/ICASSP.2005.1415214

4. Solewicz Y., Koppel M. Considering Speech Quality in Speaker Verification Fusion [Electronic resource]. Access mode: http://www.isca-speech.org/archive/archive_papers/interspeech_2005/i05_2189.pdf

5. Richiardi J. A probabilistic measure of modality reliability in speaker verification / J. Richiardi, P. Prodanov, A. Drygajlo // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 23-23 March 2005: proceedings. – Philadelphia, PA, USA: IEEE, 2005. – P. 2154–2160. – DOI: 10.1109/ICASSP.2005.1415212

6. Richiardi J. Speaker Verification with Confidence and Reliability Measures / J. Richiardi, A. Drygajlo, P. Prodanov // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 14–19 May 2006: proceedings. – Toulouse, France: IEEE, 2006. – P. 1238–1242. – DOI: 10.1109/ICASSP.2006.1660102

7. ISO/IEC 19795-6:2012(en) Information technology – Biometric performance testing and reporting — Part 6: Testing methodologies for operational evaluation. [Electronic resource]. Access mode: https://www.iso.org/obp/ui/#iso:std:iso-iec:19795:-6:ed-1:v1:en

8. The NIST Year 2012 Speaker Recognition Evaluation Plan [Electronic resource]. Access mode: http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf

9. Navratil J., Klusacek D. On linear DETs [Electronic resource]. Access mode: http://www.research.ibm.com/CBG/papers/icassp07_navratil.pdf

10. 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. – Хмельницький. – 2017. – № 3(249). – С. 222–228.

11. Ковтун В.В. Метод представлення ознак у автоматизованій системі розпізнавання мовця критичного застосування / М.М. Биков, В.В. Ковтун, М.С. Фурман // Вісник Хмельницького національного університету, серія: Технічні науки. – Хмельницький, 2017. – № 5(253). – С. 112–120.

References

1. Kovtun V.V. Research of neural network classifier in speaker recognition module for automated system of critical use / Mykola M. Bykov, Viacheslav V. Kovtun, Andrzej Smolarz, Mukhtar Junisbekov, Aliya Targeusizova, Maksabek Satymbekov // SPIE 10445, Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments. – 2017. – 1044521. – DOI: 10.1117/12.2280930.

2. Huggins M.C., Grieco J.J. Confidence metrics for speaker identification [Electronic resource]. Access mode: https://pdfs.semanticscholar.org/613d/7b60da94100d152f38611cd1ea9fd42056e3.pdf

3. Campbell W. M. Estimating and evaluating confidence for forensic speaker recognition / W. M. Campbell, D. A. Reynolds, J. P. Campbell // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 23-23 March 2005: proceedings. – Philadelphia, PA, USA: IEEE, 2005. – P. 4117–4120. – DOI: 10.1109/ICASSP.2005.1415214

4. Solewicz Y., Koppel M. Considering Speech Quality in Speaker Verification Fusion [Electronic resource]. Access mode: http://www.isca-speech.org/archive/archive_papers/interspeech_2005/i05_2189.pdf

5. Richiardi J. A probabilistic measure of modality reliability in speaker verification / J. Richiardi, P. Prodanov, A. Drygajlo // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 23-23 March 2005: proceedings. – Philadelphia, PA, USA: IEEE, 2005. – P. 2154–2160. – DOI: 10.1109/ICASSP.2005.1415212

6. Richiardi J. Speaker Verification with Confidence and Reliability Measures / J. Richiardi, A. Drygajlo, P. Prodanov // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 14–19 May 2006: proceedings. – Toulouse, France: IEEE, 2006. – P. 1238–1242. – DOI: 10.1109/ICASSP.2006.1660102

7. ISO/IEC 19795-6:2012(en) Information technology – Biometric performance testing and reporting — Part 6: Testing methodologies for operational evaluation. [Electronic resource]. Access mode: https://www.iso.org/obp/ui/#iso:std:iso-iec:19795:-6:ed-1:v1:en

8. The NIST Year 2012 Speaker Recognition Evaluation Plan [Electronic resource]. Access mode: http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf

9. Navratil J., Klusacek D. On linear DETs [Electronic resource]. Access mode: http://www.research.ibm.com/CBG/papers/icassp07_navratil.pdf

10. 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. – Khmelnytskyi. – 2017. – № 3(249). – S. 222–228.

11. Kovtun V.V. Metod predstavlennia oznak u avtomatyzovanii systemi rozpiznavannia movtsia krytychnoho zastosuvannia / M.M. Bykov, V.V. Kovtun, M.S. Furman // Visnyk Khmelnytskoho natsionalnoho universytetu, seriia: Tekhnichni nauky. – Khmelnytskyi, 2017. – № 5(253). – S. 112–120.

Рецензія/Peer review : 2.10.2018 р. Надрукована/Printed :22.11.2018 р.
Рецензент: д.т.н., проф. Бісікало О.В.